Neutral Tone in Mandarin:

Representation and Interaction with Utterance-level Prosody

by ZHANG YIXIN

Selwyn College

This thesis is submitted for the degree of Doctor of Philosophy of

Theoretical and Applied Linguistics

at July 31, 2021

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text.

It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text.

It does not exceed the prescribed word limit for the Modern and Medieval Languages and Linguistics Degree Committee.

Neutral Tone in Mandarin: Representation and Interaction with Utterance-level Prosody

Zhang Yixin

In Standard Mandarin, there are syllables that do not carry any of the four citation tones (T1: High-level tone, T2: Mid-rising tone, T3: Low-convex tone and T4: High-falling tone), and they are said to have a neutral tone (NT). These syllables are usually shorter, lighter, prosodically grouped with the preceding CT-bearing syllables. These characteristics of NT have led to a prevailing view that it has no underlying phonological specification. However, research has focused more on how the surface pitch variations of NT are realized rather than the underlying representation of NT.

In contrast, morphological, sociolinguistic and diachronic work on NT has suggested that NT may not be a homogeneous entity. In this thesis, I provide acoustic and psycholinguistic evidence that there are two types of NT, Intrinsic NT and Derived NT. Intrinsic NT refers to morphemes that were lexicalized as tone-deleted, unstressed syllables even before the formation of the four CTs of modern Mandarin. Derived NT refers to morphemes derived from the CTs via stress-related tone-deletion.

In Part A, the phonological representation of Intrinsic and Derived NT is explored through two production and two processing experiments. The results show that Intrinsic NT is likely to have an underspecified tonal target while Derived NTs are underlyingly CTs. In addition, both subtypes of NT are metrically light, unlike heavy CTs.

Part B explores the interaction between NTs and utterance-level prosody in production and perception experiments. NT-bearing syllables have lengthening patterns under focus similar to CT-bearing syllables, in contrast to the realization of unstressed syllables in English. In perception, the identification of intonation (Statement vs. Question) on Intrinsic NT was similar to Derived NT. When compared to CTs, the NTs elicit less bias towards question than T4, and higher accuracy than T2, which may result from their simpler surface representations.

To Those Who Helped Me

To Those Who Read This

Acknowledgments

First of all, my deepest gratitude goes to my supervisors, Prof. Brechtje Post and Dr. Elaine Schmidt, who have guided me in every step of my thesis work, trained me in various research techniques, helped me in my experimental design and academic writing, and supported me through the difficult times during my PhD. I feel honored to have them as my supervisors.

In addition, I would like to thank Prof. Francis Nolan, who was my Mphil supervisor, and has given me many useful comments and suggestions on my PhD work.

I also wish to thank my examiners, Prof. Henriette Hendriks and Prof. Yiya Chen. Their comments and questions, which are very inspiring, have helped me to improve the quality of my thesis.

I would like to thank the other members in the phonetic lab as well, from who I have learned a lot. In particular, I wish to thank Yang Li, who have trained me in statistics and the use of R, Jasper Sim, who have done the recordings for my first experiment, and Kirsty E. McDougall, who have helped me with the statistical comparisons of tone shapes.

Finally, I would like to thank my families for their consistent support, especially my parents and my cousins Zhengshuo Yan and Liyan Li.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Overview

This thesis explores the phenomenon of neutral tone in Standard Mandarin. Standard Mandarin is a tone language that exploits phonologically contrastive pitch variations (i.e., lexical tones) to distinguish lexical meanings. In Standard Mandarin, there are four citation tones (CTs) with phonemic status that can distinguish two morphemes of the same segmental structure (**Table 1.1.a**).

Table 1.1.a Citation tones in Standard Mandarin

Tone	Pitch Contour	Chinese Example	Tone Letter	Chao Tone-Numerals[1]	Gloss
1	High-level	妈	ma˥	ma55	mother
2	Mid-rising	麻	ma˧˥	ma35	hemp
3	Low-dipping	马	ma˨˩˧	ma214	horse
4	High-falling	骂	ma˥˩	ma51	scold

There are also syllables that do not carry any of these four CTs and they are said to have a neutral tone (NT). The syllables that bear NT are commonly recognized as weak elements as they are usually shorter, lighter and prosodically grouped with the preceding CT-bearing syllables. In other words, the NT-bearing syllables could not appear in the word-initial positions or on their own but must be attached to other CT-bearing syllables.

In the traditional view, NT is often seen as a tone neutralization phenomenon in which the

[1] With five numbers, Chao's numeral system distinguishes 5 different tone heights (1 is the lowest and 5 is the highest); tone contours are represented by combinations of different numbers (Chao, 1930; 1965). This system enables a convenient transcription of auditory impressions of tones and hence is widely adopted in tone studies. It should be noticed, however, although Chao allows combination of numbers, he still sees tone as an indivisible entity. In other words, [55] describes a high-level tone rather than a tone with two high targets. This view is fundamentally different from those proposed later (e.g., Yip, 1980 in Section 2.3.2).

contrasts between the four CTs neutralize due to the loss of stress (Chao, 1965). However, different from the other tone change phenomena in Mandarin, the assignment of NT in Mandarin cannot be neatly predicted by any grammatical rule. There is not yet a general consensus on which words involve NT. In fact, it makes more sense to describe NT words in Mandarin as an open class. In addition, except for a small number of NT-bearing morphemes, the NT realization of the majority of them is optional, that is, these morphemes could also be realized with the CT alternates (i.e., the CTs those morphemes would bear when enunciated in isolation), and this 'optional realization' trend is increasing.

From a phonetic perspective, both the early impressionistic studies and the following instrumental research have reported a strong influence of the tonal context, especially of the preceding CT, on the surface f_0 realization of NT (for detailed review, see Section 3.1). The association with unstressed syllables and the high contextual dependency of NT have led to a prevailing view that there are no phonological specifications for NT; correspondingly, how the surface pitch variations of NT are realized has become the focus in most data-driven phonological analyses, rather than its underlying representation and the potential differences there (for detailed review, see Section 3.32).

In contrast, in the sociolinguistic and morphological studies conducted at different times, different types of NT have been defined to record or to guide the realization norm of NT in Standard Mandarin. The categorization of NT ranges from binary to multiple classifications, but none of these classifications alines with the phonological distinction that I am proposing here.

In the present thesis, I bring together insights from the phonological, morphological and

diachronic literatures in order to answer the following **three research questions** through experiments.

RQ1: Is NT a homogeneous phonological entity, or instead, are there different types of NT with different underlying representations?

RQ2: What is/are the underlying representation(s) of Mandarin NT, and is/are Mandarin NT(s) metrically lighter than CT?

RQ3: How would the representation(s) influence the interaction between NT and utterance-level prosody?

1.2 Structure of the Thesis

This thesis is divided into two parts. Part A (Chapter 4 and 5) explores the first two research questions about the underlying representation(s) of NT, and Part B (Chapter 6 and 7) explores the third question about the interaction between NT and utterance-level prosody. The remainder of this thesis is organized as follows.

Chapter 2 introduces the development and grammar of spoken Standard Mandarin. Relevant theories like autosegmental phonology and underspecification theories are also reviewed in the context of Mandarin in this chapter. Chapter 3 zooms in on **Neutral Tone**, the focus of the thesis. The existing phonetic, phonological, morphological and diachronic studies on Mandarin NT are reviewed in this chapter. At the end of this chapter, two types of NT, Intrinsic NT and Derived NT are proposed. Chapter 4 and 5 presents experiments investigating RQ1 and RQ2 from an acoustic and a processing perspective respectively. Chapter 6 presents an experiment on the durational adjustment of Intrinsic NT, Derived NT and CT words under focus. Chapter 7 presents an

experiment on the perception of intonation on NT words of either type in comparison to the representative CTs, T2 and T4. The differences and similarities between Intrinsic NT and Derived NT in their interaction with utterance-level prosody, i.e., RQ3, are investigated, shedding further light on the representations of these two types of NT. Chapter 8 concludes the findings of this study.

Chapter 2 Background

This chapter provides the background information of Standard Mandarin and the theoretical frameworks needed for the rest of the thesis. The chapter is organized as follows. Section 2.1 gives an overview of Standard Mandarin from both synchronic and diachronic perspectives. Section 2.2 zooms into the citation tones in Mandarin, introducing both the phonetic features and phonological representations of Mandarin citation tones as well as the tone sandhi phenomena. The relevant phonological theories will be introduced alongside. Section 2.3 introduces stress in Mandarin, a highly controversial topic, which is closely related to the phenomenon of NT.

2.1 Standard Mandarin: The Common Language

Chinese, the largest language group in the Sino-Tibetan family, is also the largest tone language group in the world[2]. It is spoken by numerous people of different ethnicities across a large area, and hence has a wide variety of prestige dialects. The dialectological literature usually divides Chinese dialects currently spoken in China into 9 main dialect districts, Mandarin, Cantonese (including Pinghua and Tuhua spoken in Guangxi province), Wu, Jin, Min, Hakka, Xiang, Hui and Gan. Each dialect district consists of many small sub-dialects belonging to the same dialectal family (Language Atlas of China, 1987). Mandarin district, also known as Northern Mandarin district, is the largest dialect with over 1.1 billion speakers and 8 subareas (**Figure 2.1.a**). The Mandarin dialects spoken in different subareas have different accents but are mutually understandable. However, the speech between Mandarin and non-Mandarin districts is hardly

[2] Tone languages refer to the languages that use pitch in addition to vocalic quality and consonantal place of articulation to convey lexical meanings. Different from the typical stress languages in which pitch may be used to indicate lexical stress, prosodic focus or other post-lexical pragmatic meanings, the pitch patterns of the words or morphemes in tone languages remain relatively robust; otherwise the core meaning of the words or morphemes will change.

mutually understandable. Moreover, dialects belonging to the same non-Mandarin dialectal family such as Wu dialect are of very low mutual intelligibility as well.

Figure 2.1.a Distribution of Chinese dialects and Mandarin dialects (adapted from Centre for the Protection of Language Resources of China, n.d.)

Chinese dialects differ from each other in their phonology, especially in their tones. Yue dialects usually have 8-10 tones, Wu and Min dialects usually have 7-8 tones and Mandarin dialects usually have 4 tones[3]. The shapes of tones and the phonological representations of tones vary correspondingly. For instance, Wu dialects are believed to have word tones rather than morpheme tones, because polysyllabic words in Wu dialects "are learned as discrete entities associated with a particular overall tone contour, and not as combinations of individually

[3] Some dialectological studies also report Mandarin varieties with 3 or 5 tones. 3-tone Mandarin varieties mainly exist in the north-west of China, near Jin dialect districts; 5-tone Mandarin varieties are mainly found in Jiang Huai subarea, a relatively southern subarea of Mandarin located at the middle and lower reaches of Yangtze River. The five tones do not include neutral tone but usually the *Ru* tone, namely, the checked tone.

identifiable syllables" (Sherard, 1972: 116). Although the major dialects also differ from each other in certain morphological and syntactic aspects, it is mainly the pronunciation differences that lead to the low mutual intelligibility between Chinese dialects.

Throughout history, this large dialectal variation has led to a strong need for a pronunciation norm that could be used across China. Nevertheless, unlike the early success achieved in the unification of writing systems by the first emperor Shi Huang Di of the *Qin* Dynasty (221–206 BC), a pronunciation norm was not successfully established until the second half of last century, when Standard Mandarin was promoted by the government of the People's Republic of China.

Standard Mandarin, officially known as *Putonghua* or Standard Chinese, is a dialect of Mandarin established as common language and pronunciation norm among the speakers of Chinese dialects now. Its history, however, can be dated back to at least the 14th century AD. The dialect name Mandarin is a Portuguese word meaning 'counselor' or 'minister'. It was first used by the Portuguese as the direct translation of *Guanhua*. From the *Ming* Dynasty (1368 -1644 AD) onwards, *Guanhua,* 'official speech' replaced *Tongyu,* 'common speech'[4] and became the pronunciation norm in China.

Guanhua was initially developed based on Nanjing dialect. Later on, the dialect of Beijing became more and more influential as the second emperor of the *Ming* Dynasty, Emperor Yongle, moved the capital city from Nanjing to Beijing in 1421. By 1909, the *Qing* dynasty had announced Beijing Mandarin as the national language. It is worth mentioning that the *Qing* dynasty was established by the Manchurian so that the development of Beijing Mandarin and Mandarin in

[4] *Tongyu* is a cover name for the pronunciation norms used in different dynasties before *Ming*. The norms were usually developed based on the dialects spoken in the capital cities. Therefore, *Tongyu* differs from dynasty to dynasty. In a sense, *Guanhua* is the *Tongyu* in Ming dynasty, but has been given another name.

general are influenced considerably by Manchurian, an East Asian Tungusic language. Beijing Mandarin remained the 'unofficial' standard language until 1956, when the Law of the People's Republic of China on the Standard Spoken and Written Chinese Language (Order of the President No.37) was published:

> Pǔtōnghuà is the standard form of Modern Chinese with the Beijing phonological system as its norm of pronunciation, and Northern dialects as its base dialect, and looking to exemplary modern works in báihuà 'vernacular literary language' for its grammatical norms. (Law of the People's Republic of China on the Standard Spoken and Written Chinese Language, 2000)

According to this definition, although the phonological system of Standard Mandarin is based on Beijing Mandarin, the colloquial words of Beijing Mandarin are excluded from the vocabulary of Standard Mandarin (e.g., 局气 /tɕy2 tɕʰi0/[5], 'generous and frank'). In fact, some words which hold prestigious statuses in Standard Mandarin are from Nanjing Mandarin. For instance, 老鼠 /lau3 ʂu3/, the colloquial word in Nanjing Mandarin referring to rats and mice, is considered more formal than 耗子 /xau4 tzi0/, the colloquial word in Beijing Mandarin of the same meaning among Mandarin speakers.

In addition, the phonetic samples illustrating the pronunciation of Standard Mandarin were not recorded in the city of Beijing, but in Chengde (Luanping Country, Hebei Province), a city about 160 kilometers from the city of Beijing[6]. In other words, the 'Received Pronunciation' of

[5] All the transcriptions in the thesis use International Phonetic Alphabet (IPA) unless otherwise specified.

[6] The population in Luanping consists mainly of the former immigrants from Nanjing as Emperor Yongle moved the capital city from Nanjing to Beijing. In the *Qing* Dynasty, Chengde City had become an imperial summer resort for Manchurian aristocrats. Therefore, in terms of the dialect development, Luanping accent has a very similar history to Beijing accent spoken in the city of Beijing, and Luanping belongs to the Beijing Mandarin

Standard Mandarin was established as the Luanping accent. Luanping accent was chosen because its tone realization is more straightforward than the Beijing accent. Furthermore, Luanping accent does not involve as many [ɚ] suffixes or syllable consolidation in comparison to the colloquial Beijing Dialect (Secretariat of the Academic Conference on Modern Chinese Normals, 1955). In a word, it is worth to keep in mind that from the very beginning, Standard Mandarin is different from Beijing Mandarin spoken on the city's streets, and both Mandarin dialects are constantly evolving with time.

After establishing Standard Mandarin as the official language, a national promotion of Standard Mandarin began. Since the Chinese writing system is a logographic system showing no clear cues of the pronunciation, an official romanization system *Hanyu Pinyin* (Chinese spelling alphabet) had been developed. *Hanyu Pinyin*, also known as *Pinyin*, is often used to teach and type Standard Mandarin. However, due to the large discrepancies in Chinese dialectal pronunciations, *Pinyin* is not a convenient tool for annotating the pronunciation of other non-Mandarin and Mandarin dialects. The promotion of Standard Mandarin and the national movement against illiteracy starting from 1950s have accelerated each other. As a result, the educated population in their 20-40s in China now are all Mandarin speakers, though their Mandarin may be influenced by the local accents. Since 1994, as yearly examination, *Putnghua Shuiping Ceshi* 'National proficiency Test of Standard Mandarin' has been put forward to test and certificate the proficiency of speaking Mandarin used by individuals. For convenience, in the following discussion, Mandarin refers to Standard Mandarin only unless otherwise specified.

It is worth clarifying that *Pinyin* has remained a pronunciation annotation system and never

subarea.

replaced the logographic writing system in China. A special feature of Chinese is the one-to-one correspondence between syllables, morphemes and characters. To be specific, a monosyllabic morpheme in Chinese is still written as a graphic unit, *zi* (字) 'logographic character'. A logographic character usually has a standard shape[7] and meaning(s) that can be recognized by all literate Chinese users regardless of his or her own dialectal background, but the pronunciation varies between dialectal areas.

According to historical materials, the majority of ancient Chinese words are monosyllabic words. The disyllabificiation of Chinese only started in mid-ancient time (c.600 A.D.) and it took considerable time for the disyllabic words to replace the mono-morphemic words and become the majority (for detailed introduction, see Huang & Yang, 1990). In other words, for a long period in the history, there were more monosyllabic words in formal Chinese than disyllabic or polysyllabic words so that Chinese does not annotate the boundaries between words like other languages may do. Words in China are often referred to as *zici* (字词) 'character and words' or *cizu* (词组) 'words and compounds', indicating great ambiguity in the boundaries between morphemes, words, compounds and sometimes phrases in Chinese (for a review, see Duanmu, 2007: Chapter 5). This point will be further pursued in Chapter 3 when the morphological analyses of NT words are introduced. In the present Chapter, 'words' are used as a cover term for words, compounds and pseudo compounds for convenience.

[7] This is not to say that the logographic characters have not changed over time. The most recent change is the simplification of the traditional characters launched in 1956 and mainland China now uses simplified characters as the official writing characters.

2.2 Mandarin Tones

2.2.1 Phonetic Correlates of Tones

Tone is a prominent phonological entity in all tone languages. The primary acoustic correlate of tones is the fundamental frequency (f_0), determined by the vibration of vocal cords. The perceived f_0 is referred to as pitch. Although f_0 and pitch are sometimes used interchangeably in the literature, in the present thesis, f_0 is used to describe the physical changes in tones while pitch to describe the perceptual changes.

From the articulatory perspective, Halle and Stevens (1971) proposed that tone articulation was mainly controlled by the tension of the vocal cords which has three statuses, stiff, slack and not stiff, or slack, each corresponding to three tone levels, high, mid and low. Zemlin (1981) identified two mechanisms that controlled the status of the vocal cords, the cricothyroid and the vocalis muscles; the former controlled the thickness of the cords and the latter controlled the tension of the cords, though the possible combinations of the status and the corresponding tones were not illustrated in detail. The vocalis mechanism is worth a bit more attention as it also influences another important correlate of tone, voice quality. Voice quality, known as 'murmur', 'creakiness' or 'breathiness', interacts with f_0 closely in various tone languages. According to the cross-linguistic data, it is often the murmured, creaky tones that have low f_0s (Keating & Garellek, 2015). Duanmu (2007) cited these works and further related the two mechanisms to the phonological features of tones. I will return to this point when I discuss the phonological representations of tones.

In Standard Mandarin, voice quality does not play as a contrastive role as in some other tone languages, but T3, the low convex tone, is associated with creaky voice in most cases. In addition

to voice quality, other linguistic and paralinguistic factors like consonant voicing, consonant aspiration, vowel height, prosodic prominence, phrasal position and speech rate all influence the surface f_0 patterns. It has been found that aspirated obstruents and high vowels tend to raise the f_0 while voiced obstruents and low vowels tend to lower it (for examples in Chinese dialects, see Yue-Hashimoto, 1984; for an overview of consonant-tone interaction, see Bradshaw, 1999); phrase-final tones are often lower than their initial counterparts in neutral utterance (i.e., downdrift). In fact, the voicing of the initial consonants, an important laryngeal distinction, is believed to have motivated tone genesis. It will be demonstrated in Section 2.2.3 that Middle Chinese tones and the tone changes afterwards were largely influenced by the existence and loss of the voiced initials and coda consonants.

2.2.2 Citation Tones and Tone Sandhi

Mandarin is a typical tone language. Each syllable in Mandarin bears either a citation tone (CT) or a neutral tone (NT). In this background section, I will focus on information about the four CTs (**Table 1.1.a**). The studies on NT will be reviewed in a separate chapter, Chapter 3 with more details.

When enunciated in isolation, the f_0 contours of the four CTs can be sketched as **Figure 2.2.a**. In real-time speech, however, great contextual variations have been observed.

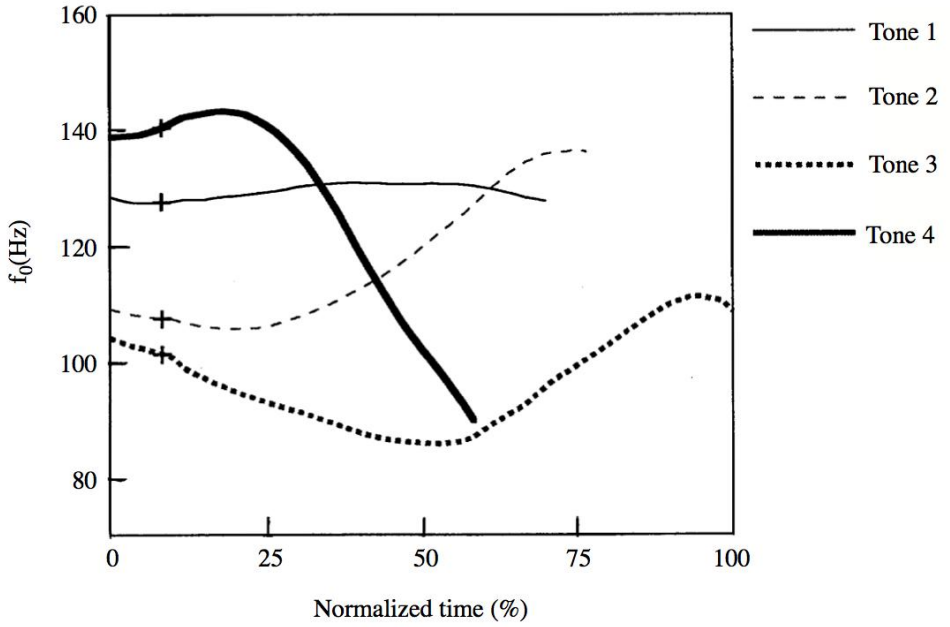


Figure 2.2.a Mean f_0 contours of four Mandarin tones in the monosyllable /ma/ produced in isolation. The time is normalized, with all tones plotted with their average duration proportional to the average duration of Tone 3 (Xu, 1997: Figure 2)

Since it is impossible to jump from a tonal target to another without a transition in articulation, the realization of tones shows carry-over effects. The onset value of a tone is usually assimilated to the offset of the preceding tone, and the whole transition can take up as much as one-third of the duration of the following syllable in Mandarin (Xu, 1994, 1997). In addition, a low onset value will also raise the height of the preceding tone (i.e., the anticipatory dissimilation). The two co-articulating effects in Mandarin are shown in **Figure 2.2.b**, and both effects are found in other tone and even non-tone languages (for Cantonese, see Li and Lee, 2002; for Thai, see Gandour et al., 1994; for Mizo, see Sarmah, 2015; for the Post-L raise in English, see Pierrehumbert, 1980).

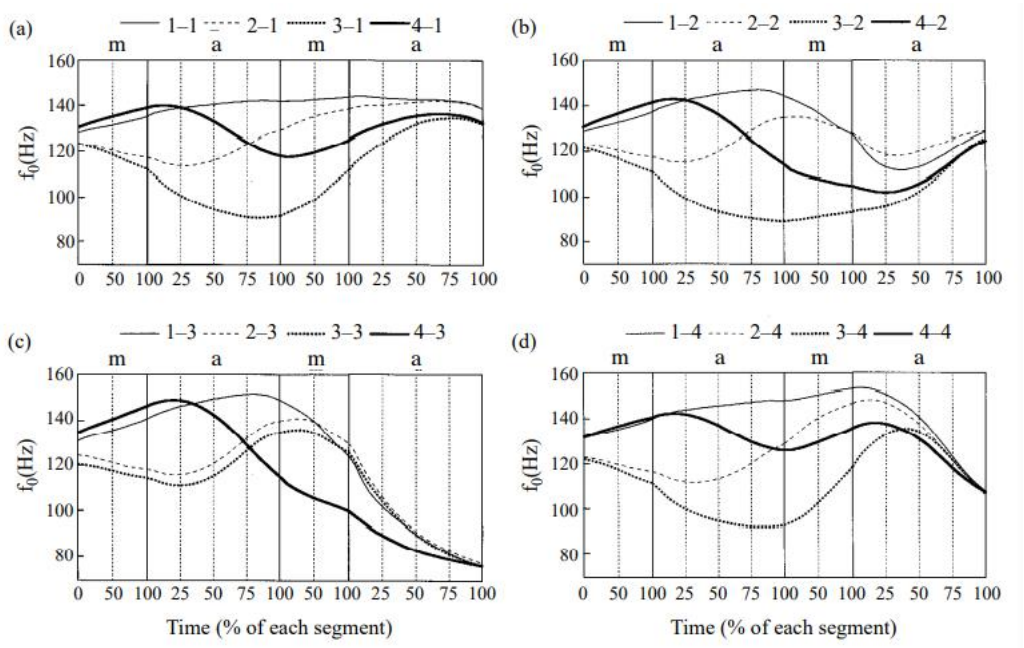


Figure 2.2.b Effects of preceding tone on f_0 contour of following tone in /ma ma/ sequences in Mandarin. In each panel, the tone in the second syllable is held constant (Tones 1-4 in (a) to (d) respectively), and the tone of the first syllable is varied (Xu, 1997: Figure3)

Moreover, in continuous speech, Mandarin contour tones T2, T3 and T4 are often not fully realized. To be specific, when T3 precedes another T3, it has a low-falling contour rather than the low convex contour found in isolated or pre-pausal positions. Much like T3, the falling contour of T4 is truncated when preceding other tones, especially when preceding another T4; described in Chao's numerals, the non-final T4 is [53] rather than [51] (Zhang, 1988; Chao, 1965). The 'half-realization' of T2 is slightly more complex. In daily speech, if a T2 in the middle positions of words or compounds is preceded by a T1 or T2, it will become a T1, with a high-level rather than a mid-rising contour (Duanmu, 2007; Zhang, 1988; Hyman, 1975; Chao, 1965).

Compared to T2 and T4, the half-realization of T3 has received the most research attention and it has triggered controversies over the underlying representations of T3. Some phonologists take the low-falling contour as the underlying form and the pre-pausal or final rising in isolation as a variant resulting from the floating high tone (e.g., Duanmu, 1999; Wang, 1997; Yip, 1980). The

others, in contrast, regard low-convex as the underlying form of T3 and low-falling as the contextual variants (e.g., Zhang, 1988; Shen, 1992). I would like to point out that in connected speech, the rising part of T3 has not disappeared. Instead, it may be seen as realized on the following T1, T2 or T4. It can be seen from the dotted lines on Figure 2.2b that there is a clear rising in every tone that follows T3, except in Figure 2.2b (c) where T3 sandhi happens. Although T3 reaches a very low point at the end of the syllable that carries it, it quickly catches up to the similar height as the other three tones in the first half of the following syllables. Therefore, from my point of view, it is more reasonable to treat all these half-realized tones as either phonological or phonetic problems, rather than to treat the half-realization of T3 as a phonological representation issue but the half-realized T2 and T4 as phonetic variations. The boundary between phonetics and phonology would be further discussed in Section 2.3.1.

More systematic tone changes caused by mainly but not only the pronunciation of the adjacent words or morphemes also widely exist in tone languages including Mandarin. This phenomenon is referred to as *Tone Sandhi*, and the most famous tone sandhi in Mandarin is T3 sandhi. When preceding another T3, Mandarin T3 would have a surface form very similar to T2 (i.e., having a rising rather than convex contour). Although this sandhi is often described in phonological literature or language teaching materials as 'T3 becomes T2 before another T3', the phonetic differences between sandhi T3 and authentic T2 have been consistently reported. For instance, Yuan and Chen (2014) found in a corpus study that the sandhi-raised T3 often exhibited a smaller f_0 range and a latter onset of rising than the authentic T2. Such evidence demonstrates, to some extent, the traces of the underlying low T3.

There are also morphological tone sandhi in Mandarin applying to only a few morphemes

including numeral morpheme *yi* ‘one’, negative adverb *bu* ‘not’ and duplicated morphemes in adjectives and adverbs. The numeral morpheme *yi* ‘one’ bears T1 in isolated and pre-pausal positions, T2 when preceding a T4 and T4 elsewhere; *bu* ‘not’ bears T2 before T4 but T4 in the other situations; the final morphemes in adjectives and adverbs, regardless of the CTs they otherwise bear, would bear T1 when duplicated. For instance, the mono-morphemic word *man* ‘slow, slowly’ carries T4, but the tones realized on the bimorphemic adverb *man man* ‘slowly’ are T4 and T1.

It is worth noting that the morphological tone sandhi is disappearing in Standard Mandarin. In addition to *yi*, numeral morphemes *qi* ‘seven’ and *ba* ‘eight’ also carry T2 when preceding T4 but T1 elsewhere. However, these rules are not followed by the native speakers anymore and hence are no longer obligatory in Standard Mandarin. The T1 in duplication morphemes is also disappearing. During the first decade of the 21st century when I received my primary school education, this rule was listed in dictionaries and text books as compulsory. However, although some speakers still follow this rule, it is no longer enforced in the text books or the national proficiency test.

To summarize, the four CTs demonstrate great phonetic and phonological variations when realized in continuous speech. Here, I would like to clarify that the term *Citation Tone* is only employed in the present thesis as a direct contrast to *Neutral Tone* which cannot be in initial or isolated positions. Hence, in the following discussion, the term ‘NT words’ refer to prosodic words that contain at least one NT syllable but also at least one CT syllable in word-initial position rather than words made up entirely of NT-bearing syllables. The term ‘CT words’, instead, refers to words which are entirely made up of CT-bearing syllables. Due to the correspondence between

tones, syllables and morphemes in Mandarin, phonologists tend to take CTs as the underlying tones from which the surface tones derive in tone sandhi. This assumption holds firm in Mandarin, but not in other word tone languages and dialects like Wu dialects.

2.2.3 Tone in History

Most Chinese dialects, except Min, are believed to have the same ancestor, Middle Chinese (c.600 A.D.). This protolanguage had four tonal categories, *Ping, Shang, Qu, Ru*, describing the level, ascending, departing and entering contours of tones respectively (e.g., Mei 1970 and K. Chang 1975). The last three tones are also called *Ze* tones, meaning oblique, as the opposite to *Ping* tone in poetic meter. The height of the tones was conditioned by the voicing status of the initial consonants, i.e., the *Yin* and *Yang* registers. It is generally believed that voiceless initials (*Yin* register) led to high pitch values of tones while voiced initials (*Yang* register) to low pitch values. The eight tones in Middle Chinese are summarized in the following table.

Table 2.2.a Eight tones in Middle Chinese

Tone Categories	Estimated Contour	Register	Tone
Ping	Long, level	Yin (Voiceless initial consonants)	Yinping
		Yang (Voiced initial consonants)	Yanping
Shang	Relatively short, level or rising	Yin (Voiceless initial consonants)	Yinshang
		Yang (Voiced initial consonants)	Yangshang
Qu	Slightly drawn and therefore falling	Yin (Voiceless initial consonants)	Yinqu
		Yang (Voiced initial consonants)	Yangqu
Ru	Short, checked, realized on syllables ending with voiceless stops	Yin (Voiceless initial consonants)	Yinru
		Yang (Voiced initial consonants)	Yangtu

The loss of the voiced obstruents led to the first great tone change in the history of Chinese

and the loss of *Ru* tone (i.e., checked tone) in the northern dialects led to the second. The different derivations of the dialects in each tone change have resulted in the dialect differences nowadays. For instance, Cantonese dialects are believed to have kept intact the lexical contrasts that used to be signaled by *Yin* and *Yang* registers so that they have a much larger tonal inventory than Mandarin dialects. It is worth mentioning that both tone changes are believed to have taken place during the regimes of ethnic minorities speaking Altaic languages[8] (L. Wang, 1980; Chang 1975). Therefore, language contact may have had great influence on the development of tone, which implies that it should be taken into consideration when exploring the phenomenon of NT (Section 3.3.2 and Section 8.4).

Most diachronic phonological information on Chinese, including the historical development of syllable structure and tonal and segmental inventories, was recorded in the official and unofficial rhyming books compiled by scholars in different dynasties. The pronunciation annotation method employed in these books is *fanqie*. Two different Chinese characters are used to annotate the pronunciation of the given character, the first character having the same consonant as the given character and the second having the same rhyme and tone. *Fanqie* remained the only annotation method for hundreds of years in China, the other annotation systems including the prototype of *Pinyin* not being developed until the early 20th century.

Although the original aim of these books was to guide poetry creation, the meticulous classification of syllables and segments in these books have made them the most comprehensive material for reconstructing the ancient Chinese pronunciation system. The 'products' of the

[8] The notion that there is an Altaic language family has been rejected by many linguists (e.g., Normal, 2009). Instead, Altaic language family is divided into Turkic languages, Mongolian languages and Manchurian-Tungus languages. Here, Altaic language is used as a cover name rather than a rigorous typological term.

rhyming books, namely, poetic works including poems and lyrics of songs or theatres, are also important resources, in particular for tracing the lexicalization of NT throughout its history (see Section 3.3.2).

2.3 Representations of Tone

2.3.1 Phonological Representations

In principle, phonetics studies the speech sounds as they are produced and perceived, but phonology deals with the sound system of language, namely, how sounds form speech patterns in a language and how they are combined to convey linguistic meanings. As noted by many, speech sounds in connected speech are influenced by many paralinguistic and non-linguistic factors and hence contain components that vary and are unrelated to the core meanings to be conveyed. Therefore, in the early phonological theories that laid the foundation for the more recent models (e.g., Prague School phonemic theory, Jakobson, 1949; Trubetzkoy, 1939), two levels of representation - phonetic and phonemic (i.e., phonological), were distinguished. Phonetic representations concern the properties of actual speech events (in perception or production) and phonemic representations only involve distinctions between sounds or sound patterns that are contrastive in the language, and hence, form phonologically distinctive categories. Since then, the question of whether we should assume that there are distinct levels of representations, as well as how many should be distinguished, has become a key aspect of phonological theorizing.

As empirical data accumulated, the mapping between the mental activities that generate or encode linguistic meanings and the articulatory activities or acoustic sounds are proved to be rather complex processes. The two-level representations treating phonemes as primitives were not enough to capture this complexity. Therefore, Structuralist Phonology developed in 1950s

introduced another underlying level, the morphophonemic level, to account for morphological alternations (i.e., allomorphs). In this way, the model does not require separate phonemic representations to account for the allomorphs in languages, but can account for them by the mapping between the morphophonemic and phonemic representations (for a review of representations in Structuralist Phonology, see Harris, 1963). As will be shown in Section 3.2, some tonal analyses of NT still follow this three-level model.

In contrast, Generative Phonology, the most widely adopted framework in the field, which was originally developed in the 1960s, argues against the phonemic level between the morphophonemic and the phonetics levels (Chomsky and Halle, 1968). Instead, it proposes an underlying representation as a combination of morphophonemic and phonemic representation of the structuralist approach. At the underlying level of the generative framework, every non-suppletive morpheme has its own phonological representation, and the phonological grammar serves as a function that maps the underlying representations to the surface forms. The mapping process is referred to as phonological derivation, guided by universal and language-specific phonological rules. It should be noted that during the derivation the rules apply in grammatically imposed orders, which implies that the generative approach in fact does allow for multiple levels. However, since special roles or properties are associated with these in-between levels, these levels are not grammatically designated levels the underlying or the surface representations.

Although traditional generative phonologists lend no formal status to morphology in the grammar, Lexical Phonology (Pulleyblank, 1986; Kiparsky, 1985), a theory following the sequential rule-based derivation proposed in the generative framework, emphasizes the importance of morphophonology in order to capture the interactions between morphology and

phonology in word building processes. In Lexical Phonology, derivation is divided into lexical and post-lexical modules (i.e., strata). The output of the post-lexical stratum corresponds to the surface representation in Generative Phonology, but the output of the lexical stratum, the lexical representation, also has special phonological properties.

Unlike the models discussed so far which try to distinguish between the abstract, (usually) categorical phonological domain and the varying, (usually) gradient phonetic domain, a different view gained traction with the development of corpus and psycholinguistic research: all sounds are part of a single system or processing mechanism. A typical example is Exemplar Phonology (e.g., Johnson, 2007; Beckman & Pierrehumbert, 2004; Pierrehumbert, 2001). Exemplar phonologists propose that representations are abstracted directly from the phonetic details a language user has experienced and therefore are sensitive to word frequency. Correspondingly, there is no need to distinguish levels of representation but only a patchwork that links all information together (e.g., Cole, 2009; Johnson, 2007; Pierrehumbert, 2001). This theory offers a better explanation of the phonetic effects on phonology such as the word frequency effect on sound changes, which are also found in the development processing of Mandarin NT.

Most phonological analyses of lexical tone in the existing literature follow the more traditional models and recognize at least a continuous phonology-phonetics line. Therefore, it is often up to the phonologists to decide whether and how certain phenomena should be accounted for from the phonological perspective or as a phonetic implementation of phonological representations. In Mandarin, for example, it is relatively uncontroversial that the T3 sandhi and the morphological tone sandhi are phonological changes, but the grammatical status of the half-realization of the contour tones is less clear (Section 2.2.2). As will be shown later, Mandarin

NT also lies in the ambiguous area. How much surface variation of NT should be accounted for in phonology has caused long-lasting debates and motivated different analyses (Sections 3.2).

Meanwhile, in modern phonological theories, phonemes are no longer seen as the most basic unit in phonological analyses, but they are decomposed into distinctive features. Following the Prague School, Jakobson, 1949; Trubetzkoy, 1939). 21 distinctive binary features were first proposed by Jakobson, Fant and Halle in 1951, attempting to capture all the phonological systems of natural languages. In this initial proposal, vowels and consonants are classified with the same features. Later, Chomsky and Halle (1968) established distinctive place features for vowels and consonants by focusing on the articulatory than the acoustic and perceptual aspects[9]. The phonological studies afterwards started to group phonemes with the same features into natural class and investigate phonetic and phonological phenomenon according to natural classes rather than to individual sounds indiscriminately. The decomposition proposal also influences the phonological analyses of tones. Before the 1950s, most of the tonal analyses like the Chao's numerals see tone as an indivisible entity like phonemes in Prague School phonemic theory, but from the 1960s, tone phonologists like SY.Wang (1967), Woo (1969), Yip (1980) and Bao (1999) also started to represent lexical tones with distinctive features. Wang (1967) and Woo (1969) proposed features for tone height and tone contours, but the later models involve only features for tone height, and also focused more on the articulatory perspective, much like the development in the feature systems for segments.

[9] It is worth attention, however, that it is possible to locate acoustic cues to certain segments, but not to certain distinctive features (e.g., Lahiri, Gewirth, & Blumstein, 1984; Blumstein & Stevens, 1980).

2.3.2 Representations of Mandarin CTs

In the present thesis, I will follow the autosegmental framework and the tonal feature system developed by Yip (1980), which are widely adopted in the tone phonology literature on Mandarin. The phonological representations of Mandarin CTs and the related theoretical background are reviewed in the present section to provide the background for the phonological analyses of NT reviewed in Chapter 3.

Autosegmental Phonology was developed by Goldsmith (1976) as a framework to represent tone and other suprasegmental phenomena, namely, the phenomena that extend beyond individual consonants and vowels. A segment in this framework is defined as "an indivisible unit, ultimately a mental unit of organization" (Goldsmith, 1990:10) and the autosegmental framework posits "two or more parallel tiers of phonological segments. Each tier itself consists of a string of segments, but the segments on each tier differ with regard to what features are specified in them." (Goldsmith, 1990:8).

In tone languages, tones are represented on a tonal tier where each segment is specified for tonal features and associated with the segments on the other tiers through universal and language-specific conventions and conditions (e.g., Association Convention and Well-Formedness Condition). This framework provides effective explanations for the special properties found in tones like tone stability (i.e., tones survive after the loss of the hosting segments) and mobility (i.e., tone spreading, tones move away from the hosting segments). Hence, it has been widely adopted in the phonological literature on tones. Although the autosegmental framework was initially developed based on data from African tone languages, tone phonologists soon started to

successfully apply the framework to analyses of tonal phenomena in Chinese dialects including Standard Mandarin (e.g., Yip, 2002, 1980; Bao, 1999; Wang, 1997).

Another important improvement autosegmental phonology brought to the phonology of tone is the strutural representations of phonological features (i.e., feature geometry) it proposes, which helps to reduce the complexity of the feature system of tones.

In the past, much effort has been put into the development of a feature system that can account for the large tone inventories, tone sandhi and assimilation phenomena found across the world. In particular, the system needs to distinguish between four or more pitch levels[10] and to characterize contour tones as well as the pitch height contrasts between contour tones of the same shape (e.g., *Yinping* vs. *Yangping* in Middle Chinese). Systems predating the autosegmental framework required three features to distinguish between the four or more pitch levels because the combination of [+high] and [+low] is meaningless. Drawing on the feature geometry proposed in autosegmental phonology, Yip (1980) employed two tonal features with internal structures *Register* [±Upper] and *Tone* [±high] to represent tones. The *Register* feature phonologically divides the pitch range into two halves and the *Tone* feature divides each register further into two levels; in this way, four tone levels are distinguished with only two features. Later studies have made refinements to this model, but the key assumption of a structural relation between the two features, *Register* and *Tone*, has never been abandoned.

An important improvement is to relate the two tonal features to the two articulatory mechanisms of tones (Section 2.2.1). According to Duanmu, the two contrasts in *Register* are

[10] Most tone languages have 2-4 distinctive level tones; languages with five level tones have been reported but the empirical evidence is inconclusive (Yip, 2002).

"stiff (non-murmured) and slack vocal cords (murmured)" and the two contrasts in *Pitch* (i.e *Tone* in Yip's model) are "thin vocal cords (high pitch, or H) and thick vocal cords (low pitch, or L)" (Duanmu, 2007: 231). Interestingly, Duanmu (1990, 2007) and Bao (1990, 1999) suggested exactly the opposite relations as Bao (1999) and related *Register* to vocal cord thickness and *Contour* (i.e *Tone* in Yip's model) to vocal cord status[11]. Either way, *Register* does not correspond to pitch height strictly as in Yip's model, and hence allows the overlapping between contour tones. This improvement generates nice results especially regarding Mandarin data. For instance, the falling tone T4 is often considered as specified with [+Upper] for *Register* while the low convex tone T3 with [-Upper] (Figure 2.1.2). According to Yip's original model, T4 should be higher than T3 because it has a higher *Register*, but acoustically the falling end of T4 is often lower than the rising end of T3 in isolation - as indicated in Chao's numerals, T4 is [51] while T3 is [214]. Yip also adopted this more abstractly defined *Register* feature in her later work (e.g., Yip, 2002).

The structural relation between the two features is also frequently discussed. Two relations have been suggested, *Register* and *Tone* are sisters (e.g., Duanmu, 2007; Bao, 1999; Yip, 2002), or *Register* dominates *Tone* (e.g., Wang, 1997; Yip, 1980). Nevertheless, if the articulator model proposed by Duanmu or Bao is followed, the two features can only be in sister relation because strictly speaking, features can only be dominated by articulators (e.g., laryngeal node) rather than other features (Duanmu, 2007; Yip, 1989; Sagey, 1986).

In the feature system proposed by Yip (1980, 2002), no contour features like [±Falling] are proposed. Instead, contour tones are represented as sequences of level tones. Moreover, due to the

[11] As mentioned in Section 2.2.1, vocal cord status also influences the voice quality. However vocal cord status does not equal to voice quality. The production of tones is a complex process and even in Duanmu's or Bao's models, the two articulatory features cannot fully cover the complex articulatory movements involved in f_0 adjustment.

autosegmental nature of Yip's model, the values of *Register* can remain constant when the values of *Tone* change, enabling complex tonal phenomena like the reduplication of contour tones found in Chinese Jin dialects. The underlying representations of the four CTs in Mandarin in the existing literature can be summarized as Figure 2.3.2.a. Henceforth, *Pitch* is used instead of Tone to avoid confusion.

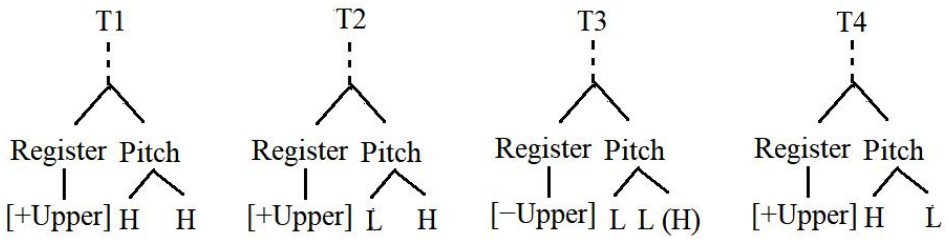


Figure 2.3.a Representations of Mandarin CTs[12]

Studies differ in the structural relations they adopt (i.e., having *Pitch* dominated by *Register* like in Wang, 1997 others to be added) and the underlying representation of T3 (i.e., having T3 represented as [LL] rather than [LLH]), but not in other aspects of the geometries.

It is worth mentioning that the decomposition of contour tones is mainly supported by data from Africa. In African tone languages, consolidations of tonal contours (e.g., [H] and [L] realized as [HL] when there is a loss of syllable in Hausa) and restrictions on the positions of contour tones (e.g., contour tones only occur in the final positions as phonetic interpolation between two opposite tones in Benue-Congo languages) are found. In Mandarin, however, no compelling evidence has been found to support or reject this assumption. Although representations like Figure 2.3.3 are constructed following the prevailing conventions in the tone phonology literature, the

[12] As discussed, there is not yet a general consensus on whether T3 is represented as a low tone with LL or a low-rising tone with LLH

possibility that contour tones in Mandarin are primitives has never been completely ruled out. The decomposition of contour tones will be further discussed in Section 3.2.2 when the association between metrical units and tonal targets is introduced.

2.3.3 Underspecification

Underspecification is a crucial aspect of any traditional feature system. In particular, the underspecification and value filling rules are of special importance to the phonological analysis of toneless elements (Pulleyblank, 1986) and therefore, to NT in Mandarin (Wang, 1997). Another Mandarin tone described as 'underspecified' is Mandarin T3, but its undersepcification has caused more controversies than the undersepcification of NT. In the present section, I will mainly review the theoretical background and the controversies over the underspecification of T3. The underspecification of NT will be reviewed in Section 3.2 in the following chapter, and revisited with the experimental results.

The underspecification theory was proposed alongside generative grammar, although the notion that phonological representations may be underspecified in some way predates this (see Cole & Hualde, 2011). In the generative framework, an evaluation metric is required to facilitate the selection of the grammar for a particular language (Chomsky & Halle, 1968). According to Chomsky and Halle, such a metric prefers the grammar that lists only the idiosyncratic properties and has other predicable ones derived. The core of underspecification theory in phonology, therefore, is to differentiate which features are marked and which are predicable by rules (Archangeli, 1988; Keating 1988). Different sub-theories have been proposed: Inherent Underspecification, Contrastive Underspecification, and Radical Underspecification .

Inherent Underspecification refers to underspecification due to the properties of the features themselves. Inherent Underspecification theories fall into two schools, Inherent Monovalent Underspecification and Node-dependent Underspecification (Archangeli, 1988). The former underspecification depends on the binary or unary nature of the feature. A unary feature inherently allows underspecification. For instance, some phonologists take [voice] as unary so that voiceless segments are unspecified for [voice] rather than specified with [-voice]. The latter underspecification is developed based on the feature hierarchy (Sagey, 1986); the segments unspecified for a node feature, e.g., [Coronal], will also be unspecified for features under it, e.g., [anterior] and [distributed].

Contrastive Underspecification and Radical Underspecification can be seen as two different complementary theories to Inherent Underspecification Theory. Since not all underspecification phenomena are inherent, binary representations are still needed for some features. These two theories are developed to decide which value of a binary feature is underspecified. Contrastive Underspecification only assigns specific values to the features that distinguish segments (Halle, 1959; Clement, 1987) while Radical Underspecification assigns specific values to the features that cannot acquire 'predictable rules' by context-free or context-dependent rules (Archangeli and Pulleyblank, 1993; Kiparsky, 1982). The FUL (Featurally Underspecified Lexicon) model proposed by Lahiri and her colleagues (see Lahiri & Reetz, 2008 for an overview) is also in line with Contrastive Underspecification as it assumes that "a segment is lexically represented by sufficient features to separate it from any other segments in the phonology of a particular language" (Lahiri and Reetz, 2008:637). This model is supported by evidence ranging from language change to language processing, offering an efficient explanation of how contextual and

individual variations in speech production and recognition are dealt with (e.g., Kotzor et al, 2020; Friedrich et al, 2006; Eulitz & Lahiri, 2004).

According to this model, speech recognition involves a ternary logic of match (i.e., incoming speech is the same as the mental representations with specified features), mismatch (i.e., incoming speech is different from the mental representations with specified features), and no-mismatch (i.e., there are underspecified features in the mental representations so that incoming speech is neither different from nor the same as the mental representations). As a result, the phonetically similar acoustic phenomena may still be processed in different manners (for a review, see Lahiri and Reetz, 2008, 2010). Take the underspecified feature [Coronal] as an example. Friedrich et al (2008) find in an electroencephalographic study that that pseudo words with [-coronal] consonants (e.g., *Brachen) activate German words with initial coronal place (e.g., Drachen [dragon]) but pseudowords with [+coronal] (e.g., *Drenze) do not as effectively activate non-coronal words (e.g., Grenze [border]). Interpreted by FUL model, the coronal feature is underspecified so that it cannot activate other words, either coronal or non-coronal. In a passive oddball paradigm using German vowel pairs [o] ([+Coronal, +Labial]) and [ø] ([+Dorsal, +Labial]), earlier and larger mismatch-negativity (MMN) is found when [ø] is the deviant than when [o] is the deviant. as the former creates a clear mismatch condition while the latter is a non-mismatch condition (Eulitz and Lahiri, 2004).

In terms of underspecification in tonal features, contour tones ask for specific distinction in voice quality and pitch height so that binary tone features are required if linguists try to keep the universality of the system (Yip, 2002). Therefore, the only option is to decide which value of a tonal feature is specified universally and the opposite values are filled by default rule-filling rules.

Based on African data, Pulleyblank (1986) proposed that only [+Upper] and [L] are specified for *Register* and *Pitch* and the other two values are filled with default rules. Wang (1997) adopted this model in her analysis of Mandarin and refined the underlying representation of four CTs in Mandarin as follows:

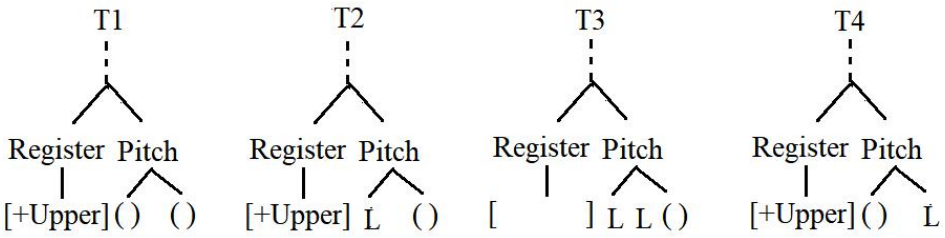


Figure 2.3.b Underspecified Representations of Mandarin CTs with underspecified features (adapted from Wang, 1997, though Register dominates Pitch in Wang's model)

Again, not much positive or negative evidence has been found for or against the underspecification model in Mandarin. In fact, the underspecified model does not suggest any more differences between tones than the alternatives unless *Register* is given more priority than *Pitch*. If so, T3 would be more underspecified than the other CTs. The acoustic data furnished in past studies offer no ground to further test these differences. Processing studies conducted more recently provide more but somehow controversial findings.

T3 has been found to perform distinctively in several electroencephalographic studies, and such special performances are interpreted as evidence for its underspecified representations. Others, however, argue that the non-underspecified representation is of higher processing economy based on results of priming experiments. In Event-related potential (ERP) studies using oddball paradigms[13] of single syllables, it is widely found that T3 tends to weaken the contrast

[13] Oddball paradigm is an experimental design widely used in psychological and psycholinguistic research. It

between the standards and the deviants (i.e., causing smaller magnitude of mismatch negativity, MMN), much like the vowel [o] in Eulitz and Lahiri's study (2004). Such asymmetry was not only found in paradigms with T2 and T3, the two tones that correlate with each other in T3 sandhi but also in paradigms with T3 and the other CTs (Politzer-Ahles et al., 2016; Li and Chen, 2015). In an oddball paradigm of disyllabic words, single T3 did not cause an MMN when the standards were disyllabic words with T3 sandhi or a mix of T3 and T3 sandhi words. More importantly, the omission MMN caused by single T3 syllables in a disyllabic T2 word condition is significantly larger than in the other conditions, suggesting that there must be more mismatch in the disyllabic T2 word condition other than the number of syllables (Chien et al, 2020).

However, in priming studies conducted by Meng et al (2021), T2 did not prime the target with T3 onset in non-sandhi contexts. For either T2 or T3 to activate lexical entries, appropriate contextual information is required, that is, the input T2 will only activate the underlying representation of T3 in the context of T3 sandhi, i.e., before another T3. If T3 is more undersepcified than T2, it should be primed by T2. Therefore, Meng and others argue that T3 is not an underspecified tone; instead, the observed differences are more likely to be due to phonological inferencing "where auditory inputs are analyzed in the context of phonological rules or constraints that specify the phonological contexts in which the alternation can take place" (Gaskell & MarslenWilson, 1998, cited in Meng et al, 2021:17). Also, Meng believes that the non-undersepcified proposal is more economical in terms of lexical access if words starting with T3 are activated by T2 unless followed by T3. It is unreasonable as well for the most complex CT,

presents the participants with sequences of repetitive stimuli are infrequently interrupted by a deviant stimulus. The deviant stimulus is the so-called 'oddball', and researchers are often interested in the participants' reaction to these oddballs.

T3, to be underspecified in Mandarin, especially if the weak NT is regarded as underspecified.

2.3.4 Tone-bearing Units

The units with which tonal features are associated are referred to as tone bearing units (TBUs). Inspired by structuralist and generative phonology, traditional autosegmentalists also tend to regard tones as the underlying properties of morphemes and hence to regard morphemes as the universal TBUs in the underlying representation. Nevertheless, as the cross-linguistic data accumulated, it was found what consists of a TBU in the underlying and surface representations may vary from language to language.

There is also a small number of analyses challenging the morpheme tone proposal in Mandarin due to the existence of mono-morphemic and multi-syllabic words (e.g., Zhang, 1988). The examples given by Zhang (1988) are either native disyllabic bi-morphemic words where the one or both morphemes lost their original meaning in history or multi-syllabic foreign words in the transliteration like 蚯蚓 or 不列颠. 蚯蚓 originally was written as 丘蚓 in which 丘 'hill' refers to the small mound of soil the earth worms make and 蚓 'worm' refers to the worm itself. However, both morphemes have more or less lost their own semantic meanings in the history and 丘蚓 become one mono-morphemic word; the fact that the radical 虫 'insects and worms' is added to 丘 shows that 丘 no longer holds its original meaning nor stands as a modifying morpheme (Section 3.3.1 gives more examples of such mono-morphemic disyllabic words in Mandarin). Britain, '不列颠 /pu2 liɛ4 tian1/', is made up of 不 /pu2/ 'no, not', 列 /liɛ4/ 'list', 颠 /tian1/ 'bump', but none of the three characters carry their semantic meanings in this transliterated word. These words, in a sense, resemble the words in word tone languages like Chinese Wu dialects, Lhasa Tibetan or Tamang (a language spoken in Nepal), where it is

unnecessary or even impossible to decompose the pitch pattern assigned to a lexical item into morphemic tones; the tone a mono-morphemic word carries (i.e., a citation tone) is not the underlying tone from which the complex word tones derived from. There even have been some sociolinguists (e.g., Gao, 1980) suggesting that the non-initial characters of such transliterated words in Mandarin should be realized with NT, but this norm is not followed now by Mandarin speakers.

I do not intend to deny the existence of such words or see them as exceptions, but I also would like to point out that these words are not that different from the other disyllabic words in Chinese. Firstly, words like 蚯蚓 are resulted from the shift of semantic emphasis rather than concentration of tones like in some African tone languages, that is, two static tonal targets realized on the same morpheme and result into a contour tone in the surface form. In addition, the transliteration of names or places is not random but also take into consideration the meaning of the characters. For instance, female names with syllable /nɪ/ are conventionally translated with the character 妮, 'girl' ; characters with negative meanings are avoided as much as possible. Thus, I would not pay further attention to the relatively subtle differences between syllable tone or morpheme tone in the underlying representation of Mandarin, and regard Mandarin as a morpheme tone language.

On the surface level, syllables, rhymes and mora have been argued to be TBUs in Mandarin. In most languages, a syllable can be divided into an onset (made up of consonants), a nucleus (made up of vowels), and a coda (made up of consonants). So are Mandarin syllables. The nucleus and the coda together are often referred to as rhyme, which plays an important role in lexical prosody. Mandarin has a relatively simple syllablic structure, $C^{G}VC$ (C: Consonant, G: glide, V:

Vowel or Diphthong)[14], but where the rhyme starts causes debates in the theoretical literature. The phonological status of the glide is debatable. In some literature, the syllable structure in Mandarin is presented as CGVC, which consider the middle glide as the start of rhyme rather than part of the onset consonant. This argument hugely reduces the size of the consonant inventory in Mandarin. However, from the ancient phonological analysis method, *fanqie,* to the modern phonetic studies, the glide G has been defined and proved to be structurally closer to the consonant C (for a more detailed review, see Duanmu, 2007: Chapter 4). Nevertheless, from ether point of view, the syllable proposal (i.e. syllables are TBUs in Mandarin) overlaps largely with the rhyme proposal (i.e. rhymes are TBUs in Mandarin). Here, I would like to emphasize that the rhyme proposal is better supported acoustically as the onset consonants do not have regular pitch patterns, and the pitch they have may results from carry-over effects in some cases. The morae proposal will be discussed in Sections 2.4 and 3.2.2 due to its close relation with stress organization and NT.

2.4 Mandarin Stress

2.4.1 Metrical Phonology and Stress

Metrical phonology is concerned with stress, which by nature is a study of relative prominence of lower- and higher-level units. The relative prominence requires at least two beats to manifest so that rhythm, the alternating prominence, is of great importance to the formation of stress patterns. In traditional models, rhymes rather than syllables are taken as the phonological materials relevant to stress and the assignment of stress is believed to be sensitive to syllable weight, manifested in the number of morae, which is the segmental slots in the rhymes (Goldsmith, 1990; Duanmu, 2007). To be specific, if a syllable is heavy, that is, being bimoraic with long

[14] The initial and coda consonants and the glides are all optional; the coda, i.e. final consonant can only be nasal consonants.

vowels, diphthongs or a short vowel plus a consonant (i.e., CV: , CVV or CVC), it is stressed (Weight-to-Stress Principle). Nevertheless, even within typical stress languages like English, there are exceptions violating this principle so that it requires refinements like assuming extrametricality for syllables at boundary positions (Goldsmith, 1990).

The basic unit in which an alternation between the strong and weak is completed is referred to as the metrical foot. Correspondingly, an ideal metrical foot should be made up of two beats, which are usually two syllables (rhymes) in stress languages (Foot Binarity Requirement). If the strong beat occurs on the left of a foot, the foot is believed to be left-headed (trochaic) and if on the right, the foot is believed to be right-headed (iambic). Not all linguists agree that a language can have both trochaic and iambic feet in the stress system, but it is generally agreed that stress-clash (e.g., an iambic foot followed by a trochaic foot with two stressed syllables occurring next to each other) should be avoided. Nevertheless, two other types of feet must be allowed to account for real cases: degenerate (unary) feet with only one syllable but carrying stress and unbounded (super) feet with one stressed syllable and two adjacent unstressed syllables (Goldsmith, 1990; Halle et al, 1994).

In typical stress languages like English, the organization of feet (e.g., being trochaic or iambic; being weight sensitive or not) has laid the foundation to distinguish stress on higher levels, e.g., expressing the primary stress, namely, the relative prominence within a word; the implementation of utterance focus (i.e., sentence stress) largely relies on stressed syllables while languages with no clear foot structures may have all the focused components lengthened (Section 6.1). Therefore, the expression of stress is cumulative in stress languages, but not so universally.

2.4.2 Metrical Structure in Mandarin

Stress is notoriously difficult to distinguish in Mandarin except when NT is present. This is not to say that Mandarin cannot assign prominence to certain parts of an utterance for pragmatic purposes (e.g., for correction), but the relative prominence of CT-bearing syllables in Mandarin words, especially in disyllabic words, is hardly distinguishable. As pointed out by several phonologists and phoneticians, whether metrical feet are employed as the building blocks to distinguish prominence in Mandarin is so controversial that it almost becomes a matter of belief (Hsieh, 2021).

The controversy is fundamentally rooted in the stress patterns disyllabic words demonstrate in Mandarin. It is widely accepted that there are heavy-heavy patterns and heavy-light patterns[15] in Mandarin, roughly corresponding to CT-CT words and CT-NT words. However, it remains unsettled whether there are more subtly different stress patterns in Mandarin disyllabic words. The most complex proposal distinguishes four patterns: heavy-heavier, heavier-heavy, heavy-light, heavy-lighter (Wang and Feng, 2006). Others proposing three patterns either have the heavy-heavy pattern divided or the heavy-light pattern divided (for overview, see Xu, 2016). These further divisions are frequently challenged because native speakers are typically not aware of them (Cao, 2007). In addition, they are also confounded by factors like speech rate, utterance positions or the tone combinations of the words (Cao, 2007). The correspondence between stress patterns and NT also becomes ambiguous when more than two patterns are distinguished (Section 3.2.2).

Those who refer to metrical feet in the phonological analyses in Mandarin differ from each

[15] Here, 'heavy' does not refers to the weight of the syllables but auditory prominence only.

other in the foot structure they propose for Mandarin. Early studies regard Mandarin feet as iambic, motivated by the final lengthening effects found in most CT-CT words; NT words are either seen as trochaic exceptions or as a tonal exception in these analyses (Zhang, 1988; Yip, 1980; Chao, 1965). This view helps to explain the aforementioned half-realization phenomenon of the contour tones, namely, tones on the foot-initial, weak syllable are not fully realized (Zhang, 1988; Yip, 1980).

However, with the development of instrumental and corpus studies, more recent studies tend to regard Mandarin feet as trochaic. In continuous speech, disyllabic words do not necessarily have longer second syllables than first. Therefore, the existence of NT makes it more reasonable to assume a trochaic rather than iambic foot in Mandarin (Yang, 2008; Duanmu, 2007). In addition, Hsieh (2021) has found through a corpus study that there are more disyllabic words starting with syllables of more phonological prominence, e.g., carrying T1 or T2 or with nasal codas. In fact, even if the two morphemes in the words are of the same meaning, the reversed combination is much less preferred if the initial syllables have more prominent phonological features. For instance, the word 门户 /men2 hu4/ 'door, portal' in which both 门 and 户 mean doors could not be used as 户门 /hu4 men2/. Nevertheless, from my point of view, Hsieh's findings may serve as indirect and rather compelling evidence for the trochaic proposal but they do not end the controversy. To me, a native speaker of Northern Mandarin /men2 hu4/ sounds heavy-heavy if not heavy-heavier. Even if put into compounds like 门户网站 /men2 hu4 waŋ3 ʈʂan4/ 'portal web', the second syllable /hu4/ sounds more prominent than the third syllable /waŋ3/, which is of more metrical prominence according to Hsieh.

In addition, to what extent stress assignment in Mandarin is sensitive to the segmental

structure of a syllable is also controversial. Studies have found acoustic reduction in NT-bearing syllables (Section 3.1), but the difficulty in stress judgement already indicates that there is not a strict correspondence between syllabic complexity and auditory prominence in Mandarin like it is in English. In normal CT-bearing syllables, the rhymes of simpler structure are not shorter than those with more complex structures. For instance, in words like 胰脏 /ji2 zaŋ4/ 'pancreatic', /zaŋ4/ does not sound more prominent than /ji2/ at all. It could be argued that the phonetic realization of the word may in fact be /ji:2 za˜4/. However, in this way, the correlation between syllabic weight in a segmental sense and stress becomes circular - indeed, heavy syllables attract stress, but reduced stress also leads to reduced segments and hence light syllables.

There are two key factors that influence the production and perception of relative prominence in Mandarin words, lexical tone and morphological structure. Although it is widely recognized that duration rather than pitch is recognized as the primary correlate to stress in tone languages, the increase in pitch height or range still influences the realization and perception of prominence in tone languages just like in non-tone languages. In Mandarin, T1 and T2 often sound more prominent than T3 and T4 due to their high endings[16]. To what extent this impact of tones should be taken into consideration in Mandarin metrical phonology invites further thinking. On the one hand, it demonstrates that the lack of auditory prominence does not necessarily mean the absence of foot structure. On the other hand, the metrical analyses that cannot reflect real-time perception are less convincing. In Chapter 4 and 6, the interaction between Mandarin NT words and focus will be explored, shedding further light on this problem.

The one-to-one correspondence between morphemes and syllables also leads to the unclear

[16] This may explain why /hu4/ is more prominent than /waŋ3/ in 门户网站 /men2 hu4 waŋ3 ʈʂan4/ 'portal web'.

stress patterns. In general, the correspondence between syllables and morphemes makes most disyllabic words in Mandarin resemble English compounds like ‘greenhouse’, of which the stress pattern in neutral utterances is also less clear than words like ‘elephant’. The loss of prominence often co-occurs with the loss of meanings or functions of morphemes, though not in every case. As introduced at the end of Section 2.1, Chinese may not be seen as a language with many disyllabic words, but a language with many mono-syllabic words and disyllabic compounds; the heavy-heavy pattern may be seen as two unary degenerate feet. From this point of view, Mandarin is a language concerned with de-stressing rather than stressing. The lexicalization of some NT morphemes is also related to historical morphological change. This point of view will be elaborated at length in Section 3.3.2 and in the final discussion.

To summarize, since word stress is such a controversial topic in Mandarin, I will not refer to it in the following discussion; instead, I will focus on the metrical strength of the syllables carrying different tones, and possible metrical structures of different words.

Chapter 3 Neutral Tone

The phenomenon of NT was first put under the spotlight by Chao Yuen Ren in the 1920s when he attempted to develop a romanization sound annotation system for spoken Chinese. In *On Gwoyeu Romatzyh*, Chao pointed out the existence of such weak elements in spoken Beijing Mandarin that were short, light and do not carry CTs, and referred to them as *qingyin*, 'light sound'. Later, he re-named these elements *qingsheng,* 'light tone', in his work on the intonation of Beijing Mandarin (Chao, 1929) and the lexical tones and intonation of Mandarin Chinese (Chao, 1933). The name *qingsheng* is still used in the Chinese literature now. Chao (1965) subsequently introduced this phenomenon to the English-speaking world as Neutral Tone. The term, to some extent, reflects Chao's view on this phenomenon, namely, that it concerns the neutralization of the contrasts between CTs when the carrier syllable is unstressed.

Chao did not only notice the phenomenon of NT, but also gave impressionistic descriptions of the phonetic features of NT and summarized the pronunciation norm of NT in Beijing Mandarin spoken in his time (Sections 3.1 and 3.3). Since Chao saw NT as a tone neutralization phenomenon, he focused on contextual variations of NT realization. This view has had a profound influence on the phonetic and phonological studies of NT that followed Chao's seminal work. However, as more distinctive characteristics of NT have been found[17], it has become clear that Chao's view is insufficient to account for some unique properties of NT, in particular, the great grammatical unpredictability and the two T3 sandhi patterns in NT words.

In terms of grammatical unpredictability, unlike the other tone sandhi phenomena in

[17] These characteristics may be newly developed and did not exist in Chao' time, e.g., the optional realization of NT.

Mandarin (Section 2.2.2), the occurrence of NT is not predicable from either a phonological or morphological perspective. NT-bearing elements are of various morphological types, ranging from grammatical morphemes to particular lexical items. There is large discrepancy in NT word lists between different dictionaries and different editions of the same dictionary as some have pointed out (e.g., Chen-Chung, 1984). Moreover, compared to the NT words in Beijing Mandarin a century ago (i.e., roughly in Chao's time), the NT realization of many words in Standard Mandarin now is becoming more and more optional, a tendency intensified also by the language policies in the second half of the 20th century (Section 3.3). What complicates the situation more is that some NT morphemes trigger T3 sandhi while others do not. There is no consistent rule based on morphology or syllable structure that can account for this variability.

The present chapter is organized as follows. Section 3.1 reviews the varying phonetic features of NT, especially the great influence of the preceding CTs. Section 3.2 reviews the phonological analyses of NT, in which the surface f_0 variation has attracted more attention than the underlying representation(s) of NT. Section 3.3 reviews the sociolinguistic, morphological and diachronic literatures on NT; it will be demonstrated that the discussion of the realization norm of NT embraces the subcategorization of NT, but proposed different types of NT that do not necessarily correspond to the hypothesis proposed in the current thesis, namely, Intrinsic NT and Derived NT.

3.1 Phonetic Studies

The phonetic features of NT-bearing elements have been widely explored from both the acoustic and the perceptual perspectives. With different experimental design and varying research emphases, different studies have reported slightly different results. As it will be reviewed in this

section, the controversies mainly lie in the f_0 values of NT, and the most important perceptual cues to NT-bearing morphemes.

Duration

The duration of NT-bearing syllables has been compared to that of CT-bearing syllables in several acoustic studies. Many studies report that the duration of a NT-bearing syllable is about half the duration of any CT-bearing syllable (Chen & Xu, 2006; Lin & Yan, 1980), and Cao (1986) reports a slightly longer duration of about 60% of the duration of the preceding CT-bearers. In contrast, Jin (2001) reports a varying duration ratio between NT and CT, ranging from 1/3 to 3/5.

f_0

Chao also provided the first impressionistic description of the pitch patterns that characterized NT. According to him, NT-bearing syllables surface with short duration and an extremely compressed pitch range, and pitch height varies according to the preceding CT, being 'half-low' when following T1, 'middle' when following T2, 'half-high' when following T3 and 'low' when following T4 (Chao, 1965). As a result of this, later studies have paid special attention to the contextual variations of NT when examining pitch patterns.

Acoustic studies confirm the critical influence of the preceding CTs observed in the impressionistic studies above and describe the f_0 variations of NT with more and slightly controversial details**.**

Table 3.1.a Results of acoustic studies on NT in the late 20th century (results converted into Chao Tone-Numerals)

Preceding CT (Contour)	Pitch Pattern of NT (Dreher & Lee, 1966)	Pitch Pattern of NT (Lin and Yan, 1980)	Pitch Pattern of NT (Gao,	Pitch Pattern of NT (Wang, 1996)

			1980)[18]	
T1 (high-level)	41	41	3	41
T2/T3 after T3 sandhi (mid-rising)	31	51	3	52
T3 (low-convex)	23	33 (male)/44 (female)	4	33
T4 (high-falling)	21	21	1	21

In addition to the studies summarized in Table 3.1a, Cao (1986) also reports that the starting pitch height of NT is the highest when the preceding CT is T2, followed by T1, T3 and T4; the ending height of NT is the highest when the preceding CT is T3, followed by T2, T1 and T4. Cao's findings (1986) are closer to Wang's findings (1996). Jin (2001) reports in further detail the pitch pattern of NT in four specific words produced by 6 individuals. Despite the individual differences, Jin's findings are similar to Lin and Yan (1980) Cao (1986), and Wang (1996). Nevertheless, studies conducted in the 1980s or earlier are usually limited in scale, including the number of participants, the number of stimuli and the number of measurement points of the f_0 contour.

The influence of tonal contexts was further examined in later studies. By analyzing the f_0 curves of one, two and three NT-bearing syllables produced consecutively in different tonal contexts, such as between different CTs or in sentence-final position, Li (2003) found a clear impact of the tonal contexts on the f_0 curves of NT. Li did not summarize in detail the pitch patterns of NTs like Cao or studies in **Table 3.1.a**, but pointed out that there was a gradual f_0 declination till the last NT when multiple NT-bearing syllables were produced in sequence. The

[18] Gao (1980) also reported that NT had a falling contour in general, but chose to calculate the average pitch value due to the short duration of NT.

gradual declination in the f_0 curves of 0-3 consecutive NT-bearing syllables was also found by Chen and Xu (2006).

Lee and Zee (2008) described at length the f_0 variations of NT in different tonal contexts, especially the influence of T1 and T3 as the following CTs, by investigating 24 disyllabic and 32 trisyllabic words, compounds or words groups in which NT occurred on the last or middle syllables (**Table 3.1.b**). It is worth mentioning that although Lee and Zee (2008) distinguished falling and level contours, they only provided general descriptions rather than accurate f_0 values. However, the differences between mid-falling and mid-level, or between mid-level and high mid-level are often very small, and it is unclear whether the differences were statistically significant.

Table 3.1.b f_0 contours of NT in different tonal contexts relative to the preceding CT contours ('-' indicates no following CT)

Preceding CT (Contour)	**NT Contour**	**Following CT**
T1 (high-level)	mid falling	-/T2/T3/T4
T2/T3 after T3 sandhi (mid-rising)	high falling	-/T2/T3/T4
T3 (low-falling)	mid-level	-/T1/T2/T4
T4 (high-falling)	low falling	-/T1/T2/T3/T4
T1/T2/T3 after T3 sandhi (high-level/mid-rising)	high level	T1
T3 (low-falling)	high mid-level	T3

Other Acoustic Features

In NT-bearing syllables, vowels and glides are centralized to a large extent, diphthongs tend to be reduced and coda nasals sometimes are lost (Lin and Yan, 1980; Cao, 1986; Chen, 1986). Some voiceless onsets, especially voiceless unaspirated affricates, may become voiced in

NT-bearing syllables (Cao, 1986; Zhou, 2018). However, these changes are much influenced by speech rate and segmental context. Therefore, phonologists like Jin (2001) suggest that the centralization or voicing should not be seen as a stable feature of NT.

Intensity, however, is not always lower in NT-bearing syllables than in CT syllables. According to Jin (2001), NT-bearing syllables can be louder, softer or of the same intensity as the preceding CT-bearing syllables, while both Jin (2001) and Cao (1986) found that NT after T3 tended to be realized with larger intensity. Lee and Zee (2008) found that the intensity of NT-bearing syllables was also influenced by the tonal contexts they were in, and there was in general a gradual declination in the intensity of the NT-bearing syllables.

Perceptual Cues

Pitch and duration are both of great importance to the perception of NT, but which of them is more important remains relatively unsettled, contributing to long lasting debates over the nature of NT, i.e., whether NT is a tonal or a stress phenomenon.

By cutting the duration of the second syllable in 7 minimal pairs of CT-CT and CT-NT words, Lin (1983) found the shorter the second syllable of a disyllabic word was, the more likely it would be judged as a NT word. Wang (2004) criticized Lin's design as it did not take into consideration that pitch contours co-vary with duration and improved the experimental design accordingly. By using Pitch Synchronous Overlap Add (PSOLA), Wang managed to keep the pitch height of the key points when changing duration. Unlike T.Lin (1983), Wang (2004) found pitch to be a more important cue to NT perception than duration. Both reduced pitch height and a falling pitch curve led the participants to judge the stimuli as having NT, and the falling contour was of special

importance to the perception of NT after T2. More importantly, Wang found that the pitch curves of NT preferred by the listeners have a lower onset pitch value than the natural production when the preceding tone is T1 and T2, indicating that these high onsets may only be a redundant phonetic phenomenon. Later studies conducted by Li and her colleagues also showed a stronger effect of pitch than duration on NT perception, though the weight of the two cues was conditioned by various factors like the position of the target word in relation to the focus of the whole utterance (Gao and Li, 2017; Li and Fan, 2014).

3.2 Phonological Analyses

The high contextual dependence of NT has led to the prevailing view that NT is not specified for tones in its phonological representation (e.g., Duanmu, 2007). Many phonologists adopt Chao's view that NT is derived from CTs, but remain vague on whether the CTs are still the underlying representations of NT or not. Instead, more research attention has been paid to the derivation of the varying surface tonal contours NT demonstrates, or the association between NT and unstressed syllables. In either case, NT in Standard Mandarin is often treated as a homogeneous entity in phonological analyses, despite the morphophonological variation it demonstrates (Section 3.3)[19].

Yip (1980), who adopts and Autosegmental approach in her account, regards NT as a toneless element at the phonological level but not as register-less. She proposes that in the underlying representation (i.e., lexical entry), NT in Mandarin is specified with [-Upper] in the *Register* tier

[19] Motivated by the grammatical unpredictability of NT and the two T3 sandhi patterns, Chen-Chung (1984) proposes that there are toned and toneless elements in the underlying representation of Mandarin in a discussion over the norm issue of NT; both elements can carry different degrees of stress. However, Chen-Chung's empirical research was conducted in Taiwan, and Taiwan Mandarin is different from Standard Mandarin spoken in many prosodic aspects; in particular, NT in Taiwan Mandarin has a relatively robust tonal contour, and is not necessarily less loud or shorter than CTs (Huang, 2018). Therefore, stressed toneless elements may sound natural to Taiwan Mandarin speakers, but not Beijing Mandarin speakers as Chao and others have pointed out (Section 3.3).

but unspecified in the *Tone* tier (i.e., *Pitch* tier). The specification of NT on the *Tone* tier comes from target spreading of the preceding CT. The surface realization then is a consequence of the interplay between [-Upper] and the target from the preceding CT, resulting in a [-Upper, H] realization which sounds mid after T1, T2 and T3 but a [-Upper, L] realization which sounds low after T4 (**Figure 3.2. a**).

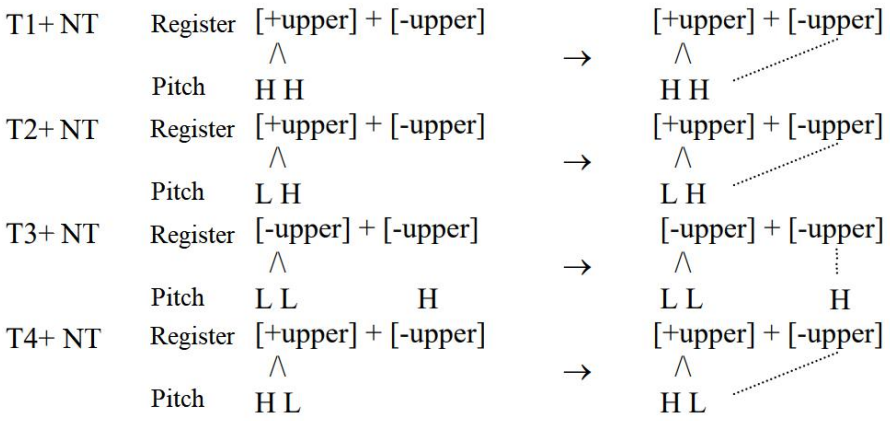


Figure 3.2.a Derivation of NT after different CTs (adapted from Yip, 1980:253)

Nevertheless, when explaining the T3 sandhi caused by NT in some words, Yip also referred explicitly to a tone-deleting (i.e., T3-deleting) process in the derivation of NT (Yip, 1980) (**Figure 3.2.b**).

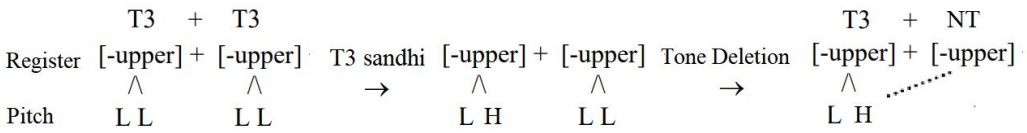


Figure 3.2.b Derivation of NT after different CTs (adapted from Yip, 1980:253)

What remains unclear is whether the *Register* features of the deleted CTs are also lost. If so, where does the [-Upper] specification of NT come from? If not, NT derives from T1, T2 and T3 should be specified for [+Upper] rather than [-Upper], meaning that these NTs would surface as

high tones (i.e., [+Upper, H]) when following T1 and T2. However, these 'results' are different from the empirical data (Section 3.1) so that the *Register* features of CTs must also be lost. Then, it becomes a question that from where NT gets the [-Upper] feature.

Despite these ambiguities, the tone-spreading proposed by Yip has been adopted by some researchers. For instance, Shen (1992) proposes that the surface target of NT is the final target of the preceding CT in citation form, though she has not followed Yip's hierarchical target system (1980) but applies a simple binary tonal feature which gives the two values (i.e., H and L) to analyze Mandarin tones (**Figure 3.2.c**).

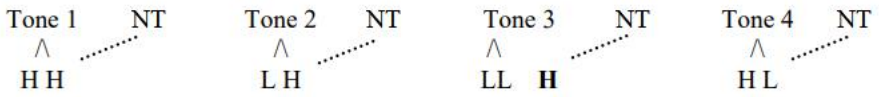


Figure 3.2.c Derivation of NTafter different CTs (based on Shen, 1992)

Unlike Yip (1980), Wang (1997) explicitly posited CTs as the underlying tones of NT in Standard Mandarin by distinguishing three levels of representation, which we could refer to as underlying phonological representation, surface phonological representation and phonetic implementation. In the underlying representation, morphemes are specified with tone targets but the targets are lost on the phonological level due to lexical tone-deletion or stress-triggered tone-deletion. The former applies before the application of the T3 sandhi rule and the later applies after T3 sandhi rule, resulting into the two sandhi patterns observed in the surface forms.

Wang (1997) also provides an explanation for the mid surface tone of NT through reference to the default value-filling rules proposed by Pulleyblank (1986). Pulleyblank (1986) states that the phonologically underspecified tones are in fact [-Upper, H] tones, and the [-Upper] and [H]

features are inserted before the application of all the other phonetic rules. The other complex surface variations observed for NT like the high realization after T3 arise from the sequential applications of the other phonetic processes like *Insertion of the [+Upper] register* and *High Target Spreading.*

As the experimental data accumulated, a target independent from the preceding CT in NT-bearing syllables has been proposed in several studies, though some still argue that the tone target is not a target of NT its own but a boundary tone target. For example, based on more detailed empirical data (See Section 3.1), Li (2003) argues that the mid target does not belong to NT but is in fact a boundary tone. According to him, the preceding CT and the following NT(s) would form a prosodic word, in which the surface realization of NT was a result of the interpolation between the preceding CT and the boundary tone of that prosodic word.

Chen and Xu (2006) interpreted a research result that was similar to Li's (2003), but provided a more phonetic-based explanation following the Target Approximation Model (Xu & Wang, 2001). According to Xu & Wang, 2001, NT has a mid-low target of its own despite the surface realizations that have been observed, while the relatively large amount of variation is attributed to the ineffectiveness of NT in overcoming the influence of the preceding CT as the weak element in its realization. The thinking is that, if allowed more time (i.e, when several NT-bearing syllables are realized consecutively in Chen and Xu's design), NT should be able to approach its proper target regardless of speech rate and focus. On empirical grounds, Chen and Xu (2006) posited a mid-low target for NT because phonetically the average pitch height of NT was not as low as T3.

Lin (2006) expresses a similar view to Chen and Xu (2006) that NT has its own tonal target

which is independent of the surrounding tones. However, he argues that NT is a phonologically low tone by interpreting the raise of NT after T3 as the universal resistance of two low targets in sequence (i.e., the Polarity rule). Regarding Chen and Xu' view, Lin (2006) interpreted the mid target Chen and Xu proposed as 'purely phonetic'. (Lin, 2006:127). However, Lin (2006) did not offer an explanation of why low NT does not cause T3 sandhi but has its own low target raised instead, which also raises question marks over this particular phonological account.

Some evidence found in sociolinguistic studies also suggests that NT in Standard Mandarin may have a relatively low target like in Lin (2006) view. Firstly, some studies report that there is a tendency for NT to be realized as T4, the falling tone (Wang, 1992; Jin, 2001). In addition, Zhou (2018) has found in a corpus study that NT morphemes can also cause T3 sandhi in words like 主**意** /tʂu3->2 **ji0**/ 'idea', 指**头** /tʂi3->2 **tʰʊɔ0**/ 'finger', 骨**头** /ku3->2 **tʰʊɔ0**/ 'bone', 脊**髓** /tɕi3->2 **suei0**/ 'spinal marrow', 打**扮** /ta3->2 **pan0**/ 'dress up', and 佐**料** /tsʊɔ3->2 **li ɑ u0**/ 'seasoning' among some speakers. Although these words only account for a small portion of the NT-bearing words and not every Mandarin speaker adopts the sandhi pronunciation (Zhou, 2018), this phenomenon may be of phonological importance as it indicates that NT may be developing a phonologically low target of its own as Lin (2006) suggested. The NT-bearing morphemes in these words (e.g., 意, 头, 髓, 扮, 料) are special because they are unrelated to T3 in any aspect, but they can raise the preceding T3 in the same way the low convex tone T3 can do.

Another group of phonologists place more emphasis on the stressless nature of NT. Linguists like Duanmu (1990, 1993, 1994) propose morae as TBUs in Mandarin (Section 2.3.4). Following the Universal Association Convention strictly, namely, all tone-bearing units are associated with at least one tone (Pulleyblank, 1986), Duanmu (1993) also suggests a very strict one-to-one mapping

between the tonal target and the morae. According to him, heavy syllables in Mandarin are bimoraic and therefore they could keep both tone targets, but light syllables are monomoraic so that they could only keep one target. The loss of morae would cause the deletion of a tone target. Combined with Duanmu's arguments against contour tones as a whole (Section 2.3.2), this assumption actually indicates a strong influence of the underlying tone targets the unstressed syllables have not preserved. However, Duanmu reckoned that this indication is not supported by data, though in reality NT has rarely been explored according to the potentially different underlying CTs. In his later work, Duanmu simply argued that the unstressed monomoraic syllables are not able to keep their underlying tones and there is no compelling evidence showing that NT has a unique target of its own (see discussion in Duanmu, 2007). This view is also shared by many who do not argue for on the association between mora and tone targets but still associate NT with the loss of stress like Chao does (e.g., Wang, 1997; Shen, 1992; Zhang, 1988).

The two T3 sandhi patterns are also attributed to the different degrees of stress, namely, a neutral tone is a neutral tone but will trigger T4 in the preceding syllable unless it is slightly stressed (Chao, 1965). Duanmu also argues that the morphemes causing T3 sandhi is weakly stressed and therefore carry reduced T3 [21] rather than the full [214]. In this way, an intermediate level of stress of phonemic function is established. Nevertheless, Chao still tries to avoid giving phonemic status to this intermediate stress, a highly controversial topic in Mandarin. After all, there is no evidence other than T3 sandhi differences helping to distinguish weak stress from non-stress.

The metrical analyses of NT words are also highly controversial since researchers have different views on the metrical organization of Mandarin words (Section 2.4). For those who

follow Chao's proposal (1965) and regard CT-CT words as iambic feet, CT-NT words are seen as special trochaic feet (e.g., Yip, 1980; Zhang, 1988). Among those who regard Mandarin as having trochaic feet only, two general proposals are made. Some analyze a CT-NT word as a single trochaic foot and a CT-CT word as two unary degenerate feet (e.g., Cheng, 1973). Others choose to remove NT-bearing syllables from feet and mark them with special features like extrametricality (e.g., Hsieh, 2021). However, these metrical analyses are based on observational data, which may be confounded by phonological factors like tone combination and paralinguistic factors like individual differences (Section 2.4). In this thesis, the metrical representation of NT is probed by manipulating the utterance focus in experimental settings (Chapter 4 and 7). The metrical weight of NT and the metrical structure of Mandarin disyllabic words would be revisited with the experimental data.

3.3 NT in Sociolinguistics and Historical Linguistics

3.3.1 The Norm and the Morphology

As introduced earlier, the realization of NT is not predictable from a phonological perspective, unlike T3 sandhi. Nevertheless, linguists still try to find as much regularity as possible in this complicated phenomenon and hence turn to morphology, attempting to give either a general rule guiding the use of NT or to summarize the NT realization norm in an exhaustive manner. Again, Chao is the first linguist who tried to summarize the norm of NT realization in Mandarin. Based on the predictability of NT, he divided the realization of NT into regular and irregular situations (1929). According to him, in situations 1)-6) NT could be regularly found:

1) On auxiliary particles, such as 阿/a/, 啊 /a/, 吧 /pa/, 的 /tə/, 得 /tə/, 着/ʈʂə/, 了/lə/, etc;

2) On empty morphemes in word final position, such as 这个 /ʈʂə4 **kə0**/ 'this', 什末[20] /ʂən2 **mə0**/ 'what', 这是 /ʈʂə4 **ʂi0**/ 'this is', 里头 /li3 **tʰəu0**/ 'inside', 我们 /wo3 **mən0**/ 'we, us', etc;

3) On positional auxiliary verbs, such as 回来 /xweɪ2 **laɪ0**/ 'come back', 拿回来 /na2 **xweɪ0 laɪ0**/ 'take back', 弄掉 /nʊŋ4 **tjɑʊ0**/ 'get rid of, remove', 低下来 /ti1 **ɕia0 laɪ0**/ 'low down', etc;

4) On positional postpositions, such as 上 /ʂɑŋ/ 'on', 里 /li/ 'in', etc;

5) On pronouns in utterance final position, such as 他 /tʰa/, etc;

6) On verbs or adjectives after '不 /bu/' in phrases like '要不要 /jɑʊ4 **bu0** jɑʊ0/'.

In the situations not captured by one of these six rules, that is, on the second or third syllables in certain lexical items, the occurrence of NT is unpredictable and item-specific (i.e., irregular). Chao (1929) made a very general rule to guide the realization of the irregular NT, that is old words often involve NT while new (loan) words do not. However, Chao himself realized that this generalized characterization was too general to be a rule. Therefore, he suggested that such item-specific occurrences of NT should be recorded and memorized individually for each lexical item (Chao, 1929, 1968).

Despite the unpredictability of the irregular NT words, Chao (1929, 1933, 1968) still regards NT as a critical component in Mandarin phonology. In his opinion, NT is important not only because it distinguishes lexical meanings as CT in certain word pairs (though in a much more limited way), but more importantly, it is an important feature of spoken Mandarin. Substituting NT with its CT-alternates (i.e., the CTs that the NT-bearing morphemes would be pronounced with in isolation) in even the 'irregular' NT words, would not be accepted by native speakers, or at least, be judged as southern Mandarin (Chao, 1929). In other words, in Chao's definition, the irregularity only refers to the morphological unpredictability of the NT morphemes, but the

[20] 末 is written as 么 now in Standard Mandarin.

realization of either the regular or irregular NT is not optional.

However, since the language situation may have changed since 1929, this analysis may no longer apply integrally to the Standard Mandarin spoken today. Another major change is that the realization of NT is becoming less obligatory including situations in 1) - 6).

Unlike Chao who conflated morphological and phrasal de-stressing (e.g., situation 5), Xu Shi Rong (1956) focused only on the morphological aspect when summarizing the realization norm of NT. Xu linked the phenomenon of NT closely to de-stressing and then to the morphological structure of words, especially to the functional load of the second or third morphemes in the word and their relation with the preceding CT-bearing morphemes. In other words, the lighter the functional load of a morpheme and the closer it is to the preceding morpheme, the more likely it is realized without stress (i.e., as NT)[21]. The de-stressing would also be conditioned by word class and how 'old' the items have been in the spoken language, that is, how early the items have come into being in Chinese. Xu (1956) proposed that there are 10 morphological structures in Mandarin disyllabic lexical words: disyllabic mono-morphemic words, coordinative words with two morphemes of similar meanings, coordinative words with two morphemes of different meanings, words made up of a modifier and a head, words made up of a verb and an object, verb compounds, reiterative words, diminutive words, words with prefixes and words with suffixes. Both heavy-light (i.e., NT word in his proposal) and heavy-heavy patterns could be found in words of all these 10 structures, but in the following situations the words are likely to be realized as heavy-light (i.e., as NT words):

[21] Xu (1956), like many Chinese scholars who first studied Chinese philology regards NT as a de-stressing phenomenon (Section 3.3.2). Stress in their studies, however, does not correspond to word stress in English, but describes the syllabic prominence.

1) Most mono-morphemic words, especially old ones that have existed in the spoken language for a relatively long time like 蜻**蜓** /tɕʰiŋ1 **tʰiŋ0**/ ‘dragonfly’;

2) Most coordinative words with two morphemes of similar meanings like 鸳**鸯** /juan1 **jɑŋ0**/ ‘mandarin duck’;

3) Coordinative words with two morphemes of different meanings but have lexical meanings changed, like 买**卖** /maɪ3 **maɪ0**/ ‘business (buy and sale)’, 来**往** /laɪ2 **wɑŋ0**/ ‘contact, communication (come and go)’, 是**非** /ʂi4 **pʰei0**/ ‘trouble, troublesome (yes and no)’.

4) Modifier-head in the following situations:

a. With empathized modifiers, especially when the modifiers describe unexpected features like 香**椿** /ɕiɑŋ1 **tʂʰʊn0**/ ‘Chinese toon (fragrant-toon)’, 长**虫** /tʂʰ ɑ ŋ2 **tʂʰɔŋ0**/ ‘snake (long-worm)’;

b. With reduced accuracy, realness, or importance:

a) Metaphoric items like 木**耳** /mu4 **ɚ0**/ ‘Auricularia (wood-ear)’;

b) Generic items of reduced accuracy like 玉**米** /y4 **mi0**/ ‘corn (jade-rice)’;

c) Items of which the heads have changed meanings like 鸡**眼** /tɕi1 **jɛn0**/ ‘heloma (chicken-eye)’;

d) Matronymic terms using a part to refer to the whole entity like 烧**锅** /ʂɑu1 **kʊɔ0**/ ‘clay pot stew (boil-pot)’;

e) Items with unimportant head: 烧**饼** /ʂɑu **biŋ0**/ ‘bread (bake-bread)’, 臭**虫** /tʂʰou4 **tʂʰɔŋ0**/ ‘bug (foul-bug)’.

5) Items with close internal structures:

a) Items with pseudo suffixes like 韭**菜** /tɕiou3 **tsʰai0**/ ‘leek’, 眉**毛** /mei2 **mɔ0**/ ‘eye-brow’;

b) Items that have been lexicalized for a long time and hence almost becomes mono-morphemic words like 老**爷** /lao3 **jɛ0**/ ‘master, lord’ 老**鼠** /lao3->2 **ʂu0**/ ‘rat, mouse’;

c) Items of which the second morphemes could be used alone in the past but not now like 泥**鳅** /ni2 **tɕʰiou0**/ ‘loach’, 刺**猬** /tɕi4 **wei0**/ ‘hedgehog’, 石**榴** /ʂi2 **liou0**/ ‘pomegranate’.

6) Items with generic morphemes referring to things like 物 /wu/ ‘thing’ 器, /tɕʰi/ ‘utensil’, 类 /lei/ ‘type’, people like 人 /ʐən/ ‘people, person’, 士 /ʂi/ ‘people, person (with high prestige)’, 工 /kɔŋ/ ‘worker, craftsman’, form or properties like 式 /ʂi/ ‘kind’, 色 /sɛ/ ‘kind, colour’, 声 /ʂəŋ/ ‘sound’, time like 季/tɕi/ ‘season’, 月/yɛ / ‘month’, 天/tʰian/ ‘day’, place like 处/tʂʰu/ ‘place’, 上/ʂɑŋ/ ‘onside’, 里 /li/ ‘inside’, events or work like 事 /ʂi/ ‘event’, 物 /wu/ ‘matter’, 业 /jɛ/ ‘industry, business’

7) Verb compounds with the second verb reduced like 来 /laɪ/ ‘come’, 去 /tɕʰy/ ‘go’

8) Reiterative verbs with diminutive second morphemes like 看**看** /k^{h}an4 **k^{h}an0**/ ‘look’

9) Diminutive nouns like 妈**妈** /ma1 **ma0**/ ‘mother’

10) Words ending with real suffixes like 子 /tsɨ/, 着/ʈʂə/, 了/lə/

The number of NT-bearing items included in Xu’s study (1956) is very large and the rules have been criticized for over-differentiating between NT and CT words. However, Xu’s proposal may better be seen as a predictive description of the development of NT in the spoken Mandarin, rather than concrete rules guiding the realization of NT from the developmental perspective as follows: morphemes with a lower functional load and a closer relationship with the preceding morpheme can be assumed to be less prominent over time and finally such lack of prominence is lexicalized as NT-bearing morphemes. The lexicalization of NT is also conditioned by the frequency of the word - the more frequently a word that matches the above conditions is used, the more likely it would be lexicalized as NT words.

A study based on the Contemporary Beijing Spoken Language Corpus[22] echoes Xu’s observation (Zhou, 2018). It found that among the 1000 most frequent words, 23.73% of them were NT words, and the percentage dropped as the word frequency falls (Zhou, 2018). The influence of morphological structure was also observed in the same corpus study. By dividing Mandarin morphemes into 5 levels based on their functional load (Level 1 are notional morphemes with the largest functional load and Level 5 are empty affixes), the combination of two Level 2 morphemes generated most NT words (38.25%), followed by the combination of a Level 1 morpheme and a Level 5 (empty) morpheme (31.75%). The percentage of NT words in the other morphological combination was smaller than 13%. Nevertheless, Zhou’s study shows

[22] This corpus was established in 1980s-1990s, involving 374 Beijing dialect speakers of various social backgrounds.

that morphological structure, functional load of the morpheme, or the frequency of the word could only indicate the likelihood that a word is an NT word, rather than predict this.

Following either Chao's or Xu's template, the realization norm of NT has been regularly revisited by different studies (e.g., Lu, 2001; Shi, 1984; Gao, 1980). Despite the differences between studies conducted at different times, a distinctive trend is that the optional realization of NT is becoming more and more tolerable, especially regarding the phrasal de-stressing situations. However, by examining the language policy literature in the 19th century, I propose that the optional realization of NT words may not be, or at least, may not *entirely* be an intrinsic phonological feature of Beijing Mandarin (based on which Standard Mandarin is developed, Section 2.1), but influenced greatly by language policy.

The most important language policy proposal was made by Li Jin Xi, who initiated and led the compilation of *Guoyu Cidian* (The Official Dictionary of Mandarin Chinese) in the first half of the 19th century and laid the foundation of the *Pinyin* system used currently. To establish a national language that is easy to spread and learn, Li proposed that lexical items involving NT should be separated into two types, 'light and toneless' and 'light but with tone'. The second type accounted for the majority of NT words and should be annotated as CTs with an optional weak stress symbol '(·)' (Li, 1948, cited in Compilation Office of Dictionaries in China, 2011:7.3). It turns out that Li's proposal is quite in line with the hypotheses made in the present thesis. However, instead of being a phonological argument, Li's proposal is more like a compromise made between the goal to keep the authenticity of Beijing Mandarin and the goal to facilitate the promotion of Standard Mandarin, the national language.

As a native speaker of a non-Mandarin dialect, Xiang dialect, Li may have more understanding of the difficulty in NT acquisition if no native judgement is available. According to him, it would be unnecessary to force the authentic Beijing pronunciation onto the whole population at the cost of a heavy memory burden (Li, 1948, in Compilation Office of Dictionaries in China, 2011:7.3). Although Li's motivation has been widely shared by many later researchers, the criteria of categorization (if exist) have been even more blurry due to language changes, reflected in the discrepancies between NT word lists in different dictionaries as well as in the different editions of the same dictionary.

Nevertheless, as an important part of the national pronunciation norm, the norms of NT realization have been consistently revisited by different scholars, and no general agreement has been reached. Since the chaos in NT words could not be ended by grammatical rules or morphological categorization, some linguists tried to regulate the use of NT by introducing arbitrary rules into Standard Mandarin. The two most representative views may be summarized as *to cancel NT in the formal reading or speaking situations, especially when it does not distinguish lexical meanings* (Xie, 1998) or *if the contours of the CT-alternates are not as very different from NT (i.e., T4 the falling tone in most cases and the half realized T3 in non-final situations, the words should no longer be annotated as NT words* (Song, 1990). However, these proposals have not been widely accepted as it interrupts the natural development of language in a sudden and arbitrary manner and the forced CT realization still sounds unnatural to northern speakers (e.g., Lu, 2001).

Since language changes could not be stopped by regulations, the prevailing attitude on NT in the recent two decades has changed from trying to regulate or normalize it to tracking and

recording the changes in lexical NT by conducting sociolinguistic surveys and renewing dictionary lists accordingly. Studies comparing different versions of Mandarin dictionaries in the recent 2-3 decades demonstrate that there is a decrease in the total number of NT words, but a small group of monosyllabic functional morphemes and suffixes have remained NT-bearers in a relatively stable manner (Wang, 2012; Lin & Li, 2017). These morphemes include modal particles, structural particles like 的 /tə/, verb particles like 着 /ʈʂə/, 了 /lə/, and suffixes like 子 /tsɨ/, 头/tʰəu/, 们 /mən0/. In fact, the 105 new NT words with the obligatory NT realization added to the 6th edition are mostly made up of these NT-bearing morphemes, meaning that they are also productive in word formation now. These stable NT-bearing morphemes carry light semantic or functional load compared to the other NT-bearing morphemes, namely, they are of less morphological importance and hence may of less prosodic prominence.

In next section, I will show that different NT morphemes may have been lexicalized at different times, and this may also lead to the observed differences in realization norm and productivity among NT morphemes. To be specific, some NT morphemes may have been lexicalized as NT or unstressed syllables earlier in the history of the Chinese language. Hence, they are shared more widely with Chinese speakers rather than (northern) Mandarin speakers only, more stable and have little chance to be connected with the modern four CTs.

3.3.2 The Historical Development of NT

Arguably, the phonetic contrast between the heavy and the light driven by articulatory factors or conveying pragmatical meanings universally exist, but the phonologicalization of the contrast is believed to root in morphological or syntactic changes rather than phonetics (Wang, 1980). The deduced derivation of NT in the traditional exegetical and historical linguistic literature may be

summarized as follows (e.g., Li, 2000; Jiang, 1994; Yuan, 1992; Wang, 1980; Chen, 1960): a morpheme with reduced semantic meaning is more likely to be realized with less stress (i.e., being phonetically light), surfacing with acoustic reductions including the compression or even deletion of tonal characteristics. When the de-stressing became lexicalized, the prototype of NT came into being. This view of the emergence of NT may be one of the fundamental reasons for NT to be treated as a phonologically homogeneous tone-deleting phenomenon by Chao Yuen Ren, who studied Chinese philology rather than phonology first. However, I shall argue in this section that even if all the NT-bearing morphemes share the same de-stressing derivation, they could still be of different phonological representations, especially tonal representations.

The deduced derivation introduced at the beginning of the section indicates that the genesis of NT is closely related to the increase of disyllabic words, which involves the split of the empty morphemes and the cliticization of the notional morphemes (e.g., Li, 2000; Jiang, 1994; Yuan, 1992). The disyllabicalization and cliticization process can be dated back to *Han* Dynasty (202 BC - 220 AD) or before according to the written documents. However, it is not until *Tang* Dynasty (618 - 906 AD) that evidence can found for the existence of contrasting stress in the dialect spoken in the north-western part of China[23]. According to Chen (1960), in the *Vajracchedika-prajna-paramita Sutra*[24] transcribed with Brahmi alphabets, an alphabetic transcription system that distinguishes segmental duration, the morphemes 人, 次, 子, 袒, 度 in disyllabic words 女人 'women', 复次 'again', 男子 'man', 偏袒 'favoritism', 灭度

[23] It is reasonable to hypothesize this is the common speech in Tang Dynasty as the capital city of Tang, Chang'an, is located in the north-western part of China.

[24] This is the Sutra transcribed into Chinese by Kumārajīva. The Brahmi transcription is from the Chinese rather than original Indian version.

‘(Buddhists’ belief) escape from a life of misery’[25] were transcribed with short vowels rather than long vowels like in the other situations. It is worth noting that this transcription may be rather phonetic than phonemic as the coda syllables of the long adjective words and the verbs at the objective positions or the utterance-second positions are also transcribed with short, allophonic vowels. It is therefore unknown how stable this de-stressing was and whether it triggered tone-deletion or not. Unfortunately, no other evidence has been found to cross-validate Chen’s findings either. Therefore, modern scholars need to deduce the pronunciation of some cliticized morphemes based on indirect evidence. The frequently used indirect evidence includes the characters these morphemes are transcribed as, and their positions in the poetic meter.

Many texts written in late *Tang* Dynasty already contained the grammaticalized morphemes like structural particle 的, empty or generic suffixes 子，头，们/么，当，生，道, and aspect particle 着，了，过, which are morphemes with obligatory NT realization and productivity in modern Mandarin (for examples see Lv, 1955; Yuan, 1992) though there is only indirect evidence for the unstressed (NT) pronunciations of some of these morphemes in Middle Chinese.

When they first emerged, morphemes 的，们/么，头 were transcribed with inconsistent characters, and then with fixed characters that differ from all the other frequently used notional characters (for examples see Lv, 1955). More importantly, most of the characters ‘lent’ to these morphemes at the beginning normally carried Ru tone (i.e., Checked tone), the fourth tone in the mid-ancient Chinese that was short, sharp and falling (Section 2.2.3). Therefore, it is reasonable to deduce that these morphemes were short and falling at that time already, much like the phonetic

[25] IPA transcription was not given here because these words were pronounced differently in Middle Chinese, and so were the other morphemes presented in this discussion.

description of NT today. The association between the grammaticalized morphemes and the new characters become relatively stable in late *Song* Dynasty or *Yuan* Dynasty (the 13th century - 14th century AD), but the inconsistent writings had not disappeared until about 20th century. For instance, in a famous vernacular novel *Jin Ping Mei (The Golden Lotus)* written in late *Ming* Dynasty (about 1567 -1620 AD), the empty suffix 头 in 骨**头** /ku3 **tʰou0**/ were sometimes written as 透 (realized as /tʰou4/ in Standard Mandarin) and sometimes written as 突 (realized as /tʰu1/ in Standard Mandarin) . The borrowed characters may be reflecting the falling contour and the vowel reduction in 头.

In *Song* and *Yuan* Dynasty (the 10th century - 14th century AD), more empty suffixes like 当, 生, 道, aspect particles 着, 了, 过 and model insertions like 不 became popular. These characters have not been transcribed with other characters but some of them have left traces of sound changes in poetic works. For example, by examining the lyrics of theaters written in *Yuan* Dynasty, Li (1992) found that though written as the same character 着, the verb 着 always occurred in the position that requires *Ping* tones, 'level tones', while the aspect particle 着 occurred in the position that requires *Ze* tones, 'oblique tones', indicating that aspect particle 着 may already have had a falling contour by then (Section 2.2.3). Unfortunately, verb 了 and 过 also carried *Ze* tones so that there is no evidence suggesting their pronunciation in ancient time. Nevertheless, in the modern Mandarin, 了 is of a lexicalized different pronunciation /lə0/ when serve as the aspect particle, but pronounced as /liɑʊ3/ otherwise. There are also studies arguing that the modal particles 咯 /lo/ and 啦 /la/ are different stages of the reduced 了 /liɑʊ3/ (see Shi, 1986). Although the aspect particle 过 has its tonal realization hard to trace, it is reasonable to expect that it also bears NT just like its peers 着 and 了.

The emergence of some other generic or empty suffixes like /tʊɔ/ (掇/夺) ，/ta/ (哒/达/搭)，/tɕiɛ/ (价/介) and the irregular disyllabic NT words (e.g., 告**诉** /kʰɑu4 **su0**/ 'tell') could not be reliably traced until late *Ming* Dynasty (16th - 17th century AD) when vernacular novels became popular (e.g., Zhang, 1953; Li, 1987). Only in these novels could their inconsistent written forms be frequently found, which indicate that these morphemes may not be realized with a stable tone. It is worth special attention that there is still no consistent writing forms for the majority of these suffixes even now because most of them have not been included in Standard Mandarin but remained dialectal or are spoken only.

The history of the modal particles, in general, is longer and more complex than the other grammatical particles in Chinese languages (e.g., Wang, 1955; Lv, 1995; Shi, 1986) . Until now, they have been active not only in Mandarin dialects but also the other Chinese dialects without proper NT like in Cantonese. Modal particles like 吗 /ma/, 呢 /nə/, 吧/pa/ that are frequently used in the modern Mandarin may have come from the ancient ones like 没/mɔ/, 无 /wu/, 尔, /er/, 罢 /pa/ and have been transcribed with various characters for a long period until special characters with radical 口 (indicating mouth or speech) have been created for them in the late *Yuan* Dynasty (Shi, 1986).

The historical development of the potential NT syllables with textual authentication may be summarized as follows:

1. De-stressed syllables existed as a phonetic phenomenon: *Tang* Dynasty or earlier
2. Grammaticalization of the structural particle '的，地，得', empty suffixes like '子，头，当，道'[26], some generic suffixes like '们/么', aspects particles '着，了，过' and some modal

[26] The ancient use of 生 as an empty suffix has disappeared in history. 生 now actives as a different suffix in the modern Mandarin meaning 'person (especially male)', like in 先生，学生，后生 and has optional NT realization.

particles with concrete evidence for their de-stressed realization: *Tang, Song, Yuan* Dynasty

3. Grammaticalization of other suffixes and disyllabic NT words: late *Ming* Dynasty or later

A key time point to be marked in this process is the *Yuan* Dynasty, the dynasty established by the Mongolian, when the disappearance and re-distribution of the checked tone was completed (Section 2.2.3). Therefore, it may be reasonable to hypothesize that the morphemes lexicalized as unstressed NT bearers before the *Ming* Dynasty (i.e., about the mid-14th century or earlier) are not able to have the modern four CTs as the underlying tones. In other words, if 'being toned in the underlying representation' is defined as 'being underlyingly represented as the modern CTs', these morphemes could only be 'toneless'. What remains to be explored, however, is whether these tones have developed a tonal target of their own in the long years.

It may be oversimplified for now to argue that the other NT-bearing morphemes or words became prevalent after *Yuan* Dynasty are specified with CTs in the tonal representations, especially those without fixed orthographic characters. It is plausible that they derived from the full tones just like the NT lexicalized earlier. However, whether this process has been completed, that is, whether the CTs which they derived from influence the realization and processing of these NT-bearing morphemes requires further investigation. Without empirical data, no firm statements on the tonal representations of these NTs can be reached.

3.4 Summary

The sociolinguistic, morphological and diachronic evidence all suggest that NT is not a homogeneous entity. In particular, by comparing the diachronic emergence of NT and the realization norm of NT words in Standard Mandarin, it becomes obvious that the morphemes with more stable NT realization overlap with the functional morphemes intrinsically developed in the

Chinese language early in the history while those with optional realization now were lexicalized later, motivated probably by the language contact with Manchurian as well.

More importantly, the NT morphemes came into being in the *Ming* dynasty or later co-developed with the Mandarin dialects, and therefore were more likely to be related to the CTs in modern Mandarin now. This differences in emerging time may indicate tonal differences between NT-bearing morphemes in Mandarin.

Give the classifications reviewed in Section 3.3, I speculate that there are two different types of NT which I will call *Intrinsic* NT and *Derived NT*. *Intrinsic* NT is carried by structural particles, aspect particles, suffixes, and modal particles. It intrinsically belongs to the Chinese language, and there is little possibility for this type of NT to be derived from the modern CTs, as discussed above. In contrast, *Derived NT* is mainly carried by the second or third morphemes in certain words or frequently used compounds. It is very likely to be derived from the CTs and the derivation may be driven, or at least, accelerated by external factors. What remains unclear is whether these two types of NT represent two distinctive phonological categories. In the following two Chapters, two series of experiments were conducted to answer two key questions, RQ1 and RQ2.

Part A Representations of Intrinsic NT and Derived NT

Chapter 4 Acoustic Realization of Intrinsic NT and Derived NT

4.1 Overview

The present chapter explores the potential tonal and metrical differences between Intrinsic NT, Derived NT and CT through two production experiments, in order to answer the research questions proposed at the end of Chapter 1, namely, whether Intrinsic NT and Derived NT have different phonological representations. As shown in Section 3.1, there have been a number of acoustic studies investigating the f_0 realizations of NT but focusing on the influence of the preceding CTs rather than the potential differences in the underlying representations of NT. To explore the acoustic realization of NT by their provisional underlying tones, the present study investigated the f_0 realization and the metrical pattern of disyllabic Intrinsic NT words (CT-Intrinsic NT words) and Derived NT words (CT-Derived NT words) according to their (provisional) underlying tones with CT words (CT-CT words) as the comparison. The stimuli were elicited in a neutral realization without focus as well as with corrective focus on the second syllables of the stimuli to probe the underlying tonal contours of the target tones, the metrical representation and the interaction between tonal and metrical representations of Intrinsic NT words, Derived NT words and CT words. This carefully elicited data with narrow focus "may well provide the supreme arbiter, much in the way that squishing and stretching complex objects may reveal the dimensions and elements of their structure" (Chen and Gussenhoven, 2008:725). It is expected that a comparison of NTs realized with and without focus may reveal whether the NT-bearing syllables may have underlying tonal targets and if so, what they may be:

In typical stress languages like English or Dutch, the expression of focus involves not only

the general increase in the duration and intensity of the focused components, but also the assignment of pitch accents (e.g., Gussenhoven, 1983, 2008; Ladd, 1980, 1996). In fact, it has been argued that the production and perception of focus in English rely more on f_0 and pitch cues rather than duration and other cues. However, in tonal languages like Mandarin, f_0 is primarily used to distinguish lexical meaning, and presumably, this imposes restrictions on the amount of variation in f_0 induced by other function, such as focus and intonation. It is found that in Mandarin CT-bearers, focus is usually manifested through exaggeration in the tonal contours and the syllable duration rather than through extra pitch accents that may induce distortion of the CT patterns (Chen, 2002, 2006; Chen and Gussenhoven, 2008). For instance, when CTs of low targets like T3 and T4 are focused, they are realized with lowered f_0 values, resulting in an increase in f_0 range in these two CTs.

In the existing literature, the tonal realization of NT on corrective focus has not been investigated directly. The existing studies exploring the interaction between focus and NT have studied NT words or phrases as a whole (Chen and Xu, 2006; Liu and Xu, 2007). Focus was elicited on and before the whole words or phrases containing NT rather than on the NT-bearing syllables themselves. By analyzing the f_0 contour of the whole word, both Chen and Xu (2006) and Liu and Xu (2007) found that focus directly impacts the f_0 realization of the preceding CT and hence indirectly influences the f_0 realization of NT. How the f_0 contours of Intrinsic NT and Derived NT may change when the NT-bearing syllables are on focus themselves has not been examined so far, and this change may reveal the underlying tonal representations of NT of either type.

In addition, an acoustic study conducted in Dutch, a typical stress language, suggests that

corrective focus may be an effective prosodic tool to explore lexical stress patterns (Sluijter and Heuven, 1995). Sluijter and Heuven found that speakers were able to put emphasis on unstressed syllables. More importantly, although the relative temporal structure of a word changes when the unstressed syllable carries narrow-focus accent, the original metrical structure of the word can still be seen through a significant residual effect of stress position. To be specific, Sluijter and Heuven (1995) calculated the relative syllable and rhyme duration of both VC-VV and VC-VVC words in Dutch, namely, dividing the absolute duration of each syllable or rhyme by the absolute duration of the word or the rhymes in the word. They found that all the lexically stressed but unaccented (i.e. unfocused) syllables and rhymes retained a significantly longer relative duration than the lexically unstressed and unaccented syllables and rhymes. They interpreted such duration difference as the "residual effect of abstract stress" (Sluijter and Heuven, 1995: 83) and this effect was more clearly seen in the rhymes rather than in the whole syllables, as rhymes are the relevant part for stress assignment in typical stress languages. They also added a control study which demonstrated that this residual effect cannot be explained by the differences in inherent segment duration or the lexical types, and hence concluded that the lexically stressed components "preserve some of their metrical prominence relative to the unstressed, unaccented rhymes" (Sluijter and Heuven, 1995: 84).

Therefore, by eliciting narrow focus on the 'unstressed' NT-bearing syllables, the underlying metrical structure of disyllabic NT words may become clearer.

In the following two experiments, I applied this approach (i.e., eliciting focus on NT-bearing syllables) to test the first two hypotheses.

H1: Intrinsic NT and Derived NT have different tonal representations, namely, Intrinsic NTs are underspecified for tone while Derived NTs are underlyingly specified as their corresponding CTs.

H2: Intrinsic NT and Derived NT are metrically light compared to CTs.

4.2 Experiment 1: On-focus production

4.2.1 Methodology

Participants

16 Northern Mandarin speakers (7 males, 9 females) aged between 18-30 (mean age 25.06) participated in the experiment. All participants were current students or employees at the University of Cambridge, but had completed their pre-university education in the Huabei region and reported Standard Mandarin as the main language they used in school. Moreover, no dialects that substantially differ from the Standard Mandarin (e.g., Tianjin Mandarin) were used at home[27]. The study (and the other studies in this thesis) was approved by the Ethics Committee of the Faculty of MMLL, and informed consent was obtained prior to the experiment (and the other experiments in this thesis). Some of the participants received a small fee in compensation for their participation.

Materials and Recording

8 Intrinsic NT morphemes, 8 Derived NT morphemes and 8 CT-bearing morphemes were chosen and each was combined with two different preceding CT-bearing morphemes, resulting in

[27] All the participants employed in the thesis are Mandarin-English bilinguals, since English is a compulsory subject in China since primary school. Since it is very hard to find mono-lingual Mandarin speakers under 40, in the studies reported in the present thesis, I could only control the language background of the participants to avoid impacts from strong local dialects other than the Standard Mandarin.

3 tone conditions with 48 final disyllabic stimulus words (**Table 4.2.a**; complete list of stimuli in Appendix A). Each of the four CTs occurred twice as the preceding tone in each condition and the distribution of the provisional underlying tones in the Derived NT condition and the CT condition was also balanced. In this way, the potential influence of the preceding CTs is balanced out. The relative complexity of syllables between conditions is also controlled. Their relative complexity was indexed by dividing the number of segments in the second syllable by that in the first syllable. By this measure, the average complexity in the Intrinsic NT condition was 0.78, in the Derived NT condition 0.97, and in the CT condition 0.89. *Anova* showed that they were not statistically different ($F[2]=2.58$, $p>0.05$). Since Intrinsic NT usually occurs on empty bound morphemes, I used grammatical elements like localizers and classifiers rather than notional morphemes in the Derived NT condition and the CT condition to minimize the effect of morphological status.

Table 4.2.a Examples of stimuli (number in transcription indicates tone)

Condition	Bound morpheme	IPA Transcription	Meaning of the morpheme	Two preceding morphemes	IPA Transcription	Meaning of the preceding morpheme	Stimulus	IPA Transcription
Intrinsic NT	吗	/ma/	Question marker	走	/tsou3/	walk	走吗	/tsou3 ma0/
				看	/kʰan4/	take	看吗	/kʰan4 ma0/
Derived NT	家	/tɕia (1)/	Family (like The Lees')	李	/li3/	Li (family name)	李家	/li3 tɕia (1)/
				赵	/ʈʂao4/	Zhao (family name)	赵家	/ʈʂao4 tɕia (1)/
CT	棵	/kʰɤ1/	A classifier for trees	三	/san1/	three	三棵	/san1 kʰɤ1/
				十	/ʂi2/	ten	十棵	/ʂi2 kʰɤ1/

5 native Northern Mandarin speakers who were not informed about the experiment confirmed that the Intrinsic NT morphemes used in the experiment do not have CT alternates with the same semantic meaning, while the 8 Derived NT morphemes can either be realized as NT or CTs without changing the meaning of the words, though the NT realization is preferred.

16 different disyllabic CT-CT words were added as fillers, and they were balanced in the tone combination (4 CTs × 4 CTs). The stimuli and the fillers were all recorded by a phonetically trained male 29-year-old who is a native speaker of Singaporean Mandarin. The primary aim for using a Singaporean speaker is not to test the perception of NT, but to elicit the corrective focus more naturally. The speaker was instructed to correctly pronounce each stimulus once and mispronounce the same stimulus once. Mispronunciations in the target words always occurred in the second syllable and were always tonal, namely, the NT-bearing syllables were realized with an unrelated CT and CT-bearing syllables were realized with a different CT. In contrast, fillers were mispronounced in either the first or the second syllable and the error was either tonal or segmental. The distribution of mispronounced tones in the target stimuli was also balanced.

The recording was carried out in a sound-treated room at the Phonetics Laboratory at the University of Cambridge. The intensity of all sound stimuli was scaled to 75dB using *Praat* (Boersma and Weenink, 2002). The recording of one correctly pronounced filler was damaged due to a saving error so that a total of 127 experimental stimuli were used in the experiment.

Procedure

The perception experiment took place in the same sound-treated room. Participants were told that they were helping a Singaporean Chinese to learn Standard Mandarin. In each trial, they saw a

stimulus in logographs and listened to the corresponding correct or incorrect recording. They were asked to repeat the word by putting it into one of two carrier sentences, either /tuei tu (stimulus)/ 'Yes, it reads as (stimulus)' to confirm the pronunciation or /pu tuei tu (stimulus)/ (No, it reads as (stimulus)) to correct the pronunciation. The confirming repetition constituted the no focus condition. All stimuli were pseudorandomized by (provisional) underlying tones and correctness. Participants' responses were audio recorded in the same lab using a MixPre-6 112 recorder and a Sennheiser M64 microphone with the battery module K6.

In the practice phase, participants were shown a video of the investigator and the recorder doing 10 trials, followed by 8 trials which participants had to do themselves. All the 18 practice trials had the same format as the trials in the actual experiment, but contained different CT-CT stimuli to avoid the investigator's influence on the production of NT of the participants. A 5-minute break was offered after every 30 trials which they could choose to skip. The total duration of the experiment was about 30 minutes including instructions.

Data analysis

I first calculated performance accuracy, defined as 'yes' versus 'no' responses to the correctly-pronounced words, and tested the differences between the conditions using Pearson's Chi-square test. Performance accuracy is important to include because it provides us with a rough estimate of how tolerant participants are to mispronunciations, and thus whether they intend to place corrective focus on words. Only correctly answered trials were included in further statistical analyses because only in these trials the correct focus realization was elicited.

The acoustic analyses were performed in *Praat* (Boersma and Weenink, 2021). For f_0

contours, a *Praat* script was applied to extract f_0 values (converted to semitones) of the sonorous part in the second syllable of each stimulus. The f_0 contours were time-normalized by dividing the sonorous parts into 10 equal intervals, and f_0 values were extracted at each 10% step, starting from 10% rather than 0% to reduce the influence of the preceding tone on the measurement. The last value (100%) was also excluded to reduce the effect of final creakiness, which was very common in the data, across speakers and tonal patterns. Whenever creakiness across the syllable had caused data loss over 50% in a token (i.e., fewer than 4 values were extracted), that token was excluded from the final analyses of f_0 contours. However, since T3, the low convex tone, is typically realized with creaky voice and most tokens showed more than 50% data loss, I had to analyze the percentage of data loss for this tone rather than the f_0 contours like for the other tones. The f_0 contours of different tones were graphically depicted by the average f_0 values (or percentage of data loss in case of T3) at each time point across *Tone* Condition (Intrinsic NT vs. Derived NT vs. CT), *Focus* Condition (With Focus vs. No Focus) and *(provisional) Underlying Tone* (NT vs. T1 vs. T2 vs. T3 vs. T4). The details will be presented with the results in the next section.

To test H1, two main statistical analyses were conducted.

Analysis 1: To explore the underlying tonal representation of Intrinsic NT, the changes in Intrinsic NT as a function of focus were indexed by average f_0 height and range. f_0 height was calculated as the average value of the 9 f_0 values of each target syllable and f_0 range was calculated as the difference between the minimum and maximum f_0 values. The statistical significance of the focus-induced changes was tested with an one-way *Anova*.

Analysis 2: To explore the underlying tonal representations of Derived NT, the f_0 contours of

Intrinsic NT, Derived NT from different CTs, and CTs in either *Focus* condition (With Focus vs. No Focus) were compared statistically with Euclidean distance-based discriminant analyses.

Discriminant analysis is a statistical procedure that generates mathematical models for assigning tokens to groups (Lachenbruch and Goldstein, 1979). The models are established on measurements from tokens whose group membership is known. This analysis is applicable to a wide range of problems, not only limited to the classification of unknown units to known classes, but also to evaluate the similarity of distinct populations, and to summarize the differentiation between groups (Irigoien et al, 2016). Euclidean distance-based discriminant analysis was chosen because it performs as well as or superior to linear discriminant analysis in many practical situations (Irigoien et al, 2016; Marco et al, 1987), for instance, the class-unbalanced situation of the present case (i.e., the number of tokens contained in each class is unbalanced). The Euclidean distance-based discriminant analyses were conducted following three steps.

Step 1: Choose the measures for the models. In the present study, the measurements are f_0 height, f_0 range, percentage of data loss and the coefficients of quaduration ratioatic equations, and the known group memberships are Intrinsic NT and CTs. f_0 height and range are calculated as in Analysis 1. The coefficients of quaduration ratioatic equations are employed because they can sketch the tonal contours in a more refined manner, since phonologically different tones can still show similarities in f_0 height and range (Andruski and Costello, 2004). The two coefficients of quaduration ratioatic equations were measured by establishing a quaduration ratioatic equation for the f_0 values of each token. A quaduration ratioatic equation (i.e., $y=a + bt+ct^2$, where y is the f_0 value and t is the time point on the normalized scale from 1 to 9, and the coefficients b- and c- provide a relatively accurate description of the contour) was estimated for each f_0 contour (**Table**

4.2.b).

Table 4.2.b The relationship between values of the b- and c-coefficient and resulting curve shapes (adapted from Andruski and Costello, 2004: Table 1)

Coefficient	**Curve Shape**
b<0 and c>0	Concave and Falling
b< 0 and c=0	Straight and Falling (not quaduration ratioic)
b≤0 and c<0	Convex and Falling
b>0 and c<0	Rising then Falling
b≥0 and c≥0	Level or Rising

Step 2: Establish the groups of which memberships are known. Six groups were established based on *Underlying Tone* and *Focus:* Intrinsic NT and T1 with focus, Intrinsic NT and T1 with no focus, Intrinsic NT and T2 with focus, Intrinsic NT and T2 with no focus, Intrinsic NT and T4 with focus and Intrinsic NT and T4 with no focus. I then classified the Derived NT by the corresponding models and calculated the percentage of classification. All six models could make classifications with accuracy over 60%. The analyses were conducted through the WeDiBaDis package (Irigoien et al, 2016) in R (R core team, 2020). Some Kappa values of the present models were low, mismatched with the high percentage of agreement due to the substantial imbalance in the table's marginal totals (i.e., Kappa paradox, see Feinstein and Cicchetti, 1990). Therefore, I evaluated the classification reliability using Gwet AC1, a weighted measurement of agreement, with the irrCAC package (Gwet, 2019). The Gwet values of the four models established on Intrinsic NT and T1 and Intrinsic NT and T2 fell between 0.6 and 0.8, indicating good agreement of these models with the data (Landis and Koch, 1977; Gwet, 2019). However, the Gwet values of the two models established on Intrinsic NT and T4 only fell between 0.5 and 0.6, indicating good

moderate agreement of these models with the data. This slightly lower agreement was due to the similar falling contours Intrinsic NT and T4 have.

Step 3: Classify the Derived NT data according to its underlying tones. The Derived NT results were separated into 8 groups: Derived NT from T1 with focus, Derived NT from T1 with no focus Derived NT from T2 with focus, Derived NT from T2 with no focus, Derived NT from T3 with focus, Derived NT from T3 with no focus, Derived NT from T4 with focus and Derived NT from T4 with no focus. Except for the two groups with Derived NT from T3, the other 6 groups were classified according to the models established in Step 2 using the WeDiBaDis package. Regarding Derived NT from T3, the average percentage of data loss over the 9 f_0 values was calculated for this tone, as well as for Intrinsic NT and T3. The effects of *Tone Condition* and *Focus* on the average percentage of data loss were tested by *Anova*.

To test H2, the metrical prominence of the second syllable was examined by measuring duration and intensity. The change in relative duration was indexed by duration ratio (= Duration of the 2nd syllable/ Duration of the 1st syllable) and relative intensity by intensity ratio (= Average intensity of the sonorous part of the 2nd syllable/ Average intensity of the sonorous part of the 1st syllable). The absolute duration and intensity of the two syllables were also calculated to illustrate the dynamic between them under focus. The effects of *Tone Condition* and *Focus* on duration and intensity ratio of the syllables as well as the absolute duration and intensity were evaluated by Linear mixed effects (LME) models using lmer in the lmerTest package (Kuznetsova et al, 2017) in R (R core team, 2020). Log transformation was done to ensure normal distribution of the data. Compared to the relative duration Sluijter and Heuven (1995) used, the ratios can more directly reflect the relative length between the NT-bearers and the preceding CT-bearers, and enable

cross-condition comparisons. Nevertheless, the relative duration of the preceding CT-bearing syllables in *With Focus* condition (= Duration of the 1st syllable/ Duration of the word) was also calculated and tested against the relative duration of the second (NT-bearing) syllables in *No Focus* condition (= Duration of the 2nd syllable/ Duration of the word) using T-test within each *Tone Condition* as Sluijter and Heuven did for a direct comparison with their results. In the present case, the preceding CT-bearing syllables in *With Focus* condition are the equivalent of the lexically stressed but unaccented syllables in Sluijter and Heuven (1995) while the second (NT-bearing) syllables in *No Focus* condition are the equivalent of the unstressed but unaccented syllables.

I selected the optimal fixed structure by using stepwise comparisons from the most complex effect (two-way interaction) to the simplest (main effect) and the optimal random effect structure according to the smallest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). The *ANOVA* served to compare different models to determine whether excluding factors from the analysis led to a better fit (Field, Miles, and Field, 2012). The details of the model will be presented with the results. Tukey post-hoc comparisons were carried out with the lsmeans function in lsmeans (Lenth, 2016).

Predictions

P1: If the data support H1, I expect to find

a) the f_0 of Intrinsic NT is changed significantly by focus,

b) Derived NTs show large similarities to the CTs they derived from.

P2: If the data support H2, I expect to find the duration and intensity ratios of both Intrinsic

NT and Derived NT words are not increased by focus.

4.2.2 Results

Performance Accuracy

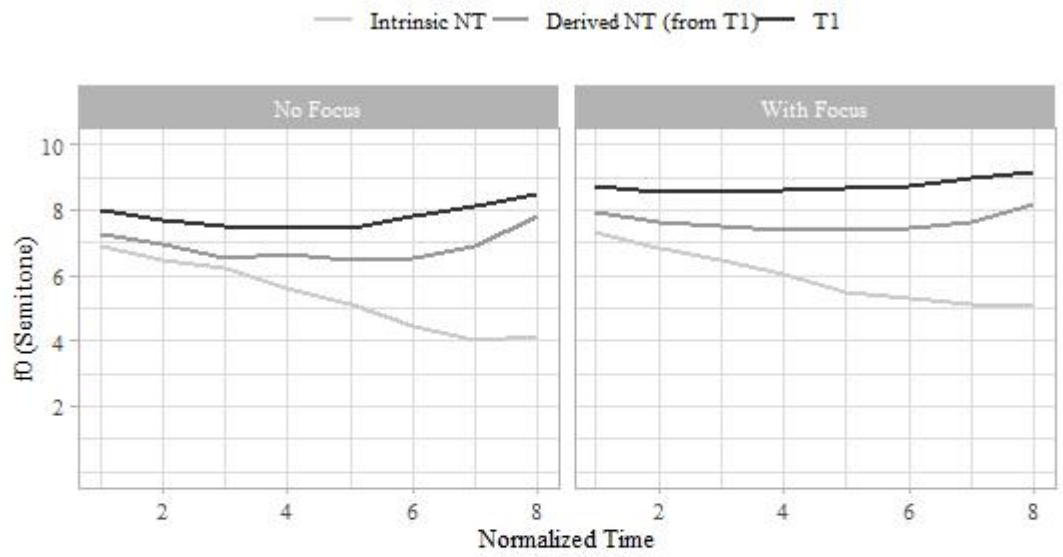
Performance accuracy measures the percentage of tokens that were correctly identified as mispronunciations or correct pronunciations. Results show that the effect of *Tone Condition* on performance accuracy was statistically significant [Chi-square test: X^2 (2, N = 4197) = 245.28, p <0.001]. Participants demonstrated the highest accuracy in *CT* tokens (98%), followed by the Derived NT condition (92%), with the lowest accuracy in the Intrinsic NT condition (84%). However, a closer analysis by focus showed that only *Intrinsic NT With Focus* words had a low performance accuracy (64.84%) while the accuracy in the unfocused condition is above 90%. In other words, participants tended to incorrectly accept the mispronunciation of Intrinsic NT words as CTs, but correctly rejected all other mispronunciations.

f_0 Contour

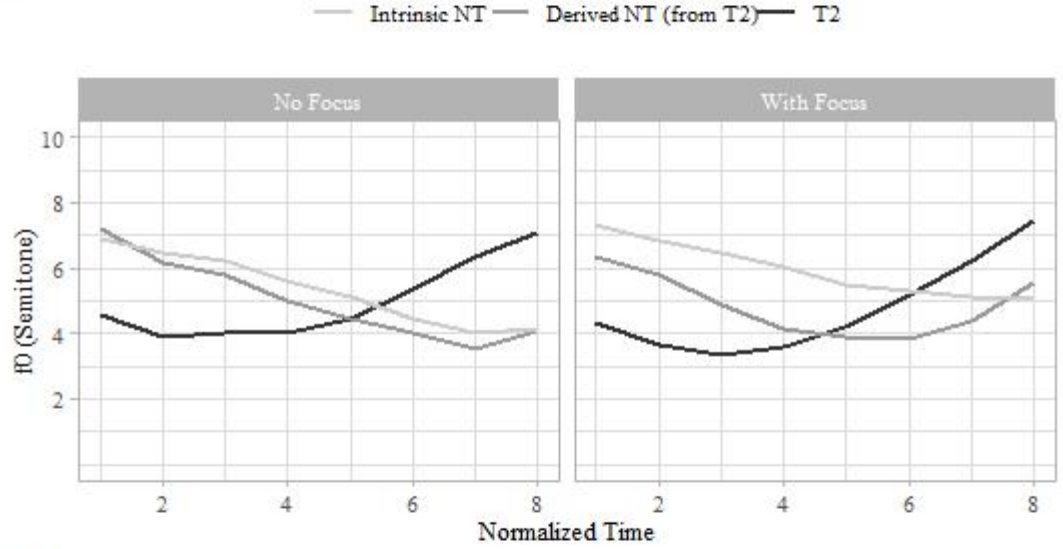
CTs were realized with somewhat hyper-articulated f_0 contours when under focus. Corrective focus led to an expansion of the f_0 span in T1, T2 and T4 **(Figure 4.2.a** and **Table 4.2.c**) as well as a higher percentage of data loss in the peak of creakiness in T3[28] (**Figure 4.2.b**). In contrast, Intrinsic NT was generally realized with a falling contour, focused or not (**Figure 4.2.a** and **Table 4.2.c**).

[28] As mentioned in Section 4.2.1, T3 is associated with creakiness and hence has too much data loss. Correspondingly, the graphic f_0 contours of T3 and Derived NT from T3 were not accurate.

I


Intrinsic NT
Derived NT (from T1)
T1
No Focus
With Focus
fD (Semitone)
Normalized Time

II


Intrinsic NT
Derived NT (from T2)
T2
No Focus
With Focus
fD (Semitone)
Normalized Time

III

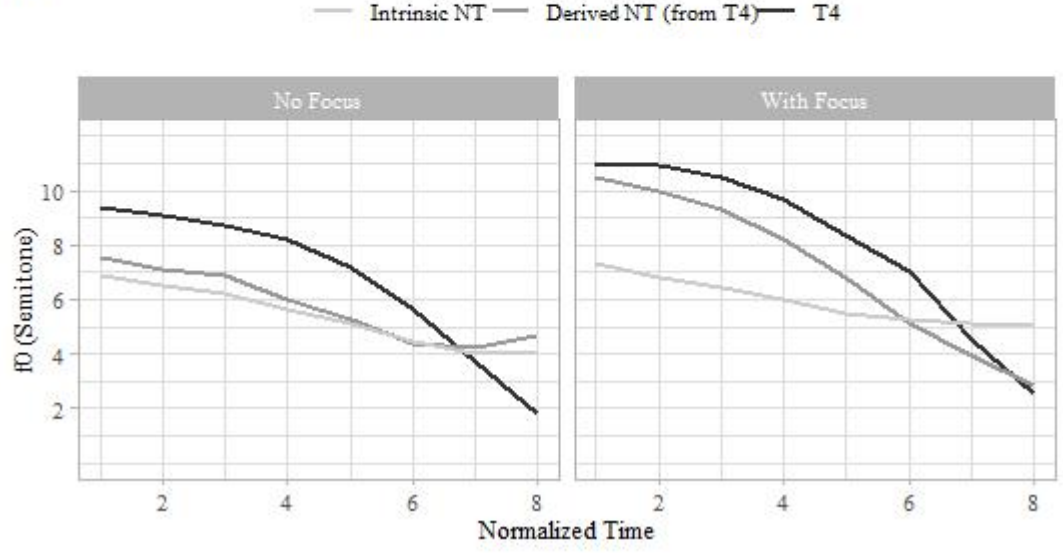

Intrinsic NT
Derived NT (from T4)
T4
No Focus
With Focus
fD (Semitone)
Normalized Time

Figure 4.2.a f_0 contours of Intrinsic NT, Derived NTs and CTs with or without focus (T3 excluded for creakiness)

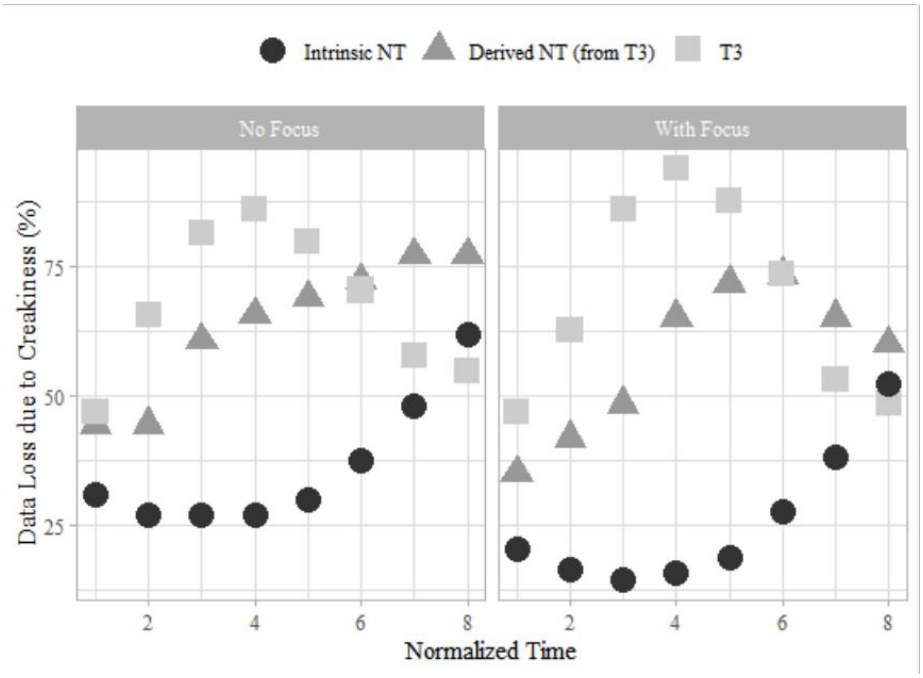


Figure 4.2.b Data loss due to creakiness of Intrinsic NT, Derived NT from T3 and T3 with and without focus (peaks indicate very low voice)

Anova test showed that corrective focus did not trigger any significant change in the average f_0 height or range in Intrinsic NT condition.

Table 4.2.c Average f_0 height and range by *Tone Condition* and *Focus*

Focus	**Focus**	**Underlying Tone**	**f_0 Height**		**f_0 Range**	
			Average (Semitones)	SE	**Average (Semitones)**	SE
Intrinsic NT	No Focus	0	5.53	0.59	3.9	0.33
	With Focus	0	5.97	0.57	4.12	0.35
Derived NT	No Focus	1	7.06	0.6	1.53	0.11
	With Focus	1	7.99	0.66	1.13	0.09
	No Focus	2	5.16	0.68	4.59	0.49
	With Focus	2	4.85	0.73	5.14	0.44
	No Focus	4	6.12	0.74	4.47	0.42
	With Focus	4	7.87	0.67	6.67	0.49
CT	No Focus	1	7.64	0.58	1.22	0.08
	With Focus	1	8.69	0.68	1.27	0.23
	No Focus	2	4.89	0.62	3.76	0.25
	With Focus	2	4.73	0.6	4.41	0.26

	No Focus	4	7.2	0.63	5.89	0.44
	With Focus	4	8.63	0.63	7.05	0.46

It can be observed in **Figure 4.2.a** that without focus, Derived NTs were more like Intrinsic NT in f_0 contours, but when corrective focus was added, they became similar to the CTs they derived from, except for NT derived from T1 (**Figure 4.2.a** I): with and without focus, NT derived from T1 demonstrated a high and level contour just like T1. The Euclidean distance-based discriminant analyses supported these observations (**Table 4.2.d**). Without focus, the majority of Derived NT from T2 or T4 carrying focus showed larger similarities to Intrinsic NT than to T2 or T4. When carrying focus, the majority of Derived NT from T2 or T4 resembled the corresponding CTs. For Derived NT from T1, there were always more tokens similar to T1 than to Intrinsic NT regardless of the focus status. However, the focus did increase the percentage of Derived NT classified as T1. In summary, focus increased the similarity between Derived NT and T1, T2, and T4.

Table 4.2.d Classification rates and posterior probabilities of Derived NT tokens (Posterior probabilities index how likely the Derived NT tokens are classified as the corresponding CTs or Intrinsic NT) .

Provisional Underlying tones	**Focus Status**	**Percentage of Classification**		**Total Number**	**Posterior Probabilities**	
		CT	Intrinsic NT		CT	Intrinsic NT
T1	No Focus	72.58%	27.42%	62	79.28%	20.72%
	With Focus	55.32%	44.68%	47	89.39%	10.61%
T2	No Focus	57.14%	42.86%	49	70.71%	29.29%
	With Focus	0.00%	100.00%	47	36.16%	63.84%
T4	No Focus	68.52%	31.48%	50	65.76%	34.24%
	With Focus	32.00%	68.00%	54	28.22%	71.78%

The data loss patterns in **Figure 4.2.b** showed that without focus, Derived NT from T3 was

more like Intrinsic NT which showed most loss at the end of the syllable, whereas when carrying corrective focus, Derived NT from T3 had the highest data loss percentage in the middle as T3. Results of the *Anova* showed that the effect of *Tone Condition* on average data loss was significant [$F(2, 12629.40) = 30.37$, $p < 0.0001$]. Post hoc comparisons showed that the differences between Intrinsic NT (30.74%) and the other two conditions were significant ($p < 0.0001$), but the difference between Derived NT (60.61%) and T3 (68.36%) was not.

Relative Duration

Regardless of whether NT was on focus or not, the average duration ratio of Intrinsic NT words was smaller than average duration ratio of Derived NT words than that of CT words.

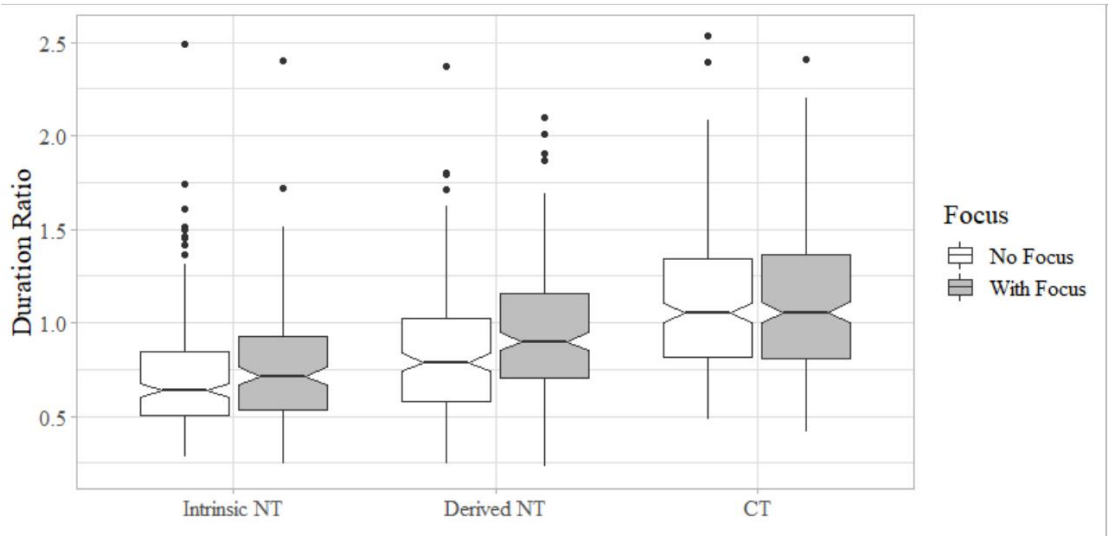


Figure 4.2.c Duration ratio in each tone condition with or without corrective focus

The LME model shows that *Tone Condition*, *Focus* and their interaction all significantly affected duration ratio of the target words (**Figure 4.2.e**). Pairwise comparisons showed that the increase in duration ratio elicited by corrective focus is only significant in the Derived NT condition ($p < 0.0001$) but not in the Intrinsic NT or CT condition. This means that the relative duration increase in Intrinsic NT words and CT words was highly constrained, and that they were

influenced significantly less by the effect of corrective focus. In addition, the cross-condition difference was always significant between CT and Intrinsic NT (both *p*s <0.001). However, only when not carrying focus was the duration ratio of Derived NT words significantly lower than that of CT words (p <0.05). These results show that in terms of relative duration, the two types of NT did not differ significantly from each other, but that Intrinsic NT always behaved like a distinctive category from CTs, while Derived NT was somewhere in the middle; when carrying focus, the boundary between Derived NTs and CTs was blurred.

Table 4.2.e Linear mixed-effects model of the effects of *Tone Condition*, *Focus* and their interaction on duration ratio

Final model Analysis	Duration Ratio~ Tone Condition + Focus + Tone Condition: Focus + (1\Token)			
	SS	df	F	*p*
Tone Condition	0.20836	2	12.3497	<0.0001***
Focus	0.17827	1	21.1325	<0.0001***
Condition: Focus	0.10195	2	6.0429	0.002***

Significance levels * =.05 ** = .01 *** = .001

When examined closely, the stability in the relative duration pattern of Intrinsic NT across the focus and non-focus conditions resulted from the significant and synchronous increase in the duration of both syllables. The stability found in CTs, however, comes from the restricted increase in duration of the second syllable (**Figure 4.2.d** and **Figure 4.2.e**).

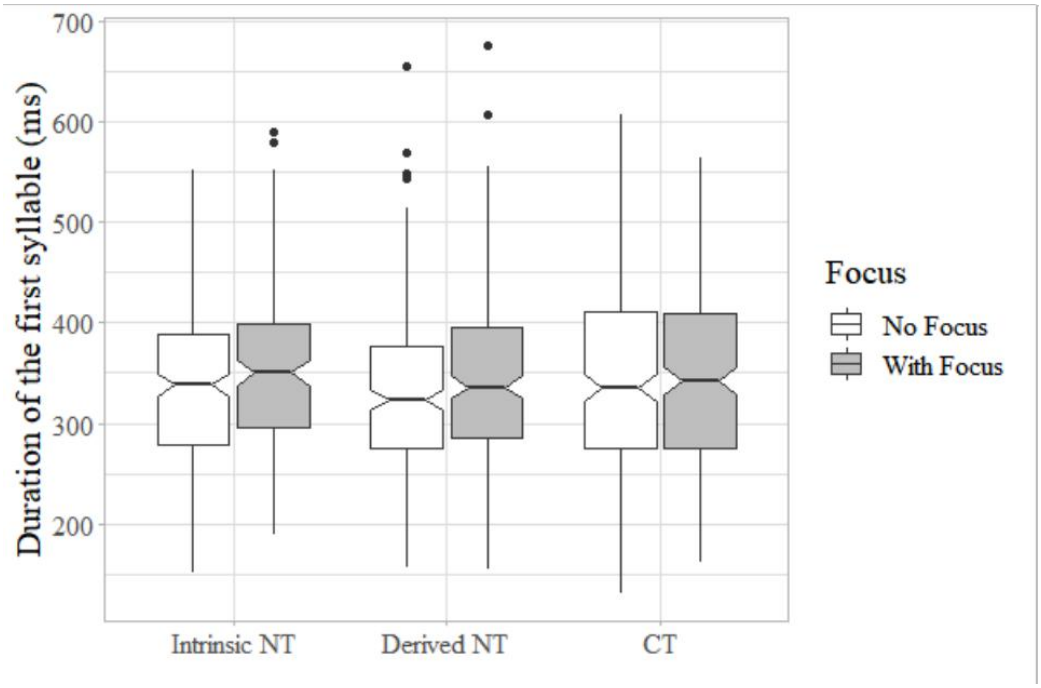


Figure 4.2.d Duration of the 1st syllables in each tone condition with or without corrective focus

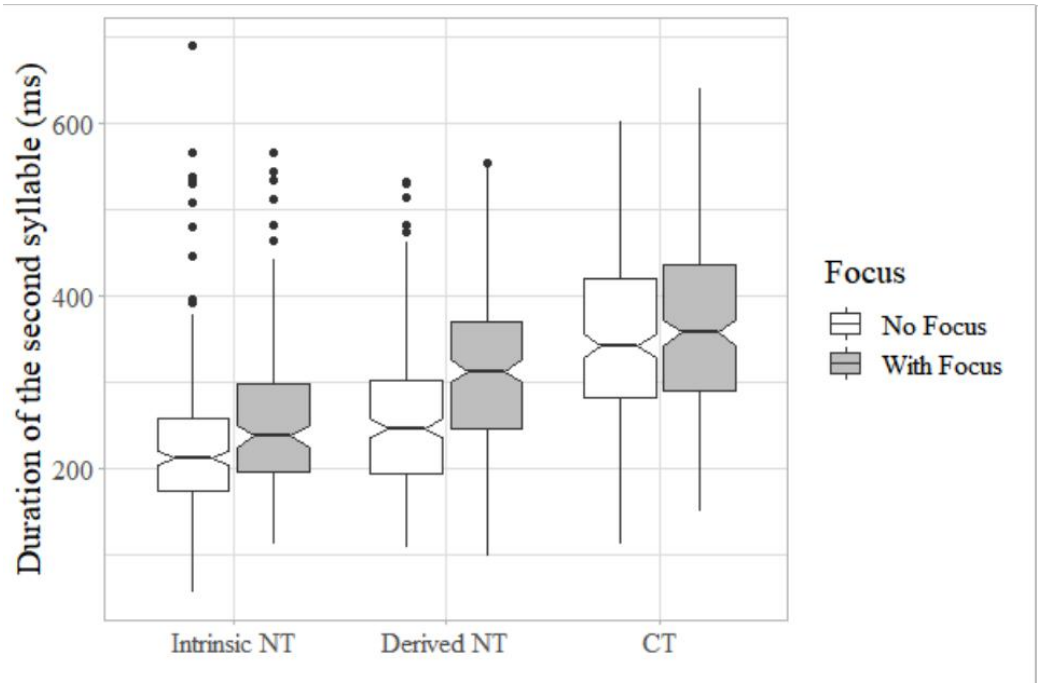


Figure 4.2.e Duration of the 2nd syllables in each tone condition with or without corrective focus

The LME model shows that *Tone Condition*, *Focus* and their interaction all significantly affected absolute duration of the second syllables, but only *Focus* affected absolute duration of the first syllables (**Table 4.2.f**). Post-hoc comparisons further demonstrated that in Intrinsic NT and

Derived NT words, corrective focus induced significant lengthening in both syllables (p <0.01), while in CT words, it only induced significant lengthening in the first syllable.

Table 4.2.f Linear mixed-effects models on the absolute duration of the two syllables

Final model Analysis	The 1st Syllable Duration ~ Tone Condition + Focus + Tone Condition: Focus + (1\Token)			
	SS	df	F	*p*
Focus	538.61	1	19.477	<0.0001***
Final model Analysis	The 2nd Syllable Duration~ Tone Condition + Focus + Tone Condition: Focus + (1\Token)			
	SS	df	F	*p*
Tone Condition	280.68	2	30.249	<0.0001***
Focus	288.28	1	62.134	<0.0001***
Condition: Focus	103.46	2	11.15	<0.005***

Significance levels * =.05 ** = .01 *** = .001

Table 4.2.g Duration ratio, relative duration and absolute duration of the syllables by *Tone Condition* and *Focus*

Focus	Duration Ratio		The 1st Duration				The 2nd Duration			
	Average	SE	Average (ms)	SE	Relative (%)	SE	Average (ms)	SE	Relative (%)	SE
No Focus	0.70	0.02	338.42	2.85	60.11	0.66	227.72	2.89	36.68	0.98
With Focus	0.76	0.02	351.90	2.78	58.22	0.57	258.39	3.03	39.08	0.83
No Focus	0.82	0.02	329.37	2.85	56.55	0.62	256.54	2.94	40.38	0.82
With Focus	0.94	0.02	346.14	2.89	52.95	0.58	311.43	3.03	51.35	1.03
No Focus	1.10	0.02	344.22	3.12	49.18	0.54	354.86	3.12	56.61	1.12
With Focus	1.11	0.02	346.25	2.98	48.82	0.54	364.02	3.11	58.67	1.16

T-tests comparing the relative duration of the 1st CT-bearing syllables in *With Focus* and the 2nd (NT-bearing) syllables in *No Focus* showed that the relative duration of the 1st CT-bearing syllables was significantly longer in both Intrinsic NT [F (1, 3.04) = 237.42, p <0.001] and Derived NT conditions [F (1, 0.03) = 1.80, p <0.001]; in contrast, the 1st syllables CTs words in *With Focus* were significantly shorter than the 2nd syllables in *No Focus* [F (1, 1.18) = 56.74, p <0.001].

Relative Intensity

Unlike the findings for the average duration ratio, the average intensity ratio was less than one in all three tone conditions, indicating an on average 'weaker' second syllable, in this respect, in all words regardless of focus status (**Figure 4.2.f**).

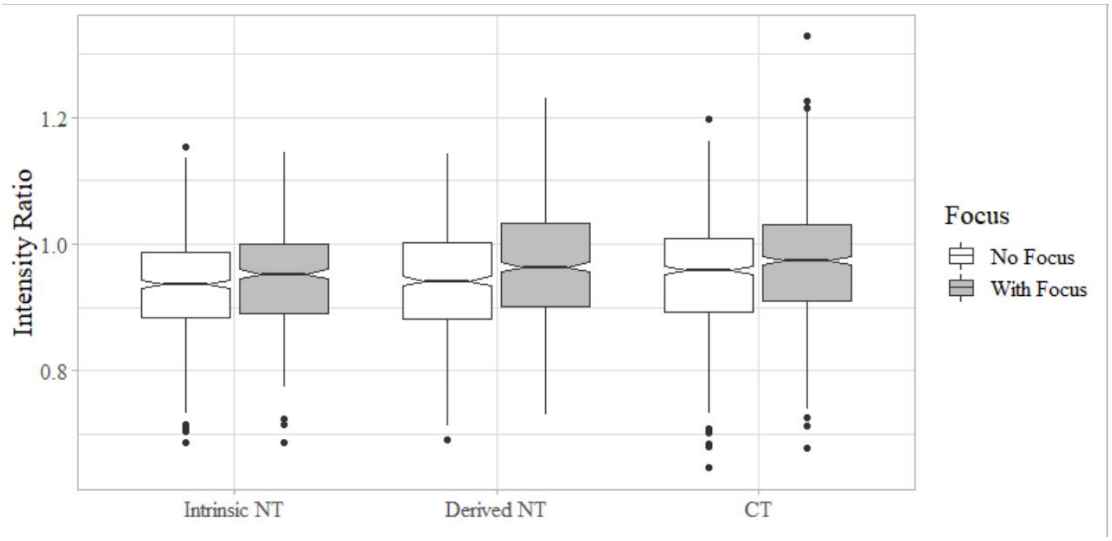


Figure 4.2.f Intensity ratio in each tone condition with or without corrective focus

An LME with intensity ratio as the independent variable showed that only the effect of *Focus* reached significance (**Table 4.2.h**). In other words, target syllables carrying any type of tone all significantly increased in the relative intensity when carrying corrective focus.

Table 4.2.h Linear mixed-effects model on intensity ratio

Final model Analysis	Intensity Ratio ~ Tone Condition + Focus + (1\Token)

	SS	df	F	*p*
Tone Condition	0.001586	2	0.248	0.7814
Focus	0.146212	1	45.743	<0.0001***

Significance levels * =.05 ** = .001 *** = .0001

However, no main effect for tone category was found. Similarly, only *Focus* increased significantly the absolute intensity of the two syllables in the words, regardless the *Tone* conditions (**Table 4.2.i**).

Table 4.2.i Linear mixed-effects models on the absolute intensity of the two syllables

Final model Analysis	The 1st Syllable Intensity~ Tone Condition + Focus + Tone Condition: Focus + (1\Token)			
	SS	df	F	*p*
Focus	86.125	1	22.188	<0.0001***
Final model Analysis	The 2nd Syllable Intensity~ Tone Condition + Focus + Tone Condition: Focus + (1\Token)			
	SS	df	F	*p*
Focus	0.053861	1	19.477	<0.0001***

Significance levels * =.05 ** = .01 *** = .001

4.2.3 Discussion

This experiment investigated the acoustic realizations of Intrinsic NT and Derived NT in comparison with CTs, with and without corrective focus.

With regard to the tonal representations of Intrinsic NT, the prediction P1 a) is supported as no significant increase was found in the average f_0 range and height of Intrinsic NT between unfocused and focused situations. NT did demonstrate a general falling contour regardless of focus situation, but the fall was not comparable to the falling tone T4 ending with a low target.

The data loss under focus in Intrinsic NT caused by creakiness was not comparable to the low convex tone T3 either. The focus has not steepened the falling contour of Intrinsic NT. In fact, it has not brought much change in the f_0 or the creakiness-triggered data loss in Intrinsic NT. Therefore, I argue against all accounts that assign a low tonal representation or a low target to Intrinsic NT, whether it is the target of NT itself (e.g., Lin, 2006) or a boundary tone (e.g., Li, 2003). Intrinsic NT may have a mid-level phonetic target (Chen and Xu, 2006; Wang, 1997). However, since corrective focus has not significantly flattened the falling contour of Intrinsic NT, it is unlikely for this target to be phonologized.

With regard to the tonal representations of Derived NT, predictions P1 b) is supported, as the f_0 contours of Derived NTs varied according to the CTs they derived from as well as the focus status. When not carrying focus, the similarities between Derived NTs and CTs were found mainly on the Derived NT from T1 and T3. In terms of NT derived from T1, it still demonstrated a level contour similar to high-level T1 instead of the falling Intrinsic NT regardless of the focus status. Even when not carrying focus, the Euclidean distance-based discriminant analysis classified more Derived NT from T1 as T1 rather than Intrinsic NT. In terms of Derived NT from T3, its average percentage of data loss in Derived NT from T3 was as high as that in T3, significantly higher than that in Intrinsic NT, regardless of the focus status. Corrective focus, furthermore, enlarged the similarities between Derived NT and CTs. The Euclidean distance-based discriminant analyses showed that the proportion of Derived NT tokens were classified as the corresponding T1, T2 and T4 rather than Intrinsic NT when carrying focus. I would like to point out that it may not be that Derived NTs from T1 and T3 are special as they retain more contour feature of the underlying CTs, because NT and T4 share a similar falling contour (Section 3.1 and 3.3.1). Instead, it is the

Derived NT from T2 is special as it does not retain much contour feature of the rising tone at all, probably due to the fact that rising pitch is marked universally compared to the level or falling pitch contour.

Overall, the results on f_0 realization support H1, suggesting that Intrinsic NTs are underspecified for tone in the underlying representation while Derived NTs are underlyingly specified as the corresponding CTs.

With regard to the metrical structure of Intrinsic NT and Derived NT, P2 is only partially supported as focus did not trigger a significant increase in the duration ratio of Intrinsic NT whereas the duration ratio of Derived NT increased when Derived NT syllables were on focus.

The significant difference in the duration ratio between Intrinsic NT words and CT words in either focus status demonstrated that disyllabic Intrinsic NT words show a distinctive heavy-light pattern from the heavy-heavy CT. In addition, the not significant increase in the duration ratio brought by corrective focus demonstrates the stability of this heavy-light pattern and hence supports the hypothesis that syllables bearing Intrinsic NT are metrically light. It is worth noting that in Intrinsic NT words, corrective focus still triggered a significant increase in the absolute duration of the NT-bearing syllables. It was the simultaneous significant lengthening in the preceding CT-bearing syllables that kept the heavy-light pattern of Intrinsic NT words. In contrast, the CT-bearing syllables preceding Derived NT did not increase as much so that the heavy-light pattern of Derived NT words was not as stable.

When not carrying focus, the duration ratio of disyllabic Derived NT words was not significantly different from the ratio of the Intrinsic NT words but from that of CT words.

However, when carrying focus, the differences between Derived NT words and CT words also became not significant because the duration ratio of Derived NT words was raised significantly by corrective focus, making the metrical weight of Derived NT inconclusive. On the one hand, Derived NT may not be as light as Intrinsic NT since the lengthening of syllables bearing Derived NT under focus was much larger than that of Intrinsic NT syllables. On the other hand, evidence suggesting that Derived NT has lighter metrical weight than the CT is also found. The average duration ratio of Derived NT words with NT on focus was still smaller than 1, and it was not significantly different from the ratio of Intrinsic NT words either, indicating a longer first syllable in disyllabic Derived NT words in general, unlike the CT words. In particular, similar residuals of the original temporal structure to Sluijter and Heuven's studies (1995) were also observed here, namely, the preceding CT-bearing syllables without focus in either Intrinsic NT or Derived NT condition still had longer relative duration than the NT-bearing syllables without focus. The metrical weight of Derived NT will be further investigated in the following experiments.

In all three tone conditions, the initial syllables had higher intensity than the second syllable, though the second syllables had a larger increase in intensity than the first syllables when focused. Therefore, intensity results cannot shed light on differences between the different tones.

It is worth further attention that the present experiment indicates that final- and focus-induced lengthening seem to be additive in syllables bearing Intrinsic and Derived NT but not in CT-bearing syllables, echoing the previous findings on the non-additive lengthening in Mandarin CT-bearing syllables (Chen, 2006). The lack of restriction on lengthening found in either NT-bearing syllables may be related to the under-realized or even absent tone targets. This point will be re-visited with the results of Experiment 2.

To summarize, Production Experiment 1 demonstrated that Intrinsic NT, Derived NT and CTs are realized with acoustic properties that systematically differ from each other. In both their f_0 and durational pattern, Derived NTs seem to be at a middle position between Intrinsic NT and CTs. The clear differences between the Intrinsic and Derived NT conditions revealed under corrective focus provide strong supporting evidence that their underlying tonal representations are different. However, with regard to the metrical representations of Derived NT, whether Derived NTs are metrically light as Intrinsic NT requires further investigation (Experiment 2 and 5).

4.3 Experiment 2: On focus production with emphasis

In this experiment, two 'degrees' of focus, corrective focus and corrective focus with emphasis, were elicited to bring more variation to the f0 realization and relative duration of Intrinsic NT, Derived NT and CT-CT words to test to what extent the findings of Experiment 1 are systematic, in particular, whether the 'middle position' of Derived NT between Intrinsic NT and CTs in terms of f0 and durational pattern under focus is phonological. In other words, when given more emphasis, will the f0 and durational differences between Derived NTs and the corresponding CTs disappear completely?

Previous studies indicate that the distinction between the unfocused and focused is phonological whereas the distinction between the focused and the focused with more emphasis is phonetic. Focus in typical stress languages like English is implemented through assigning pitch accents to the components on focus (Section 4.1), while the focus of higher degree is implemented through f_0 expansion of the pitch accents rather than introducing new pitch accents (e.g., Gussenhoven and Teeuw, 2007; Ladd and Morton, 1997). In Mandarin CT words, the f_0 expansion from focused to focused with emphasis is very limited compared to the expansion from unfocused

to focused due to the priority of keeping the shapes of CTs, but the absolute duration of the focused syllables increases steadily from the unfocused to the focused and to the focused with emphasis (Chen and Gussenhoven, 2008). Therefore, investigating whether and how the f_0 realization of Mandarin NTs changes when given more emphasis will help to reveal whether the tonal and durational differences between Intrinsic NT, Derived NT and CT are phonological or phonetic, and hence shed more light on the underlying representations of the two types of NT.

4.3.1 Methodology

Participants

8 Northern Mandarin speakers (2 males, 6 females) aged between 18-24 (mean age 22.00) participated in the experiment. All participants were undergraduate or post-graduate students at the Shanghai Jiao Tong University[29]. Like participants of Experiment 1, they all completed their pre-university education in the Huabei region and reported Standard Mandarin as the main language they used in school and at home. Informed consent was obtained prior to the experiment.

Materials

The same materials used in Experiment 1 were used here. The stimuli were equally split into two sets, set A and set B.

Procedure

The production experiment took place in quiet rooms in Shanghai Jiao Tong University. Like

[29] Although some experiments in the thesis were conducted in Shanghai Jiao Tong University, the language background of the participants has been carefully controlled to avoid the influence of the local dialects. None of the participants had lived in Shanghai for more than three years as they all completed their pre-university education in the Huabei region. In addition, the experiments were conducted on the campus in Minhang district, which is a semi-closed campus far away from the city centre. The teaching staff and students there also have a highly diverse language background. Therefore, there is not a main dialect used on campus and the participants were unlikely to be influenced strongly by the local Shanghai Wu dialect.

in Experiment 1, the participants were told that they were helping a Singaporean to learn Standard Mandarin. In each trial, they will first see a stimulus in logographs and listen to the recording of an incorrect pronunciation. They were asked to correct the pronunciation using the carrier sentences /pu tuei tu (stimulus)/ 'No, it reads as (stimulus)' first. Then, the recording of the incorrect pronunciation will be played again for half of the trials and the correct pronunciation played again for the other half to imitate the shadowing repetition in a real teaching scenario. The participants were asked to confirm the pronunciation by putting the word into the carrier sentence /tuei tu (stimulus)/ 'Yes, it reads as (stimulus)' or to correct the word again by putting it into the carrier sentence /pu tuei tu (stimulus)/ 'No, it reads as (stimulus)'. The instructor made it clear to the participants that the first recordings they heard always needed to be corrected; in addition, if the second recording was not different from the first one, it needed to be corrected as well.

Half of the participants were asked to correct again the stimuli in set A and the others to correct again the stimuli in set B. In other words, *With Emphasis* and *No focus* were elicited on two different sets of stimuli but *With focus* on both sets of stimuli for each participant. Participants' responses were recorded using a Zoom H1 handy recorder at 96000Hz/26Bit.

In the practice phase, the participants were shown a video of the investigator doing 8 trials, followed by 6 practice trials they completed themselves. All practice trials had the same format but contained different CT-CT stimuli as the trials in the actual experiment. After 32 trials participants were offered a 10-minute break. The total duration of the experiment was about 30 minutes.

Data Analysis

Very few incorrectly answered trials were found in the present experiment compared to Experiment 1 (13 trials out of 768), and they were excluded as the focus was not correctly elicited in those conditions.

The acoustic analyses were performed in *Praat* (Boersma and Weenink, 2021). The time-normalization of tones was conducted in the same way as Experiment 1. The focus-induced expansion in f_0 in Intrinsic NT, Derived NT and CT conditions were measured by average f_0 height and range. Average height and range were calculated as in Experiment 1. Due to the creakiness in T3 and Derived NT from T3, the f_0 height and range of these two tones were not analyzed.

The effects of *Focus* on the f_0 height and range in Intrinsic NT condition were tested using *Anova*. The effects of *Tone Condition* (Derived NT vs. CT), *Focus* (No Focus vs. With Focus vs. With Emphasis) and *Underlying Tone* (T1 vs. T2 vs. T4) on the f_0 height and range in the other two conditions were evaluated by an LME model as in Experiment 1. Intrinsic NT was analyzed separately because it had only one underlying tone rather than 3.

Relative duration was still indexed by duration ratio. The effects of *Tone* Condition (Intrinsic NT vs. Derived NT vs. CT) and *Focus* Status on duration ratio and the absolute duration of the syllables were evaluated by an LME model as in Experiment 1, and log transformation was done to ensure the normal distribution of the data.

Predictions

Based on results of Experiment 1, I make further predictions as follows.

P1: In Intrinsic NT, the increases in f_0 height and range from the unfocused to the focused to the focused with emphasis are all marginal due to the lack of a strong underlying tone target.

P2: In Derived NT, the increase in f_0 height and range (or the average percent of data loss) from the unfocused to the focused is significant but the increases from the focused to the focused with emphasis are marginal, in line with previous findings on CTs.

P3: In Intrinsic NT, the increases in duration ratio from the unfocused to the focused to the focused with emphasis are all marginal due to its underlying metrical structure.

P4: In Derived NT, the increases in duration ratio from the unfocused to the focused are significant but the increases from the focused to the focused with emphasis are marginal; the average duration ratio is still under 1 when Derived NT carries focus with emphasis.

4.3.2 Results

f_0 Expansion

The results on f_0 expansion echoed the findings of Experiment 1. Corrective focus triggered f_0 expansion in all *Tone* conditions, regardless of the underlying tones.

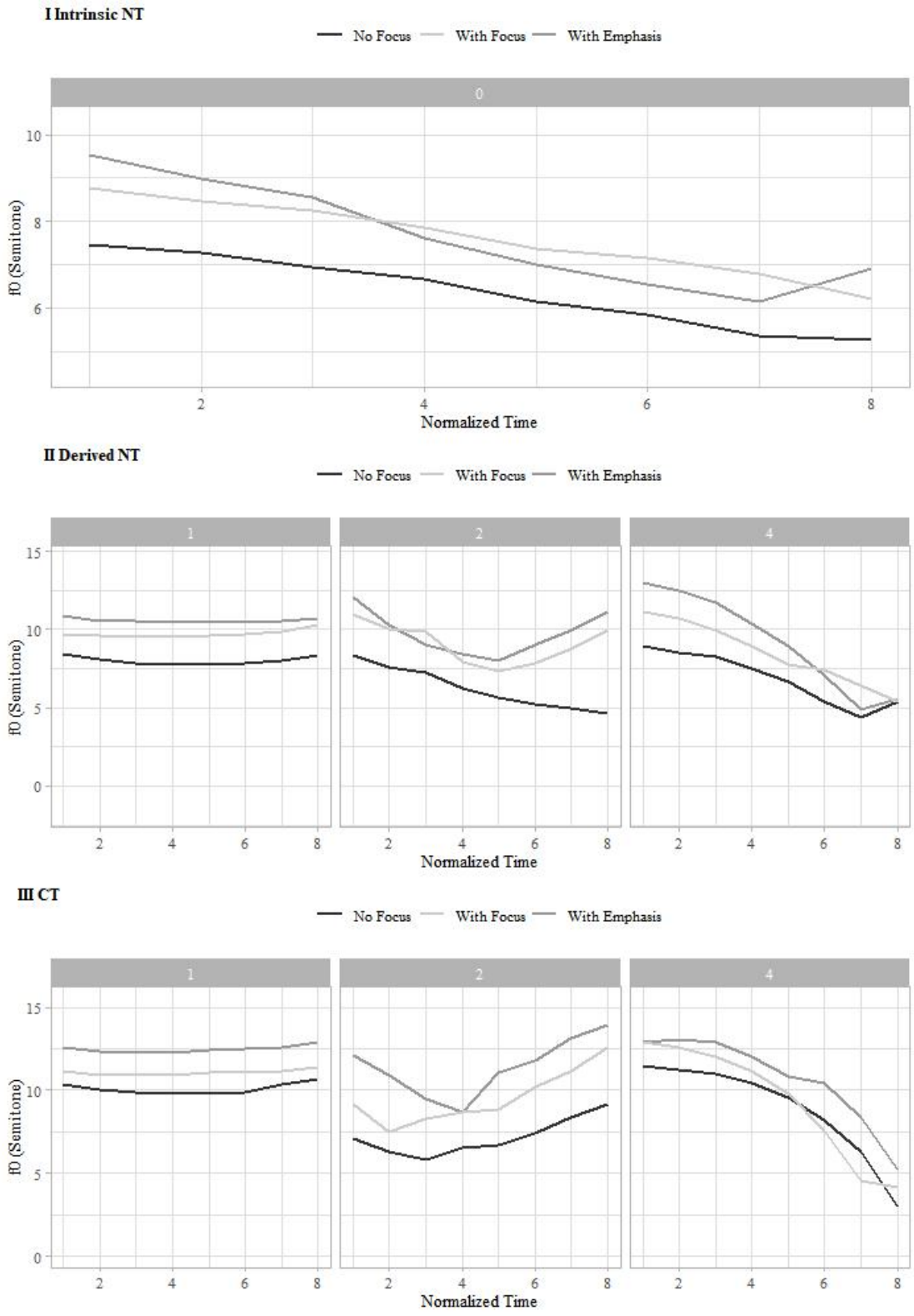


Figure 4.3.a f_0 contours of each tone condition by underlying tone and focus (numbers in subtitles indicate the underlying tones)

Table 4.3.a Average f_0 height and range by *Tone Condition*, *Focus* and *Underlying Tone*

Tone Condition	Focus	Underlying Tone	f_0 Height		f_0 Range	
			Average (Semitones)	Standard Error	Average (Semitones)	Standard Error
Intrinsic NT	No Focus	0	6.49	0.56	4.06	0.34
	With Focus		7.78	0.34	3.71	0.26
	With Emphasis		7.47	0.53	3.70	0.39
Derived NT	No Focus	1	7.91	0.53	1.65	0.12
	With Focus		9.73	0.50	1.19	0.04
	With Emphasis		10.57	0.57	0.91	0.06
	No Focus	2	6.38	0.58	4.78	0.45
	With Focus		8.80	0.49	3.66	0.16
	With Emphasis		9.54	0.51	4.54	0.33
	No Focus	4	7.52	0.56	4.19	0.35
	With Focus		9.20	0.42	4.68	0.15
	With Emphasis		10.21	0.67	6.57	0.40

	No Focus		9.95	0.52	1.49	0.11
	With Focus	1	11.08	0.52	1.10	0.05
	With Emphasis		12.47	0.57	1.16	0.07
	No Focus		7.04	0.51	4.33	0.18
CT	With Focus	2	9.39	0.35	5.28	0.13
	With Emphasis		11.32	0.54	5.16	0.20
	No Focus		9.53	0.53	5.90	0.35
	With Focus	4	10.31	0.43	7.10	0.32
	With Emphasis		10.98	0.64	6.83	0.30

It can be observed from **Table 4.3.a** that the differences between the unfocused stimuli and focused stimuli in each condition were relatively large, but the differences between the two focused conditions were marginal. However, one-way *Anova* conducted in the Intrinsic NT condition showed that the effect of *Focus* condition was not significant on average f_0 height nor range.

LME models show that in the other two conditions, on f_0 height, the effects of *Tone Condition* and *Focus* were significant but on f_0 range, only the effect of *Underlying Tone* was significant (**Table 4.3.3.b**).

Table 4.3.b Linear mixed-effects model on f_0 height and range in Derived NT and CT conditions

Final model Analysis	f_0 height ~ Tone Condition + Focus + Underlying Tone + Tone Condition: Focus+ Focus: Underlying Tone+ Tone Condition: Underlying Tone + (1\Subject) + (1\Token)			
	SS	df	F	*p*
Tone Condition	28.39	1	6.8659	<0.05*
Focus	327.52	2	39.6031	<0.0001***
Final model Analysis	f_0 range ~ Tone Condition + Focus + Underlying Tone + Tone Condition: Focus+ Focus: Underlying Tone+ Tone Condition: Underlying Tone +(1\Subject) + (1\Token)			
	SS	df	F	*p*
Underlying Tone	305.470	2	40.0783	<0.0001***

Significance levels * =.05 ** = .01 *** = .001

Post hoc comparisons demonstrated that the average f_0 height of CTs was significantly higher than that Derived NTs (p <0.05). The average f_0 height of the unfocused tokens was significantly lower than the focused tokens (p <0.0001) than the focused tokens with emphasis (p <0.0001), but the difference between the two focused statuses was not significant. With regard to the underlying

tones, it was T1 that was significantly higher than T2 (p <0.05).

With regard to the average f_0 range, the tokens with T1 as the underlying tone (i.e., T1 and Derived NT from T1) showed significantly smaller ranges than the tokens with T2 and T4 as the underlying tones (ps <0.0001).

Relative Duration

In terms of relative duration, the LME model showed that *Tone Condition*, *Focus* and their interaction all significantly affected the duration ratio of the target words (**Table 4.3.c**).

Table 4.3.c Linear mixed-effects model on duration ratio

Final model Analysis	Duration Ratio ~ Tone Condition + Focus + Tone Condition: Focus + (1\Subject) + (1\Token)			
	SS	df	F	*p*
Tone Condition	2.2256	2	15.1021	<0.0001***
Focus	2.0305	2	13.7779	<0.0001***
Condition: Focus	1.4650	4	4.9706	0.0005***

Significance levels * =.05 ** = .01 *** = .001

Post hoc comparisons showed that the corrective focus significantly increased the duration ratio of *No Focus* in Derived NT condition (p <0.0001) and CT condition (p <0.05), but the duration ratio difference between *With Focus* and *With Emphasis* were not significant in either condition (**Figure 4.3.b** and **Table 4.3.e**). In Intrinsic NT condition, no significant change in the duration ratio was found.

In addition, like in Experiment 1, when not carrying focus, no significant difference was found between the duration ratio of the Intrinsic NT words and Derived NT words, but between

both of them and CT words (*ps* <0.001); when focused or focused with emphasis, Derived NT became significantly different from Intrinsic NT (*p*s<0.001) but not from CTs (**Figure 4.2.b**).

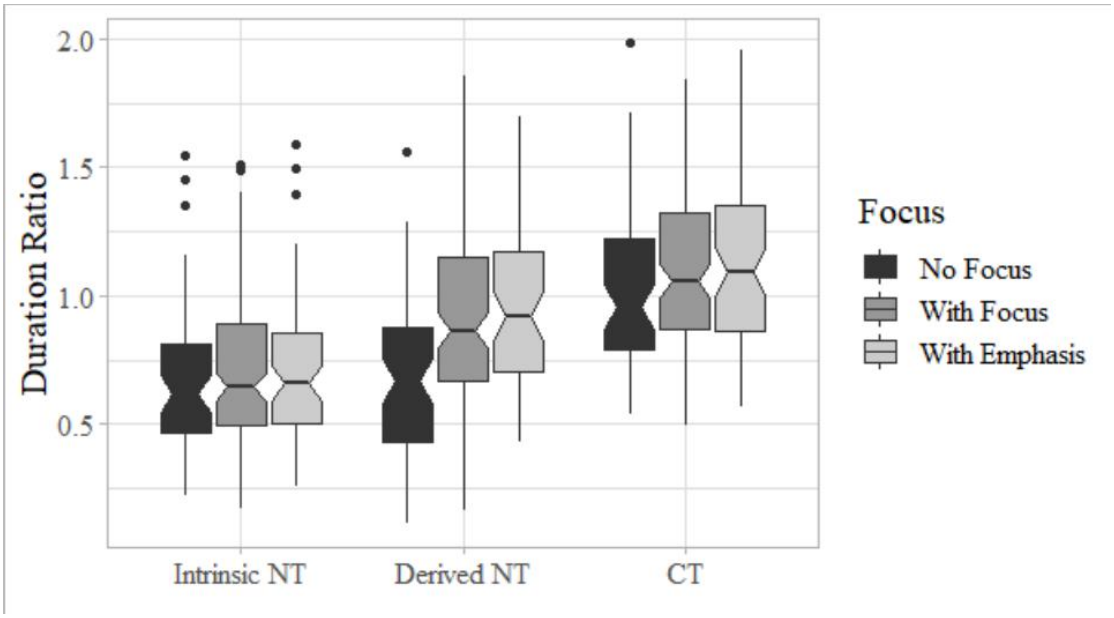


Figure 4.3.b Duration ratio in each tone condition by focus status

In terms of the absolute duration of the first syllables, only the effect of *Focus* was significant (**Table 4.3.d**). Post-hoc comparisons show that the increase from *No Focus* to *With Focus* (*p*s<0.001) as well as that from *With Focus* to *With Emphasis* was significant (*p*<0.05) (**Figure 4.3.c**).

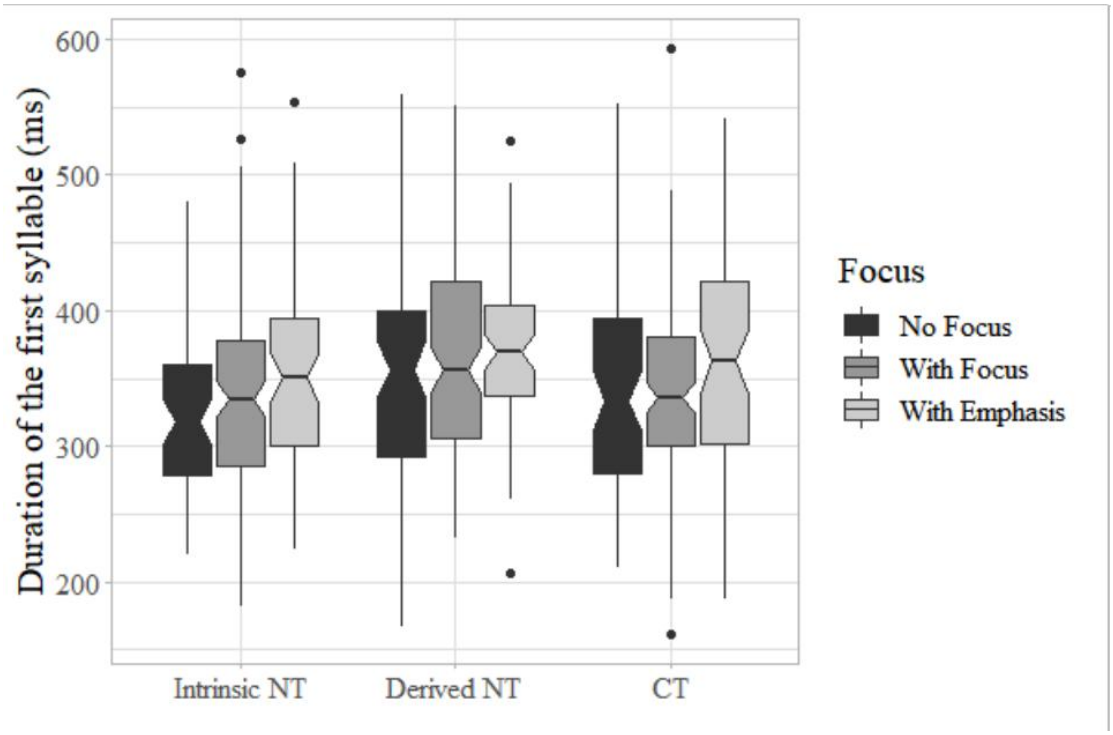


Figure 4.3.c Duration of the 1st syllable in each tone condition by focus status

In contrast, the effects of *Tone Condition*, *Focus* and their interaction on the absolute duration of the second syllables were all statistically significant. The results of post-hoc comparisons were slightly different from Experiment 1. Duration differences of Intrinsic NT between *No Focus and With Focus* and between *With Focus* and *With Emphasis* were not significant, unlike in Experiment 1 where corrective focus induced significant lengthening in Intrinsic NT. The duration increases of Derived NT from *No Focus* to the other two focus statuses, *With Focus* and *With Emphasis* were significant (*ps* <0.0001) but not the increase between the two focus statuses. The duration increases of the 2nd CT from *No Focus* to *With focus* was marginally significant (p = 0.053) but the increase from *With Focus* to *More Focus* was not significant (**Figure 4.3.d** and **Table 4.3.e**).

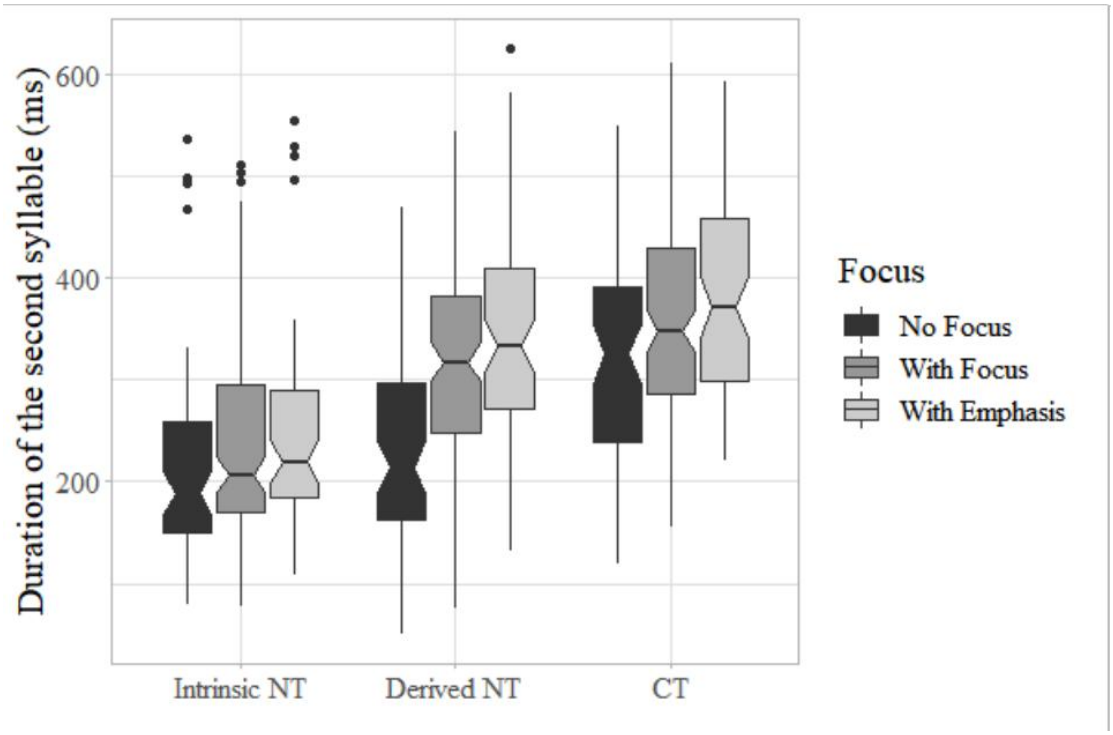


Figure 4.3.d Duration of the 2nd syllable in each tone condition by focus status

Table 4.3.d Linear mixed-effects model of effects of *Tone Condition* and *Focus* on absolute duration

Final model Analysis	The 1st Syllable Duration ~ Tone Condition + Focus + Tone Condition: Focus + (1\Subject) + (1\Token)			
	SS	df	F	*p*
Focus	44.75	2	7.9634	<0.001***
Final model Analysis	The 2nd Syllable Duration~ Tone Condition + Focus + Tone Condition: Focus + (1\Subject) + (1\Token)			
	SS	df	F	*p*
Tone Condition	303.47	2	24.6216	<0.0001***
Focus	550.76	2	44.6847	<0.0001***
Condition: Focus	178.08	4	7.2239	<0.0001***

Significance levels * =.05 ** = .01 *** = .001

Table 4.3.e Duration ratio and absolute duration of the syllables by *Tone Condition* and *Focus*

Tone Condition	Focus	Duration Ratio		Duration of the First Syllable		Duration of the Second Syllable	
		Average	SE	Average	SE	Average	SE
Intrinsic NT	No Focus	0.70	0.59	325.60	2.43	225.49	3.16
	With Focus	0.73	0.57	333.22	2.75	246.90	3.13
	With Emphasis	0.74	0.59	355.47	2.71	251.00	3.15
Derived NT	No Focus	0.67	0.55	357.62	2.84	223.48	3.34
	With Focus	0.91	0.57	371.56	2.77	328.56	3.12
	With Emphasis	0.96	0.56	386.31	2.45	341.55	3.35
CT	No Focus	1.01	0.55	341.77	2.84	322.10	3.25
	With Focus	1.14	0.60	343.52	2.72	364.51	3.16
	With Emphasis	1.12	0.58	367.56	3.70	383.58	3.22

4.3.3 Discussion

The results of this supplementary experiment demonstrated that the effects of focus with emphasis on the f_0 realization and the relative duration of Intrinsic NT, Derived NT and CT were relatively marginal. Even though there were no differences with regard to duration ratio and f_0 range, f_0 height was significantly higher in CT stimuli than in Derived NT stimuli when the second syllables were on focus (with or without emphasis).

With regard to the tonal realization, P1 is supported as neither corrective focus nor focus with emphasis triggered significant changes in the average f_0 height and range of Intrinsic NT. This finding echoes the results of Experiment 1 and supports the hypothesis that Intrinsic NT is underspecified in its underlying tonal representation. The lack of an underlying tone does not allow a large degree of f_0 expansion but constrains the tonal interaction between Intrinsic NT and focus either.

P2, the predication regarding the tonal realization of Derived NT, is only partly supported. As predicted, both corrective focus and focus with emphasis increased f_0 height significantly, but the increase from the focused condition to the focused-with-emphasis condition was not significant. However, significant focus-induced increase was not found in f_0 height, unlike what was predicted in P2.

With regard to the metrical structure of NT, P3 and P4 supported. As predicted in P3, neither focus nor focus with emphasis triggered significant an increase in the duration ratio of Intrinsic NT words, suggesting that the metrical lightness of Intrinsic NT is stable. As predicted in P4, the duration ratio of Derived NT increased significantly from the unfocused to the focused, but not

from the focused to the focused with emphasis; the average duration ratio was also under 1 when focused with emphasis. However, this time, when focused, significant differences were found between Intrinsic NT and Derived NT words.

Taking into consideration the changes in absolute duration, a slightly different result from Experiment 1 was found. This time, the restriction of lengthening was not only found in CTs, but also in Intrinsic NT. The stability found in the relative duration of Intrinsic NT words did not come from synchronous increase in both syllables, but restricted increase in the second NT-bearing syllables on corrective focus or focus with emphasis. The difference may result from individual difference or experimental design. In contrast, the absolute duration of Derived NT still increased significantly when on focus, but the emphasis did not trigger any more significant change.

Compared to the previous findings on CTs, steady increases in the absolute duration of the focused CT-bearing syllables have not been found in the present study (Chen and Gussenhoven, 2008). Corrective focus only triggered marginally significant increase on the final CT-bearing syllables while no significant increase was found from the focused to the focused with emphasis. Since in Chen and Gussenhoven (2008), the target stimuli were elicited in non-final positions, the lack of significance in durational increase found in the present may result from the non-additive manner in which the final- and focus-induced lengthening are realized in Mandarin (Chen, 2006). Moreover, previous studies also find that focus domain (i.e. which components are under focus) and morpho-phonological constraints influence the acoustic realizations, especially the durational realization of focus in Mandarin (Chen, 2006; Chen, Lee, and Pan, 2016). This may also explain the differences between the present results and Chen and Gussenhoven (2008), because disyllabic rather than monosyllabic words were used and focus was elicited on a single morpheme in a word

rather than on a whole mono-syllabic word. The influence of focus domain and stimulus length on NT realization will be further explored in Experiment 5 (Chapter 6).

4.4 Summary of Experiment 1 and 2

The present chapter explored the representations of Intrinsic NT and Derived NT from an acoustic perspective. By examining the two NTs in comparison with the CTs in both typically-realized unfocused, focused and more focused situations, systematic differences have been found between Intrinsic NT and Derived NT(s) and CTs.

The results confirm H1 that Intrinsic NT and Derived NT are different in the underlying tonal representation, namely, Intrinsic NT is underspecified for tone but Derived NTs are specified as the CTs they derived from. Regarding the metrical structure, the evidence supports the hypothesis of a clear and stable light metrical structure for Intrinsic NT. The differences between Derived NT and CTs that Ⅰ found in the unfocused condition disappeared when Derived NT was on focus. Therefore, it seems plausible that Derived NT is metrically light compared to CTs, but since the duration of syllables carrying Derived NT can become significantly longer on focus, unlike syllables that carry Intrinsic NT, Derived NT's 'lightness' is somehow not as stable as that of in Intrinsic NT. The lack of stability may suggest that the tone overrides the metrical structure in the implementation of focus in Mandarin; possibly, the realization of tones requires longer second syllables, and hence results in the large increase observed in the absolute and relative duration of Derived NT on focus. It is also possible that there is a three-way distinction between Intrinsic NT, Derived NT and CT in the metrical structure. The durational changes of NT under focus will be further explored in Chapter 6, and the possible metrical structures of these words will be discussed with the experimental results in Chapter 8.

To summarize, I propose that in the Standard Mandarin that is spoken now, there are two types of NT that differ from each other in their underlying tonal representations, and they are both different from CT in the metrical weight. However, the phonological status of Derived NT, namely, whether it is another category of NT or is a result of NT sandhi remains to be further investigated. It is reasonable to propose that Derived NT is light and toned as Li has suggested, which directly explains the traces of the underlying T1 and T3 found in unfocused Derived NT from T1 and T3. However, it is also possible that these tone-specific traces are there because NT sandhi is an incomplete neutralization between contrastive CTs. This point will be revisited with further evidence from the following chapters.

Chapter 5 Processing of Intrinsic NT and Derived NT

5.1 Overview

Based on the results of the production experiment, in this chapter, I explore additional evidence for the representational difference by examining how the two types of NT are processed cognitively. The starting assumption is that, if intrinsic and derived NT have different phonological representations, these differences may not only manifest themselves in differences in their acoustic realization, but also in the way they are perceived and processed in online speech comprehension. For instance, although previous studies showed systematic differences in the phonetic realization of the rising T3 after sandhi and the authentic rising tone, T2, Mandarin native listeners did not perceive these two tones categorically (e.g., Peng, 2000). However, a recent n eye-tracking study tapping into the automatic processing shows that the Mandarin natives tend to process the two tones differently (Tu and Chien, 2020). Specifically, the rising T3 was first identified as T3 while T2 was first identified sometimes as T2 and sometimes as T3. The acoustic differences between the two tones were only used when the sandhi context was available and then, the target word was activated. Tu and Chien (2020) attributed their findings to markedness, namely, the rising T3 is the marked (allo)tone while T2 is the regular tone. However, the fact that they have neither controlled for complexity of the characters nor word frequency of stimulus pairs may undermine their findings.

Regarding the perception of NT, past perceptual studies did not pay much attention to the possibility that there may be underlying tonal differences between different types of Mandarin NT, much like past acoustic studies (as reviewed in Section 3.1). The only two studies that take the underlying tone into consideration were done by Li (Li and Fan, 2015; Li et al, 2014). However,

instead of examining the influences of the underlying tonal differences on the perception of NT, Li and her colleagues still focused on the weightings of duration and pitch, and only controlled for the underlying tones of the stimulus pairs they used, i.e., Li et al (2014) used 蘑菇 /mo2 ku0(1) / 'mushroom' and 魔箍 /mo2 ku1 / 'magic ring' in and Li and Fan (2015) used 舌头 /ʂɤ2 tʰəu 0 / and 蛇头 /ʂɤ2 tʰəu2/[30] instead of varying the underlying tones systematically or separating Intrinsic NT and Derived NT. Therefore, although these studies found that listeners were able to differentiate between CTs and NT as well as between minimal pairs of CT and NT words, to what extent the perception of NT words is influenced by the underlying tonal representations still remains unexplored. In fact, some linguists argue that the meaning distinguishing function of (Derived) NT from the CT with the same underlying tone is context dependent and hence is limited (e.g., Chen-Chung, 1984; Jin, 2001).

The experiments in Chapter 4 demonstrate that Intrinsic NT and Derived NT are likely to be of similar metrical weight, but that they have different underlying tonal representations despite the surface phonetic similarities. Moreover, both are distinctively different from CTs when realized in neutral unfocused speech. Based on these findings, I hypothesize that:

H3: Intrinsic NT and Derived NT are processed differently due to their underlying tonal differences, in spite of the surface similarities they have.

To test this hypothesis, the processing of Mandarin NTs is explored with regard to their underlying tonal representations. The chapter reports a CT/NT word identification experiment using a visual-world eye-tracking paradigm (Experiment 3) and an auditory only AX word

[30] According to the present proposal, 舌头 /ʂɤ2 tʰəu 0 / is an intrinsic NT word and therefore does not share the same underlying tones with /ʂɤ2 tʰəu 2 /.

discrimination experiment with reaction time (Experiment 4). The two experiments complemented each other. Experiment 3 recorded the eye-movements and therefore tap into the automatic online processing of NT words. However, there is a possibility that the logographic characters presented in Experiment 3 serve as specific contexts and hence slightly improve participants' performance. This is why the auditory AX discrimination task was added to avoid this potential confound.

5.2 Experiment 3: Identifying NT words

5.2.1 Methodology

Participants

20 native northern Mandarin speakers (11 males, 9 females) aged between 18-30 (mean age 25.4) participated in the experiment. Their dialectal backgrounds were strictly controlled as in the production experiments in Chapter 4. All participants were current students or employees at the University of Cambridge, but had completed their pre-university education in the Huabei region and reported Northern Mandarin as the main language they used in school and at home. All participants were right-handed with normal or corrected-to-normal vision, and no participants reported any hearing or speech impairments. Participants were reimbursed for their participation.

Stimuli and Design

Three conditions were created based on the underlying tonal representations (**Table 5.2.a**; complete list of stimuli in Appendix B). 93 pairs of disyllabic NT and CT words were chosen as the final stimuli from 140 pairs of candidates based on an internet-based norming study of the familiarity of each item. 24 participants with a similar language background and age range to the participants of this experiment participated in the norming study. They were given a questionnaire online to rate how familiar they were with each word from 1, extremely unfamiliar, to 9,

extremely familiar. Familiarity was defined as how frequently they encountered the word and whether they knew its meaning and usage. They were also asked to mark the NT words if they thought they were not NT words, and to provide the meaning of words they were unsure about. Only minimal pairs with a familiarity score of 5.5 or higher, (SD 3.0 or smaller), and familiarity difference between the pairs of 2.5 or smaller were chosen as the final stimuli.

I aimed to balance the distribution of CTs in the first syllable, namely in the preceding CTs, but excluded T3 to avoid T3 sandhi effects. Due to the limitation of the natural language, there were more Derived NT and CT pairs with underlying T3 and T4 than T1 and T2.

Table 5.2.a Experimental conditions and stimulus examples

Condition	**Description**	**Example**	**Number of Stimulus Pairs**
Intrinsic NT	Distinguishing between Intrinsic NT words vs. CT words	蚊子(/wən2-tsi0/) vs. 文字(/wən2-tsi4/)	31
Derived NT (Different second tones)	Derived NT words vs. CT words (the second syllables have different underlying tones)	道理(/dɑu4-li0(3)/) vs. 倒立(/dɑu4-li4/)	31
Derived NT (Same second tone)	Distinguishing between derived NT words vs. CT words (the second syllables share the same underlying tones)	园里(/juan2-li0(3)/) vs. 原理 (/juan2-li3/)	31

The complexity of the visual stimuli was indexed by the number of strokes of the two-character words. The average difference between the CT and NT pairs was 2.41 in Intrinsic NT Condition, -0.46 in Derived NT (Different second tones) and 0.32 in Derived NT (Same second tone). A One way *Anova* showed that the difference between conditions was not significant [$F(2,112) = 0.78$, $p = 0.46>0.05$][31].

A 28-year-old female Beijing Mandarin speaker who was trained in broadcasting recorded the stimuli in a sound treated room at the Phonetics Laboratory at the University of Cambridge. The recording of the words was done in minimal pairs to facilitate the following manipulation. The recordings of the stimuli were spliced to standardize the first syllables of the stimuli using *Praat*. For 15 of the stimulus pairs in each condition the first syllables of the CT words were used (i.e. in these pairs, both NT and CT words had the first syllables of the CT words). For the other 16 stimulus pairs, the first syllables of the NT words were used (i.e. in these pairs, both NT and CT words had the first syllables of the NT words). The average intensity of the recordings was scaled to 75dB. The digitally edited recordings were judged as natural by the speaker and two further native speakers who did not participate in the study.

The stimuli were split into two groups, each consisting of 46 or 47 pairs of stimuli which were used, respectively, in Experiment 3 and 4. The 46-pair group was used in Experiment 3 for half of the participants and in Experiment 4 for the other half. Experiment 3 and 4 were conducted one after another for each participant but the order of the two experiments was alternated between participants.

[31] -0.46 means that in Derived NT (Different second tones), it was the NT words slightly more complex than the CT words. The about 2-stoke differences between Intrinsic NT condition and the other two Derived NT conditions are not as large as it seems to be as a stoke can be as complex as 乚 or as simple as 丶.

Procedure

The experiment was programmed in Experiment Builder. Participants were presented with auditory stimuli which they had to match to logographs in a visual world paradigm. The eye movements were recorded at 1000 Hz using an EyeLink 1000 plus.

The eye-tracker was calibrated with a 5-point calibration procedure. Viewing was binocular, but eye-movements were recorded only from the dominant eye. In each trial, participants were first presented with a word pair in logographic characters (Font: Songti, Size: 36) horizontally located in the midline of the screen (1920 ×1080) with a fixation cross located in between. On the screen, one word appeared on the right side of the screen and the other one left with the order pseudorandomized by experimental condition. When the participants felt ready, they needed to fix at a fixation cross in the center of the screen to activate an auditorily presented word (i.e., the auditory target). They had to match to this auditory target to one of the visually presented words in the word pair, and indicated their choice by pressing the left/right key. Areas of interest (AoI) were defined as 520-pixel high and 720-pixel wide quadrants around each logographic word.

The experiment consisted of 92 or 94 test trials (depending on which subset of the stimuli were used) and 26 filler trials with 2 breaks in between. The fillers were CT-CT words with the same segmental structures but different second tones. The whole procedure took around 30 minutes. Participants were given 9 practice trials which were in the same format as the experimental trials, and the practice trials were also CT-CT words like the fillers.

Data Analysis

The target AoI was defined as the AoI around the word corresponding to the auditory target

in each trial; the other AoI contained the competitor (Competitor AoI). Only eye gaze positions within these two AoIs in each trial were analysed. Trials with over 25 percent track loss were excluded as well as trials with extreme key-pressing times using the Interquartile Rule (Tukey, 1977). The accuracy of word-end identification was tested with a Pearson's chi-square test.

Correctly answered trials were further analyzed for time clusters of divergence between the two AoIs as well as the proportion of switched looks in each condition. However, the analysis of the whole data set including incorrect items demonstrated a similar pattern of difference between conditions but the difference was larger.

The time clusters of divergence were defined as the periods when the proportion of looks into the Target AoI became significantly higher than the proportion of looks into the Competitor AoI. The starting points of these clusters are the points of divergence. A bootstrapped cluster-based permutation test (CBPT) (Maris and Oostenveld, 2007) within subjects was conducted through eye-tracking R (Dink and Ferguson, 2015) to calculate the time clusters of divergence. The time-bin width of the test was defined as 20ms and the test was repeated 150 times on the shuffled data and corrected for multiple comparisons. The effects of *Condition* and *Auditory Target* (i.e., the target words to be identified, NT words or CT words) and their interaction were tested by *Anova*. Proportion of switched looks was also calculated within each time-bin using the onset switches function in eye-tracking R to illustrate how much the participants switched between the two AoIs in each time bin.

Predictions

Results of Experiment 1 and 2 suggest that Intrinsic NT and Derived NT have different

underlying tonal representations, and Intrinsic NT is highly likely to be underspecified for tones. In this way, the processing of the three conditions in the present study corresponds to the ternary matching statuses in the FUL model proposed by Lahiri and her colleagues (Section 2.3.3). To be specific, in *Intrinsic NT*, there is no mismatch between NT stimuli and CT stimuli; in *Derived NT (Different second tones)*, NT stimuli and CT stimuli mismatch with each other while in *Derived NT (Same second tone)* they match each other. If H3 is supported by the data, namely, the underlying tonal representation plays a more important role in NT word processing than the surface realizations, the identification in the mismatched condition, *Derived NT (Different second tones)*, is expected to be easier than that in the no-mismatch condition, *Intrinsic NT*, than in the matched condition, *Derived NT (Same second tone)*, reflected in the different identification accuracy, points of divergence and average proportions of looks and switched looks of these conditions. Specified predictions are made as follows.

P1: The identification is more accurate in the *Derived NT (different second tones)* condition than in the Intrinsic NT condition than in the *Derived NT (same second tone)* condition; the identification accuracy in the *Derived NT (same second tone)* condition may only be slightly above chance level;

P2: The point of divergence is earlier in the *Derived NT (different second tones)* condition than in the Intrinsic NT condition than in the *Derived NT (same second tone)* condition;

P3: The average proportion of looks of the time cluster of divergence is highest in the *Derived NT (different second tones)* condition, followed by the Intrinsic NT condition and the *Derived NT (same second tone)* condition;

P4: Participants started to stop switching away from the target AoI earliest in *Derived NT (different second tones)* than in the Intrinsic NT condition than in the *Derived NT (same second tone)* condition.

5.2.2 Results

Identification Accuracy

Participants were most accurate 98.30% (N = 636) in the *Intrinsic NT* condition, followed by *Derived NT with different second tones*, 96.80% (N = 663), almost 10% higher than *Derived NT with same second tones,* 88.87% (N=665). The difference between *Intrinsic NT* and *Derived NT (same second tones)* was significant, $\chi^2(1, 1301) = 47.44$, $p < 0.001$, as was the difference between the two Derived NT conditions, $\chi^2(1, 1328) = 31.91$, $p < 0.001$. The difference between *Intrinsic NT* and *Derived NT (different second tones)* was not significant.

Cluster of Divergence

The conditions have the point of divergence at roughly the same time, all later than the average offset time of the stimuli. The point was at 960ms in *Intrinsic NT* condition, 980ms in *Derived NT (same second tones)* condition and 1020ms in *Derived NT (different second tones)* (**Figure 5.2.a**). The average offset time of the stimuli was 805.42 ms (SE 12.49) in *Intrinsic NT* condition, 793.76 ms (SE 12.17) in *Derived NT (same second tones)* condition and 795.56 ms (SE 12.04) in *Derived NT (different second tones)*. *Anova* test showed that the influences of *Condition* on the average offset time of the stimuli was not significant.

The average proportion of looks into target AoI over the cluster was the highest in the *Intrinsic NT* condition (84.70%), followed by *Derived NT (different second tones)*, (82.10%) and

Derived NT (same second tones) (79.30%).

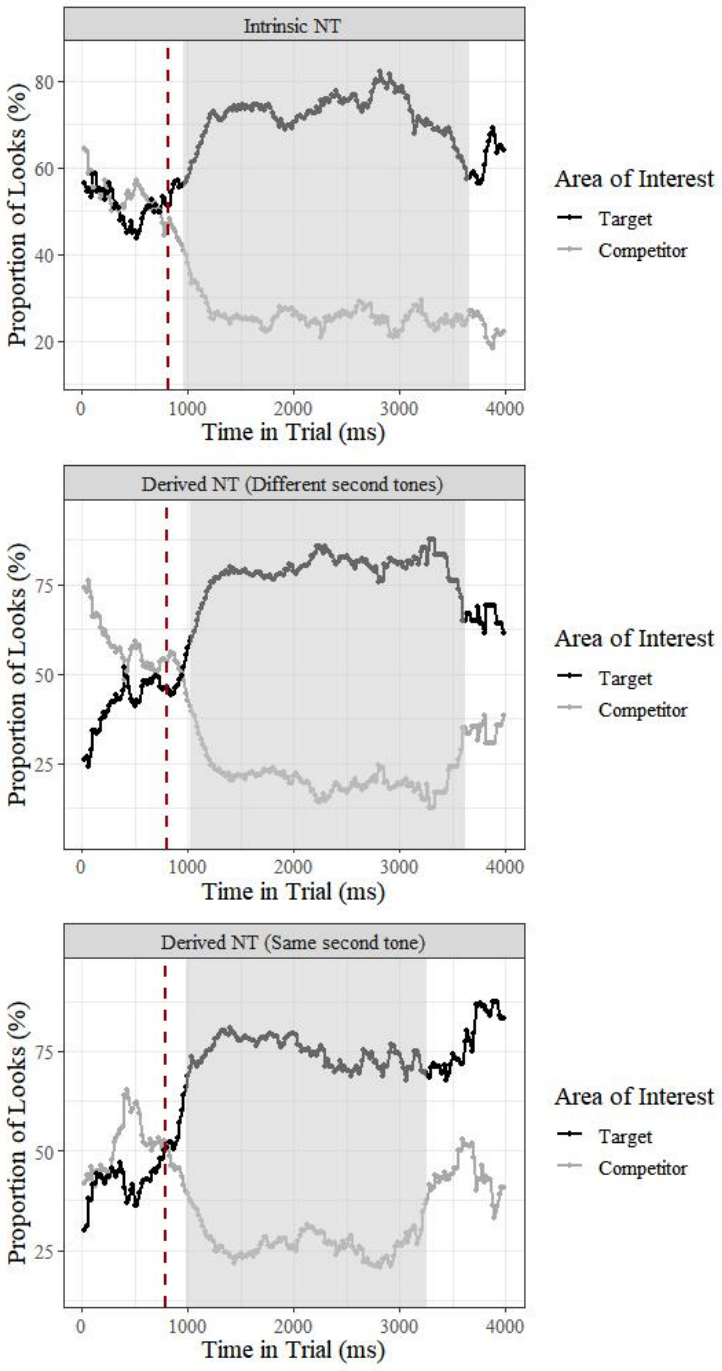


Figure 5.2.a Proportion of looks in each condition. The gray shades indicate the clusters of divergence and red vertical dash lines indicate the average sound offset times

In all conditions, when the *Auditory Target* was NT, the time cluster of divergence started much later after stimulus offset (**Figure 5.2.b** and **Table 5.2.b**). Moreover, the proportion of looks

into target AoI was also lower when the *Auditory Target* was NT in both *Intrinsic NT* and *Derived NT (different second tones)* conditions. *Anova* tests showed that only the effect of *Condition* on the proportion of looks into target AoI was significant [F(2, 1533) = 6.54, p <0.005] but not *Auditory Target.* Post-hoc comparisons showed that the proportion of looks into target AoI in *Derived NT (same second tones)* was significantly lower than that in the other two conditions (p <0.05).

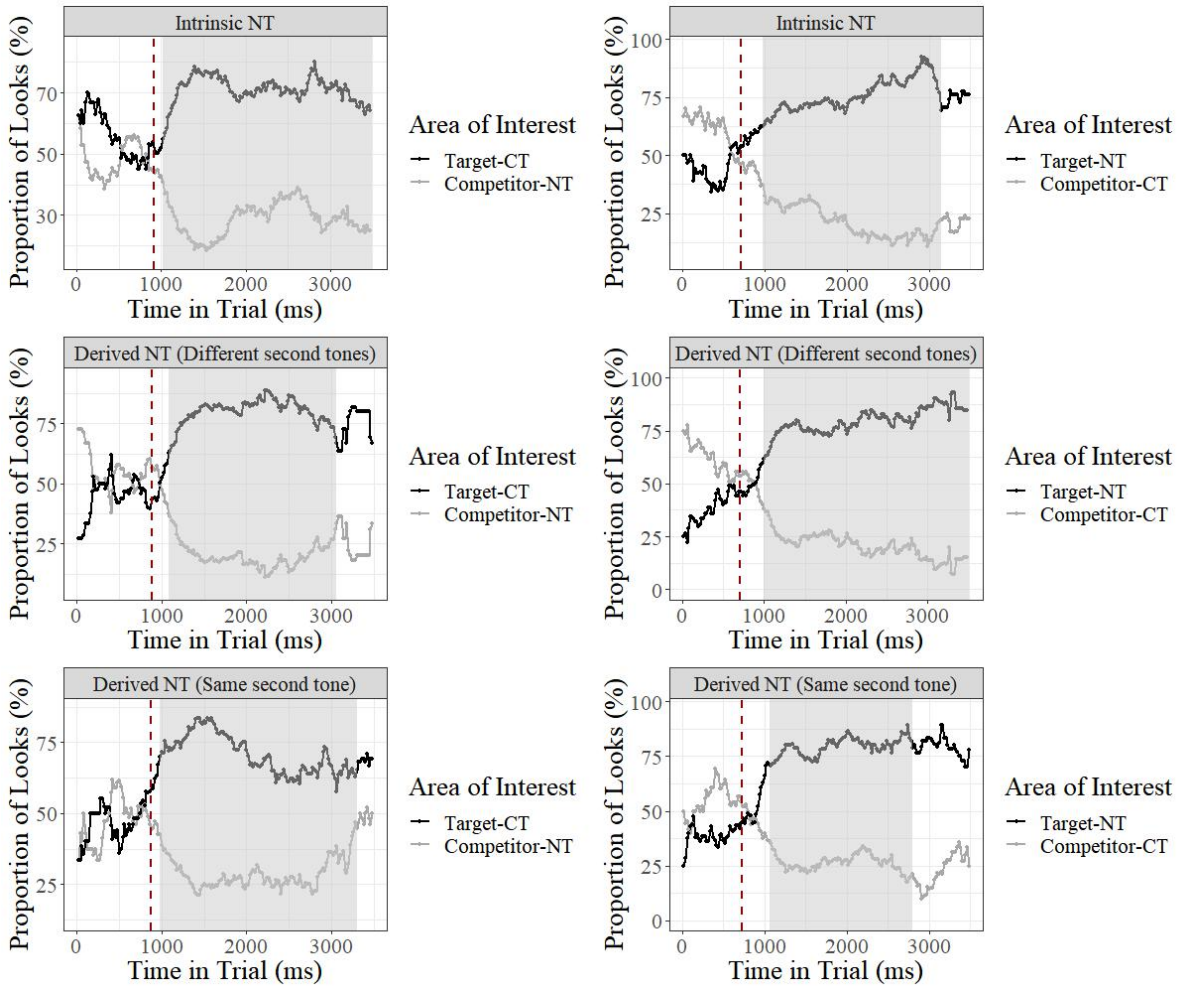


Figure 5.2.b Clusters of divergence by *Condition* and *Auditory Target*

Table 5.2.b Average proportion of looks into the target AoI by *Auditory Target* and *Condition*

Condition	Auditory Target	Start Point (ms)	Offset Time		Average Proportion of Looks into Target AoI (%)
			Average (ms)	Standard Error	
Intrinsic NT	CT	1040	913	12.289	86.64
	NT	980	714	9.179	83.67
Derived NT (different second tones)	CT	1080	883	11.49	83.63
	NT	1000	708	10.20	82.15
Derived NT (same second tones)	CT	980	868	11.14	76.60
	NT	1060	722	11.31	80.23

Proportion of Switched Looks

Proportion of switched looks demonstrates that as the trials went on, the participants switched less and less away from the target AoI but more and more towards it in all conditions (**Figure 5.2.c**). However, the crossing point of the solid and dashed lines, namely, when participants started to stop switching away from the target AoI occurred the earliest in *Intrinsic NT*, followed by *Derived NT (different second tones)* and much later in *Derived NT (same second tone).* The results showed that although the cluster of divergence started quite early in *Derived NT (same second tone)*, the participants were still uncertain about their choice and kept switching between the two AoIs.

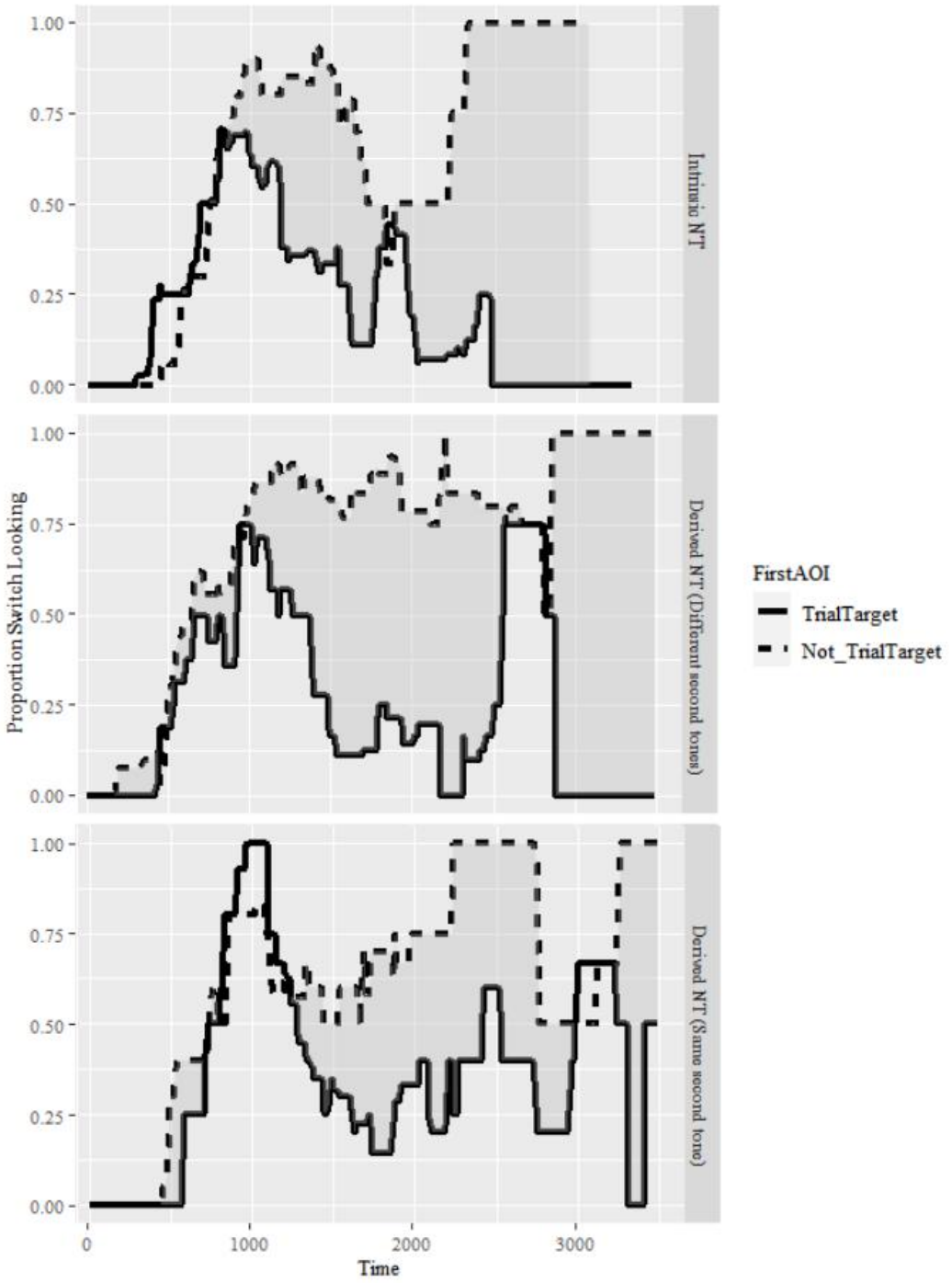


Figure 5.2.c Proportion of switched looks in each condition. The solid line indexes switching away from the target AoI and the dashed line indexes switching away from the competitor AoI.

5.2.3 Discussion of Experiment 3

In general, the findings of Experiment 3 support the hypothesis that the underlying tonal representations influence the perception and processing of Intrinsic and Derived NT words (H3).

However, all predictions were only partly supported as the processing differences reflected in the identification accuracy, points of divergence and average proportions of looks and switched looks were mainly found between *Intrinsic NT* and *Derived NT (Same second tone)* and the two Derived NT conditions but not between *Intrinsic NT* and *Derived NT (different second tones)*.

Regarding the identification accuracy, different from P1, participants were able to identify words well above chance in all conditions, and no significant differences were found between the accuracy in *Intrinsic NT* and *Derived NT (different second tones).* Nevertheless, the significant differences found between the *Derived NT (Same second tone)* condition and the other two conditions still demonstrated a strong effect of the underlying tonal representations, since identifying lexical items with the same underlying tones was proved more difficult than discriminating lexical items with different underlying tones in the present experiment.

Regarding the point of divergence, P2 is rejected because no significant differences were found between the three conditions, and *Derived NT (Same second tone)* also has the divergence started as early as the other two conditions.

Regarding the average proportion of looks within the time clusters of divergence, P3 was only partly supported by the data as significant differences were found between the *Derived NT (Same second tone)* and the other two conditions, but not between the *Intrinsic NT* condition and the *Derived NT (Different second tones)* condition.

Regarding the proportion of switched looks, P4 was partly supported as well because the participants kept switching for longer period between the targets and competitors only in *Derived NT (Same second tone)*, but quickly fixed at the targets in the other two conditions.

The differences between *Derived NT (Same second tone),* where the targets and competitors differ from each other in the metrical representations only, and the other two conditions are predicted by the FUL model (Section 2.3.3), but the lack of differences between *Intrinsic NT* and *Derived NT (Different second tones)* is surprising. Following either the FUL model or the markedness theory Tu and Chien (2010) followed, the underspecified tonal representation of Intrinsic NT should trigger a less clear 'mismatch' in the underlying representations in *Intrinsic NT* than in *Derived NT (Different second tones)*, where the targets and the competitors have different underlying tones and therefore make the identification harder. In other words, the identification in the *Derived NT (Different second tones)* condition should have been easier than in the *Intrinsic NT* condition. However, instead of questioning the underspecification of Intrinsic NT, the high accuracy and identification speed found in this condition may be attributed to a ceiling effect.

The different tendencies observed in the processing of NT and CT targets in the *Intrinsic NT* support the limited number hypothesis as well. The results show that participants intended fix on the Intrinsic NT targets slower than on the CT targets. In other words, compared to CTs, the identification of NT requires extra time. It is plausible that listeners tolerated more surface variations (or mismatches) in the identification of NT, the underspecified tone, and hence made decisions later than the identification of CTs, the fully specified tones. Such variation may enable a relatively 'fuzzy search' in the identification of words with underspecified tones and hence may increase the time taken (Chien et al, 2020). This finding might suggest that the higher accuracy and identification speed found in Intrinsic NT condition in general is due to the limited number of Intrinsic NT morphemes rather than a unique Intrinsic NT target. Nevertheless, further evidence is

required to test this hypothesis.

The mismatch between the points of divergence and the proportions of switched looks is also noteworthy. Although the participants tend to look more into the target in *Derived NT (Same second tone)* as early as in the other two conditions, they did not stop switching between the target and the competitor as soon as in the other two conditions. This mismatch may indicate that the look towards the target AoI is an unintentional orientation led by acoustic information. As Mishra et al (2013:2) pointed out, "... language-mediated eye movements are automatic rather than subject to conscious and controlled decision-making... Language-mediated eye movements appear to be for the most part unconscious and have all the hallmarks of an overlearned behavior". In this way, it is plausible that the underlying tonal representations did not influence the processing of the surface acoustic cues that much, but mainly affected the lexical retrieval processes. In other words, the metrical differences in Mandarin do not block competitors with the same underlying tones effectively in isolated spoken word recognition, as the ambiguity can be further facilitated by the context in natural communication. This finding echoes the previous findings in English that even in this typical stress language, metrical stress does not help auditory word recognition very much and the influence of the metrical stress exists only in pre-lexical rather than lexical influences in word processing (Soltano, 1997; Slowiaczek, 1990).

Another indication of the present results is that duration may be a more important cue in online processing of NT than pitch. Since in each trial, the pitch information is presented in a gradual and incremental manner and a previous study found that participants exploit tonal information in an incremental fashion (Shen et al, 2013), the participants would have looked more at the targets before the end of the second syllable if pitch was used as the primary cue in the

online processing of NT words. However, the current results demonstrated that participant did not fix onto the target AoI more until about 150ms after the average end of the stimuli (i.e., the average sound offset time) in all three conditions, when the duration contrast between the second and the first syllables was available. Therefore, the eye movements were more likely to be directed at the targets as a result of duration cues and not because of early pitch information.

5.3 Experiment 4: Discriminating NT words

5.3.1 Methodology

Participants

The same group of participants that participated in Experiment 3 also participated in Experiment 4.

Stimuli

The same two sets of stimuli used in Experiment 3 were used in Experiment 4.

Procedure

This AX discrimination experiment was programmed in E-prime 2.0 (Psychology Software Tools, Pittsburgh, PA). In each trial, the participants were presented auditorily with a stimulus pair with an inter-stimulus interval of 30ms. A question appeared right after the end of the second stimulus asking the participants to judge whether they heard the same words or different words. The question appeared in English to prevent any priming effects. Participants were instructed to press the 'same' key if they thought both stimuli were the same word, even if they sounded slightly different and 'different' if they thought the two stimuli were two different words. The 'same' button was on the left for half of the participants and on the right for the other half to

mitigate possible effects of handedness. The question corresponded to the position of the keys: when the ‘same’ button was on the left, the question appeared as ‘Same or Different?’ and when the ‘same’ button was on the right, the question appeared as ‘Different or Same?’ After the participants responded, the question disappeared automatically, and the next trial started. The next trial would not start until they gave an answer. The key-pressing before the appearance of the question was also recorded in case participants made their choice before the end of the second stimulus. The order of the stimuli was pseudo randomized by tone conditions, underlying tones as well as whether the trial contained two same stimuli.

The experiment consisted of 94 test trials and 26 filler trials with 2 breaks in between. Half of the test trials were with CT-NT words and CT-CT counterparts (both sequences presented), and the other half with the same sounds played twice. Before the start of the test trials, participants were given 9 practice trials that were in the same format as the test trials but with different words. Explicit explanation was given to make sure that participants were not reacting to acoustic differences but to lexical distinctions. The participants’ responses as well as their reaction times were recorded. The whole procedure took about 25 minutes.

Data analysis

Discrimination accuracy and the average reaction time of the correctly answered trials in each condition was calculated. The extreme outliers were excluded using the Interquartile Rule (Tukey, 1977). The accuracy differences were tested using Pearson’s Chi-square test. The effects of *Condition, Participant*, and *Trial type* (Different or same) were evaluated by a Linear mixed effects (LME) model using lmer in the lmerTest package (Kuznetsova et al, 2017) in R (R core team, 2020). The model building process was the same as in Experiment 1 and 2.

Predictions

P1: Discrimination is most accurate in the *Derived NT (different second tones)* condition followed by the Intrinsic NT condition, and the *Derived NT (same second tone)* condition;

P2: The lack of the logographic context in this experiment might make the discrimination in the *Derived NT (same second tone)* condition hard so that the accuracy may be at chance-level;

P3: Reaction time is shortest in the *Derived NT (different second tones)* condition followed by the Intrinsic NT condition, and then the *Derived NT (same second tone)* condition.

5.3.2 Results

Identification Accuracy and Bias Analysis

The discrimination accuracy was the highest in *Derived NT (different second tones)* (96.05%), followed by *Intrinsic NT* (93.14%) and *Derived NT (same second tone)* (83.76%). Chi-square tests showed that the accuracy differences between conditions were all significant: between *Intrinsic NT* and *Derived NT (same second tone)*, $\chi^2(1, 1292) = 26.62$, $p < 0.001$, between *Intrinsic NT* and *Derived NT (different second tones),* $\chi^2(1, 1336) = 5.03$, $p < 0.05$, and between *Derived NT (different second tones)* and *Derived NT (same second tone),* $\chi^2(1, 1374) = 56.76$, $p < 0.05$.

Reaction Time

The LME model showed that the effects of *Condition*, *Trial type* and the interaction of *Participant* and *Trial type* on reaction time were significant (**Table 5.3.a**). The post-hoc comparison further demonstrated that the reaction time in *Derived NT (same second tone),* 998.32 ms, was significantly lower than that in *Intrinsic NT* ($p < 0.05$), 888.34 ms, and *Derived NT*

(different second tones), 833.37 ms (p <0.005). The difference between *Derived NT (different second tones)* and *Intrinsic NT,* however, was not significant. In addition, participants were significantly quicker discriminating two same stimuli than two different ones (p <0.005).

Table 5.3.a Effects of *Condition*, *Trial type* and the interaction of *Participant* and *Trial type* on reaction time

Final model Analysis	Reaction Time~ Condition + Trial type + Participant + Condition: Trial type+ Trial type: Participant + Condition: Participant + Condition: Trial type: Participant			
	SS	df	F	p
Condition	1458277	2	1.0656	<0.05*
Trial type	13145486	1	19.2122	<0.0001***
Trial type: Participant	7386403	1	10.7953	<0.01***

Significance levels * =.05 ** = .01 *** = .001

5.3.3 Discussion of Experiment 4

The results of Experiment 4 confirm that Intrinsic NT and Derived NT are processed differently in Mandarin, echoing the findings in Experiment 3. The satisfying discrimination accuracy found in all three conditions suggests again that the underlying tone is important but not decisive to the processing of Mandarin NT.

As predicted in P1, the discrimination accuracy in the *Intrinsic NT* and *Derived NT (different second tones)* conditions were all above 90%. The accuracy in the *Derived NT (same second tone)* condition was lower but still much higher than chance-level, so that P2 is rejected. However, different from the results of Experiment 1, the discrimination in *Derived NT (different second tones)* was significantly more accurate than the discrimination in *Intrinsic NT* condition. Therefore, although the absence of logographic visual context does not reduce the absolute performance

accuracy much, the perception of *Intrinsic NT*, compared to *Derived NT (different second tones)*, became more difficult without the logographs.

In terms of processing speed, slightly different form the eye-tracking data, the shortest reaction time was found in the *Derived NT (different second tones)* condition. However, since no statistical significance was found between *Derived NT (different second tones)* and *Intrinsic NT*, P3 is still rejected.

5.4 Discussion of Experiment 3 and 4

These two experiments investigated the lexical processing of Mandarin NTs with regard to their underlying tonal representations. Both the eye-movement and the discrimination data show that Intrinsic NT and Derived NT are processed differently from one another due to the different underlying tonal representations. The similar surface forms (i.e., the short duration and the varying pitch contours) still enable the native speakers to distinguish NT words of either type from CT words with satisfying accuracy, but a clear prohibitive effect of the same underlying tonal representations was seen in the *Derived NT (same underlying tone)* while a clear faciliative effect of the different underlying tonal representations was found in the *Derived NT (different underlying tones)*. The processing speed in *Intrinsic NT* was also lower than in *Derived NT (different underlying tones)* in Experiment 5. Therefore, the processing similarities between *Intrinsic NT* and *Derived NT (different second tones)* found previously is likely to be due to a ceiling effect. The identification in *Intrinsic NT* may have been facilitated by the existence of the logographic contexts in Experiment 4.

Part B Interaction between NT and Utterance-Level Prosody

Chapter 6 NT and Focus

It has been established that the implementation of utterance-level prosody, like prosodic focus and intonation, is often influenced by more localized prosodic features like lexical tone and stress. In Mandarin, the phonetic and phonological exploration of the interaction between localized and utterance-level prosody has mainly focused on CT words and phrases. However, since the experiments in Part A provide evidence from different perspectives that Intrinsic NT, Derived NT and CT differ in their tonal and metrical representations, it is reasonable to expect that they may interact with the realization and perception of the utterance prosodies in different ways (RQ3). This possibility is explored in this second part of the thesis by investigating:

1) whether NT- and CT-bearing words and phrases behave differently under corrective focus given their different underlying representations;

2) whether the perception of intonation (Question vs. Statement) is facilitated in words that are un(der)specified for tone.

1) is investigated in this chapter with a focus on the distribution of focus-induced lengthening.

6.1 Introduction: Focus-induced lengthening

The experiments in Chapter 4 found that in neutral speech without focus, Intrinsic NT and Derived NT words demonstrated similarly small duration ratio, despite underlying tonal differences. However, when focused, Intrinsic NT, Derived NT and CTs demonstrated a three-way distinction in duration ratio, and the differences between Derived NT and CT were blurred. I

speculated at the end of Chapter 4 that Intrinsic NT, Derived NT and CT might be different from each other in the metrical strength they are associated with, or there might be tonal differences between Intrinsic NT and Derived NT that somehow led to their different lengthening patterns when on focus. In this chapter, this question is further pursued by eliciting focus on different parts of Intrinsic NT, Derived NT, and CT words and phrases. Assuming that the three differ from each other in their metrical structure, the following hypothesis can be made.

H4: Intrinsic NT and Derived NT behave differently from one another and from CTs when there is focus on the words or phrases they are in.

Previous studies show that although focus is phonologically associated with a focus-bearing unit, its durational correlates may extend beyond this unit. As introduced later, the focus-bearing unit as well as the durational correlates involved show language-specific patterns due to their different metrical organization cross-linguistically. To facilitate the explanation of relation between focus-bearing unit and the durational correlates, I adopt the terms from Chen (2006) and separate *pragmatic focus domain* from *durational adjustment domain. Pragmatic focus domain*, indicated by its name, refers to the components in the utterance that are on focus while *durational adjustment domain* refers to the linguistic units which actually undergo lengthening. These two domains do not necessarily correspond to each other. As demonstrated in Experiment 1 and 2, although the focus was elicited on the second syllables of the stimulus words, the preceding syllables were also lengthened though the magnitude was influenced by the word type (Intrinsic NT vs. Derived NT vs. CT), focus type and individual differences. In other words, in Experiment 1 and 2 (Chapter 4), although the *pragmatic focus domain* was the second syllable of the disyllabic stimuli, the *durational adjustment domain* seemed to be the whole stimulus word.

In typical stress languages like English, the existing studies have found that when polysyllabic words are on focus, it is the syllables that carry the primary stress (i.e., the word stress) that undergo most of the lengthening, and that the lengthening on the other syllables in the same words result from the spreading of the lengthening on the word stress syllables, and therefore show a smaller lengthening magnitude (Turk and White, 1999). Furthermore, influenced by the trochaic nature of these languages, the rightward spreading of lengthening is also found to be stronger than the leftward spreading. When sub-word level constituents are on focus, the lengthening still spreads to the other syllables within the same word, triggering focus-induced lengthening of the whole word in English. Therefore, it is argued that in English, the durational adjustment domain is the prosodic word; the attenuating effects of syllable boundary and foot boundary are weaker than the prosodic word boundary (Cambier-Langeveld, 2000).

In pitch-accented languages such as Swedish, it is also the pitch-accented syllables that undergo the lengthening most when the words are on focus, but the lengthening only extends to the unstressed syllables in the same feet rather than in the same words (Heldner and Strangert, 2001). Therefore, the metrical foot rather than the prosodic word is arguably the durational adjustment domain in Swedish. Moreover, the phonologically longer segments within the stressed syllables are lengthened significantly, but the phonologically short vowels are hardly changed (Heldner and Strangert, 2001). In other words, the contrast between the short and the long vowels in stressed syllables is sharpened rather than reduced, showing that the distribution of focus-induced lengthening is also sensitive to the phonological system of a language.

In tone languages, the situation becomes more variable. In Vietnamese, a tone language with no system of culminative word stress but accentual prominence at the phrasal level, little spillover

of lengthening is found on the syllables adjacent to the syllables on focus even within the same prosodic words or feet (Nguyễn, 2010). When the whole constituents are on focus, all the syllables are lengthened, though the syllables in even positions are lengthened more than the odd syllables, indicating that Vietnamese has a binary iambic foot structure. In Shanghai Wu dialect, a word tone language with checked tone, corrective focus on words or phrases does not induce lengthening on the short syllables, even if they are at the initial positions (i.e being the metrical heads); instead, the second syllables tend to show a compensatory lengthening effect when the initial syllables are short (Chen, 2009). Overall, the tone sandhi domain (i.e., prosodic words) plays an important role in focus realization in the Shanghai dialect.

In Mandarin polysyllabic CT words, the whole word is lengthened when sub-word constituents are on focus; clear spillover effects have been found on the preceding and the following syllables outside the pragmatically focused domain within the same words (Chen, 2006). Prosodic words are the durational adjustment domain in Mandarin, at least of Mandarin CT-bearers. However, when whole words are on focus, the CT words demonstrate a non-uniform distribution of lengthening. Instead of having gradual downgraded lengthening from the most prominent edge or lexically-stressed syllable, the four-syllable CT words under corrective focus show the following lengthening pattern: the lengthening of the final (fourth) syllable is of the largest magnitude, followed by the initial syllable, then the third syllable and finally the second syllable (Chen, 2006). Chen (2006) offered two interpretations of this non-uniform lengthening, based on a strong (right) edge-prominence and a (debatable) trochaic disyllabic foot or as a result of a strong edge-effect only. According to the first interpretation, the lengthening is still sensitive to the metrical structure of the focused words in addition to the strong (right) edge-prominence so

that the third syllable, which is metrically stronger than the second syllable is lengthened more. According to the second, more phonetic-based interpretation, although both syllables in between undergo a gradient spread of lengthening from the edge syllables, the third syllables may have lengthened even more because last syllables carry more edge prominence than initial syllables. These two accounts are based on different interpretations of the metrical organization in Mandarin, the first regarding Mandarin as having binary trochaic feet and the second as having no systematic word-level stress – instead, CT-bearing syllables may form unary degenerated feet themselves.

How focus is realized on NT is not as well-explored as how it is realized on CTs in Mandarin. The only study that explores the effects of both focus and morpho-syntactic boundaries on neutral tone (and hence prosodic domains) was done by Lin and Chen (2019). Although the focus of that study is on tonal realization rather than durational adjustment and used Tianjin Mandarin, a Mandarin dialect with four slightly different tones from Standard Mandarin. Li and Chen still found that the relative position of the disyllabic NT words on focus (On focus vs. Pre-focus) and the morpho-syntactic phrasing (Subject-predictate vs. Below-NP) of the NT words influenced both the tonal and the durational realization of NT-bearing syllables in Tianjian Mandarin. In particular, they found NT-bearing syllables in Below-NP condition to be shorter than in Subject-predict condition so that the boundary-related (or final) lengthening also affects NT-bearing syllables in Tianjin Mandarin. In addition, both the focus position and the morpho-syntactic phrasing affects the realization of NT, especially the raising effect of the following Tone 1, a low-falling rather than high-level tone in Tianjin Mandarin.

In the present experiment, I will look at focus-induced lengthening in NT words and phrases. This issue is worth investigation as unlike CT words, words involving NT demonstrate a clearer

heavy-light pattern, and possibly involve a stronger metrical contrast, much like the words in stress or pitch-accented languages. It then becomes an interesting question to explore to what extent the distribution of focus-induced lengthening in Mandarin NT words show similarities to English words due to this clear prominence contrast, and whether the short-long contrast between NT and CT may be intensified like in Swedish or Shanghai Wu dialect when the whole words are on focus. The results will not only improve our understanding of the metrical structure of NT words, but also typological patterns of focus-induced lengthening.

6.2 Methodology

Participants

17 Northern Mandarin speakers (7 males, 10 females) aged between 18-29 (mean age 24) participated in the experiment. All participants were current students at the Shanghai Jiao Tong University who had completed their pre-university education in the Huabei or the most northern part of Huadong region and reported Northern Mandarin as the main language they used in school and at home. All of them were right-handed and none of them reported any hearing impairments. Informed consent was obtained prior to the experiment. One female participant did not emphasize on any part of the stimuli in the second half so her recording was not used in the final analyses, leaving a total of 16 speakers.

Stimuli

Disyllabic and trisyllabic Intrinsic NT, Derived NT and CT words or conventional phrases were used as stimuli. For disyllabic words, half of the stimuli used in Experiment 1 were used here, namely, 8 intrinsic NT morphemes, 8 derived NT morphemes and 8 CT-bearing morphemes that were each combined with one CT. Each of the four CTs occurred once as the preceding tone in a

tone condition and the distribution of the underlying tones in the Derived NT Condition and the CT Condition was also balanced. For trisyllabic stimuli, 12 trisyllablic Intrinsic NT phrases with tone combination as T1+ Intrinsic NT + CTs, 4 trisyllabic Intrinsic NT phrases and 4 trisyllabic Derived NT phrases with tone combination as CTs +NT +NT were used; 8 trisyllablic CT phrases with tone combination as T1+ CT + CT were added as the base line (**Table 6.2.a**; complete list of stimuli in Appendix C). The following CTs, and the syllable structures of the preceding and following syllables were balanced.

Table 6.2.a Examples of stimuli (numbers in the transcription indicates tones) [32]

Stimuli Condition	Bound morpheme	IPA Transcription	Examples of Stimuli	Translation	IPA Transcription
One Intrinsic NT	的\子\了	\tɤ0\ \zi0\ \lɣ0\	灰的书	Grey-attributive particle-book ‘Grey book’	\xuei1-tɣ0-ʂ‘u1\
Two Intrinsic NT	的了	\tɤ0 lɣ0\	灰的了	Grey-nominalizer-aspect particle ‘Turned into grey’	\xuei1 tɤ0 lɣ0\
Two Derived NT	下去	\ɕia0(4) lai0(4)\	飞下去	Fly-down-come ‘Fly downwards’	\fei1 ɕia0 tɕʰy0\
CT	棵\幅\首\倍	\khɣ1\ \fu2\ \ʃəu3\ \pei4\	三棵树	Three-classifier-three ‘three trees’	\san1 khɣ1 ʂ‘u4\

[32] Limited by natural language, there is no meaningless Intrinsic NT or Derived NT words or phrases like transliterated foreign names available (e.g.,不列颠 /pu2 liɛ4 tian1/ ‘Britain’ in which none of the three morphemes have concrete semantic meanings). Therefore, compared to the previous studies on CT words, the stimuli used in the present study are not as semantically neutral.

Mispronunciations in the stimuli were necessary in order to elicit corrective focus from participants like in Experiment 1 and 2. Since in Experiment 1, the participants were relatively insensitive to the mispronunciation of Intrinsic NT, the mispronunciation in the present experiment was on tone as well as the onset consonants across all the conditions. The mispronounced words were not semantically meaningful to avoid misinterpretation of the pragmatic focus domain.

Disyllabic words were recorded correctly once, once with mispronunciation on the initial syllable, once with mispronunciation on the second syllable, and once with mispronunciation on both syllables; the trisyllabic stimuli were recorded once correctly, once with mispronunciation on the initial syllable, once with mispronunciation on the second syllable, once with mispronunciation on the third syllable, once with mispronunciations on the first two syllables, and once with mispronunciations on all the syllables. The stimuli were recorded by the author herself in a quiet room with Zoom H1 handy recorder at 96000Hz\26Bit.

Procedure

Some participants were recorded in the psycholinguistic center and others in quiet meeting rooms in Shanghai Jiao Tong University. The experiment consisted of two sessions using the two sets of stimuli. In each trial, the participants' task was to judge whether the recording they heard in each trial was correct or not according to the stimulus word presented in logographs. In one session, the participants were asked to repeat the word by putting it into one of two short carrier sentences, either \tuei, tu (stimulus) \ 'Yes, it reads as (stimulus)' to confirm the pronunciation or \pu tuei tu (stimulus)\ 'No, it reads as (stimulus)' to correct the pronunciation. In the other session, the carrier sentences were longer: \tuei tu (stimulus) ʈʂɤ4 kɤ4 ɕi2\ 'Yes, it reads as (stimulus) this

word' and \pu tuei tu (stimulus) ʈʂɤ4 kɤ4 ɕi2\ 'No, it reads as (stimulus) this word'. Half of the participants responded to stimuli in set A with the short carrier sentences and the other half responded to stimuli in set B with short carrier sentences. In addition, half of the participants responded first with the short sentences and the others responded first with the longer sentences. The trials were pseudo randomized within each session.

To familiarize themselves with the procedure, the participants were shown a video of the author and another speaker who did not participate in the experiment doing 10 trials. Then, the participants were asked to do 8 practice trials that were in the same format but with different CT-CT stimuli as the trials in the actual experiment.

The sessions were recorded on a Zoom H1 handy recorder at 96000Hz\26Bit. The participants were offered a break every 20 trials as well as between sessions, and they could choose to skip the break if they wanted. The total duration of the experiment was about 45 minutes including the instructions and breaks.

Data analyses

Performance accuracy, defined as saying 'no' to the mispronounced words and 'yes' to the correctly-pronounced words, was calculated first. All participants showed very high accuracy. Only 7 participants made incorrect responses, but the total number for all the participants was smaller than 10. The incorrect trials were removed from the analyses, leaving 4215 trials in total.

The word and syllable boundaries in the recordings were marked using *Praat* (Boersma and Weenink, 2021). Segmentation was based on both spectrograms and waveforms, supplemented by evaluation of the audio recording. The duration of each word and syllable was extracted using a

self-written *Praat* script.

Due to the large amount of data collected, they were divided into three groups based on the different tonal and metrical structures for the purpose of the analyses: the disyllabic words, the trisyllabic CT phrases and trisyllabic NT phrases with two NTs, and the trisyllabic CT phrases and phrases with one Intrinsic NT. In each group, the effects of *Tone Condition* (Intrinsic NT vs. Derived NT vs. or CT), *Pragmatic Focus Domain* (i.e., which syllable(s) were carrying focus), and *Sentence Position* (stimulus words or phrases being sentence-medial or final) on the duration of the whole word and on each syllable were evaluated through LME models. Since the focus of the present study is on comparing the lengthening patterns between Intrinsic NT, Derived NT and CT, the effects of *Pragmatic Focus Domain* and *Sentence Position* are only briefly reported if they did not interact with *Tone Condition.*

The duration ratio between the second syllable and the preceding CT syllable in disyllabic words was further calculated, and the effects of *Tone Condition, Pragmatic Focus Domain* and *Sentence Position* were also evaluated by LME models. The model-building process was the same as that in Experiment 1 and 2.

Predictions

If H4 is supported, I expect to find different lengthening patterns of Intrinsic NT, Derived NT and CT in various pragmatic focus domains. Specified predictions are made as follows.

P1: Intrinsic NT stimuli differ from CT stimuli, but resemble the accentual lengthening patterns found in English words. Specifically, syllables bearing Intrinsic NT are lengthened by the spill-over lengthening from the preceding CTs rather than being lengthened

themselves when the whole constituents are under focus. Additionally, there is little spill-over lengthening towards adjacent syllables when syllables bearing Intrinsic NT are focused.

P2: Derived NT shows larger similarity to CT, namely, syllables bearing Derived NT are lengthened themselves when the whole constituents are under focus. When the syllables bearing Derived NT are focused, the adjacent syllables were also lengthened due to spill-over effects.

P3: Final- and Focus-induced lengthening are additive in Intrinsic NT and Derived NT, but not in CTs.

6.3 Results

6.3.1 Stimulus Duration

Significant effects of the interaction between *Tone Condition* and *Pragmatic Focus Domain* and between *Tone Condition* and *Sentence Position* on stimulus duration were only found in the disyllabic stimuli, but not in the trisyllabic stimuli with two NTs or the trisyllabic stimuli with Intrinsic NT in the middle (**Table 6.3.a**).

Table 6.3.a Linear mixed-effects models on stimulus duration

Final model	Duration ~ Tone Condition + Pragmatic Focus Domain + Sentence Position + Tone Condition: Pragmatic Focus Domain + Tone Condition: Sentence Position + Pragmatic Focus Domain: Sentence Position + Tone Condition: Pragmatic Focus Domain: Sentence Position + (1\Participant) + (1\Stimulus)			
	SS	df	F	*p*
Tone Condition	862.50	2	298.8631	<0.0001***

Pragmatic Focus Domain	344.43	3	79.5655	<0.0001***
Tone Condition: Pragmatic Focus Domain	440.80	6	5.0914	<0.0001***
Tone Condition: Sentence Position	180.81	2	6.2638	0.001955**
Pragmatic Focus Domain: Sentence Position	18.90	3	0.4358	0.727379
Final model	Duration~ Tone Condition + Pragmatic Focus Domain + Sentence Position + (1\Participant) + (1\Stimulus)			
	SS	df	F	*p*
Tone Condition	784.20	2	5.3168	0.005001 **
Pragmatic Focus Domain	439.11	3	32.0651	<0.0001***
Sentence Position	496.30	1	50.055	<0.0001***
Final model	Duration~ Tone Condition + Pragmatic Focus Domain + Sentence Position + (1\Participant) + (1\Stimulus)			
	SS	df	F	*p*
Tone Condition	784.20	2	5.3168	<0.0001***
Pragmatic Focus Domain	439.11	3	32.0651	<0.0001***
Sentence Position	496.30	1	50.055	<0.0001***

Significance levels * =.05 ** = .01 *** = .001

Disyllabic Stimuli

When examined closely (**Table 6.3.b** and **Figure 6.3.a**), the lengthening effects of *Pragmatic Focus Domain* on the whole word were similar in all three *Tone Conditions*. However, post-hoc comparisons showed some statistical differences between Derived NT and the other two, i.e., Intrinsic NT and CTs. In Derived NT condition, the differences between the pragmatic focus

domains were all significant (*p*s<0.001). In Intrinsic NT and Derived NT, *No Focus* was significantly shorter than the other domains (*p*s<0.001), and the other differences were not significant. In addition, although sentence final stimuli were longer than the medial ones in all three conditions, post hoc tests showed that the effect of *Sentence Position* was only marginally significant in CT condition (*p*=0.053) but not significant in the other two NT conditions (**Table 6.3.b**).

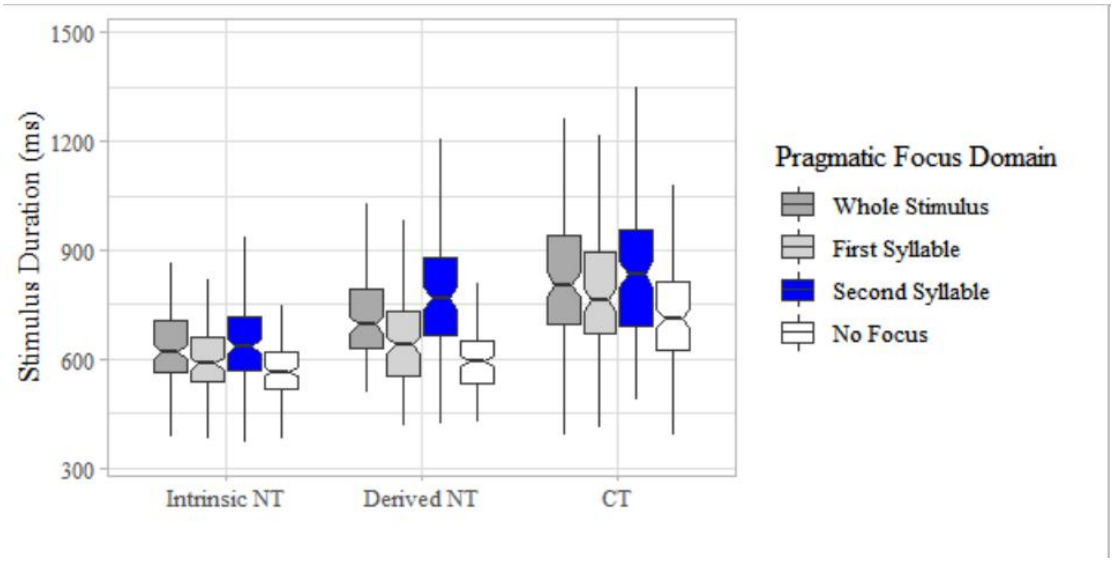


Figure 6.3.a Duration of disyllabic stimuli by *Tone Condition* and *Pragmatic Focus Domain*

Table 6.3.b Average duration of disyllabic stimuli

Tone Condition	**Pragmatic Focus Domain**	**Duration**		**Sentence Position**	**Duration**	
		Average (ms)	SE		Average (ms)	SE
Intrinsic NT	**Whole Stimulus**	570.11	3.02	**Medial**	598.9	11.25
	First Syllable	608.75	3.32			
	Second Syllable	650.62	3.62	**Final**	633.71	11.15
	No Focus	630.87	3.15			
Derived NT	**Whole Stimulus**	605.36	3.2	**Medial**	670.32	14.76

	First Syllable	657.9	3.74			
	Second Syllable	788.91	4.16	**Final**	717.85	15.91
	No Focus	723.23	3.64			
CT	**Whole Stimulus**	722.96	3.89	**Medial**	755.26	17.74
	First Syllable	794.75	4.26			
	Second Syllable	844.16	4.39	**Final**	841.14	18.04
	No Focus	831.8	4.29			

Trisyllabic Stimuli with Two NTs

Post-hoc comparisons showed that the stimuli in the two NT conditions were significantly shorter than the stimuli in the CT conditions (*p*s<0.001), but the duration differences between themselves were not significant. Overall, *No Focus* was significantly shorter than the other pragmatic focus domains (*p*s<0.0001); *First Syllable* was significantly shorter than the others except *No Focus* (all *p*s<0.05); *First Two Syllables* was also significantly lower than *Second Syllable* and *Third Syllable*. Sentence final stimuli were significantly longer than sentence medial stimuli (*p*<0.0001).

Table 6.3.c Average duration of CT-NT-NT stimuli

		Duration(ms)	**SE**
Tone Condition	**Intrinsic NT**	800.35	39.33
	Derived NT	898.20	44.10
	CT	1072.75	45.42
Pragmatic Focus Domain	**Whole Stimulus**	1026.92	50.82
	First Two Syllables	986.31	46.99
	First Syllable	935.90	43.67

	Second Syllable	988.33	45.51
	Third Syllable	993.41	46.17
	No Focus	835.32	44.89
Sentence Position	**Medial**	918.42	45.72
	Final	1003.75	48.13

Trisyllabic Stimuli with Intrinsic NT

Post hoc comparisons showed that Intrinsic NT stimuli were significantly shorter than CT ($p < 0.0001$), unfocused stimuli were shorter than the focused regardless of where the focus was located ($ps < 0.0001$) and sentence-final phrases regardless were significantly shorter than the medial ones ($p<0.0001$).

Table 6.3.d Average duration of CT-NT-CT stimuli

		Duration (ms)	**SE**
Tone Condition	**Intrinsic NT**	936.1	45.46
	CT	1099.13	46.25
Pragmatic Focus Domain	**No Focus**	982.33	46.58
	First Two Syllables	1032.27	47.98
	First Syllable	997.19	43.09
	Second Syllable	1046.54	52.07
	Third Syllable	1043.44	47.11
	Whole Stimulus	905.61	42.05
Sentence Position	**Medial**	948.9	44.43
	Final	1053.61	48.62

6.3.2 Syllable Duration

Disyllabic Stimuli

In disyllabic stimuli, the effects of the interaction between *Tone Condition* and *Pragmatic Focus Domain* and between *Tone Condition* and *Sentence Position* were only significant in the duration of the second syllables (**Table 6.3.e**).

Table 6.3.e Linear mixed-effects models on syllable duration in disyllabic stimuli

Final model	Duration of the First Syllable~ Tone Condition + Pragmatic Focus Domain + Sentence Position + (1\Participant) + (1\Stimulus)			
	SS	df	F	*p*
Tone Condition	61.47	2	5.3168	0.005001 **
Pragmatic Focus Domain	556.07	3	32.0651	<0.0001***
Sentence Position	289.35	1	50.055	<0.0001***
Final model	Duration of the Second Syllable ~ Tone Condition + Pragmatic Focus Domain + Sentence Position + Tone Condition: Pragmatic Focus Domain + Tone Condition: Sentence Position + Pragmatic Focus Domain: Sentence Position + Tone Condition: Pragmatic Focus Domain: Sentence Position + (1\Participant) + (1\Stimulus)			
	SS	df	F	*p*
Tone Condition	672.78	2	514.5092	<0.0001***
Pragmatic Focus Domain	126.40	3	64.4443	<0.0001***
Sentence Position	205.90	1	45.7556	<0.0001***
Tone Condition: Pragmatic Focus Domain	208.10	6	5.2499	<0.0001***
Tone Condition: Sentence Position	218.02	2	15.9156	<0.0001***

Significance levels * =.05 ** = .01 *** = .001

Post-hoc comparisons showed that the duration of the first syllables in different tone conditions was significantly different from each other (*p*s<0.0001); the differences in the duration of the first syllables between different pragmatic focus domains were also significantly different

except for the difference between *No Focus* and *Second Syllable*; the first syllables in sentence final stimuli were significantly longer than those in sentence medial stimuli (p<0.0001).

Table 6.3.f Duration of the first syllable in disyllabic stimuli in each pragmatic focus domains

		Duration (ms)	**SE**
Tone Condition	**Intrinsic NT**	36.24	8.28
	Derived NT	37.63	8.93
	CT	37.38	9.33
Pragmatic Focus Domain	**No Focus**	38.16	2.97
	First Syllable	38.41	3.11
	Second Syllable	37.86	2.98
	Whole Stimulus	33.81	2.65
Sentence Position	**Medial**	35.80	2.94
	Final	38.37	2.99

With regard to the lengthening of the second syllable, similar to the lengthening of the whole stimuli, the Derived NT condition was slightly different from the other two conditions when there was focus on the second syllable. In Derived NT condition, the differences between the pragmatic focus domains were all significant (ps<0.0001). In Intrinsic condition, *No Focus* was significantly shorter than the others (ps<0.0001), and *First Syllable* shorter than *Second Syllable* (p<0.01). In CT condition, differences between pragmatic focus domains were all significant (ps<0.01) except the difference between *Second Syllable* and *Word* **(Figure 6.3.b** and **Table 6.3.g)**. It seems that the differences in the lengthening of the whole stimuli found in Section 6.3.1 were mainly due to the lengthening pattern of the second syllables in disyllabic stimuli, especially in the two NT conditions.

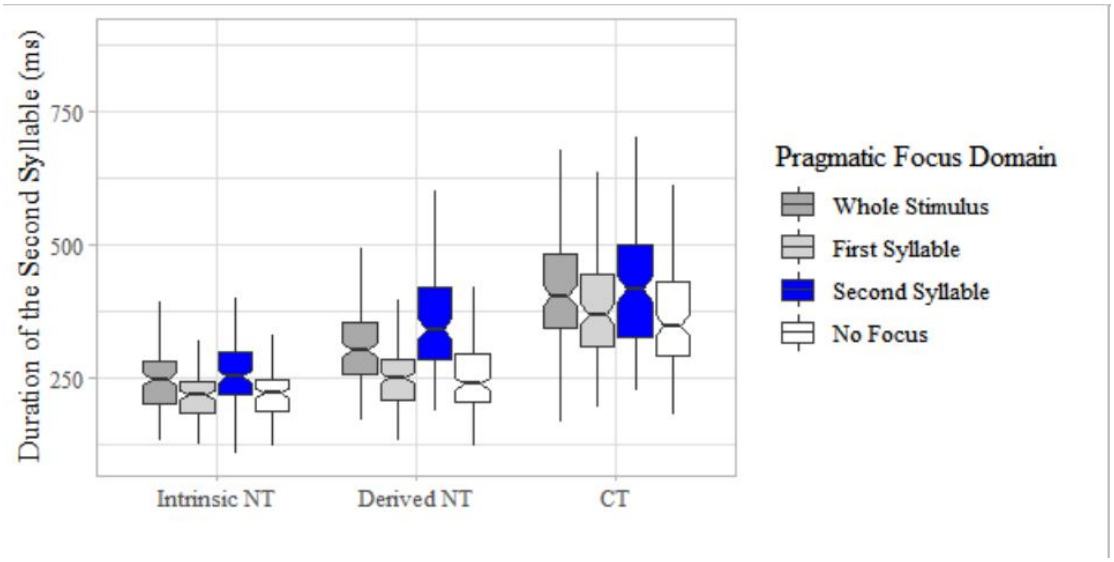


Figure 6.3.b Duration of the second syllable in disyllabic stimuli by *Tone Condition* and *Pragmatic Focus Domain*

Table 6.3.g Duration of the second syllable in disyllabic stimuli

Tone Condition	**Pragmatic Focus Domain**	**Duration**		**Sentence Position**	**Duration**	
		Average (ms)	SE		Average (ms)	SE
Intrinsic NT	**Whole Stimulus**	246.67	2.61	**Medial**	233.71	2.36
	First Syllable	218.51	2.16			
	Second Syllable	257.35	3.41	**Final**	239.53	2.43
	No Focus	221.72	2.49			
Derived NT	**Whole Stimulus**	313.46	6.72	**Medial**	286.37	2.98
	First Syllable	258.82	4.93			
	Second Syllable	358.77	11.03	**Final**	305.61	3.27
	No Focus	252.53	5.24			
CT	**Whole Stimulus**	415.00	10.64	**Medial**	366.27	3.13
	First Syllable	378.37	8.54			

	Second Syllable	429.91	13.94	**Final**	426.90	3.41
	No Focus	363.27	8.63			

In addition, a post-hoc comparison shows that the second syllables in sentence-final CT stimuli were significantly longer than sentence-medial ones (p<0.0001), different from what has been found previously in Experiment 1 and 2. The differences between sentence positions in the two NT conditions, however, were not significant.

Trisyllabic Stimuli with Two NTs

In trisyllabic stimuli with two NTs, the effects of the interaction between *Tone Condition* and *Pragmatic Focus Domain* were found on the duration of the initial and the final syllables, but not the second syllable (**Table 6.3.h**). The syllables in stimuli at sentence-final position were also significantly longer than the medial ones.

Table 6.3.h Linear mixed-effects models on stimulus duration in CT-NT-NT and CT-CT-CT stimuli

Final model	Duration of the First Syllable ~ Tone Condition + Pragmatic Focus Domain + Sentence Position + Tone Condition: Pragmatic Focus Domain + Tone Condition: Sentence Position + Pragmatic Focus Domain: Sentence Position + Tone Condition: Pragmatic Focus Domain: Sentence Position + (1\Participant) + (1\Stimulus)			
	SS	df	F	*p*
Pragmatic Focus Domain	1715.55	6	79.7366	<0.0001***
Sentence Position	187.35	1	43.5379	<0.0001***
Tone Condition: Pragmatic Focus Domain	133.85	10	3.1106	<0.001***
Final model	Duration of the Second Syllable ~ Tone Condition + Pragmatic Focus Domain			

	+ Sentence Position + (1\Participant) + (1\Stimulus)			
	SS	df	F	*p*
Tone Condition	92.30	2	13.793	<0.001***
Pragmatic Focus Domain	930.11	5	55.597	<0.0001***
Sentence Position	104.75	1	31.308	<0.0001***
Final model	Duration of the Third Syllable ~ Tone Condition + Pragmatic Focus Domain + Sentence Position + Tone Condition: Pragmatic Focus Domain + Tone Condition: Sentence Position + Pragmatic Focus Domain: Sentence Position + Tone Condition: Pragmatic Focus Domain: Sentence Position + (1\Participant) + (1\Stimulus)			
	SS	df	F	*p*
Tone Condition	87.99	2	13.9631	<0.001***
Pragmatic Focus Domain	1039.29	5	65.9705	<0.0001***
Sentence Position	469.06	1	148.8728	<0.0001***
Tone Condition: Pragmatic Focus Domain	334.13	10	10.6049	<0.0001***
Tone Condition: Sentence Position	81.58	2	12.9465	<0.0001***

Significance levels * =.05 ** = .01 *** = .001

Regarding the duration of the first syllables, in all three conditions, significant differences were found between *No Focus* and the other pragmatic focus domains, except for *No Focus* and *Third Syllable* in the CT condition (p <0.0001); the results of the post-hoc comparisons between the other pragmatic focus domains were summarized in **Table 6.3.i**. The lengthening patterns of the three tone conditions were similar, but of different magnitude. In other words, the initial CTs in Intrinsic NT and Derived NT conditions were lengthened as well when the other parts of the stimuli were on focus; in the CT condition, focus on the other parts, did not lengthen the initial

CTs as much as when the initial CTs were on focus themselves. However, unlike in the Intrinsic NT condition, lengthening of the initial CTs induced by focus on the final Derived NT was not comparable to the focus on the initial CTs themselves (**Table 6.3.i**, **Table 6.3.j** and **Figure 6.3.c**).

Table 6.3.i Post comparisons between pragmatic focus domains on the duration of the initial syllables within each tone condition in CT-NT-NT and CT-CT-CT stimuli

Pragmatic Focus Domains	**Intrinsic NT**	**Derived NT**	**CT**
Initial Syllable vs. Second Syllable	not significant	not significant	<0.01
Second Syllable vs. Final Syllable	not significant	not significant	not significant
Initial Syllable vs. Final Syllable	not significant	<0.0005	<0.0001
Initial Syllable vs. First Two Syllables	not significant	not significant	not significant
Second Syllable vs. First Two Syllables	not significant	not significant	not significant
Final Syllable vs. First Two Syllables	not significant	<0.05	<0.05
Initial Syllable vs. Whole Stimulus	not significant	<0.0001	<0.0001
Second Syllable vs. Whole Stimulus	not significant	not significant	not significant
Final Syllable vs. Whole Stimulus	not significant	<0.0001	<0.0001
The first two syllable vs. Whole Stimulus	not significant	<0.05	<0.0005

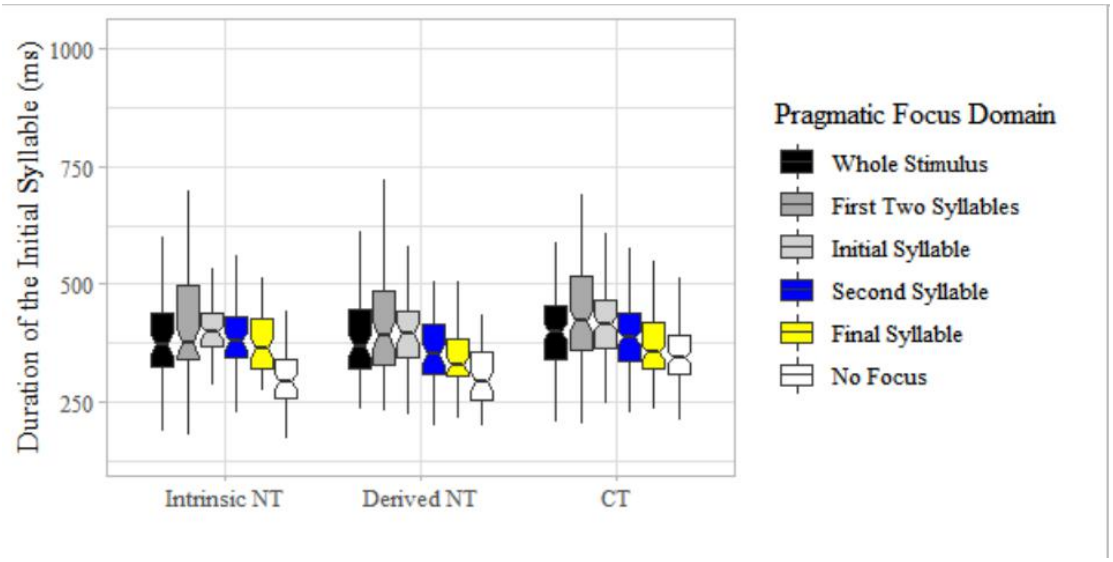


Figure 6.3.c Duration of the initial syllables in CT-NT-NT and CT-CT-CT stimuli by *Tone condition* and *Pragmatic Focus Domain*

Table 6.3.j Duration of the initial syllables in CT-NT-NT and CT-CT-CT stimuli

Tone Condition	Pragmatic Focus Domain	Duration	
		Average (ms)	SE
Intrinsic NT	Whole Stimulus	412.33	3.33
	First Two Syllables	388.41	3.03
	Initial Syllable	409.00	2.4
	Second Syllable	393.00	2.9
	Final Syllable	381.62	3
	No Focus	301.62	2.5
Derived NT	Whole Stimulus	430.71	3.88
	First Two Syllables	387.23	2.96
	Initial Syllable	399.12	2.96
	Second Syllable	366.11	2.81
	Final Syllable	342.90	2.57
	No Focus	304.11	2.46
CT	Whole Stimulus	443.25	3.51
	First Two Syllables	402.71	2.85
	Initial Syllable	420.33	2.86
	Second Syllable	386.81	2.7
	Final Syllable	370.44	2.6
	No Focus	348.79	2.45

Regarding the duration of the second syllables, post-hoc comparisons showed that the second syllables in Intrinsic NT condition was significantly shorter than those in CT (p<0.001) condition and so were the second syllables in Derived NT condition (p<0.05); the differences between pragmatic domains were all significant (ps<0.05), except for the differences between *Second Syllable* and *First Two Syllables*, between *Second Syllable* and *Whole Stimulus* and between *First*

Two Syllables and *Whole Stimulus* (**Table 6.3.k**).

Table 6.3.k Average duration of the second syllables in CT-NT-NT and CT-CT-CT stimuli

		Duration (ms)	SE
Tone Condition	**Intrinsic NT**	200.27	2.38
	Derived NT	24.80	2.67
	CT	31.28	3
Pragmatic Focus Domain	**Whole Stimulus**	28.97	3.25
	First Two Syllables	28.38	3.06
	Initial Syllable	24.56	2.8
	Second Syllable	29.78	3.08
	Final Syllable	26.53	2.75
	No Focus	22.86	2.75
Sentence Position	**Medial**	26.12	2.9
	Final	27.56	3.13

Regarding the duration of the third syllables, Derived NT demonstrated similar lengthening patterns to Intrinsic NT across pragmatic focus domains, but the differences between pragmatic focus domains were of much larger magnitude (**Table 6.3.l**). In Intrinsic NT stimuli, the third syllables were not lengthened much except when they were on focus themselves, and so were the CT stimuli. In Derived NT stimuli, the final Derived NT syllables were not lengthened only when the first syllables were on focus so that *First Syllable* also significantly shorter than the other pragmatic focus domains except *No Focus* (**Table 6.3.l** and **Table 6.3.m** and **Figure 6.3.d**).

Table 6.3.l Post comparisons between pragmatic focus domains on the duration of the final syllables within each tone condition in CT-NT-NT and CT-CT-CT stimuli

	Intrinsic NT	**Derived NT**	**CT**

No Focus vs. Initial Syllable	not significant	not significant	not significant
No Focus vs. Syllable	not significant	<0.0001	not significant
No Focus vs. Final Syllable	<0.05	<0.0001	<0.0001
No Focus vs. First Two Syllables	not significant	<0.0001	not significant
No Focus vs. Whole Stimulus	not significant	<0.0001	not significant
Initial Syllable vs. Second Syllable	not significant	<0.0001	not significant
Second Syllable vs. Final Syllable	not significant	<0.0005	<0.0001
Initial Syllable vs. Final Syllable	<0.01	<0.0001	<0.0001
Initial Syllable vs. First Two Syllables	not significant	<0.0001	not significant
Second Syllable vs. First Two Syllables	not significant	<0.01	not significant
Final Syllable vs. First Two Syllables	not significant	not significant	<0.0001
Initial Syllable vs. Whole Stimulus	not significant	not significant	not significant
Second Syllable vs. Whole Stimulus	not significant	not significant	not significant
Final Syllable vs. Whole Stimulus	not significant	<0.01	<0.0001
The first two syllable vs. Whole Stimulus	not significant	not significant	not significant

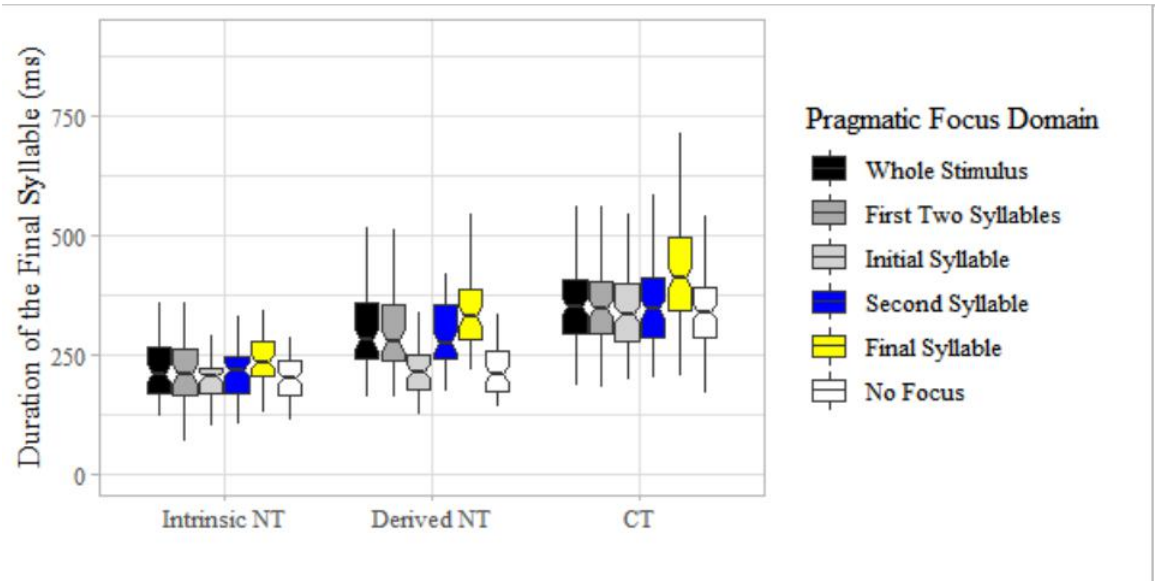


Figure 6.3.d Duration of the final syllables in CT-NT-NT and CT-CT-CT stimuli by *Tone condition* and *Pragmatic Focus Domain*

Table 6.3.m Duration of the final syllables in CT-NT-NT and CT-CT-CT stimuli

Tone Condition	Pragmatic Focus Domain	Duration		Sentence Position	Duration	
		Average (ms)	SE		Average (ms)	SE
Intrinsic NT	**Whole Stimulus**	214.11	6.29	**Medial**	205.6	2.24
	First Two Syllables	220.81	6.13			
	Initial Syllable	200.75	4.24			
	Second Syllable	222.68	6.65			
	Final Syllable	246.38	6.18	**Final**	232.12	2.58
	No Focus	208.42	5.61			
Derived NT	**Whole Stimulus**	295.72	7.87	**Medial**	257.11	2.77
	First Two Syllables	298.76	7.76			
	Initial Syllable	226.13	6.24			
	Second Syllable	293.01	6.71			
	Final Syllable	341.1	7.99	**Final**	299.87	2.86
	No Focus	216.66	4.92			

CT	**Whole Stimulus**	352.93	9.58			
	First Two Syllables	353.12	8.58	**Medial**	334	3.01
	Initial Syllable	342.72	8.62			
	Second Syllable	355.98	8.91			
	Final Syllable	429.82	11.63	**Final**	393.72	3.15
	No Focus	348.22	9.48			

In terms of the interaction between *Tone Condition* and *Sentence Position*, post-hoc comparisons showed that in all three conditions, sentence-final final syllables were significantly longer than the medial ones (p<0.005)

Trisyllabic Stimuli with Intrinsic NTs

In trisyllabic stimuli with Intrinsic NT, the duration of all three syllables were only significantly affected by *Tone Condition*, *Pragmatic Focus* and *Sentence Position*, but not the interaction between them (**Table 6.3.2.n** and **Table 6.3.2.o**).

Table 6.3.n Linear mixed-effects models on stimulus duration in CT-NT-CT and CT-CT-CT stimuli

Final model	Duration of the First Syllable ~ Tone Condition + Pragmatic Focus Domain + Sentence Position + (1\Participant) + (1\Stimulus)			
	SS	df	F	*p*
Tone Condition	83.40	1	22.214	0.0001948 ***
Pragmatic Focus Domain	2101.20	5	111.932	<0.0001***
Sentence Position	314.71	1	83.82	<0.0001***
Final model	Duration of the Second Syllable ~ Tone Condition + Pragmatic Focus Domain + Sentence Position + (1\Participant) + (1\Stimulus)			
	SS	df	F	*p*
Tone Condition	201.22	1	77.89	<0.0001***
Pragmatic Focus Domain	552.94	5	42.808	<0.0001***
Sentence Position	132.20	1	51.171	<0.0001***
Final model	Duration of the Third Syllable ~ Tone Condition + Pragmatic Focus Domain + Sentence Position + (1\Participant) + (1\Stimulus)			

	SS	df	F	*p*
Tone Condition	0.62	1	0.1428	0.7102
Pragmatic Focus Domain	1391.35	5	63.9441	<0.0001***
Sentence Position	1261.85	1	289.9632	<0.0001***

Significance levels * =.05 ** = .01 *** = .001

Table 6.3.o Syllable duration on average in CT-NT-CT and CT-CT-CT stimuli

		The 1st Syllable Duration (ms)	**SE**	**The 2nd Syllable Duration (ms)**	**SE**	**The 3rd Syllable Duration (ms)**	**SE**
Tone Condition	**Intrinsic NT**	356.33	3.12	185.50	2.18	352.58	3.14
	CT	395.42	2.98	312.81	3	363.66	3.15
Pragmatic Focus Domain	**Whole Stimulus**	421.71	3.59	249.22	3.24	348.10	3.1
	First Two Syllables	383.77	2.95	245.43	3.12	351.40	3.03
	First Syllable	397.78	2.91	225.81	2.83	334.03	2.97
	Second Syllable	372.33	2.83	259.33	3.14	347.51	3
	Third Syllable	338.84	2.72	228.68	2.85	416.21	3.35
	No Focus	317.13	2.6	208.33	2.79	344.84	3.03
Sentence Position	**Medial**	359.76	3	229.52	2.92	328.66	2.93
	Final	384.12	3.16	242.74	3.13	385.42	3.21

6.3.3 Duration Ratio

The duration ratio was significantly influenced by *Tone Condition*, *Pragmatic Focus Domain*, the interaction between *Tone Condition* and *Pragmatic Focus Domain* as well as the interaction between *Tone Condition* and *Sentence Position* (**Table 6.3.p**).

Table 6.3.p Linear mixed-effects model on duration ratio

Final model	Duration Ratio ~ Tone Condition + Pragmatic Focus Domain + Sentence Position + Tone Condition: Pragmatic Focus Domain + Tone Condition: Sentence Position + Pragmatic Focus Domain: Sentence Position + Tone Condition: Pragmatic Focus Domain: Sentence Position + (1\Participant) + (1\Stimulus)			
	SS	df	F	*p*
Tone Condition	0.94	2	11.6575	<0.0.0005***
Pragmatic Focus Domain	5.67	3	47.1107	<0.0001***
Tone Condition: Pragmatic Focus Domain	0.51	6	2.1357	<0.05*
Tone Condition: Sentence Position	1.17	2	14.5396	<0.0001***

Significance levels * =.05 ** = .01 *** = .001

In Intrinsic NT and CT conditions, the duration ratio of *First syllable* was significantly lower than the others (*p*s<0.05), but the ratio differences between the others were not significant. In contrast, the focus on the first or second syllable, significantly changed the duration ratio in Derived NT conditions from *No Focus* and *Whole Stimuli* (*p*s < 0.005) (**Table 6.3.q** and **Figure 6.3.e**). *Whole Stimuli* increased the duration ratio of Derived NT words with marginal significance (*p* = .051), whereas the increase of duration ratio by *Whole Stimuli* in Intrinsic NT words was not significance. It is worth noting that except when focus was the second syllable, significant differences were found between duration ratio between Derived NT and CT words (*p*s<0.05.), but when Derived NT was on focus, the differences between Derived NT and CT disappeared,

echoing Experiment 1 and 2. No significant differences were found between Intrinsic NT and Derived NT in focused or unfocused situations, regardless of pragmatic focus domains. In addition, *sentence-final words* had a significantly larger duration ratio than sentence-middle words but only in the CT condition, (p <0.001).

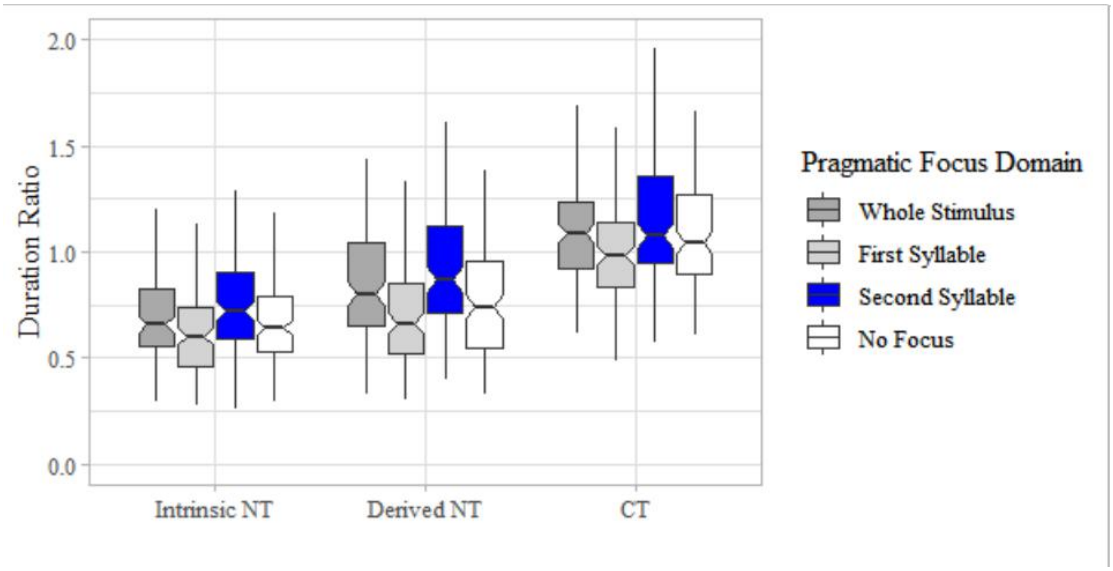


Figure 6.3.e Duration ratio in disyllabic stimuli by *Tone Condition* and *Pragmatic Focus Domain*

Table 6.3.q Duration ratio in disyllabic stimuli

Tone Condition	**Pragmatic Focus Domain**	**Duration Ratio**		**Sentence Position**	**Duration Ratio**	
		Average	SE		Average	SE
Intrinsic NT	**Whole Stimulus**	0.70	0.02	**Medial**	0.81	0.02
	First Syllable	0.62	0.02			
	Second Syllable	0.75	0.02	**Final**	0.83	0.02
	No Focus	0.69	0.02			
Derived NT	**Whole Stimulus**	0.85	0.03	**Medial**	0.71	0.02
	First Syllable	0.71	0.02			
	Second Syllable	0.93	0.02	**Final**	0.67	0.01
	No Focus	0.78	0.03			

CT	**Whole Stimulus**	1.10	0.02	**Medial**	1.04	0.02
	First Syllable	0.99	0.02			
	Second Syllable	1.17	0.03	**Final**	1.13	0.02
	No Focus	1.09	0.02			

6.4 Discussion

The present experiment finds that corrective focus on any part introduces robust lengthening of the whole constituent, regardless of whether it involves Intrinsic NT, Derived NT, or CTs only. The distribution of lengthening within the stimulus was similar in patterns across conditions, but the magnitude of lengthening was impacted by the tones involved and the size and location of the pragmatic focus domain.

With regard to P1 and P2, namely, whether the distribution of lengthening in Intrinsic NT stimuli resemble English words and whether Derived NT differs from NT, the current findings suggest that the two types of NT were similar to each other, and to CT in general.

Unlike the unstressed syllables in English or Swedish, the current findings suggest that NT-bearing syllables are more likely to be lengthened by the focus on the larger domain rather than be lengthened through the spill-over effect of the lengthened preceding CTs. The lengthening of the initial CT-bearing syllables triggered by the focus on the whole constituents was roughly the same as the lengthening triggered by the focus on these syllables themselves in all disyllabic and trisyllabic stimuli. However, the spill-over lengthening from the focused preceding CTs onto the following Intrinsic NT syllables was not comparable in magnitude to the lengthening of these syllables induced by the focus on the larger prosodic constituents or on these NT-bearing syllables

themselves. In fact, the spill-over effect from the focused initial CTs onto the final NT-bearing syllables within the same trisyllabic phrases with two NTs was not significant. Compared to Intrinsic NT, Derived NT was lengthened more by the focus on a larger domain, probably due to the existence of the underlying tones.

In addition, when a larger domain is on focus, the durational contrast within Intrinsic NT and Derived NT was not intensified either. When the whole stimuli were on focus, the duration ratio of disyllabic Intrinsic NT stimuli was unchanged and the duration ratio of disyllabic Derived NT stimuli were only marginally increased. Again, the results demonstrated that syllables bearing either NT were lengthened by focus on the larger domain, just like CTs.

Regarding the spill-over effect, the results are complex at the first glance. Although the rightward spreading from NT onto the following NTs was found in some situations, but the magnitude was relatively small. Moreover, even when the syllables on the left carried CTs, the NT syllables on the right were not lengthened much when the left syllables were on focus; in fact, instead of lengthening the following syllables that bear Intrinsic NT, the focus on the preceding CT syllables in disyllabic Intrinsic NT words tended to shorten them. In contrast, when the second or third NT-bearing syllables were on focus, the initial CT-bearing syllables were lengthened significantly. Moreover, in trisyllabic stimuli the initial CT-bearing syllables were not significantly longer when they were on focus themselves than when the following NT, either Intrinsic NT or Derived NT, was on focus. This finding contradicts what was predicted in P1, i.e., there was little spill-over lengthening on the syllables around Intrinsic NT, but seemed to confirmed P2, i.e., the syllables neighboring Derived NT were lengthened due to the spill-over effect. This tendency to lengthen the CT-bearing syllable before NT was also observed in the two experiments in Chapter 4,

but did not reach statistical significance.

However, if there was a leftward spreading in Mandarin, the most obvious lengthening should be observed on CTs preceding CTs rather than CTs preceding NTs, as CT-bearing syllables assumed to be of more metrical strength. Since this was not observed, I proposed that it is better to see the leftward spreading in either Intrinsic NT or Derived NT condition as a joint effect of 1) the implementation of focus on the syllable that is pragmatically focused and 2) the requirement to keep the heavy-light pattern of the NT words rather than a strong leftwards spill-over effect.

This dependence on the preceding CT-bearing syllables in the realization of focus-induced lengthening found in NT echoes NT's dependence on the preceding CT in the tonal realization, and confirms the metrical lightness of both types of NT. Even though Derived NT was lengthened to a large extent when on focus, it still triggered larger lengthening on the preceding CT than the second-syllable CTs did.

Regarding the final lengthening, P4 is rejected because the present results found longer syllables at the utterance final rather than the utterance medial position in both disyllabic and trisyllabic stimuli; in particular, it was often in CT stimuli that the final syllables were significantly longer than the medial ones. This finding is also different from Experiment 1 and Chen's results (2006).

In fact, an additive tendency was already observed in Experiment 2. There, when there was focus with emphasis on the final CT-bearing syllables, the absolute duration of the syllable increased; the duration was significantly longer than the unfocused final CT-bearing syllables (Section 4.3). This non-uniform finding needs to be explained with the potential mechanism

behind the non-additive interaction between lengthening caused by different linguistic factors.

A widely used account is the constraint of expandability on lengthening (Cooper et al's, 1985), which is assumed to correlate with language-specific phonological features. For instance, Cambier-Langeveld (2000) suggests that it is the phonemic contrasts of vowel length of Dutch that prevents the Dutch speakers to implement lengthening in an additive manner, or else, ambiguity may be caused. The lexical tonal systems in Mandarin, similarly, may reduce the room for multiple durational contrasts as well, but not as much as the vocalic contrast in Dutch. After all, it is still possible to expand the f_0 height and range over a longer duration without distorting the contour much, as Experiment 1 and 2 and other studies have demonstrated. Therefore, there may be a tendency towards a non-additive interaction between final and focus-induced lengthening in Mandarin, but not a compulsory constraint of expandability.

To summarize, this experiment presents two key findings. Firstly, within the same language, the distribution of focus-induced lengthening demonstrates a relatively uniform pattern; although NT words show heavy-light patterns like the words with lexical stress in stress languages, they still demonstrate larger similarity of CTs. Secondly, it suggests that both Intrinsic NT and Derived NT are likely to be associated with lighter metrical weight compared to CTs, reflected in the lengthening they caused in the preceding CTs when on focus.

Chapter 7 Intonation on NT

7.1 Intonation and Tone

In tone languages, tone and intonation are the two most significant prosodic features that use f_0 as the primary acoustic correlate. At a lexical level, f_0 is employed to distinguish word meanings, and at utterance level, to convey post-lexical information such as discourse function (e.g., Interrogatives vs. Declaratives). The ways in which lexical and utterance-level prosody interact and the mechanisms behind the interaction have attracted much research attention.

Compared to non-tone languages, there is more restriction on intonation realization in tone languages, but the degree of restriction may not be as large as commonly expected. The restriction comes mainly from the need to retain a recognizable tonal contour of each lexical item. Therefore, the complete obliteration of lexical tones by intonational tone is not permitted, especially in morpheme tone languages (Yip, 2002). Instead, the intonation is realized through the change of the pitch register of the whole utterance, the insertion of boundary tones (e.g., model particles in Cantonese), register re-set, and suspension of the downdrift (Gussenhoven, 2018; Yip, 2002; Pulleyblank, 1986; Chao, 1965), and different languages use these mechanisms in different ways under the influence of syntactic and pragmatic factors. Although some linguists have put forward universal hypotheses like the Frequency Code (e.g., accounting for the fact that declarative intonation tends to be realized with falling pitch cross-linguistically while interrogatives and especially yes/no questions are typically realized with a rising pitch; Ohala, 1983; Gussenhoven, 2002), our knowledge and understand of what is universal in intonation realization is still very limited (Gussenhoven, 2004, 2016; DiCanio et al, 2018). Potential counter-evidence against the Frequency Code has been found in a number of Polynesian languages such as Hawaiian and a

number of West-Atlantic languages. There, yes/no-questions are realized with general falling contours (Rialland, 2007).

Research has put forward different theories to model the realization, perception and processing of intonation in tone and non-tone languages, such as *the small ripples riding on larger waves* metaphor by Chao (1965), statistical models based on corpus data (e.g., Chen et al, 1992), static target and interpolation model (Shih, 1998; Duanmu, 2007), physiology-based Soft TEMplate Mark-up Language (Kochanski and Shih, 2003) and articulation-oriented parallel encoding and target approximation model (Xu and Wang, 2001; Xu, 2004). Some of these models assume that intonation determines the pitch register in which lexical tones are realized but no other explicit interaction between tone and intonation (e.g., Chen et al, 1992). However, empirical data demonstrate that CTs and intonation interact with each other in realization and perception in Mandarin as well as in other tone languages (e.g., Xu and Mok, 2014; Ren et al, 2013; Yuan, 2006, 2011; Shih, 2004).

Since no large distortion of the tone contours is found in declarative sentences without utterance focus, intonation contours in Mandarin tend to be interpreted as statements by default unless the prosodic 'question markers' are present. According to Yuan (2004), questions can only be identified when the actual question mechanisms are heard like a high peak, late peak, or terminal rise (House, 2003; Gosy and Terken, 1994). From an acoustic perspective, such a terminal rise is manifested as the higher articulatory strength of sentence final CTs and sometimes the raising of the overall f_0 of the utterance; hence changes the surface f_0 contours of the final tones (Yuan, 2006; Shih, 2004; Shen, 1989). The tonal effect on intonation, in contrast, is mainly determined by the strength of CTs and therefore is also correlated to the length and syntactic

structure of the utterance (Shih, 2004). To be specific, when there is corrective focus on certain CT-bearing syllables, f_0 differences in paired question-statement is not the same as the differences in sentence pairs without focus.

As expected, questions are also harder to be perceived than statements in general. There is often a bias towards declaratives rather than interrogatives in intonation identification (Ren et al, 2016; Yuan, 2006, 2011; Yuan and Shih, 2004), and there are tonal effects on question intonation perception. Yes/no questions are found to be the easiest to be identified on a sentence ending with falling T4 but the hardest on a sentence ending with rising T2 by Yuan (2006, 2011), because there is more distortion of the falling contour of T4. According to Yuan (2006), question intonation flattens the falling contour of T4 but exaggerates the rising contour of T2. Ren et al (2016) further demonstrate that question-statement contrast elicits a clear mismatch negativity on monosyllables carrying falling T4, but not on those carrying rising T2. These findings suggest that the identity of (sentence-final) tone affects the process of mapping f_0 contours to the corresponding intonation types. In other words, the tonal category is required in if not before, the processing of the intonation type. Therefore, tone and intonation are likely to interact at the phonological rather than the mere phonetic level (Yuan, 2011).

Liu et al (2020), however, reports slightly different results. They only find a large asymmetry in highly constrained semantic contexts between the perception of interrogatives between T2 and T4; in semantically neutral contexts questions with T2 and T4 are equally difficult to perceive. This perceptual asymmetry suggests lexical tone and intonation interacts with each other, and semantic meaning can help listeners separate intonational information from tonal information when the lexical tone is a falling tone. Therefore, it appears that the perception of intonation relies

more on post-lexical information.

It should also be noted that in Liu's study (2020), questions ending in T4 do not only reach a higher f_0 than statements ending with T4, but also showed a distinctly higher f_0 at the initial. In other words, in Liu's study, the f_0 curve of T4 was not flattened as much as in Yuan's study. Therefore, it may also be that large psychoacoustic changes are of great importance to intonation perception, namely, the distortion to the opposite direction to the phonological tone may facilitate intonation identification. To make concrete arguments on the interaction between intonation and tone, further investigation is still required.

All the above-mentioned studies have focused on CTs. How interrogative intonation is realized on NT words is only investigated by Liu and Xu (2007) from an acoustic perspective. They explored the f_0 trajectory of interrogative and declarative sentences ending with 5 NTs pronounced in a row, in comparison with sentences ending with 3 NTs and 2 high-level tones (i.e., T1). The NTs involved in this study were all Intrinsic NT by the definition in this thesis. Liu and Xu (2007) find that sentence-final NTs still have a gradual falling contour even in questions, differing from the sentence-final T1s which have a slightly rising contour. The effects of intonation and focus have been mainly manifested on the final CTs in utterance rather than NTs. The tonal changes of NT-bearing syllables still come from the changes intonation caused in CTs, resulting into indirect interactions between NT and intonation. Liu and Xu (2007) interpret these results as the underspecified tonal target NT does not facilitate the realization of interrogative intonation; instead, the major interfering effects are still found between intonation and CTs. This interpretation echoes the tone-strength argument Shih has proposed (2004) and further indicates that the linguistic units that cannot keep underlying lexical tonal targets cannot bear intonational

tones either.

Slightly different from Liu and Xu's findings, Li et al (2018) report that questions with or without the interrogative marker /ma/ at the end all have a higher f_0 overall than declarative statements or rhetoric question with an ascending tendency, regardless of the identity of the final tone. These results may indicate that there is a rising contour of question-final NT, a tendency that is relatively universal, though this was not mentioned explicitly by Li et al (2018). The differences between Liu and Xu's results and Li et al's results may result from two reasons. Firstly, the sentences in Liu and Xu's study were much longer than those in Li et al's. The longer the utterance is, the harder it is to raise the overall f_0 so that Li et al (2018) have found a more obvious register change than Liu and Xu (2007). Moreover, the 5 consecutive NT in Liu's study together with the preceding CT has resulted into a super foot (or a super prosodic word) that needs to be realized in one breath without pauses. This is not only slightly unnatural to me, a native speaker of Mandarin, but may also trigger breathy voice and hence the lowering of the pitch.

Liu and Xu's results seem to suggest that the un(der)specified tonal target of Intrinsic NT does not facilitate the implementation of the terminal rising in interrogative questions much. Instead, it is the preceding CTs that get raised by interrogative intonation; the falling Intrinsic NTs still demonstrate a gradual falling contour when given more time (i.e., realized in a consecutive manner). This finding may further indicate that although intonation in Mandarin is sensitive to the local prosodic mechanism, it does not interfere with the phonological derivation of the lexical tones.

Nevertheless, to what extent the underlying un(der)specification of Intrinsic NT influences

the perception of intonation on NT words remains untested. The interaction between Derived NT and interrogative questions has not been explored neither. Based on the different underlying tonal differences, I make the following hypothesis with regard to the perception of intonation on NTs.

H5: The perception of intonation on Intrinsic NTs is easier than on Derived NTs and CTs due to the absence of underlying tone.

Testing H5 will not only shed further light on the representations of NTs, but also improve our understanding of the perceptual mechanism of intonation in Mandarin.

7.2 Methodology

Participants

30 Northern Mandarin speakers (13 males, 17 females) aged between 18-29 (mean age 21.3) participated in the experiment. All participants were current students at the Shanghai Jiao Tong University, but had completed their pre-university education in the Huabei or the most northern part of Huadong region and reported Northern Mandarin as the main language they used in school and at home. All of them were right-handed and none of them reported any hearing impairments. Informed consent was obtained prior to the experiment.

Stimuli

Five Intrinsic NT morphemes, five T2-bearing morphemes and five T4-bearing morphemes were chosen, and each was combined with eight different preceding T1-bearing morphemes, resulting in 120 words with Intrinsic NT and CT. Due to the limitation of the natural language, only six words with T1 and Derived NT with underlying T2 are available. Therefore, another six words with T1 and Derived NT with underlying T4 were chosen, resulting in a total of 132 total

stimulus words (complete list of stimuli in Appendix D). 18 T1-T1 words and 18 T1-T3 words were added as fillers. T1 was used as the preceding tone for two reasons: to create an experiment of reasonable length, only one CT could be used and the high-level contour of T1 enabled the most natural manipulation. To minimize the effect of morphological status, all T2 and T4 morphemes were localizers and classifier.

The stimuli were recorded by myself, a native northern Mandarin speaker, and another female speaker with very similar language background. Both speakers hold Level 1 (the top level) certificate of the National Mandarin Test. Two speakers rather than one were recorded to ensure that the present experiment was not just testing the varying surface acoustic differences.

The recording was done separately by each speaker in a quiet room with Zoom H1 handy recorder at 96000Hz/26Bit. Each speaker did the recording twice, once by stimulus (i.e., recording each stimulus twice, one in question intonation and one in statement intonation consecutively) and once by intonation type (i.e., recording all stimuli in question intonation and all stimuli in statement intonation). However, it turned out that the recording by intonation resulted in an undesirable ascending pitch pattern in interrogatives and down-drifting in declaratives between different stimulus words recorded adjacently. Therefore, only the recordings by stimulus were used in the perception experiment.

Special attention was taken to produce clear and accurate stimuli to ease the manipulation later. Stimuli that were unclear, or of too much co-articulation to be further manipulated were re-recorded. The naturalness of the stimuli was cross-examined by the two speakers as well as a naive male northern Mandarin speaker by asking them to pick out the unnatural stimuli.

The recordings were cross-spliced to neutralize the acoustic parameters of the preceding T1-bearing syllables between the two intonation conditions (declarative and interrogative) using *Praat* (Boersma and Weenink, 2021). For each recording, the initial syllable and second syllable were separated into two sound files. The pitch height, pitch range and duration of the initial syllable were manipulated into the average pitch height, pitch range and duration of the statement and the question versions of the same stimulus word produced across all words by the same speaker. The second syllables were then spliced onto the manipulated initial syllable realized with the intonation type, that is, the second syllables in statement intonation were attached to the initial syllables in question intonation of the same word and speaker and vice versa. Stimuli that could not be naturally separated or manipulated were re-recorded by stimulus and the same three people also examined the naturalness of the re-recorded stimuli.

In total, 37 stimuli were re-recorded. The average intensity of the recordings was scaled to 75dB. The digitally edited recordings were judged as natural by the speaker other than the first author and two native speakers who did not participate in the study. The stimuli were separated into two sets of 238 word-pairs with different intonations (i.e., 476 stimuli in total). Each set had half of the stimulus words from one speaker and the other half from the other speaker. The order of the stimuli was pseudorandomized by tone, intonation and speaker. 15 participants were tested with one set and 15 tested with the other.

The acoustic analyses of the manipulated stimuli were performed in *Praat* (Boersma and Weenink, 2021). For f_0 contours, a self-written *Praat* script was applied to extract f_0 values (converted to semitones) of the sonorous part in the second syllable of each stimulus. The f_0 contours were time-normalized by dividing the sonorous parts into 20 equal intervals and f_0 values

were extracted at each 10% step. The last value was excluded to reduce final creakiness. It should be noticed, however, there are still much final creakiness in NT and T4 realized with statement intonation so that their averaged f_0 contours demonstrated some final abnormalities in **Figure 7.2.a**.

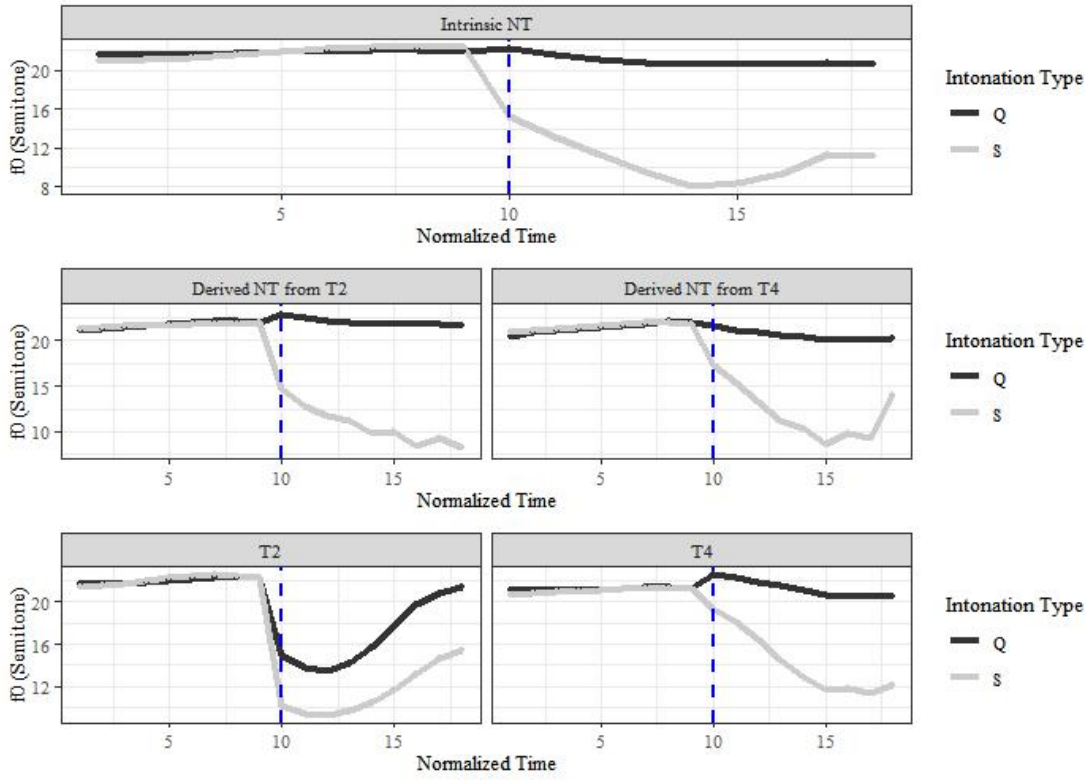


Figure 7.2.a Averaged f_0 contour of all the stimuli by *Tone* and *Intonation* (the blue dashed line separates the first and the second syllable)

I further analyzed the f_0 height and range (i.e., the difference between the minimum and maximum f_0 values) of the second tones of the stimuli with creakiness less than 40%. Linear mixed effect (LME) models were built to test the effects tone, intonation, speaker as well as their interaction on f_0 height and range. The models were built through a process similar to Experiment 1, 2 and 5. LME showed that tone, intonation, the interaction between tone and intonation and the interaction between tone and speaker had significant effects on both the f_0 height and range of the

second syllables (**Table 7.2.a)**. Post-hoc comparison showed that the average pitch height of all tones in the second syllables was significantly raised by question intonation; the average ranges of Intrinsic NT, Derived NT from T2 or T4 and T4 were significantly reduced in question while the average range of T2 was significantly increased (**Table 7.2.b**). Furthermore, the interaction between speaker and tone did show systematic differences, namely, when realizing the same tone, the two speakers did not differ from each other in either the f_0 height or range.

Table 7.2.a Liner mixed effect models on the average f_0 height and range of the second syllable

Final model	Average f_0 Height ~ Tone + Intonation + Speaker + Tone:Intonation + Tone:Speaker + Intonation:Speaker + (1\Stimuli)			
	SS	df	F	*p*
Tone	635.77	4	68.9	<0.0001***
Intonation	1192.5	1	516.91	<0.0001***
Tone:Intonation	187.77	4	20.35	<0.0001***
Tone:Speaker	62.36	4	6.75	<0.0001***
Final model	Average f_0 Range ~ Tone + Intonation + Speaker + Tone:Intonation + Tone:Speaker + Intonation:Speaker + (1\Stimuli)			
	SS	df	F	*p*
Tone	559.97	4	28.41	<0.0001***
Intonation	449.65	1	91.25	<0.0001***
Tone:Intonation	1572.5	4	79.78	<0.0001***
Tone:Speaker	73.03	4	3.71	<0.01**

Significance levels * =.05 ** = .01 *** = .001

Table 7.2.b Average f_0 height and range of second-syllable Intrinsic NT, Derived NT from T2, Derived NT from T4, T2 and T4 in questions and statements

Tone	**Intonation**	**Height**			**Range**		
		Average (Semitones)	SE	Post-hoc Comparison results	Average (Semitones)	SE	Post-hoc Comparison results
Intrinsic NT	**Q**	20.90	0.33	$p < 0.0001$	2.05	0.22	$p < 0.0001$
	S	11.11	0.50		8.07	0.68	
Derived NT2	**Q**	20.94	0.75	$p < 0.0001$	1.55	0.16	$p < 0.0001$
	S	14.60	0.73		6.81	0.60	
Derived NT4	**Q**	21.57	0.08	$p < 0.0001$	1.56	0.03	$p < 0.0001$
	S	13.79	0.31		7.60	0.31	
T2	**Q**	16.59	0.28	$p < 0.0001$	10.48	0.55	$p < 0.0001$
	S	11.50	0.20		6.68	0.24	
T4	**Q**	21.29	0.24	$p < 0.0001$	2.43	0.28	$p < 0.0001$
	S	14.59	0.26		10.21	0.60	

Duration differences were not further examined in some previous studies as they had a larger temporal scope of tone-intonation interaction by using sentences. Given that f_0 did not start to increase significantly until the pre-last syllable in their study, their investigation was more sympathetic towards the final-rising theory. However, the present study used a much smaller temporal scope so that duration ratio in each tone condition and intonation of the stimuli (= duration of 2nd syllable/ duration of the neutralized 1st syllable) was calculated. Linear effect models were established using tone, intonation, speaker and their interactions through the same process as experiment 1, 2 and 5. The models demonstrated that duration ratio was significantly influenced by tone and the interaction between tone and intonation (**Table 7.2.c**). Post-hoc comparison results were shown in **Table 7.2.d**.

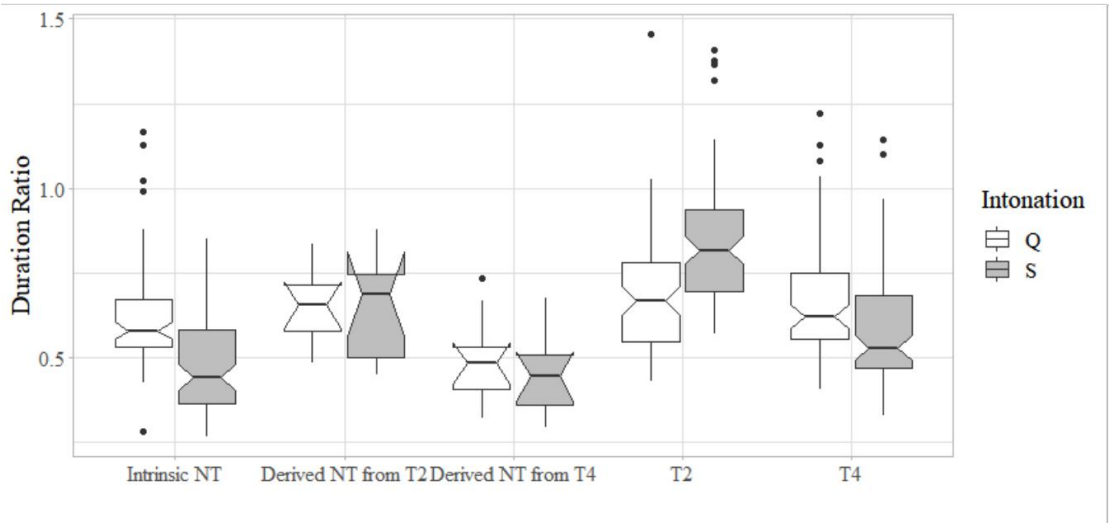


Figure 7.2.b Duration ratio of the stimuli by *Tone* and *Intonation*

Despite the differences in duration ratio, Intrinsic NT and Derived NT had similar absolute duration, which was much shorter than the second syllables bearing T2 and T4 (**Figure 7.2.c**). The linear effect model demonstrated that the duration of the second syllables was also significantly influenced by tone and the interaction between tone and intonation (**Table 7.2.c**). In addition, the

duration differences between Intrinsic NT and Derived NT from T2 or T4 were not significant, but the differences between the NTs and T2 or T4 were all statically significant (*ps* <0.005) (**Table 7.2.d**). These findings were slightly different from Yuan (2006) as no duration differences have been found between interrogative T2 and declarative T2 in that study.

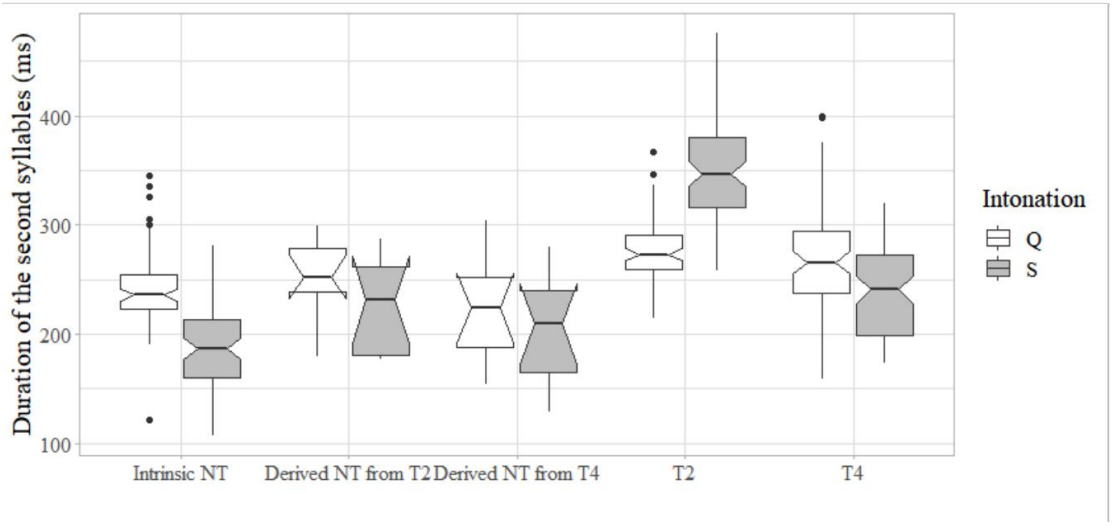


Figure 7.2.c Duration of the second syllables in all stimuli by *Tone* and *Intonation*

Table 7.2.c Liner effect models on the average duration ratio and the 2nd syllable duration of Intrinsic NT, Derived NT from T2 and T2, T2 and T4 in question and statement

Final model	Duration Ratio ~ Tone + Intonation + Speaker + Tone:Intonation + Tone:Speaker + Intonation:Speaker + (1\Stimuli)			
	SS	df	F	*p*
Tone	1.21	4	20.78	<0. 0001***
Tone:Intonation	0.52	4	8.82	<0. 0001***
Final model	Duration ~ Tone + Intonation + Speaker + Tone:Intonation + Tone:Speaker + Intonation:Speaker + (1\Stimuli)			
	SS	df	F	*p*
Tone	230.31	4	66.74	<0. 0001***
Tone:Intonation	100.12	4	29.68	<0. 0001***

Significance levels * =.05 ** = .01 *** = .001

Table 7.2.d Average duration ratio and the 2nd syllable duration of Intrinsic NT, Derived NT from T2 and T2, T2 and T4 in question and statement

Tone	**Intonation**	**Duration ratio**			**The 2nd syllable duration**		
		Average	SE	Post-hoc Comparison results	Aver**age (ms)**	SE	Post-hoc Comparison results
Intrinsic NT	**Q**	0.62	0.03	$p < 0.01$	239.86	6.20	$p < 0.0001$
	S	0.48	0.02		188.32	5.05	
Derived NT2	**Q**	0.70	0.03	not significant	250.90	6.87	not significant
	S	0.83	0.03		225.07	6.45	
Derived NT4	**Q**	0.68	0.03	not significant	225.46	6.99	not significant
	S	0.58	0.03		203.18	7.88	
T2	**Q**	0.66	0.02	$p < 0.0001$	275.73	7.96	$p < 0.0001$
	S	0.64	0.02		349.70	5.82	
T4	**Q**	0.50	0.02	$p < 0.0001$	269.16	7.77	$p < 0.0001$
	S	0.44	0.02		237.45	6.54	

Procedure

The experiment was programmed in PsychoPy 3.0 (Peirce et al, 2019). In each trial, two words 'Trial Starts' appeared for 70 milliseconds. Participants would then be listening to a manipulated recording while seeing a screen that contained two horizontally arranged icons, ‘?’ and ‘!’. ‘.’ was not used to index statement due to the visual imbalance between ‘.’ and ‘?’. To avoid confusion, it was made clear in the instruction that ‘!’ means statement only rather than exclamation. The participants were asked to indicate their choice by pressing the keys labelled as ‘right’ or ‘left’. After the participants made their choice, they would enter the next trial automatically. If they did not make choice within 3000 milliseconds, this trial would end and enter the next trial automatically. The '?' was on the left for half of the participants and right for the other half to avoid the interference of right-handedness.

Each experiment consisted of 304 test trials, 36 fillers and 42 repeated test trials (382 trials in total) with two breaks available in between. The participants were able to skip the break if they wanted. The 42 repeated trials were words from the other stimulus set that appeared only once either in statement or in question intonation to prevent per-expectation (i.e., expecting the word to carry in the other intonation type when appearing the second time). 9 practice trials were given at the beginning to help the participants familiarize themselves with the procedure. The first author was with the participants when they were doing the practice trials in case the participants had questions, but left them alone in the question rooms in the psycholinguistic center in Shanghai Jiao Tong University. The whole experiment took 35-40 minutes including the instruction and the practice trials.

Data analysis

Identification accuracy was calculated to measure how well an intonational function (i.e., declarative vs. interrogative) was recognized by the listeners. The response was considered as accurate only if the intonation was identified as the same intonational function that the speakers were asked to produce. The effects of *Tone*, *Intonation* and individual differences between *Speakers* on identification accuracy, a binary categorical variable (Accurate vs. Inaccurate), were evaluated by a logistic mixed effects model, using glmer in the lmerTest package (Kuznetsova et al, 2017) in R (R core team, 2020). I selected the optimal fixed structure by using stepwise comparisons from the most complex effect to the simplest and the optimal random effect structure according to the smallest Akaike Information Criterion (AIC). The *ANOVA* served to compare different models to determine whether excluding factors from the analysis led to a better fit (Field, Miles, and Field, 2012). The details of the final models are presented with the results.

According to signal detection theory (Macmillan and Creelman, 2005), identification involves not only the ability to discriminate between the two intonation conditions but also the bias towards one of them in ambiguous situations, which according to previous studies (e.g., Yuan, 2006) should be towards a declarative interpretation. Therefore, I also measured the discriminability of each intonation with A' and bias with B''_D. A' ranges between 0 and 1, 1 indicating maximum performance and 0.5 indicating chance performance. B''_D ranges from -1 and 1, 1 indicating maximum bias towards statement and -1 indicating maximum bias towards question. I pooled over tone conditions, i.e., Intrinsic NT, Derived NT2, Derived NT4, T2 and T4, to calculate A' and B''_D.

Reaction time (= the time of key-pressing *minus* the offset time of the auditory stimulus) was

collected alongside as a measure of the difficulty of identifying intonation. Outliers in the reaction time data were removed following the Interquartile Rule (Tukey, 1977), and the effects of *Tone*, *Intonation* and individual differences between *Speakers* on reaction time (a continuous numeric variable), were evaluated by a linear mixed effect model. The model was established through a similar process as Experiment 1, 2 and 5.

The acoustic correlates (e.g., duration ratio, average pitch and height of the first and second syllables) of the intonation identification were analyzed further in order to reveal what kind of changes in which acoustic correlate(s) help or prohibit intonation identification for each tone. I separated the data into 5 subsets by tone. For identification results, I arbitrarily assigned the value 0 to statement and 1 to question and established logistic regression models to explore what changes in which correlates may lead the participants to identify the stimulus as declarative statement or interrogative question, using stepwise comparison. If the coefficient of a correlate is positive, the increase in it will increase possibility for the stimuli to be identified as a question; if the coefficient of a correlate is negative, the increase in it will increase possibility for the stimuli to be identified as a statement. For reaction time, I established linear regression models to evaluate the effects of the acoustic correlates using stepwise comparison. If the coefficient of a correlate is positive, the increase in it will lead to a longer reaction time.

Predictions

Acoustic analyses of the stimuli revealed large similarities between Intrinsic NT, Derived NT, and T4 in f_0 in both types of intonation, regardless of the underlying tones of Derived NT. Therefore, even if the underlying tones interact with intonation as hypothesized in H5, the perception of intonation on Derived NT overall should still be easier than that on CTs, at least,

than on T2, due to the surface similarities of Intrinsic NT and T4. Specified predictions are as follows:

P1: If H5 is supported, I expect to find the highest identification accuracy in the Intrinsic NT condition, followed by T4, Derived NT from T4, Derived NT from T2 and T2, but this trend is more clearly found in the identification of questions rather than statements;

P2: If H5 is supported, I expect to find the least identification bias towards statements in the Intrinsic NT condition followed by T4, Derived NT from T4, Derived NT from T2, and T2;

P3: If H5 is supported, I expect to find the shortest reaction time in the Intrinsic NT condition, followed by T4, Derived NT from T4, Derived NT from T2, and T2, but this trend is more clearly found in the identification of questions rather than statements.

7.3 Results

Identification Accuracy and Bias

Results of an LMER found significant effects of *Tone*, *Intonation* and individual differences between *Speakers*, as well as an interaction between *Tone* and *Intonation* on identification accuracy (**Table 7.3.a**). The identification accuracy by *Tone* and *Intonation* is reported in **Table 7.3.b**. Post hoc comparisons showed that the accuracy differences between the two intonations were significant on T2 and T4 (ps<0.001) but not on Intrinsic NT or the two Derived NTs. With regard to the identification of question, the accuracy on T2 was significantly lower than that on the other tones (ps<0.001) as well; moreover, the accuracy on T4 was also signification lower than that on Intrinsic NT (p<0.005). With regard to the identification of statement, the accuracy on T2 was significantly lower than that on the other tones (ps<0.001). It is worth mentioning that overall

the participants had a high accuracy identifying recording from the author herself (91.39%) than from the other speaker (87.64%). However, no significant interactions between speaker and tone/intonation was found. This suggests that the inter-speaker differences in identification accuracy does not affect the effects we found in tone and intonations.

Table 7.3.a Logistic mixed-effects model on identification accuracy

Final model Analysis	Identification Accuracy ~ Tone*Intonation* Speaker + (1\Participant) + (1\Stimulus)		
Main Effects	SS	df	*p*
Intonation	47.541	1	<0.0001***
Tone	105.784	4	<0.0001***
Speaker	33.234	1	<0.0001***
Intonation: Tone	38.152	4	<0.0001***

Significance levels * =.05 ** = .001 *** = .0001

Table 7.3.b Identification accuracy by *Tone* and *Intonation*

Tone	Intonation	Identification Accuracy	Total Number
Intrinsic NT	**Q**	92.33%	1200
	S	91.83%	1200
Derived NT2	**Q**	89.17%	360
	S	91.67%	360
Derived NT4	**Q**	89.58%	360
	S	90.42%	360
T2	**Q**	80.92%	1200
	S	86.50%	1200
T4	**Q**	86.67%	1200
	S	95.92%	1200

Discriminability and bias result showed that intonation on NT and T4 are highly differentiable, more differentiable than the intonation on T2 (**Table 7.3.c**). It should be noted

however, that there was almost no bias towards statement in intonation perception on Intrinsic NT and Derived NT from T2, but a bias towards question interpretations for Derived NT from T4, a bias towards statement interpretation on T2 and an even stronger bias towards statement on T4.

Table 7.3.c Hit rates (H), False Alarm (FA), Discriminability (A') and Bias (B''D) in each tone condition

Tone	Hit Rates (H)	False Alarm (FA)	A''	B''_D
Intrinsic NT	92.34%	7.68%	0.96	0.01
Derived NT from T2	90.42%	19.17%	0.91	-0.10
Derived NT from T4	90.00%	20.00%	0.91	-0.39
T2	81.58%	13.07%	0.90	0.22
T4	87.29%	3.91%	0.95	0.57

Reaction Time

Significant effects of *Tone*, and the interaction between *Tone* and *Intonation* on reaction time were found (**Table 7.3.d**). The identification of intonation on Derived NT from T4 took the longest time on average, 64.22 ms (SE 7.65), followed by Derived NT from T2, 62.60 ms (SE 7.83). Intrinsic NT took 61.45 ms (SE 6.91) on average, T2 and T4 took even less time, 59.45ms (SE 7.78) and 58.42ms (SE 7.20). The post-hoc comparisons found statistical significance between Intrinsic NT and T4 and between Derived NT from T4 and T4 (ps <0.05). In terms of tone and intonation interaction, the post-hoc comparisions found no statistical significance between tones in the same intonation condition or the two intonations realized on the same tones. Furthermore, the other pairwise differences with statistical significance across tones and intonations were not meaningful and hence were not reported here.

Table 7.3.d Linear mixed-effects model on reaction time

Final model Analysis	Reaction Time ~ Tone + Intonation + Tone: Intonation + (1|Participant) + (1|Stimulus)			
Main Effects	SS	df	F	*p*
Intonation	945.71	4	5.3431	0.286
Tone	50.61	1	1.1438	<0.0005***
Intonation: Tone	1438.07	4	8.1248	<0.0001***

Significance levels * =.05 ** = .001 *** = .0001

When examined closely (**Figure 7.3.3a** and **Table 7.3.e**), post-hoc comparison shows only the reaction time difference between the statement and question on the Intrinsic NT as well as the statement on Intrinsic NT and the statement on the two CTs was statistically significant (*p*s<0.001).

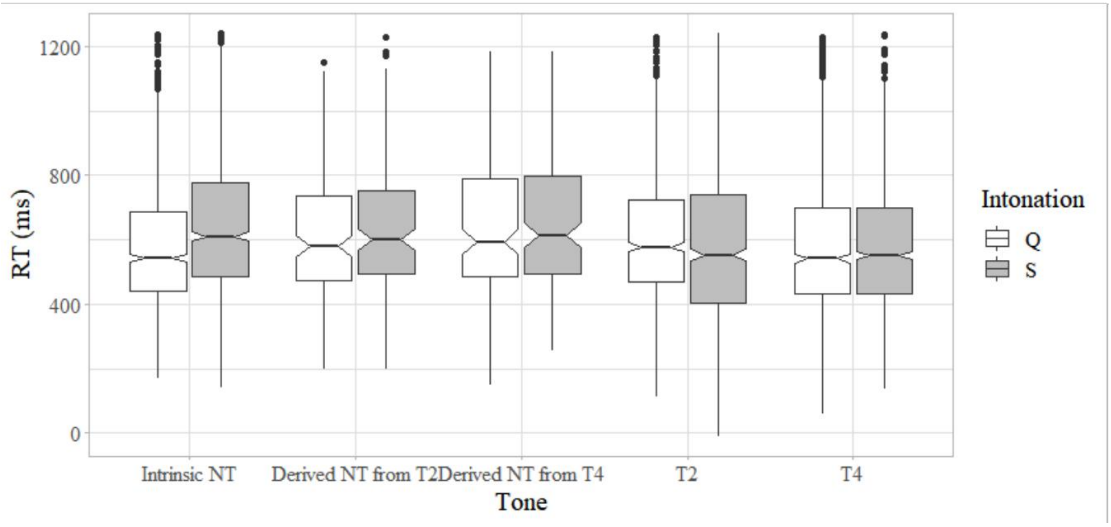


Figure 7.3.a Reaction time by *Tone* and *Intonation* of the correctly answered trials

Table 7.3.e Reaction time by *Tone* and *Intonation* of the correctly answered trials

Tone	**Intonation**	**Reaction Time**	
		Average (ms)	**SE**
Intrinsic NT	**Q**	647.71	0.63

	S	582.00	0.59
Derived NT2	**Q**	638.20	1.11
	S	613.93	1.06
Derived NT4	**Q**	648.55	1.05
	S	641.99	1.14
T2	**Q**	581.33	0.71
	S	608.27	0.59
T4	**Q**	581.60	0.59
	S	587.29	0.63

Acoustic Correlates of Intonation Perception

The acoustic correlates of intonation identification differed across tone conditions. The intonation perception accuracy related with more acoustic cues on Derived NT from T2, T2 and T4, than on Intrinsic NT and Derived NT from T4 (**Table 7.3.f**). With regard to the question identification in Derived NT from T2, the decreases in the average pitch range of the second syllables (i.e., the NT-bearing syllables), the increases in the pitch height of both syllables as well as the average range of the preceding T1 and the increases in duration ratio all significantly increased the possibility of a stimulus with Derived NT from T2 to be identified as a question. To the question identification in T2, the increases in the average pitch height and pitch range of the second (i.e., T2-bearing) syllable, the decreases in average pitch height of the first syllable and in duration of the second syllable were of statistical importance. To the question identification in T4, the increases in the pitch height of the second syllable (i.e., T4-bearing) as well as the decreases in the pitch range of the second syllable and in the pitch height of the first syllable were of statistically significance.

Question identification on Derived NT from T4 and Intrinsic NT, in contrast, was only significantly related to changes in two parameters. For Derived NT from T4, they were the increases in average pitch height of the second (i.e., NT-bearing) syllables and decreases in the absolute duration of the second syllables. For Intrinsic NT, they were the increases in average pitch height and decreases in average pitch range of the second syllables, namely, a higher and flatter realization of Intrinsic NT increased the possibility of it being identified as interrogative.

Table 7.3.f Logistic regression models on intonation identification in each tone condition

Intrinsic	Final model Analysis	Intonation ~ Duration Ratio + Duration of the second syllable + Average pitch height of the second syllable + Pitch range of the second syllable			
	Main Effects	Coefficient	SS	df	*p*
	Average pitch height of the second syllable	0.82378	186.6823	1	<0.0001***
	Pitch range of the second syllable	-0.15164	12.5132	1	<0.0005***
Derived NT from T2	Final model Analysis	Intonation ~ Duration Ratio + Duration of the second syllable + Average pitch height of the second syllable + Pitch range of the second syllable + Average pitch height of the first syllable + Pitch range of the first syllable			
	Main Effects	Coefficient	SS	df	*p*
	Average pitch height of the second syllable	0.30498	79.7615	1	<0.0001***
	Pitch range of the second syllable	-0.4099	15.49	1	<0.0001***
	Duration Ratio	-1.54342	18.05	1	<0.0001***
	Average pitch height of the first syllable	-1.17557	8.44	1	0.004 **
	Pitch range of the first syllable	0.76005	3.79	1	0.05 .
Derived NT from T4	Final model Analysis	Intonation ~ Average pitch height of the second syllable + Duration of the second syllable + Average pitch height of the first syllable			
	Main Effects	Coefficient	SS	df	*p*
	Average pitch height of the second syllable	1.40289	135.43	1	<0.0001***
	Duration of the second syllable	-0.86707	2.9264	1	0.089 .

T2	Final model Analysis	Intonation ~ Average pitch height of the second syllable + Duration of the second syllable + Pitch range of the second syllable + Average pitch height of the first syllable			
	Main Effects	Coefficient	SS	df	*p*
	Average pitch height of the second syllable	0.60428	337.85	1	<0.0001***
	Duration of the second syllable	-7.0396	20.75	1	<0.0001***
	Pitch range of the second syllable	0.05108	5.97	1	<0.05*
	Average pitch height of the first syllable	0.19613	9.81	1	<0.005 **
T4	Final model Analysis	Intonation ~ Average pitch height of the second syllable + Average pitch height of the first syllable + Pitch range of the second syllable + Duration of the second syllable			
	Main Effects	Coefficient	SS	df	*p*
	Average pitch height of the second syllable	0.88566	176.6	1	<0.0001***
	Average pitch height of the first syllable	-0.21008	5.89	1	<0.05*
	Pitch range of the second syllable	-0.10609	5.25	1	<0.05*

Significance levels * =.05 ** = .01 *** = .001; . indicate marginal difference

The main acoustic correlate of the reaction time shared by all five tones was the duration of the second syllable. The reaction time was inversely related to the duration of the second syllable, namely, the shorter the second syllable was, the longer the reaction time. Reaction time was also

inversely correlated with the pitch range of the second syllable on Intrinsic NT and T4, namely, the smaller the pitch range, the longer the reaction time.

Table 7.3.g Linear regression models on reaction time in each tone condition

Intrinsic NT	Final model Analysis	Reaction Time ~ Duration of the second syllable + Pitch range of the second syllable + Duration Ratio + Average pitch height of the second syllable				
	Main Effects	Coefficient	SS	df	F	*p*
	Duration of the second syllable	-0.648924	0.99	1	21.24	<0.0001***
	Pitch range of the second syllable	-0.004716	0.4	1	8.68	<0.005**
	Duration Ratio	0.083922	0.22	1	5.02	<0.05
	Average pitch height of the second syllable	-0.003642	0.14	1	2.97	0.084 .
Derived NT from T2	Final model Analysis	Reaction Time ~ Duration of the second syllable + Pitch range of the second syllable				
	Main Effects	Coefficient	SS	df	F	*p*
	Duration of the second syllable	-0.234268	0.26	1	4.55	<0.05
Derived NT from T4	Final model Analysis	Reaction Time ~ Duration of the second syllable + Duration Ratio				
	Main Effects	Coefficient	SS	df	F	*p*
	Duration of the second syllable	-0.20208	0.21	1	2.36	<0.05
T2	Final model Analysis	Reaction Time ~ Duration of the second syllable + Duration Ratio				
	Main Effects	Coefficient	SS	df	F	*p*
	Duration of the second syllable	-0.59164	1.07	1	19.91	<0.0001***
	Duration Ratio	0.08449	0.03599	1	5.51	<0.05 *
T4	Final model Analysis	Reaction Time ~ Duration of the second syllable + Pitch range of the second syllable + Duration Ratio				
	Main Effects	Coefficient	SS	df	F	*p*
	Duration of the second syllable	-0.0025484	0.140798	1	6.02	<0.0001***
	Pitch range of the second syllable	-0.0025484	0.0009922	1	5.54	<0.05*
	Duration Ratio	0.0729777	0.0369405	1	1.03	<0.05*

Significance levels * =.05 ** = .01 *** = .001; . indicate marginal difference

7.4 Discussion

The present experiment examined the identification of intonation on Intrinsic NT and Derived NT in comparison to two representative CTs, i.e., the rising T2 and falling T4, on which the associated intonation contours are the hardest and easiest to identify, respectively.

The acoustic analyses of the stimuli produced by two speakers demonstrated that there were no tone-specific variations in the intonation implementation on Intrinsic NT and Derived NT from T2 and T4. All NTs demonstrated a falling contour in declarative statements but had the falling flattened in interrogative questions, much like T4. The acoustic analyses of the Intrinsic NT and Derived NT stimuli indicate that the final rise of the interrogative intonation on NT is not implemented as a direct rising contour, despite the underlying un(der)specification of Intrinsic NT and the weak metrical strength associated with NT in general; instead, in line with Liu and Xu (2007), questions are realized through the flattening of a falling contour. This finding may suggest that the realization of lexical tone and intonation interacts at a post-lexical level, and both Intrinsic NT and Derived NT must have acquired a tone target before entering into the post-lexical derivation. In other words, in Derived NT realization, the expression of discourse function (interrogative vs. declarative) is prioritized over the realization of underlying tones.

Since acoustic changes were similar across conditions, the identification accuracy and A' values (indexing discriminability) in Intrinsic NT, Derived NTs and T4 were of a similar (high) level. As predicted in P1, the identification accuracy of questions was highest in the Intrinsic NT condition, higher than in T4 and T2. The seemingly intermediate accuracy values for the two

Derived NT conditions were statistically indistinguishable from that of Intrinsic NT on the one hand, and T4 on the other. The identification accuracy of statements, however, contradicted P1 as no differences were found between the Intrinsic NT, Derived NT from T2 and Derived NT from T4 conditions, neither for statements nor for questions.

The identification bias echoed the results of identification accuracy and supported P2 in general: the B''_D values of Intrinsic NT and Derived NT from T2 were smaller than 0.01, indicating a bias-free identification of intonation in these two tones; identification bias towards statement was found on both CTs and Derived NT from T4. Interestingly, however, the bias found on Derived NT from T4 was towards the question, the 'marked' intonation, rather than towards the statement. A possible explanation is that this bias towards question may be an incident due to the relatively small number of Derived NT stimuli since the smaller the total number is, the larger the ratio and the ratio-based parameters will look like. After all, the bias found in the current study was not very large (smaller than 0.5).

The findings so far suggest that Intrinsic NT facilitates intonation perception in a more balanced and equally clear way in comparison to T4; Derived NTs also show a similar tendency, and no differences are found between Derived NT from T2 and Derived NT from T4. It is plausible that due to the underlying tonal underspecification of Intrinsic NT, there is no interaction between lexical and intonational tones underlyingly. Therefore, the underspecified Intrinsic NT helps the identification of intonation structure more than Derived NT or CT which have underlying tones.

However, this interpretation indicates a strong influence of the underlying tones, which has

not been found in the two Derived NTs. Derived NT from T4 did not enable a higher identification accuracy nor a less biased identification than Derived NT from T2; on the contrary, there was a larger bias between intonation types in Derived NT from T4 than in Derived NT from T2, though the bias was not towards statement like T4 but was towards questions. Therefore, the surface falling contour must also have played an important role in intonation perception on NT as it did on T4. The analyses of the acoustic correlates of the identification provided supportive evidence to this hypothesis.

The analysis of the acoustic correlates of tone and intonation demonstrated that for syllables carrying Intrinsic NT, Derived NT and T4, the interrogative identification was significantly related with reduced pitch range, namely, the smaller the pitch range was, the more likely the participants identified the intonation as question. This acoustic pattern may represent an important marker of questions in Mandarin, facilitating intonation perception of phonetically falling tones, regardless of their underlying tones.

Furthermore, I speculate that the acoustic correlates involved in intonation identification on different tones may also indicate the difficulty in the intonation structure. For Intrinsic NT words, the reduced pitch range and raised pitch height of NT appeared to have made the participants more likely to identify the Intrinsic NT stimuli as being interrogative. The acoustic features of the preceding syllable and duration information, however, were not as important in the intonation perception on Intrinsic NT. The identification of intonation on Derived NT from T4 required pitch and duration information of the NT-bearing syllable, that is, a higher and longer Derived NT with underlying T4 signaled the interrogative intonation on it as well. In contrast, the identification on T4 did not only involve the pitch range and average height of T4 in the same way as Intrinsic NT,

but also the average pitch height of the preceding tone. A lower preceding T1 also increased the possibility of a T4 stimulus to be identified as a question, suggesting that the high T4 only may not be enough to signal questions; a T4 which is clearly higher than the previous tone was also required. The even more complex acoustic correlates found in Derived NT from T2 and T2 itself may indicate the greater perception difficulties of intonations in these two-tone conditions.

Reaction time results, however, did not show much variation between tones or intonation types. Contrary to P3, the identification of statements on Intrinsic NT in fact took more time than their identification on CTs. The linear regression model on the acoustic correlates of the reaction times in each tone provided further evidence as in all five tone conditions, the reaction time was inversely correlated to the duration of the second syllable. In other words, the shorter second syllables have resulted in a longer time between the offset of the stimuli and the key pressing in the present experiment. The opposite findings towards the accuracy and bias may be due to the very short duration of the NT-bearing syllables, as the identification and key-pressing process will always need a certain amount of time no matter how easy it may be. As shown in Table 7.2c, the second syllables of the stimuli were all shorter than 30ms, and Intrinsic NT in question intonation was even shorter than 20ms in average. It is plausible that the participants were not able to process the stimuli properly and have the key pressed right afterwards, but need extra time. Therefore, reaction time may not effectively reflect the intonation perception difficultly on disyllabic Mandarin words. Experiments with high time-resolution methods like eye-tracking are required to further explore this problem.

To summarize, H5 is supported to some extent as the identification of intonation on Intrinsic NT was more accurate and balanced than on CTs. However, the lack of statistically significant

differences between Intrinsic NT and Derived NT and between the two types of Derived NT tested indicates that intonation identification is not only facilitated by the absence of the underlying tone, but also by the flattening of the falling contour in the phonetic realization.

Chapter 8 Final Discussion

8.1 Answers to the Research Questions

At the beginning of the thesis, I posed three research questions regarding the phenomenon of Neutral Tone in Standard Mandarin. RQ1 and RQ2 about the non-homogeneous nature of NT were mainly explored in the experiments in Part A, while Part B focused on the interaction between NTs and utterance-level prosody.

RQ1: Is NT a homogeneous phonological entity, or instead, are there different types of NT with different underlying representations?

The main finding of the present thesis is that there are indeed two types of NT, Intrinsic NT and Derived NT; they are not only realized differently when on focus, but also processed differently in spoken word processing. The data obtained further suggest that these two types of NT are not only morphologically and diachronically different, but that they also have different underlying tonal representations as was seen in Experiment 1-4.

RQ2: What are the underlying representations of Mandarin NT, and are Mandarin NT(s) metrically lighter than CT?

Regarding the tonal target of Intrinsic NT, Experiment 1 shows that Intrinsic NT kept a general falling contour even when on focus. Instead of being exaggerated, the falling contour of Intrinsic NT even showed a tendency to be flattened when focused or focused with emphasis. The processing differences found between Intrinsic NT and Derived NT also suggested that Intrinsic NT is underspecified compared to CTs or Derived NTs. If Intrinsic NT has a fully specified tonal

representation, the relatively small number of Intrinsic NT morphemes should have made it easier to be identified than Derived NTs from the CTs they derived from. However, this was not found in Experiment 4 as no logographic context was provided. Therefore, Intrinsic NT is unlikely to be specified with any underlying tone, neither with a low tone as suggested by Lin (2006) and other sociolinguists, nor with a boundary tone as suggested by Cheng (1973) or Li (2003) (Section 3.2). Instead, the relatively robust falling contour (which was not changed by focus or question intonation) may indicate that Intrinsic NT has a mid-level target; the un(der)specification of Intrinsic NT will be discussed at length in Section 8.2.

Regarding the tonal representations of Derived NT, both the acoustic and the processing data suggest that Derived NTs are represented as CTs in their underlying tonal representations. From an acoustic perspective, clear similarities were found between the f_0 curves of Derived NT and the CTs when focused. From a processing perspective, a facilitating effect of different underlying tones and an inhibiting effect of the same underlying tone in CT-NT word identification and discrimination were both observed. Therefore, I conclude that Derived NTs are underlyingly CTs.

What remains to be discussed is whether and how Derived NTs have lost their underlying tones. The acoustic and processing evidence in Part A leaves two possibilities. It is plausible that Derived NT is light and toned throughout, and it is only the lack of prominence that masks the realization of the tones. This interpretation would explain the traces of the underlying CTs which are still left in unfocused Derived NTs as well as the CT-like realization of Derived NT on focus. Alternatively, it might be that the relationship between Derived NT and Intrinsic NT is like the rising T3 and T2 after T3 sandhi, while the traces of the underlying CTs may be attributed to an incomplete neutralization. My investigation of the interaction between NT and intonation

(Experiment 6), supports the second account more than the first. Acoustic analyses of the stimuli found no differences between Intrinsic NT, Derived NT from T2, and Derived NT from T2; all three tones demonstrate a falling contour in statements and had a falling contour which was flattened in questions, indicating that they interact with intonation in a similar way[33]. The differences in intonation perception between Intrinsic NT and Derived NT did not reach statistical significance either, and no clear influence of the underlying T2 or T4 was found in the Derived NT condition. Taken together, these results may suggest that intonatioonal tones are able to override the underlying tones of Derived NTs, and this point will be pursued further in Section 8.2.

In terms of the metrical weight, the heavy-light pattern of Intrinsic NT words and the heavy-heavy pattern of CT words are found to be relatively robust, but the temporal organization of Derived NT words changed significantly when the syllables bearing Derived NT were on focus (Experiment 1, 2 and 5). This is not to say that the duration ratio of Intrinsic NT and CT words did not increase at all when there was focus on the stimuli, but the increase did not reach any statistical significance. It may be argued that the robust heavy-light pattern found in Intrinsic NT words comes from the underlying tonal differences between Intrinsic NT and CT, and hence there is no need to refer to a metrical structure separately. However, the tonal differences alone may not be enough to explain the different changes in the absolute duration of the two syllables in Intrinsic NT and CT words. Experiment 1 and 2 found a stronger association between Intrinsic NT and its preceding CT, and this association was also observed in Intrinsic NT (and Derived NT) in Experiment 5.

In Experiment 1, 2 and 5, the relatively unchanged duration ratio of CT stimuli resulted from

[33] This finding requires further investigation on larger scale.

the restricted lengthening of the second CT-bearing syllables when focused. In contrast, the stable duration ratio found in Intrinsic NT stimuli came from the dynamics between the lengthening of the preceding CT syllables and the following NT syllables. When the NT was on focus, the stability in duration ratio was reflected as synchronous large duration increases in both syllables in Experiment 1 and 5; In Experiment 2, it came from the relatively restricted lengthening in both syllables in Experiment 2. Individuals seemed to choose different strategies to realize the focus on Intrinsic NT, namely, to significantly lengthen it or not, but all prioritized keeping the heavy-light structure. This tendency was also observed in the trisyllabic CT-NT-NT stimuli in Experiment 5. Focus on the second and third syllables bearing Intrinsic NT caused as much lengthening on the initial CT-bearing syllables as when these initial CTs were on focus themselves.

The metrical lightness of Derived NT is less stable than that of Intrinsic NT. When not focused, no differences were found between Intrinsic NT and Derived NT in their relative duration to the preceding CT, and both were shorter than CTs. However, when there was focus on the syllables bearing Derived NT, the durational differences between Derived NT and CT became blurred. However, some findings in Experiment 1, 2, and 5 still suggest that there is an inner structure that prevents unlimited expansion of Derived NT. The focus with emphasis in Experiment 2 did not further increase the duration or the f_0 height or range of Derived NT from the focused condition. Moreover, Derived NT syllables on focus also triggered lengthening on the preceding CT-bearing syllables. In other words, the Derived NT syllables showed dependence on the preceding CT-bearers much like the Intrinsic NT syllables. This dependence may indicate that there is a need to keep the heavy-light(-light) pattern in Derived NT words or phrases as well and therefore it is plausible that Derived NT has a lighter metrical weight compared to CT.

To summarize, the current data demonstrate that both Intrinsic NT and Derived NT have metrical weight different from CT words, but whether Intrinsic NT and Derived NT are metrically different is worth further discussion. It is plausible that they are of different metrical weight, but it is also possible that it is the distinctive underlying tonal representations that influence the duration of Intrinsic NT and Derived NT appear when on focus. Since Intrinsic NT, Derived NT and CT only showed a three-way distinction in lengthening patterns when the NT-bearing syllables were on focus rather than when there was a focus on any part of the stimuli, the second account currently appears to be the more plausible. This point will be further discussed in Section 8.2, and possible metrical structures of NT words will be proposed there.

RQ3: How will the representations influence the interaction between NT and utterance-level prosody?

In general, Intrinsic NT and Derived NT are similar to each other in the way they interact with focus on domains larger than syllables as well as discourse functions (interrogative or declarative).

Although the experimental findings appear to support the hypothesis that both Intrinsic NT and Derived NT have lighter metrical weight compared to CT, the distribution of focus-induced lengthening in NT words still resembled the distribution in CT words rather than what would be found in typical stress or pitch-accent languages. As discussed in Section 6.4, neither Intrinsic NT nor Derived NT was more likely to be lengthened directly by the focus rather than the spill-over lengthening from the preceding CT-bears (presumably, what would be the metrical head in stress and pitch-accented languages); the only difference between the two types of NT lies in the

magnitude of lengthening. Derived NT was sometimes lengthened more than Intrinsic NT, possibly due to the existence of underlying tones. The heavy-light contrast of NT words was not intensified by the focus on larger domains either, unlike the long-short tone contrast in Shanghai Wu dialect or the long-short vowel contrast in Swedish. It seems that only when the NT-bearing syllable(s) are on focus, the tonal representation will override the heavy-light(-light) patterns in the Derived NT words and cause a significant lengthening of syllables bearing Derived NT. These findings may further suggest that within the same language, there tends to be a uniform distribution pattern of focus-induced lengthening, despite the metrical differences between the syllables on focus. This hypothesis is worth further investigation.

With regard to the interaction with intonation, Intrinsic NT, Derived NT from T2 and Derived NT from T4 did not demonstrate a rising contour when being interrogative; instead, all of them had the falling contours flattened by the question intonation. Correspondingly, the participants were able to identify intonation on NT of either type with satisfying accuracy and speed and more importantly, in an unbiased manner, even when the preceding T1 was neutralized in duration and pitch. In this way, a direct interaction between both types of NT and question intonation was observed in realization as well as in perception. Unlike Liu and Xu (2007), NTs seemed to interact with intonation as individual tones rather than as parts of larger prosodic units (i.e., prosodic words). This intensive intonation-tone interaction seemed to require each syllable in Mandarin to surface with a tonal target before entering post-lexical derivation as well (Section 8.2).

Correspondingly, the perception of intonation is also largely influenced by surface changes, which explains the lack of significant differences between Intrinsic NT and Derived NT as well as the asymmetrical perceptual difficulty between questions on T2 and T4. Nevertheless, the

perceptual differences found between Intrinsic NT and T4 may still suggest that the absence of the underlying tone does not dampen the intonational information in Intrinsic NT, and hence enables a balanced and clear identification of intonation, especially question intonation. This tendency was also observed between Intrinsic NT and Derived NT, for instance, as the identification accuracy was higher in Intrinsic NT by percentage than in Derived NT from T2 or T4. Therefore, it may be concluded that intonation perception in Mandarin is affected by the underlying tones of the local elements as well as the distortion in the surface tones.

8.2 Mid Target and Underspecification

Echoing previous studies (Li, 2003; Chen and Xu, 2006; Liu and Xu, 2007), the present thesis also finds a robust falling contour in Intrinsic NT syllables. To account for the gradual falling observed in consecutive NT-bearing syllables, Chen and Xu proposed that NT "does have a specific underlying pitch target, which is likely to be static and mid" (p67); moreover, "this target is likely to be implemented with a weak articulatory strength" (p67) and therefore is slower in overcoming the influence of the preceding CTs, resulting in the large contextual variations. Whether the weak strength is also part of the underlying representation of NT has been left by Chen and Xu as a question for further research. The results of Experiment 1, 2 and 5, as discussed before, confirm that Intrinsic NT of either type is of underlying metrical lightness; in other words, the "weak strength" in Chen and Xu's terms is part of the underlying specification of NT.

The phonological status of this mid-level target is worth a closer examination. As discussed before, since no significant focus-induced change in f_0 height or range was found, this target is unlikely to be a phonologized underlying representation of Intrinsic NT. Or else, there should have been a larger increase in the f_0 height or a larger reduction in the f_0 range of Intrinsic NT when

focused, unlike the unchanged f_0 found here. If so, where this target comes from becomes an interesting question.

Some have assumed that the mid-level target is a baseline f_0 which is automatically realized through relaxing the articulatory muscles (see discussion in Yip, 2002: Chapter 9). However, since the falling movement was found in both the unfocused and focused condition, this assumption is unlikely to be true. It makes no sense to assume that the muscles are as relaxed in focused as in unfocused situations, also in view of the fact that a durational increase was found in focused Intrinsic NT syllables. Therefore, this mid-level target must have been reached by some active cooperation between the cricothyroid and the vocalis muscles (Section 2.2.1). According to the explanation Chen and Xu (2006) gave based on Hollien's works (Hollien, 1960; Hollien and Moore, 1960), this active cooperation is likely to involve "active contraction of the pitch-lowering muscles such as the thyroarytenoids (vocalis) in conjunction with the relaxation of the pitch-raising muscle such as the cricothyroids" (p68). In other words, speakers are deliberately moving their articulators to realize NT in daily speech. The motivation for them to do so is to realize a mid-level target and where this target comes from becomes an interesting question.

Yip (2002) summarized three situations in which syllables can acquire tone targets (which are usually low rather than high targets): a) specified with tone targets throughout the phonology and phonetics, b) being toneless throughout the phonology and phonetics but acquired pitch value through phonetic interpolation and c) being toneless during the phonology but specified with targets at the phonology-phonetic interface.

The traditional literature, including Yip (1980, 2002), regards Mandarin NT as situation b).

However, the results of the present experiments, together with Chen and Xu's results (2006), suggest that Intrinsic NT falls into situation c). These syllables themselves are not active in tonal activities like tone-spreading or tone sandhi, but still surface stably with a target as if they had a target underlyingly. The rule-based theory models regard these targets as filled by default rules. Yip (2002) believes that these value-filling rules and their application restrictions are language specific. Unlike Yip, Wang (1997) and Pulleyblank (1986) believe that the rules filling [-Upper] and [H] are universal. According to the model Wang (1997) presented, Mandarin NT is phonologically toneless, but acquired an [-Upper, H] specification by two phonetic value-filling rules that applied before all other phonetic rules.

However, the rule-based theory is rejected by many due to the non-derivational approach it assumes in its analysis. To avoid this rejection, Yip (2002:64) proposes an interface operation to fill the targets to the toneless elements after all the phonological derivational analyses. Yip hypothesizes that this operation is driven by a language-specific requirement for a fully specified tone for a fully specified tone to be in place at the start of the process of phonetic implementation. Standard Mandarin, a typical tone language with one-to-one correspondence between morphemes, syllables, characters and tones, is highly likely to have such a requirement, so that it is also plausible to assume that Intrinsic NT acquires a mid-level target though the interface operation. Moreover, Mandarin may also require every morpheme to be realized with at least a surface tone target to enable the realization of more global prosodies. This is backed by the results of Experiment 6 which show that both the implementation and the perception of question intonation interacts with a slightly falling contour of NT of either type (see also Liu and Xu, 2007).

According to either hypothesis (i.e., default value-filling or the phonetics/phonology interface

operation), Intrinsic NT is likely to be more underspecified than the CTs. In particular, the previous EEG studies have reported special behaviors of T3 which some of them have attributed to the underspecification of T3, though this underspecification is not supported either by the phonological models I follow or the behavioral studies (Section 2.3.4). Nevertheless, a comparison between Intrinsic NT and T3 using EEG techniques may help to further clarify the special behaviors of T3 and deep our understandings of the underspecification of tones. An oddball paradigm experiment with ERP components to further explore the underspecification status of Intrinsic NT in comparison with Derived NT and CT was planned, but had to be abandoned due to the outbreak of the COVID-19 pandemic.

Compared to Intrinsic NT, the situation with Derived NT is more complex. Derived NT may have lost its underlying tones somewhere in the derivation, as the interaction between NT and question intonation was not much influenced by underlying tonal differences. However, at which level Derived NT lost the tones remains an open question. There are two possibilities. Firstly, Derived NT lost the underlying CT target in the phonological process. At the phonological end, it is as toneless as Intrinsic NT and somehow acquires a mid-target at the phonology-phonetic interface. This interpretation is in line with many traditional models which assume that NT is toneless (Section 3.2). It is still possible, however, that Derived NT lost the underlying CT targets at the phonetic level. This would mean that at the phonological end, Derived NT is still represented as CTs. In this account, the metrical lightness of Derived NT results in the 'unrealization' or 'half-realization' of CT targets, as well as the large contextual variations found in the f_0 realization. Only when expressing interrogative meaning is pragmatic meaning prioritized over the expression of the underlying tones, leading to little acoustic differences between Derived

NT from T2 and T4 and correspondingly, a perception uninfluenced by the underlying tones. To fully explore this question, an acoustic study with more stimuli and participants is required but this falls beyond the scope of this thesis.

8.3 Metrical Representation of NT

So far, I have referred to both types of NT as metrically light without presenting detailed analyses of the metrical structure of either type of NT. With the current data, two different proposals may be made, 1) Intrinsic NT, Derived NT and CT words have different metrical structures and 2) Intrinsic NT words and Derived NT words have the same metrical structure which differs from that of CT words.

As indicated before, the empirical evidence provided in this thesis favors 2) over 1). The underlying tonal difference is enough to explain the differences in the extensibility of f_0 and the duration ratio differences found between the two types of NT when NT-bearing syllables are on focus, i.e., the tonal representation overrides the metrical lightness when syllables bearing Derived NT are on focus. This does not happen in syllables bearing Intrinsic NT because there is no underlying tonal specification of Intrinsic NT. There is no need to give Derived NT a unique metrical weight distinctive from Intrinsic NT or vice versa, as no significant differences were found between Intrinsic NT and Derived NT in other situations.

As introduced in Section 2.4, there has been a long-standing debate over the metrical organization of CT words in Mandarin. Since the data in this thesis have no direct bearing on the controversies, I simply repeat the different proposals that have been put forward by the main schools of thought, reviewed in Section 2.4.

1) CT-bearing syllables form binary iambic feet (and sometimes unary degenerated feet) and NT words form trochaic feet. A feature like [-stress] (Zhang, 1988; Yip, 1980) may be proposed for Intrinsic NT and Derived NT, which does not only trigger the loss of underlying tones (if they exist) but also changes the foot organization from iambic to trochaic.

2) CT-bearing syllables form unary degenerate feet themselves. NT-bearing syllable of either type form trochaic feet with the preceding CT-bearing syllables.

3) CT-bearing syllables form binary trochaic feet or unary degenerated feet. NT-bearing syllables of either type may be seen as the extrametrical elements cliticalized to a binary or a unary foot, forming a prosodic word with the preceding CT(s).

While my findings suggest that the syllables carrying Intrinsic NT and Derived NT are both metrically light, it is still possible that Intrinsic NT and Derived NT words have different metrical structures. One possibility is to mark Intrinsic NT with a unique feature like [+Extrametrical] or *Non-finality* as such feature limits the change in prominence relation (Hyde, 2011). Derived NT and the preceding CT-bearing syllable, in contrast, form a trochaic foot as suggested by Yip (1980) or Duanmu (2007). In this way, syllables which bear Intrinsic NT form prosodic words with the preceding unary or binary feet while the tone-deletion, or at least tone-simplification, of Derived NT may be seen as the avoidance of a stress clash within a foot. However, this analysis makes the following phenomenon hard to explain. In Mandarin, there are conventional phrases involving both Derived NT and Intrinsic NT, e.g., 走去了/tsou3 t͡ɕy 0(4) lə0/ 'walked-to' and the positions of the last two NT-bearing morphemes are interchangeable, i.e., 走了去 /tsou3 lə0 t͡ɕy 0(4)/ is also acceptable. The positional flexibility of Mandarin NT is only conditioned by syntax but not the type of NT. It is difficult to explain how the non-final 去 in 走去了 is unstressed, or how the

final 去 in 走了去 forms a foot with the initial CT with an extrametrical element in between. Therefore, it makes more sense to analyze Intrinsic NT and Derived NT words as the same in metrical analyses.

8.4 A Developmental View on Mandarin NT

In the present thesis, I have established that there are two types of NT in Mandarin, Intrinsic NT and Derived NT, that differ from each other in their underlying tonal representations. Comparing the present findings to the diachronic development of the emergence of NT (Section 3.3.2), it can be proposed that the morphemes that were lexicalized as unstressed NT-bearers before the *Ming* Dynasty (i.e., about the mid-14thcentury or earlier) are underspecified in their underlying tonal representations while morphemes that were lexicalized as NT-bearers afterwards are specified as CTs. However, it should be noted that suffixes coming in later but with no formal characters like /tʊɔ/ (掇 / 夺) could not be tested in the present design. Nevertheless, these morphemes are also listed as oral only or dialectal in most dictionaries.

The correspondence between the differences in the underlying tonal representation and the time of emergence of NTs allows me to speculate that from a developmental view, Derived NT is Intrinsic NT in development. It is assumed that (Intrinsic) NT emerged from the reduction in semantic meanings and hence the independence of the morphemes (Section 3.2). In this way, the pseudo suffixes carrying Derived NT may have the potential to lose their CTs completely in the far future, since the lexicalization of Intrinsic NT also took place at least 3 centuries ago.

The loss of tone and lexical prominence in the other more notional morphemes may be more complex. The further tone-deleting process of these morphemes requires them to be de-stressed, or

to have the preceding morpheme stressed. Chinese philological literature usually assumes a leftward movement of the semantic prominence in old, frequently used words or compounds (e.g., Lv, 1955). However, not all Derived NT words follow this step. In fact, the exact opposite examples can be found. For instance, the word 兄弟 (/ɕɥʊŋ ti/) is realized with T1-T4 when meaning 'brothers or friends' and with 'T1-NT' when meaning 'the younger brother', but the second morpheme 弟 actually refers to the younger brother while the first morpheme means the elder brother. It may be argued from a pragmatical perspective that the word 兄弟 was first used as a polite alternation of 弟 弟 (/ti4 ti0/) 'the younger brother' as in the traditional patriarchal hierarchy, being young means having a less prestigious position and 弟 is further reduced to show politeness. This account is rather complex and not available for every exception.

A more direct account is that 兄弟 becomes one lexical entity and hence a tighter metrical relation exists between the two morphemes when meaning 'the younger brother', much like the Situation 4c proposed by Xu (1956), that is, disyllabic words with a close modifier-head structure tend to have the second morpheme de-stressed. This tendency indicates that Mandarin, like English, prefers trochees over iambs. Such a phenomenon resembles the stress movement in English, a trochaic language with lexical stress. Compounds like *armchair* and *ice-cream* used to have primary stress on the last syllable but have now changed to the initial syllable (F. J. Nolan: personal communication). The weighting of semantic meanings of the two morphemes has not changed as armchair is still a chair rather than an arm, but they become closer and therefore the compounds are lexicalized into words. In this way, it is the tighter metrical relation that leads to the loss of prominence and lexical tones on the second syllables. When given enough time to

develop further, the second syllable morphemes in these words may become toneless as well.

Another factor that may have motivated the stress movement in Mandarin and the development of Derived NTs may be the language contact with Manchurian. The Manchurian language is a trochaic language with lexical stress like English: /ˈaqə/ ‘young man, son’, /ˈkala/ ‘hand’, /ˈbatuɹu/ ‘brave’, /ˈbaitʃambi/ ‘investigate-mbi (-mbi is the verb suffix of present tense verbs in Manchurian)’. The Manchurian language or Manchurian-accented Mandarin may once have held prestige because the *Qing* dynasty was established by the Manchurian. In fact, many Manchurian words became loan words in Mandarin spoken now, for example, the word 掰扯 /pai1 **tʂʰə0**/ ‘analyze, argue’ comes from /ˈbaitʃambi/ ‘investigate’. According to Zhao (1993), such contact has led to a larger number of NT words in Beijing Mandarin compared to other Chinese dialects.

However, I would like to argue against Zhao’s claim. The timeline he proposed for the emergence of most NT words is too late compared to what I have summarized in Section 3.3.2. The explosion of NT words did not happen in the *Qing* dynasty, but no later than the late *Ming* dynasty, right after the disappearance of the checked tone.

In fact, there may not have been an explosion of NT words at all. It is possible that most NT words only existed in spoken language and hence were not transcribed in written manuscripts before the *Ming* Dynasty. Formal written Chinese remained very different from spoken Chinese in many aspects including vocabulary and syntax until the Vernacular Movement in the early 20th century. The speaking-writing divergence in Chinese history is similar to, or even larger than the writing-speaking differences in non-Mandarin dialect districts. As a result, there was little chance

for most bi-morphemic words to be recorded in written texts, where mono-morphemic words were still considered more formal. Only in the *Ming* Dynasty and later did vernacular novels become popular and the language used there was informal and hence closer to spoken language rather than the formal written language at that time. Therefore, the authors needed to 'borrow' characters to transcribe the morphemes that were used only in spoken language, and hence provided evidence for the existence of certain NT morphemes.

Nevertheless, the language-contact hypothesis should not be undermined due to Zhao's mistakes as great phonological changes in the history of Chinese all happened during periods where minority regimes prospered. However, recognizing the importance of language contact with Manchurian, together with some dialectal data may suggest a different view from the above, namely, Derived NT may have (had) very little opportunity to become Intrinsic NT.

First, dialectal research supports the language-contact hypothesis. In Southwestern Mandarin dialects like Guiyang and Chongqing dialects, Intrinsic NT and some frequently used Derived NT morphemes bear a high-level tone, [55], regardless of the preceding tones and their duration is not as short as NT in Standard Mandarin (Li, 2004). Interestingly, the developmental trajectory of Southwestern Mandarin is also slightly different from Beijing and some other northern Mandarin. Mandarin has mainly been brought into Southwestern provinces in the early *Qing* Dynasty by immigrants speaking Jianghuai Mandarin, a Mandarin dialect from which Nanjing dialect was developed. Compared to Beijing Mandarin which has also developed from Nanjing dialect initially, Southwestern Mandarin dialects have not been influenced as largely by Manchurian and other northern dialects.

Furthermore, the underspecification of Intrinsic NT may be due to the tone changes afterwards rather than the simple loss of prominence. The neutralized tones in some non-Mandarin dialects like Jin, Xiang and Gan dialects either surface with a short but independent tonal contour (e.g., [21] in Shenmu Dialects, Xing, 1999) or short and reduced tones varying according to the morphemes (e.g., in Loudi dialects, Liu, 2001). Usually, it is believed that there have not been any major tone changes in these dialects such as the disappearance of the checked tone in most Mandarin dialects. With the popularization of literacy and the *Pinyin* system, major tone changes like the disappearance of the checked tone may no longer take place in the same way. Under this scenario, Derived NT may have a very slim chance to lose underlying tone(s) completely like Intrinsic NT did, but only share a similar surface form with Intrinsic NT. In fact, from a sociolinguistic perspective, Derived NT may not only lose the chance to become Intrinsic NT, but is also disappearing in Standard Mandarin spoken now.

The recent literature that compared NT words in different editions of dictionaries has found an overall decreasing trend in Derived NT words. In general, there is a decrease in the total number of NT words as many of them become CT words, but a small group of monosyllabic functional morphemes have remained NT-bearers in a relatively stable manner (Lin and Li, 2017; Wang, 2012). For instance, Wang (2012) compared *Xinhua* dictionaries published from 1957 until 2011 and found that the NT annotation of most monosyllabic empty morphemes have remained unchanged. The multiple CT annotation given to some modal particles are also canceled (e.g., the modal particle 啊 /a/ used to be annotated with T1 and T4 in the early editions). In contrast, the annotation of the NT-bearing morphemes that only occur in certain words is found to be oscillated between NT and CT over years. As the single morpheme, they have seldom been annotated with

NT, and many of them have lost the NT annotation even in the words.

Similarly, by comparing the 5th and the 6th edition of *Contemporary Chinese Dictionary*, Lin and Li (2017) found that more than 100 NT words with either obligatory or non-obligatory realizations in the 5th edition have turned into CT-only words in the 6th edition. However, none of the words in these changes involve Intrinsic NT morphemes defined in the present thesis. In contrast, although 105 new NT words with the obligatory NT realization have been added to the 6th edition, they are mostly new words made up of the clitic Intrinsic NT morphemes.

As introduced in Section 3.3.1, there is not only a decrease in the number of Derived NT words but an increase in the tolerance to the optional realization. There are more NT words with optional realization turning into CT words than the other way round in the dictionaries published more recently. This opposite trend does not necessarily falsify the observations made by Chao and Xu in early or mid-1950s, (i.e., old words are likely to become NT words), but may be brought about by the promotion of Standard Mandarin. Combined with literacy education, the norm of pronunciation in Standard Mandarin may be better described as the normalization of the pronunciation of all characters according to Beijing Mandarin. This process, compared to the normal acquisition process of the native language, is more bottom-up and therefore may discourage phonetic or even phonological changes that would have been learned in a top-down manner[34]. In addition, learning the normalized pronunciation of each character through *Pinyin*, which in most cases ties a single pronunciation to a specific character further enforces a one-to-one mapping between characters and pronunciations. This trend is also observed in

[34] Of course before children learn to read Mandarin, they have already acquired spoken language, but literacy education may lead to reanalyzing the correspondence between morphemes and tones. This might then in turn lead to a different realization of these morphemes.

characters with multiple pronunciations and other morphemic tone sandhi, like mentioned in Section 2.2.

In fact, it is not only Derived NT that is disappearing. Suffixes like /tɕi/ (叽/唧), /tuo/ (掇/夺) ， /ta/ (哒/达/搭) and /tɕiɑ/ or /tɕiɛ/ (价/介) were popularized in Beijing as well as some northern Mandarin dialects (at least) after the 16th century but do not have proper logographic transcriptions till now. They should have become Intrinsic NT like 子 /tsi/ or 头 /tʰou/, but have been considered as dialectal or informal and removed from Standard Mandarin. Therefore, it is reasonable to expect that these morphemes will be used less and less by Mandarin speakers as well.

The sociolinguistic studies based on the Contemporary Beijing Spoken Language Corpus reported the same decreasing trend (Zhou, 2018). As early as in 1980s, Beijing speakers with relatively low educational background actually used the fewest NT words, fewer than speakers with a high educational background than speakers with a medium-level educational background (Zhou, 2018). This hyper-correction phenomenon shows that NT has been avoided even by the local Beijing dialect speakers as a feature of less prestige. The disappearance of the colloquial NT words as well as the NT realization of many Derived NT words were also observed when I tried to find proper stimuli for the study. This declining tendency observed in Beijing does not only indicate that many suffixes and pseudo suffixes with strong potential to become Intrinsic NT are dying out, but more importantly the productivity of NT in Standard Mandarin has been reduced and interrupted in its headwaters.

In summary, although Derived NT may be Intrinsic NT in development, the developmental

process may be slowed down, ceased or even reversed. In the near future, the similarities may continue to reduce between Derived NT and Intrinsic NT but increase between Derived NT and CT in Standard Mandarin, or even in Beijing Mandarin.

8.5 Conclusion

In this thesis I have demonstrated through phonetic production and perception studies and psycholinguistic processing studies that in the Standard Mandarin spoken now, there are two types of NT that differ from each other in their underlying tonal representations as well as their diachronic derivation, even though they are both metrically light. The findings are summarized in **Table 8.5.a.**

Table 8.5.a Two types of NT

	Intrinsic NT	**Derived NT**
Derivation	Lexicalization of the unstressed functional morphemes	Driven by the disappearance of the checked tone or language contact
Time of Derivation	Most in Yuan Dynasty or before	Most in Ming Dynasty or later
Tonal Representation	Underspecified	As the four CTs in Mandarin
Potential Metrical Weight	Light	Light
Morphemes Involved	Structural particle, Aspect particles, Suffixes, Modal particles	Mainly the second or third notional morphemes in words or frequently used compounds
Surface Realization as NT	Obligatory	May be optional

NT, a highly controversial issue in Mandarin phonology, is a complex and evolving phenomenon. The formation of NT and its realization norm in Standard Mandarin is largely influenced by contemporary language policies. Various sub-categorizing proposals have been put forward in sociolingusitic and morphological studies, but in the traditional phonetic and phonological literature NT is still regarded as a homogeneous tone-neutralization phenomenon following Chao's proposal (1965). This has led to varying and often irreconcilable results in production and perception studies.

In this thesis, I have approached this issue from multiple different angles with a range of methodologies and explored the underlying tonal and metrical representations of different types of NT. Based on the morphological differences as well as historical differences I found, I have proposed that there are two types of NT, Intrinsic NT and Derived NT (Chapter 2 and 3). In my production experiments, I have demonstrated that narrow focus on the NT-bearing syllables does not bring much change in the f_0 contours or the metrical structures of the disyllabic words in Intrinsic NT words but enlarges the similarities between the f_0 contours of Derived NTs and the CTs they derived from and increases significantly the relative duration of the Derived NT syllables in the words (though not fully comparable with CT-bearing syllables). Such findings suggest that Intrinsic NT is underspecified for tones in its underlying representation, but Derived NT is represented as CT. Additionally, it suggests that both tones have a lighter metrical weight than CTs (Chapter 4). The processing evidence further supports these findings. The identification and discrimination of Derived NTs from their CTs are facilitated when the two stimuli have different underlying tones but prohibited when they share the same underlying tones. In contrast, the identification of Intrinsic NT is somewhere in the middle when no logographic contexts are

available (Chapter 5).

The exploration of the different phonological representation of NT is important. It improves our understanding of the tonal system of Mandarin, and lays the foundation for further research on the interaction between NT and utterance-level prosody, which is explored in the second half of the thesis.

Here, I found that despite the underlying tonal differences, Intrinsic NT and Derived NT words show similarities to each other in how they undergo focus-induced durational adjustments as well as in their interaction with question intonation. The focus on larger domains (i.e. words or phrases) does not intensify the heavy-light contrast in Intrinsic NT or Derived NT stimuli like it does in pitch-accent languages such as Swedish. Furthermore, neither Intrinsic NT nor Derived NT are lengthened by the spill-over effect of the preceding CT, unlike seen in unstressed syllables in typical stress languages. These findings do not only confirm that both types of NT have a light metrical weight compared to CTs, but also have topological significance, that is, within the same language, the focus-induced durational adjustment demonstrates a relatively uniform pattern. In other words, both the heavy CTs and the light NTs are lengthened directly by the focus (Chapter 6). The production and perception of intonation is also relatively uninfluenced by the underlying tones of either type of NT. Instead, both Intrinsic NT and Derived NT have a slightly falling rather than rising f_0 contour in questions, though the fall is not as aggressive as in statements. Correspondingly, no significant differences or perceptual difficulties have been found between Intrinsic NT and Derived NT as is the case with T2 and T4, and in general intonation identification on NT is easier than on CTs. On the one hand, these findings suggest an important role of surface f_0 and durational changes in intonation perception in Mandarin. On the other hand, they also

indicate that intonational tones override the underlying tones of Derived NT (Chapter 7).

All these results help us to re-consider the phonological representations and derivations of NT (and tone in general) in Mandarin. It seems that in addition to the underlying differences, both Intrinsic NT and Derived NT share a similar surface target, which is likely to be low and static and is not realized through target interpolation or default resting of the articulatory muscles. Following the structuralist model, I speculate that as a syllable tone language, Mandarin does not allow target-less elements in phonetic stages, and therefore every syllable, specified underlyingly or not, must have acquired a tonal target at tone-intonation interface (Section 8.2).

The underlying tonal representations of Intrinsic NT and Derived NT also correspond roughly to the time in which they have been phonologized. The NT-bearers that were lexicalized and phonologized before the major tone change, are likely to carry Intrinsic NT while those lexicalized and phonologized later under the influence of the Manchurian language are likely to carry Derived NT. Therefore, from a developmental perspective, Derived NT can be seen as Intrinsic NT in development, though the development takes a long time, and possible another major tone change, and this development has been slowed down or even reversed by the current language policy (Section 8.3).

To summarize, neutral tone is an important topic in Mandarin. It is not a homogeneous phenomenon. Instead, there are two different types of NT, Intrinsic NT and Derived NT who behave differently from one another, and from CTs. A careful separation of different types of NT helps to reduce the ambiguities in future research on this topic, e.g., on the processing and acquisition of the underspecified Intrinsic NT and the sociolinguistic revisit of Intrinsic NT and

Derived NT in several decades in the future.

Appendix

Appendix A Stimuli for Experiment 1 and 2

Condition	Stimuli	IPA Transcription	Glossary
Intrinsic NT	走吗	/tsou3 ma/	will you go
	看吗	/kʰan4 ma/	will you look
	谁呢	/ʂei2 ne/	who is that
	他呢	/tʰa1 nə/	how about him
	多么	/tʊɔ1 mə/	how
	什么	/ʂən2 mə/	what
	走了	/tsou3 lə/	have gone
	看了	/kʰan4 lə/	have looked
	吃着	/ʈʂʰi1 ʈʂə/	eating
	拿着	/na2 ʈʂə/	taking
	紫的	/tsi3 tə/	purple (thing)
	绿的	/ly4 tə/	green (thing)
	车子	/ʈʂʰɛ1 tsi/	car
	房子	/fɑŋ2 tsi/	house
	枕头	/ʈʂən3 tʰou/	pillow
	木头	/mu4 tʰou/	wood
Derived NT	李家	/li3 tɕia/	the Li's family
	赵家	/ʈʂao4 tɕia/	the Zhao's family
	桌边	/ʈʂuo1 piɛn/	on the table
	台边	/tʰai2 piɛn/	on the stage
	车钱	/ʈʂʰɛ1 tɕʰiɛn/	transportation fare
	酒钱	/tɕiou3 tɕʰiɛn/	money for drinking

	媒人	/mei2 ʐən/	matchmaker
	爱人	/ai4 ʐən/	lover
	厂里	/ʈʂʰɑŋ2 li/	in the factory
	柜里	/kuei4 li/	in the cupboard
	帮手	/pɑŋ1 ʂou/	helper
	人手	/ʐən2 ʂou/	worker
	五个	/wu3 kə/	five (classifier)
	十个	/ʈʂʰi2 kə/	ten (classifier)
	车上	/ʈʂʰɛ1 ʂɑŋ/	in the car
	地上	/ti4 ʂɑŋ/	on the ground
CT	三棵	/san3 kʰɤ1/	three (classifier for trees)
	十棵	/ʂi4 kʰɤ1/	ten (classifier for trees)
	五杯	/wu3 pei1/	five cups (of)
	四杯	/si4 pei1/	four cups (of)
	七幅	/tɕʰi1 fu2/	seven (classifier for paintings)
	十幅	/ʂi2 fu2/	ten (classifier for paintings)
	九群	/tɕiou3 tɕʰyn2/	nine flocks of
	六群	/liou4 tɕʰyn2/	six flocks of
	八匹	/pa1 pʰi3/	eight (classifier for horses)
	十匹	/ʂi2 pʰi3/	ten (classifier for horses)
	几首	/tɕi2 ʂou3/	several (classifier for poems/songs)
	一首	/ji4 ʂou3/	one (classifier for poems/songs)
	千位	/tɕʰiɛn1 wei4/	thousands of (classifier for people)
	一位	/ji2 wei4/	one (classifier for people)
	百次	/pai3 tsʰi/	hundred times
	万次	/wan tsʰi/	ten thousands times

Appendix B: Stimuli for Experiment 3 and 4

Condition	Pairwise Average difference	Pairwise SD difference	NT stimuli	Glossary	Average Familiarity score	SD	CT stimuli	Glossary	Average Familiarity score	SD
Intrinsic NT	1.6	1.26	篮子	basket	9	1.48	蓝紫	bluish violet	7.4	2.75
	0.35	0.03	垫子	cushion	8.8	1.96	电子	electronics	8.45	1.94
	0.6	2.97	虱子	bedbug	8.35	2.37	师资	teachers; faculty	8.95	1.36
	0.65	0.5	骰子	dice; rap	8.85	1.53	投资	invest	9.5	1.02
	0.35	0.67	新的	the new	9.5	1.02	心得	experience; feeling	9.15	1.35
	0.6	0	咕嘟	onomatopoeia of water	8.05	2.31	孤独	lonely; loneliness	8.65	2.31
	0.1	0.37	园子	garden	8.55	2.11	原子	atom	8.65	1.74
	0.55	3.55	风头	the trend of events	7.65	3	风投	Venture capital	8.2	1.81
	0.05	0.15	个子	size; height	8.9	1.45	各自	Separately	8.95	1.6
	0.05	0.14	案子	case	8.6	1.69	暗自	to oneself	8.55	1.83
	0.1	0.06	蚊子	mosquito	9.5	0.97	文字	word; character	9.4	0.92

	0.2	1.08	湿的	the wet	9.05	1.28	师德	teacher's ethics	8.85	1.77
	1.15	0.71	难呀	difficult	7.65	2.31	南亚	South Asian	8.8	1.6
	1.55	4.6	干巴	dry	6.65	3.05	干爸	god-father	8.2	2.25
	1.55	1.36	僵着	being numb; being stiff	7.7	2.49	江浙	Jiangsu and Zhejiang provinces	9.25	1.13
	0.45	0.03	盒子	box	9.35	1.24	合资	joint venture	8.9	1.26
	0.25	0.43	蹄子	hoof	8.3	2.37	题字	inscription	8.55	1.94
	1.1	1.52	候着	waiting	7.75	2.91	后者	the latter	8.85	1.39
	1.2	1.1	学着	to learn	7.6	2.54	学者	scholar	8.8	1.44
	1	0.89	快了	soon	8.6	1.69	快乐	happy; happiness	9.6	0.8
	0.65	0.61	帘子	curtain	8.65	1.88	莲子	lotus seed	9.3	1.27
	0.35	0.67	新的	the new	9.5	1.02	心得	experience; feeling	9.15	1.35
	0.45	0.27	直的	the straight	8.7	1.42	值得	worthwhile	9.15	1.15
	1.05	1.02	链子	chain	8	2.35	练字	practice calligraphy	9.05	1.32
	1.45	1.19	签子	stick	7.55	2.64	铅字	font type	6.1	3.19
	1.05	0.64	笛子	flute	9.15	1.53	嫡子	firstborn	8.1	2.17

	1.75	1.18	轿子	sedan	8.3	1.98	教子	godson	6.55	3.15
	1.7	1.34	看头	something worth seeing or reading	7.05	2.56	看透	see through	8.75	1.22
	1.45	1.41	柜子	cupboard	9.2	1.25	贵子	lovely baby	7.75	2.66
	0.8	0.09	栗子	chestnut	8.85	2.26	粒子	particle	8.05	2.18
	0.35	0.22	句子	sentence	9.4	1.07	巨资	huge investment	9.05	1.28
Derived NT (Different second tones)	0.75	0.52	暧昧	ambiguous (in relationship); be ambiguous (in relationship)	8.9	1.34	爱美	love the beautiful things	8.15	1.86
	0.35	1.03	过失	mistake	9.1	1.18	过时	out of date	8.75	2.21
	0.6	0.22	拉扯	pull; drag	8.4	2.01	拉车	pull the cargo	7.8	2.23
	1.9	1.25	右边	the right-side	9.35	1.11	诱变	mutagen	7.45	2.36
	0.1	0.62	台上	on the stage	8.35	1.68	台商	Taiwanese businessman	8.45	2.3
	0.3	0.05	道理	reason	9.05	1.53	倒立	headstand	8.75	1.48
	0.85	1.37	座位	seat	8.4	2.31	作为	deed	9.25	0.94
	0.2	0.04	附上	attach	8.25	1.87	富商	rich businessman	8.45	1.83

	0.7	1.44	厅里	in the hall	8.6	2.22	听力	Listening; audition	9.3	0.78
	0.15	0	出息	prospect	9.05	1.28	出席	attendance	9.2	1.29
	0.05	0.07	圣人	saint	9.15	1.06	胜任	be competent	9.1	1.14
	0.25	0.44	牌坊	memorial archway	8.5	1.88	排放	blow off	8.75	1.44
	0.5	0.31	编辑	editor; edit	9.3	1.05	边际	margin	8.8	1.36
	0.05	0.19	神气	manner	8.8	1.29	神奇	magical	8.75	1.1
	0.2	0.04	附上	attach	8.25	1.87	富商	rich businessman	8.45	1.83
	1.15	1.38	厉害	serve; awesome	7.9	1.2	利害	Advantages and disadvantages (referring to situations)	9.05	2.59
	0.65	0.17	年纪	age	9.4	0.8	年级	grade	8.75	0.97
	0.6	0.29	架势	posture	8.75	1.48	驾驶	drive	9.35	1.19
	1.85	2.33	精神	energetic	9.45	0.74	精深	profound	7.6	3.07
	0.3	0.26	背部	back	9.05	0.97	被捕	be arrested	8.75	1.24
	1.05	1.67	知识	knowledge	9.45	0.8	致使	lead to	8.4	2.48
	0.35	0.51	目的	aim	9.55	0.74	墓地	graveyard	9.2	1.25

	0.7	0.64	冰里	inside the ice	8	2.35	兵力	military force	8.7	1.71
	0.75	0.72	惦记	remember with concern	9.1	1.14	电极	electrode	8.35	1.85
	0.9	1.01	苍蝇	fly	9.25	0.94	苍鹰	eagle	8.35	1.96
	1.8	0.47	禁忌	taboo	9	1.22	晋级	promotion; be promoted	7.2	1.69
	1.85	0.42	课上	during the class	8.4	1.74	客商	travelling trader	6.55	1.32
	0.8	0.29	势利	power	8.8	1.4	事理	reason	9.6	1.69
	0.95	0.1	幅度	range; scope	8.9	1.34	服毒	take poison	7.95	1.24
	0.25	0.34	巴结	flatter	9	1.14	八戒	a famous figure in a legend	8.75	1.48
Derived NT (Same second tone)	0.15	0.17	威风	power and prestige	9.1	1.58	微风	breeze	9.25	1.41
	1.65	1.42	意思	meaning	9.35	1.11	一丝	a hint of	7.7	2.53
	2.05	1.53	程家	the Cheng's family (Cheng is a Chinese family name)	6.6	3.18	成家	get married	8.65	1.65
	0.55	0.14	大意	careless	8.55	1.66	大义	righteousness	9.1	1.51
	1.3	1	园里	inside the garden	7.95	2.09	原理	theory	9.25	1.09
	0.65	1.1	剩下	remain; left over	9.3	1.05	盛夏	midsummer	8.65	2.15

	0.7	0.79	瓶里	inside the bottle	8.55	2.01	评理	judge between right and wrong	9.25	1.22
	0.3	0.13	包涵	bear with	9.05	1.28	包含	include	9.35	1.15
	0.3	0.32	报酬	payment	9.55	0.67	报仇	revenge	9.25	0.99
	0.35	0.32	附和	chime in with	8.9	1.34	复合	compound; get back together	8.55	1.66
	0.4	0.08	仇人	enemy	9.2	1.29	愁人	distressing	8.8	1.36
	0.45	0.7	实诚	(dialect)honest	8.9	1.26	十成	fully	8.45	1.96
	0.7	0.42	瓷实	(dialect)solid	7.05	2.46	磁石	magnetic stone	7.75	2.88
	0.8	0.9	进来	come in	9.35	1.28	近来	recently	8.55	2.18
	0.9	0.77	意识	consciousness	9.35	1.06	一时	for a short while	8.45	1.83
	0.7	0.64	冰里	inside the ice	8	2.35	兵力	military force	8.7	1.71
	0	0.11	报复	revenge	9.05	1.43	暴富	become rich suddenly	9.05	1.32
	0	0.61	辈分	position in the family hierarchy	8.95	2.04	备份	copy	8.95	1.43
	0.15	0.24	实力	strength	9.35	1.01	实例	living example	9.2	1.25
	0.15	0.03	福气	luck	9.2	1.08	服气	be convinced	9.35	1.11

	0.2	0.02	服侍	attend; wait on	9.05	1.4	服饰	apparel	9.25	1.37
	0.2	0.25	成色	quality	8.5	1.77	橙色	the orange color	8.3	2.03
	0.25	0.19	战士	soldier	9.2	1.21	战事	warfare	8.95	1.4
	0.35	0.51	目的	aim	9.55	0.74	墓地	graveyard	9.2	1.25
	0.4	0.23	护住	protected	8.6	1.71	互助	help each other	9	1.48
	0.5	0.14	河里	inside the river	8.7	1.71	合理	reasonable	9.2	1.57
	0.55	0.94	行李	luggage	9.1	1.41	行礼	salute	8.55	2.36
	0.6	0.28	牢里	inside the prison	8.55	2.18	老李	nickname for Li	7.95	2.46
	0.65	0.62	雾里	inside the fog	8.55	1.99	物理	physics	9.2	1.36
	0.65	0.41	身手	skill or talent (in Kungfu)	8.75	1.48	伸手	stretch out one's hand	9.4	1.07
	0.25	0.09	市里	in the city	8.75	2.05	事理	reason	8.5	2.13

Appendix C: Stimuli for Experiment 5

Disyllabic Stimuli

Intrinsic NT	IPA Transcription	Glossary	Derived NT	IPA Transcription	Glossary	CT	IPA Transcription	Glossary
走吗	/tsou3 ma/	will you go	赵家	/ʈʂao4 tɕia/	the Zhao's family	三棵	/san3 kʰɤ1/	three (classifier for trees)
他呢	/tʰa1 nə/	how about him	桌边	/ʈʂuo1 piɛn/	on the table	五杯	/wu3 pei1/	five cups (of)
什么	/ʂən2 mə/	what	酒钱	/tɕiou3 tɕʰiɛn/	money for drinking	十幅	/ʂi2 fu2/	ten (classifier for paintings)
看了	/kʰan4 lə/	have looked	媒人	/mei2 ʐən/	matchmaker	六群	/liou4 tɕʰyn2/	six flocks of
拿着	/na2 ʈʂə/	taking	厂里	/ʈʂʰɑŋ2 li/	in the factory	十匹	/ʂi2 pʰi3/	ten (classifier for horses)
紫的	/tsi3 tə/	purple (thing)	帮手	/pɑŋ1 ʂou/	helper	一首	/ji4 ʂou3/	one (classifier for poems/songs)
车子	/ʈʂʰɛ1 tsi/	car	五个	/wu3 kə/	five (classifier)	千位	/tɕʰiɛn1 wei4/	thousands of (classifier for people)
木头	/mu4 tʰou/	wood	地上	/ti4 ʂɑŋ/	on the ground	百次	/pai3 tsʰi/	hundred times

Trisyllabic Stimuli

CT -Intrinsic NT-CT	**IPA Transcription**	**Glossary**	**CT-CT-CT**	**IPA Transcription**	**Glossary**	**CT- Intrinsic NT- Intrinsic NT**	**IPA Transcription**	**Glossary**	**CT-Derived NT-Derived NT**	**IPA Transcription**	**Glossary**
灰的书	grey book	/xuei1-tɣ0-ʂ‘u1/	三棵树	/san1 kʰɣ ʂ‘u4/	three trees	灰的了	/xuei1 tɣ0 lə/	turned grey	飞下去	/fei1 ɕia0 tɕʰy0/	fly downwards
黑的鸡	black chicken	/xei1-tɣ0-tɕi1/	三棵草	/san1 kʰɣ tsʰɑu4/	three grasses	白的了	/pai1 tɣ0 lə/	turned white	坐下去	/tsʊɔ4 ɕia0 tɕʰy0/	sit downwards
喝了粥	drink congee	/hɣ1-lɣ0-tʂou1/	三幅画	/san1 fu2 xuɑ4/	three pieces of paintings	紫的了	/tsi3 tɣ0 lə/	turned purple	读下去	/tu2 ɕia0 tɕʰy0/	read onwards
黑的旗	black flag	/xei1-tɣ0-tɕʰi1/	三幅纸	/san1 fu2 tʂi3/	three pieces of papers	绿的了	/ly4 tɣ0 lə/	turned green	走下去	/tsou3 ɕia0 tɕʰy0/	walk downwards
灰的牛	grey cow	/xuei1-tɣ0-niou2/	七倍多	/tɕʰi1 pei4 tʊɔ1/	seven times as many						
包子皮	steamed bun skin	/pɑu1-tsi0-b'i2/	七首词	/tɕʰi2 ʂou3 tsʰi2/	seven poems						
喝了酒	drink wine (perfect tense)	/hɣ1-lɣ0-tɕiou3/	七倍长	/tɕʰi1 pei4 tʂʰɑŋ/	seven times as long						

吃了米	eat rice (perfect tense)	/tʂi1-lɣ0-mi3/	七首歌	/tɕʰi2 ʂou3 kɣ1/	seven songs						
梳子齿	comb teeth	/ʂ‘u1-tsi0-tʂi3/									
梳子背	comb back	/ʂ‘u1-tsi0-pei4/									
吃了菜	ea vegetables	/tʂi1-lɣ0-ts‘ai4/									
包子铺	steamed bun shop	/pɑu1-tsi0-b'u4/									

Appendix D: Stimuli for Experiment 6

Intrinsic NT	**Glossary**	**Derived NT from T2**		**Derived NT from T4**		**T2**	**Glossary**	**T4**	**Glossary**
青的	dark green	扎实	solid	吃过	eat (perfect tense)	清除	clear up	氢气	hydrogen
片子	(informal) film	结实	substantial	喝过	drink (perfect tense)	偏食	dietary bias	捉住	nabbed
牵着	pull (progressive tense)	知识	knowledge	飞过	fly (perfect tense)	踢门	kick the door	轻度	light degree
摘了	take off (perfect tense)	车钱	transportation fare	三个	three-classifier	超级	super	摘下	take off
飞哪	fly	苍蝇	fly （the insect)	七个	seven-classifier	风格	style	飘去	float away
他的	his	姑娘	girl	八个	eight-classifier	驱除	drive away	吸气	breathe in
身子	body					积食	indigestion	夹住	clipped
掐着	pinch (progressive tense)					敲门	knock the door	风度	demeanor

开了	open (perfect tense)					分级	classification	飞下	fly down
捞哪	drag for					规格	standard	捉去	nab away
新的	new					消除	eliminate	风气	atmosphere
疯子	crazy people					猪食	pig food	包住	wrap in
披着	cloak in (progressive tense)					家门	home door (referring to family)	精度	degree of accuracy
贴了	stick (perfect tense)					三级	the third degree	抛下	give away
吃哪	eat					失格	disqualification	吹去	blow away
弯的	curved					割除	cut off	争气	try to make a good showing
鸽子	pigeon					猫食	cat foot	粘住	sticked
吹着	blow (progressive tense)					开门	open the door	八度	octave
秃了	become bald (perfect tense)					初级	primary	割下	cut down
掏哪	drag out					窗格	pane of the window	追去	chase after
金的	golden					擦除	abrase	吹气	blow

钎子	bot pick					衣食	close and food	箍住	hook together
包着	wrap (progressive tense)					朱门	red door (referring to the wealthy families)	高度	height
花了	spend (perfect tense)					高级	senior	天下	the world
哭哪	cry					升格	promote; upgrade	扔去	throw away
猜的	guess					摘除	excise	天气	weather
毡子	felt					分食	cannibalize	抓住	catch
跟着	follow (progressive tense)					专门	specialized	七度	seven degrees
低了	become low (perfect tense)					低级	junior	低下	low done
交哪	hand in					高格	high-level	交去	hand in there
亲的	real (child/parent)					根除	eradicate	呼气	breathe out
梳子	comb					吃食	food	勒住	tighten
摸着	touch (progressive tense)					高门	high door (referring to the noble family)	宽度	width
背了	shoulder (perfect					一级	the first level	吃下	eat down

	tense)								
抓哪	catch					三格	three grids	穿去	go across
宽的	wide					删除	delete	精气	vital essence
锅子	cook-pot					餐食	food	捏住	pick up or catch with the fingers/tweezers;
摊着	display (progressive tense)					天门	gate of the heavenly palace	湿度	wetness
踢了	kick (perfect tense)					升级	upgrade	喝下	drink down
喝吧	drink					七格	seven grids	踢去	kick towards

Reference

Andruski, J. E., & Costello, J. (2004). Using polynomial equations to model pitch contour shape in lexical tones: An example from Green Mong. *Journal of the International Phonetic Association*, *34*(2), 125-140.

Archangeli, D. (1988). Aspects of underspecification theory. *Phonology*, *5*(2), 183-207.

Archangeli, D., & Pulleyblank, D. (1993). The content and structure of phonological representations MIT Press.

Bao, Z. (1990). *On the nature of tone* (Doctoral dissertation, Massachusetts Institute of Technology).

Bao, Z. (1999). *The structure of tone*. Oxford University Press on Demand.

Boersma, Paul & Weenink, David (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.51, retrieved 22 July 2021 from http://www.praat.org/

Bradshaw, M. M. (1999). *A crosslinguistic study of consonant-tone interaction*. The Ohio State University.

Centre for the Protection of Language Resources of China (n.d.). *Distribution Map of Chinese Dialects*. The Collecting and Recording Platform of China Language Resources. Retrieved June, 20, 2021, from http://zhongguoyuyan.cn.

Cambier-Langeveld, G. M. (2000). *Temporal marking of accents and boundaries* (p. 171). Thesus.

Cao, J.F. (2007). Acoustic features of Neutral Tone in Standard Mandarin, *Applied Phonetics, 4,* 1-6. [曹剑芬. (1986). 普通话轻声音节特性分析. *应用声学*, *4*, 1-6.]

Cao, J.F. (2007). *Phonetic research of synchronic Chinese.* The Commercial Press. [曹剑芬. (2007). *现代语音研究与探索*. 商务印书馆.]

Chang, K. (1975). Tonal developments among Chinese dialects. *中央研究院歷史語言研究所集刊 [Bulletin of the Institute of History and Philology]*, 636-709.

Chao, Y. R. (1929). Intonation of Beijing Mandarin. *Collection of Chao Yuen Ren's Linguistic Studies.* The Commercial Press. [赵元任 (1920). 北平语调的研究, *赵元任语言学论文集*, 北京: 商务印书馆, 2002.]

Chao, Y. R. (1930). A system of tone letters. *Le maître phonétique*.

Chao, Y. R. (1929). Lexical tone and intonation in Chinese. *Collection of Chao Yuen Ren's*

Linguistic Studies. The Commercial Press. [赵元任. (1933). 汉语的字调跟语调,*赵元任语言学论文集*, 北京: 商务印书馆, 2002.]

Chao, Y. R. (1965). *A grammar of spoken Chinese*.

Chen, G. (1960). On the history of Neutral Tone in Chinese. *Zhongguo yuwen (3),* 137-137. [陈国. (1960). 汉语轻音的历史探讨. 中国语文(3), 137-137.]

Chen, S. H., Chang, S., & Lee, S. M. (1992). A statistical model based fundamental frequency synthesizer for Mandarin speech. *The Journal of the Acoustical Society of America*, *92*(1), 114-120.

Chen, Y. (2002). Accentual Lengthening of Monosyllabic-Constituents in Beijing Mandarin. In *Speech Prosody 2002, International Conference*.

Chen, Y. (2006). Durational adjustment under corrective focus in Standard Chinese. *Journal of phonetics*, *34*(2), 176-201.

Chen, Y., & Xu, Y. (2006). Production of Weak Elements in Speech – Evidence from F_0 Patterns of Neutral Tone in Standard Chinese. *Phonetica*, *63*(1), 47–75.

Chen, Y., & Gussenhoven, C. (2008). Emphasis and tonal implementation in Standard Chinese. *Journal of Phonetics*, *36*(4), 724-746.

Chen-Chung, Y. (1984). Cultural Principles Underlying English Teaching in China. *Language learning and communication*, *3*(1), 1-92.

Chien, Y. F., Yan, H., & Sereno, J. A. (2020). Investigating the Lexical Representation of Mandarin Tone 3 Phonological Alternations. *Journal of Psycholinguistic Research*, 1-20.

Chomsky, N., & Halle, M. (1968). The sound pattern of English.

Clements, G. N. (1987). Toward a substantive theory of feature specification. In *North East Linguistics Society* (Vol. 18, No. 1, p. 7).

Cole, J. (2009). Emergent feature structures: Harmony systems in exemplar models of phonology. *Language Sciences*, *31*(2-3), 144-160.

Cole, J., & Hualde, J. I. (2011). Underlying representations. *The Blackwell companion to phonology*, 1-26.

Compilation Office of Dictionaries in China. (2011). *Guo yu ci dian.* The Commercial Press. [中国大辞典编纂总处. (2011). 国语辞典. 北京: 商务印书馆.]

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question–answer contexts. Journal of the Acoustical Society of America, 77, 2142–2156.

De Marco, C., & Bergen, A. (1987). A security measure for random load disturbances in nonlinear power system models. *IEEE Transactions on Circuits and Systems*, *34*(12), 1546-1557.

Dink, J. W., & Ferguson, B. (2015). eyetrackingR: An R package for eye-tracking data analysis.

Dreher, J. J., & Lee, P. C. E. (1968). Instrumental investigation of single and paired Mandarin tonemes. *Monumenta serica*, *27*(1), 343-373.

Duanmu, S. (1990). *A formal study of syllable, tone, stress and domain in Chinese languages* (Doctoral dissertation, Massachusetts Institute of Technology).

Duanmu, S. (1994). Against contour tone units. *Linguistic inquiry*, *25*(4), 555-608.

Duanmu, S. (1993). Rime length, stress, and association domains. *Journal of East Asian Linguistics*, *2*(1), 1-44.

Duanmu, S. (1999). Metrical structure and tone: evidence from Mandarin and Shanghai. *Journal of East Asian Linguistics*, *8*(1), 1-38.

Duanmu, S. (2007). *The phonology of standard Chinese*. OUP Oxford.

Eulitz, C., & Lahiri, A. (2004). Neurobiological evidence for abstract phonological representations in the mental lexicon during speech recognition. *Journal of cognitive neuroscience*, *16*(4), 577-583.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology*, *43*(6), 543-549.

Field, A., Miles, J., & Field, Z. (2012). Discovering statistics using R.

Friedrich, C. K., Lahiri, A., & Eulitz, C. (2008). Neurophysiological evidence for underspecified lexical representations: asymmetries with word initial variations. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(6), 1545.

Gandour, J., Potisuk, S., & Dechongkit, S. (1994). Tonal coarticulation in Thai. *Journal of Phonetics*, *22*(4), 477-492.

Gussenhoven, C. (2008). Types of focus in English. In *Topic and focus* (pp. 83-100). Springer, Dordrecht.

Gao,Y. (1980) Neutral Tone in Beijing Mandarin. *Language Teaching and Linguistic Studies* (02), 82-98.

Gao, J., & Li, A. (2017, October). Production of neutral tone on disyllabic words by two-year-old Mandarin-speaking children. In *International Seminar on Speech Production* (pp. 89-98). Springer, Cham. [高玉振. (1980). 北京话的轻声问题. *语言教学与研究*(02), 82-98.]

Goldsmith, J. A. (1976). *Autosegmental phonology* (Doctoral dissertation, Massachusetts Institute of Technology).

Goldsmith, J. A. (1990). *Autosegmental and metrical phonology* (Vol. 1). Basil Blackwell.

Gósy, M., & Terken, J. (1994). Question marking in Hungarian: timing and height of pitch peaks. *Journal of Phonetics*, *22*(3), 269-281.

Gussenhoven, C. (1983). Testing the reality of focus domains. Language and Speech, 26(1), 61–80.

Gussenhoven, C. (2002). Intonation and interpretation: phonetics and phonology. In *Speech Prosody 2002, International Conference*.

Gussenhoven, C., & Teeuw, R. M. (2007). A moraic and a syllabic H-tone in Yucatec Maya.

Gussenhoven, Carlos. (2018). The Phonology of Tone and Intonation (Cambridge University Press, 2004, reprint with corrections 2008).

Gwet, K. L. (2019). irrCAC: Computing Chance-Corrected Agreement Coefficients (CAC). R package version 1.0. https://CRAN.R-project.org/package=irrCAC

Halle, M. (1959). Questions of linguistics. *Il Nuovo Cimento (1955-1965)*, *13*(2), 494-517.

Halle, M., Idsardi, W., & Goldsmith, J. A. (1994). General properties of stress and metrical structure. *DIMACS Serier of Discrete Mathematics and Theoretical Computer Science*, *17*, 37-70.

Halle, M., & Stevens, K. N. (2013). A note on laryngeal features. In *From memory to speech and back* (pp. 45-61). De Gruyter Mouton.

Harris, Z. S. (1963). Structural linguistics.

Heldner, M., & Strangert, E. (2001). Temporal effects of focus in Swedish. *Journal of Phonetics*, *29*(3), 329-361.

Hollien, H. (1960). Vocal pitch variation related to changes in vocal fold length. *Journal of Speech Language & Hearing Research, 3*(2), 150.

Hollien, H. , & Moore, G. P. . (1960). Measurements of the vocal folds during changes in pitch. *J Speech Hear Res, 3*(2), 157.

House, D. (2003, August). Perceiving question intonation: the role of pre-focal pause and delayed focal peak. In *Proc 15th ICPhS* (pp. 755-758).

Hsieh, F. (2021). On the existence of word stress in Chinese [论汉语有词重音].

Huang Zhi Qiang & Yang Jian Qiao. (1990). On the Disyllabification of Chinese Words. *Fudan University Journal (Social Science Edition)*(01), 98-101. [黄志强, & 杨剑桥. (1990). 论汉语词汇双音节化的原因 (On the Disyllabification of Chinese Words). 复旦学报(社会科学版)(01), 98-101.]

Huang, K. (2018). Phonological identity of the neutral-tone syllables in Taiwan Mandarin: An acoustic study. *Acta Linguistica Asiatica*, *8*(2), 9-50.

Hyman, L. M. (1975). *Phonology: theory and analysis* (Vol. 10). Harcourt College Pub.

Hyde, B. (2011). Extrametricality and Non-Finality. *The Blackwell companion to phonology*, 1-25.

Irigoien, I., Mestres i Naval, F., & Arenas Solà, C. (2016). Weighted distance based discriminant analysis: The R package WeDiBaDis. *The R Journal, 2016, vol. 8, num. 2, p. 434-450*.

Jakobson, R. (1949). On the Identification of Phonemic Entities; *Travaux du Cercle Linguistique de Copenhague* 5: 205-213

Jakobson, R., Fant, C. G., & Halle, M. (1951). Preliminaries to speech analysis: The distinctive features and their correlates.

Jiang, S. Y. (1994). *Introduction to Studies on Modern Chinese.* Peking University Press. [蒋绍愚. (1994). *近代汉语研究概况.* 北京大学出版社.]

Jin, S. (2001). *Dynamic study of Neutral Tone in modern Chinese*. Publishing House of Nationalities. [劲松. (2001). *现代汉语轻声动态研究.* 民族出版社.]

Johnson, K. (2007). Decisions and mechanisms in exemplar-based phonology. *Experimental approaches to phonology*, 25-40.

Keating, P. A. (1988). Underspecification in phonetics. *Phonology*, *5*(2), 275-292.

Keating, P., & Garellek, M. (2015). Acoustic analysis of creaky voice. *Poster apresentado em sessão especial sobre voz crepitante no Encontro Anual da Linguistic Society of America em Portland (OR)*.

Kiparsky, P. (1985). Some consequences of lexical phonology. *Phonology*, *2*(1), 85-138.

Kochanski, G., & Shih, C. (2003). Prosody modeling with soft templates. *Speech Communication*, *39*(3-4), 311-352.

Kotzor, S., Zhou, B., & Lahiri, A. (2020). (A) Symmetry in vowel features in verbs and pseudoverbs: ERP evidence. *Neuropsychologia*, *143*, 107474.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. (2017). lmerTest package: tests in linear mixed effects models. *Journal of statistical software*, *82*(1), 1-26.

Lachenbruch, P. A., & Goldstein, M. (1979). Discriminant analysis. *Biometrics*, 69-85.

Ladd, D. R. (1980). The structure of intonational meaning. Bloomington: Indiana University Press.

Ladd, R. (1996). Intonational phonology. Cambridge: Cambridge University Press.

Ladd, D. R., & Morton, R. (1997). The perception of intonational emphasis: continuous or categorical?. *Journal of phonetics*, *25*(3), 313-342.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, 363-374.

Lahiri, A., & Reetz, H. (2008). Underspecified recognition. In *Laboratory phonology 7* (pp. 637-676). De Gruyter Mouton.

Lahiri, A., & Reetz, H. (2010). Distinctive features: Phonological underspecification in representation and processing. *Journal of Phonetics*, *38*(1), 44-59.

Law of the People's Republic of China on the Standard Spoken and Written Chinese Language, (2000). http://english1.english.gov.cn/laws/2005-09/19/content_64906.htm.

Lee, W.-S., & Zee, E. (2014). Chinese Phonetics. In *The Handbook of Chinese Linguistics* (pp. 367–399). Wiley-Blackwell.

Lenth, R. V. (2016). Least-squares means: the R package lsmeans. *Journal of statistical software*, *69*(1), 1-33.

Li, R. (1987). Neutral Tone morphemes in the old novels. *Zhongguo yuwen* 6. [李荣. (1987). 旧小说里的轻声字. 中国语文 6.

Li, X. (1992). 'zhe' and 'zan' in *Xixiang* -- the historic evidence of Neutral Tone in Chinese. *Journal of Chongqing Normal University: Philosophy and Social Science Edition, (4),* 82-93. [黎新第. (1992).《董西厢》曲句"着""咱"二字的平仄——汉语轻声的早期历史印迹之一. *重庆师范大学学报：哲学社会科学版, (4),* 82–93.]

Li, S. (2000). The diachronic research on Neutral Tone and r-ending retroflexion -- Traditional phonology and modern phonetic studies in Chinese. *Phonetics and Phonology.* [李思敬. (2000). 现代北京话的轻音和儿化音溯源--传统音韵学和现代汉语语音研究结合举隅. *声韵论丛*.]

Li, X. (2004). Phonetic studies of Southwest Mandarin. (Doctoral dissertation, Shanghai Normal University). [李霞. (2004). 西南官话语音研究. (博士论文, 上海师范大学).]

Li, Y., Lee, T., & Qian, Y. (2002). Acoustical F0 analysis of continuous Cantonese speech. In *International Symposium on Chinese Spoken Language Processing*.

Li, Z. (2003). *The phonetics and phonology of tone mapping in a constraint-based approach* (Doctoral dissertation, Massachusetts Institute of Technology).

Lin, H. (2006). Mandarin neutral tone as a phonologically low tone. *Journal of Chinese Language and Computing*, *16*(2), 121-134.

Li, A., Gao, J., Jia, Y., & Wang, Y. (2014, December). Pitch and duration as cues in perception of neutral tone under different Contexts in Standard Chinese. In *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific* (pp. 1-6). IEEE.

Li, A., & Fan, S. (2015). Correlates of Chinese neutral tone perception in different contexts. In *ICPhS*.

Li, X., & Chen, Y. (2015). Representation and processing of lexical tone and tonal variants: evidence from the mismatch negativity. *PloS one*, *10*(12), e0143097.

Lin, Y. H., & Li, L.Y. (2017). On the Neutral Tone words in Contemporary Chinese Dictionary (the 6th edition). *Journal of Heibei Normal University: Philosophy and Social Science Edition* (40), 101-106. [林瑀欢, & 李丽云. (2017). 《现代汉语词典》(第 6 版)轻声词处理问题刍议. *河北师范大学学报:哲学社会科学版*(40), 101-106.]

Lin, M. C., & Yan, J. Z. (1980). Acoustic features of Neutral Tone in Mandarin. *Dialects,* (3), 166-178. [林茂灿, & 颜景助. (1980). 北京话轻声的声学性质. *方言*, (3), 166-178.]

Lin T. (1983). A primary test on neutral tones in Beijing Mandarin. *Papers on Linguistics,* (10). [林焘 (Lin T). (1983). 探讨北京话轻音性质的初步实验. *语言学论丛,* (10).]

Liu, L. H. (2001). *Research on Loudi dialect.* Central South University Press. [刘丽华. (2001). *娄底方言研究.* 中南大学出版社.]

Liu, F., & Xu, Y. (2007). The neutral tone in question intonation in Mandarin. In *Eighth Annual Conference of the International Speech Communication Association*.

Liu, C. T., & Chen, L. M. (2020). Testing the applicability of third tone sandhi at the intonation boundary: The case of the monosyllabic topic. *Language and Linguistics*, *21*(4), 636-651.

Lu, Y. Z. (2001). *Neutral Tone and r-ending retroflexion*. The Commercial Press. [鲁允中. (2001). *轻声和儿化.* 商务印书馆.]

Lv, S. (1955). *Collection of essays on Chinese grammar.* Science Press. [吕叔湘. (1955). *汉语语法论文集.* 科学出版社.]

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data.

Journal of neuroscience methods, *164*(1), 177-190.

Meng, Y., Wynne, H., & Lahiri, A. (2021). Representation of "T3 sandhi" in mandarin: significance of context. *Language, Cognition and Neuroscience*, 1-19.

Mishra, R. K., Olivers, C. N., & Huettig, F. (2013). Spoken language and the decision to move the eyes: To what extent are language-mediated eye movements automatic?. *Progress in brain research*, *202*, 135-149.

Nguyễn, T. A. T., & Thư, A. (2010). Rhythmic pattern and corrective focus in Vietnamese polysyllabic words. *The Mon-Khmer Studies Journal*, *39*, 1-28.

Norman, J. (2009). A new look at Altaic. *Journal of the American Oriental Society*, *129*(1), 83-89.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Ren, G. Q., Tang, Y. Y., Li, X. Q., & Sui, X. (2013). Pre-attentive processing of Mandarin tone and intonation: Evidence from event-related potentials. *Functional brain mapping and the endeavor to understand the working brain*, *6*, 95-108.

Rialland, A. (2007) Question prosody: An African perspective. In Riad, T & C. Gussenhoven. Tones and Tunes, a typological perspective. Berlin: Mouton de Gruyter. Pp. 35–62.

Sagey, E. C. (1986). *The representation of features and relations in non-linear phonology* (Doctoral dissertation, Massachusetts Institute of Technology).

Sarmah, P., Dihingia, L., & Lalhminghlui, W. (2015). Contextual variation of tones in Mizo. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Secretariat of the Academic Conference on Modern Chinese Normals. (1956). *Documents of the Academic Conference on Modern Chinese Normals.* Science Press. [现代汉语规范问题学术会议秘处. (1956). *现代汉语规范问题学术会议文件汇编.* 科学出版社.]

Shen, X. S. (1992). Mandarin neutral tone revisited. *Acta Linguistica Hafniensia*, *Vol 24*(No 1).

Shen, J., Deutsch, D., & Rayner, K. (2013). On-line perception of Mandarin Tones 2 and 3: Evidence from eye movements. *The Journal of the Acoustical Society of America*, *133*(5), 3016–3029.

Sherard, M. (1972). *Shanghai phonology*. Cornell University.

Shi, P. W. (1984). Neutral Tone and Neutral Tone Teaching. *Yu wen jian she,* (5), 50-52. [石佩雯. (1984). 轻声和轻声教学. *语文建设*, (5), 50-52.]

Shi, C. (1986). Compendium of History of Chinese Grammar. Huadong Normal University Press.

[史存直. (1986). 汉语语法史纲要. 华东师范大学出版社.]

Shih, C. (1998). Intonation. In *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach* (pp. 141-190). Kluwer Academic Publishers.

Shih, C. (2004). Tonal effects on intonation. In *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*.

Slowiaczek, L. M. (1990). Effects of lexical stress in auditory word recognition. *Language and Speech*, *33*(1), 47-68.

Sluijter, A. M., & van Heuven, V. J. (1995). Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch. *Phonetica*, *52*(2), 71-89.

Soltano, E. G., & Slowiaczek, L. M. (1997). *Lexical stress and metrical stress in auditory word recognition* (Doctoral dissertation, Acoustical Society of America).

Song, X.Q. (1990). Acoustic evidence for the realization norm of Neutral Tone in Standard Mandarin. *Yu wen jian she,* (05), 48-51. [宋欣桥. (1990). 普通话轻声词规范的语音依据. 语文建设(05), 48-51.]

SY, W. (1967). Phonological features of tone. *International journal of American linguistics*, *33*(2), 93-105.

Trubetzkoy, N. (1939). Grundzüge der Phonologie. *Travaux du cercle linguistique de Prague* 7.

Tsu-Lin, M. (1970). Tones and prosody in Middle Chinese and the origin of the rising tone. *Harvard Journal of Asiatic Studies*, *30*, 86-110.

Tu, J. Y., & Chien, Y. F. (2020, May). The processing of mandarin chinese tonal alternations in contexts: An eye-tracking study. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6244-6248). IEEE.

Tukey, J. W. (1977). *Exploratory data analysis* (Vol. 2, pp. 131-160).

Turk, A. E., & White, L. (1999). Structural influences on accentual lengthening in English. *Journal of phonetics*, *27*(2), 171-206.

Ohala, J. J. (1983). The origin of sound patterns in vocal tract constraints. In *The production of speech* (pp. 189-216). Springer, New York, NY.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. Behavior Research Methods. 10.3758/s13428-018-01193-y

Peng, S.-H. (2000). Lexical versus 'phonological'representations of Mandarin Sandhi tones. In

Papers in Laboratory Phonology V: Acquisition and the Lexicon (pp. 152–167). Cambridge University Press.

Pierrehumbert, J. B. (1980). *The phonology and phonetics of English intonation* (Doctoral dissertation, Massachusetts Institute of Technology).

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. Frequency and the emergence of linguistic structure, ed. by Joan Bybee and Paul Hopper, 137-57.

Politzer-Ahles, S., Schluter, K., Wu, K., & Almeida, D. (2016). Asymmetries in the perception of Mandarin tones: Evidence from mismatch negativity. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(10), 1547.

Pulleyblank, D. (1986). *Tone in lexical phonology* (Vol. 4). Springer Science & Business Media.

Psychology Software Tools, Inc. [E-Prime 2.0]. (2012). Retrieved from https://support.pstnet.com/.

Wang, J. (1996, October). An acoustic study of the interaction between stressed and unstressed syllables in spoken Mandarin. In Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96 (Vol. 3, pp. 1616-1619). IEEE.

Wang, J. (1997). The representation of the neutral tone in Chinese Putonghua. In *Studies in Chinese phonology* (pp. 157-184). De Gruyter Mouton.

Wang, L. (1980). Hanyu Shi Gao [An Outline of Chinese Language History]. *Beijing: China Book Company*.

Wang, X.D. (1992). The trend for Neutral Tone to become T4 in Beijing dialect and its influence. *Zhongguo yuwen, 2,* 124-128. [王旭东. (1992). 北京话的轻声去化及其影响. *中国语文*, *2*, 124-128.]

Wang, Y. J. (2004). The effects of pitch and duration on the perception of the neutral tone in standard Chinese [J]. Acta Acustica, 5.

Wang (2012). Neutral Tone Annotation in Xin Hua Dictionary from 1957 to 2011. Ludong University. [王婷婷 (2012). *1957 版至 2011 版《新华字典》注音研究*. 鲁东大学.]

Wang, Z & Feng, S. (2006). Tone Contrast and Stress Pattern of Disyllabic Words in Beijing Mandarin. *Language Science* (1): 3-22 [王志洁 & 冯胜利. (2006). 声调对比法与北京话双音组的重音类型. *语言科学*(1): 3-22]

Woo, N. H. (1969). *Prosody and phonology* (Doctoral dissertation, Massachusetts Institute of Technology).

Wurm, S. A., Li, R., & Baumann, T. (1987). *Language atlas of China*. Australian Acad. of the

Humanities; Longman Group (Far East).

Xie, R. Y. (1998). The annotation of Neutral Tone in dictionary: three important points. *Collections of Chinese Dictionary Studies.* [谢仁友. (1998). *辞书中标注轻声时应注意的三个问题.* 中国辞书学文集.]

Xing. (1999). Tone Sandhi and Neutral Tone in Shenmu Dialect. Language Research (02), 62-72. [邢向东. (1999). 神木方言的两字组连读变调和轻声. *语言研究*(02), 62-72.]

Xu, S. R. (1956). Shuangyin zhuici de zhongyin guilu. *Zhonqgguo Yuwen*, *2*. [徐世荣.(1956). 双音节词的重音规律. *中国语文, 2*.]

Xu, Y. (1994). Asymmetry in contextual tonal variation in Mandarin. *Ad ances in the study of Chinese language processing*, *1*, 383-396.

Xu, Y. (1997). Contextual tonal variations in Mandarin. *Journal of phonetics*, *25*(1), 61-83.

Xu, Y. (2004). Transmitting tone and intonation simultaneously-the parallel encoding and target approximation (PENTA) model. In *International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages*.

Xu, Y., & Wang, Q. E. (2001). Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech communication*, *33*(4), 319-337.

Xu, R. B., & Mok, P. P. K. (2014). Cross-linguistic perception of Mandarin intonation. In *Proc. Speech Prosody* (Vol. 2014, pp. 638-642).

Xu, L. (2016). Research on lexical stress in Chinese. *Foreign language research, 4*. [徐来娣. (2016). 汉语词重音若干基本理论问题研究. *外语学刊*, *4*.]

Yang, C. (2008). Hayes' metrical phonology and lexical stress in Chinese. *Modern Foreign Languages*(01), 37-48. [杨彩梅. (2008). Hayes的重音理论与汉语词重音系统. 现代外语(01), 37-48.]

Yuan, B. (1992). *Introduction of Modern Chinese*. Shanghai Education Publishing House. [袁宾. (1992). *近代汉语概论.* 上海教育出版社.]

Yip, M. J. (1980). *The tonal phonology of Chinese* (Doctoral dissertation, Massachusetts Institute of Technology).

Yip, M. (1989). Feature geometry and cooccurrence restrictions. *Phonology*, *6*(2), 349-374.

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

Yue-Hashimoto, A. O. (1987). Tone sandhi across Chinese dialects. *Wang Li memorial volumes, English volume,* 445-474.

Yuan, J. (2006, December). Mechanisms of question intonation in Mandarin. In *International Symposium on Chinese Spoken Language Processing* (pp. 19-30). Springer, Berlin, Heidelberg.

Yuan, J. (2011). Perception of intonation in Mandarin Chinese. *The Journal of the Acoustical Society of America*, *130*(6), 4063-4069.

Yuan, J. H., & Chen, Y. (2014). 3rd tone sandhi in standard Chinese: A corpus approach. *Journal of Chinese Linguistics*, *42*(1), 218-237.

Yuan, J., & Shih, C. (2004). Confusability of Chinese intonation. In *Speech Prosody 2004, International Conference*.

Zemlin, W. R. (1981). Speech and Hearing Science: Anatomy and Physiology, 2nd edn. (Englewood Cliffs, NJ: Prentice-Hall).

Zhang, Z. S. (1988). *Tone and tone sandhi in Chinese* (Doctoral dissertation, The Ohio State University).

Zhou, C. M. (2018). Phonetic changes in Beijing dialect. Peking University Press. [周晨萌. (2018). *北京话语音演变研究*. 北京大学出版社.]