SAN DUANMU

METRICAL STRUCTURE AND TONE: EVIDENCE FROM MANDARIN AND SHANGHAI*

A well-known problem in Chinese phonology is that in some dialects most regular syllables keep their underlying tones, but in others the initial syllable determines the tonal pattern of a multisyllabic domain. Mandarin and Shanghai, two of the most studied dialects, best represent the contrast. Duanmu (1993) proposes that the two dialects differ in syllable structure but otherwise obey the same phonological constraints, including moraic trochee. However, a number of problems remain, such as questions regarding the metrical counting units, the predicted weight of a syllable and its phonetic duration, the economy of underlying tones, the mechanism of tone deletion, and the relation between weight and stress. This article offers a solution to the problems. The main proposal is that Chinese is both mora-counting and syllable-counting, in that a heavy syllable forms a bimoraic trochee, which I call M-foot, yet a minimal word must be a disyllabic trochee, which I call S-foot. In addition, both Mandarin and Shanghai are subject to tonal polarity, which is independently found in African languages. I also discuss the implication of the S-foot for metrical theory and other consequences of the present analysis.

1. THE PROBLEM

There is a well-known phenomenon in Chinese phonology: in some dialects most regular syllables keep their underlying tones,[1] whereas in other dialects the initial syllable determines the tonal pattern of a multisyllabic expression. The contrast can be illustrated with Standard Chinese (hereafter Mandarin) and Mainstream Shanghai. Mainstream Shanghai (also called New Shanghai by Xu et al. (1981)) is the variety spoken by the majority of people in Shanghai City (Xu et al. (1988)). It differs from Old Shanghai, spoken by a small number of old people, to be discussed in section 5.2. (1) shows data from Mandarin, transcribed in Pinyin, and (2) shows data from Mainstream Shanghai (hereafter Shanghai), transcribed in phonetic symbols. Tones are transcribed in the system of Chao (1930), which shows the starting, the ending, and if relevant the mid pitch levels, and where 5 is the highest pitch and 1 the lowest. For example, [pan] 'plate' in Mandarin starts with 3, a mid pitch, and ends in 5, a high pitch. The tone values in Shanghai are after Xu et al. (1988). Shanghai has syllabic consonants, such as [z] in [sz] 'four'.

Journal of East Asian Linguistics **8**, 1–38, 1999.

(1) Tonal stability in Mandarin

a. san55 + bei55 → 55 55	'three cups'
b. san55 + pan35 → 55 35	'three plates'
c. si51 + bei55 → 51 55	'four cups'
d. si51 + pan35 → 51 35	'four plates'

(2) Lack of tonal stability in Shanghai

a. se52 + pe52 → 55 21	'three cups'
b. se52 + bø23 → 55 21	'three plates'
c. sz34 + pe52 → 33 44	'four cups'
d. sz34 + bø23 → 33 44	'four plates'

In (1) the output tone pattern of each expression is a concatenation of the input syllable tones. In (2), however, the output tone pattern is determined by the initial syllable. In particular, five syllable tones can be distinguished phonetically in Shanghai. But if we ignore the effect of onset voicing, which affects tonal registers, and glottal rimes, Shanghai has two syllable tones, LH and HL. If the initial syllable tone is LH, the disyllabic pattern is [L H]. If the initial syllable tone is HL, the disyllabic pattern is [H L]. Most researchers on Shanghai agree that input tones on non-initial syllables are first deleted, and then input tones of the initial syllable are spread over the first two syllables (Zee and Maddieson (1979), Yip (1980), Lu (1987), Selkirk and Shen (1990), Duanmu (1995), Zhu (1995)). (3) shows the analysis of (2a).

(3) Traditional analysis of Shanghai

HL	HL		HL			H	L	
						|	|	
se	pe	→	se	pe	→	se	pe	
input			deletion			assoication		

The question of interest is why the deletion and reassociation process takes place in Shanghai but not in Mandarin. There are two approaches. One assumes that languages are naturally different, and thus the contrast between Mandarin and Shanghai needs no explanation. The other assumes that all human languages are fundamentally similar, and the difference between Mandarin and Shanghai requires an explanation that is based on universal linguistic principles. Duanmu (1993) adopts the second position and offers an analysis, reviewed in section 2. This study assumes the same position and offers a better solution.

2. The analysis of Duanmu (1993)

The argument of Duanmu (1993) is based on a crucial observation that in Shanghai (and dialects like it) all syllables have simple rimes, namely, no diphthongs or contrastive codas. In contrast, Mandarin (and dialects like it) has both diphthongs and contrastive codas. In addition, in dialects like Mandarin, there is a distinction between full syllables and weak syllables. The observation leads to the conclusion in (4) regarding syllable structure.

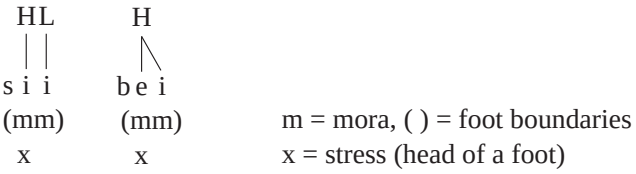
(4) Syllabic difference between Mandarin and Shanghai (Duanmu (1993))
   a. In Mandarin all full syllables are heavy (bimoraic).[2]
   b. In Shanghai all syllables are underlyingly light (monomoraic).

Given the independent syllabic difference between the two types of dialects, their tonal difference follows from the phonological constraints in (5).

(5) Constraints on Mandarin and Shanghai (Duanmu (1993))
   a. The basic metrical unit is moraic trochee.
   b. The tone-bearing unit is the moraic segment.
   c. A tonal domain is a stress domain.[3]
   d. Only a stressed syllable can retain its underlying tones.
   e. A minimal word must be bimoraic.

For illustration, consider (1c), analyzed in (6), assuming that 51 is HL and 55 is H.

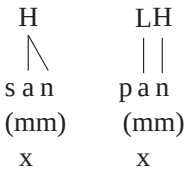
(6) Analysis of Mandarin (1c) (Duanmu (1993))



m = mora, ( ) = foot boundaries
x = stress (head of a foot)

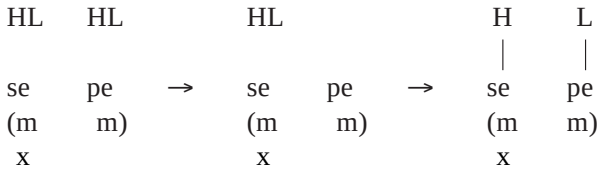
Since each syllable is heavy, it forms a bimoraic trochee whose head is on the first mora. In addition, since each syllable has two moras, it can take two tones. Next consider (1b), shown in (7), assuming that 35 is LH.

(7) Analysis of Mandarin (1b) (Duanmu (1993))



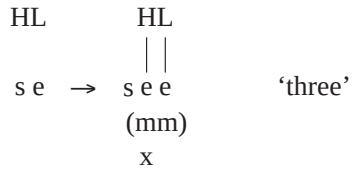
Again, each syllable forms a moraic trochee, so each syllable forms a separate tonal domain. Even though the first syllable has two tone-bearing units but just one tone, it cannot take the L of the second syllable since the latter belongs to a different domain. Let us now consider (2a) in Shanghai, shown in (8).

(8) Analysis of Shanghai (2a) (Duanmu (1993))



Because Shanghai syllables are light, a moraic trochee requires two syllables, with stress on the first. Because of (5d), the second syllable loses its underlying tones. Finally, since the first syllable is monomoraic, it can take just one tone, and the H has to be linked to the second syllable. Now when a Shanghai syllable occurs alone, it is lengthened to two moras, owing to the minimal bimoraic requirement. And because there are two moras, the syllable can take both its underlying tones. This is shown in (9).

(9) Analysis of a monosyllable in Shanghai (Duanmu (1993))



‘three’

In summary, the stability of syllable tones in Mandarin is due to the fact that every regular syllable is a bimoraic foot, which enables it to keep its underlying tones. The lack of tonal stability in Shanghai, and the left dominance effect, is due to the fact that only initial syllables have stress. Non-initial syllables lose their underlying tones for lack of stress. Besides, since the initial syllable is monomoraic, it can take just one tone and must shift its second tone to the next syllable.

3. Problems with Duanmu (1993)

The analysis of Duanmu (1993) has several problems. First, it is well known that all Chinese dialects are subject to a disyllabic requirement, by which a minimal expression should consist of two syllables. (10)–(12) show that, if a name is monosyllabic, a semantically redundant syllable (‘old’, ‘little’,

'city', 'country') is added. (13) shows words in which a semantically empty morpheme is used (italicized in the transcription and in parentheses in the gloss). (14) shows words that consist of two semantically repetitive morphemes.

(10) Personal address

Lao Zhang	Xiao Zhang	*Zhang
'Old Zhang'	'Little Zhang'	'Zhang'

(11) Place names

Wuxi	Shanghai	*Sha (Sha Shi)
'Wuxi'	'Shanghai	'Sha (Sha City)'

(12) Country names

Riben	Helan	*Fa (Fa Guo)
'Japan'	'Holland'	'France (France Country)'

(13) Empty morphemes

lao hu	*lao* shu	zhuo *zi*	mu *tou*
'(old) tiger'	'(old) rat'	'table (son)'	'wood (head)'

(14) Repetitive morphemes

shu cai	mei tan	hui hua	xue xi	zhong zhi
'vegetable'	'coal'	'to paint'	'to study'	'to plant'

It should be pointed out that although the disyllabic tendency is strong, some Chinese expressions can be monosyllabic, such as [wei] 'hello'. I will argue below that in such cases the monosyllable forms a disyllabic foot with a zero syllable.

The disyllabic requirement raises a theoretical question for dialects like Mandarin. In the theory of prosodic categories (Selkirk (1980), McCarthy and Prince (1986)), a minimal word is a foot. If a full Mandarin syllable is already a bimoraic foot, why should a minimal word need two syllables? If the Mandarin foot is not built on moras but on syllables, then a host of other questions arise: Why can every full Mandarin syllable keep its underlying tones, if not all of them are stressed? If all full Mandarin syllables are stressed, why is there no stress clash between them? Why can a full Mandarin syllable carry two tones? Why is there no tone spreading across full syllables?

The second problem with Duanmu (1993) is phonetic. According to Zhu (1995), in a disyllabic Shanghai word or compound, the first syllable is much longer than the second. Specifically, ignoring glottalized rimes, the average duration of the first rime is on the order of 200 ms whereas the average duration of the second rime is on the order of 100 ms.[4] This

may seem to support Duanmu's (1993) analysis that the first syllable is stressed and the second is not. However, a rime with an average duration of 200 ms is more like a heavy one than a light one. For example, according to Lin and Yan (1988), in heavy-light Mandarin disyllables, the average duration of the first rime is on the order of 200 ms, and that of the second is on the order of 100 ms. Similarly, according to Lin et al. (1984) and Wang and Wang (1993), in heavy-heavy Mandarin disyllables, the average duration of both rimes is on the order of 200 ms. The facts are summarized in (15).

(15) Size of rime duration in disyllabic units (normal read speech)

Dialect	First rime	Second rime
Shanghai	200 ms	100 ms
Mandarin (heavy-light)	200 ms	100 ms
Mandarin (heavy-heavy)	200 ms	200 ms

The data shows that phonetically a disyllabic unit in Shanghai matches the heavy-light pattern rather than the light-light pattern. But if the initial Shanghai syllable is heavy, one must explain why it takes just one of its underlying tones and shifts the other to the following syllable.

The third problem with Duanmu's (1993) analysis is phonological economy. Since Shanghai has only two syllable tones (ignoring tonal registers), it is sufficient to postulate just L vs. H, instead of LH vs. HL.

The fourth problem with Duanmu's (1993) analysis concerns the mechanism of tone deletion (thanks to two anonymous reviewers). The question here is whether syllable tones are prelinked before deletion or whether tone linking takes place after deletion. If syllable tones are prelinked, how can each monomoraic syllable bring in two tones (i.e., LH or HL)? If tones are not yet linked at the time of deletion, as shown in (8), how can one tell that the first two belong to the first syllable (so that they are exempt from deletion)?

The fifth problem concerns the relation between weight and stress. In many languages heavy syllables attract stress. Prince (1990) generalizes this phenomenon as the Weight-to-Stress principle. However, the reverse of the principle, i.e., all stressed syllables are heavy, appears to be too strong. For example, in the standard analysis of English, such as Halle and Vergnaud (1987), the stressed syllables in words like *city* and *apple* are light. But two recent works, Burzio (1994) and Hammond (1997), suggest that such syllables are in fact heavy.[5] This development is highly significant because English is the most studied language with regard to stress. If stressed "light" syllables in English are in fact heavy, one wonders whether the same is true in other languages. And if it is, a simpler generalization can be made with regard to weight and stress, namely, all heavy syllables are stressed,

and all stressed syllables are heavy. In this light, it is natural to question whether stressed (initial) syllables in Shanghai are indeed light.

4. The present analysis

I assume the basic argument of Duanmu (1993) that there is a syllabic difference between Mandarin and Shanghai. In Mandarin there is a contrast between full and light syllables. This fact is well known in traditional analyses. Most content words are full syllables, which have tone and longer duration and more stress than light syllables. Light syllables are mostly grammatical particles, which have no stress and whose tone depends on the full syllable before it. In Shanghai, on the other hand, there is no contrast between heavy and light syllables. All syllables have the structure CV (where V can be a syllabic consonant), which is longer in isolation and shorter in non-initial positions.

The weight difference in Mandarin must be assumed underlyingly. The syllable weight in Shanghai can be analyzed in two ways. Duanmu (1993) assumes that all Shanghai syllables are light underlyingly but are lengthened in isolation. An alternative is to assume that Shanghai syllables are unspecified for weight underlyingly, and their weight is determined by the position they occur in. In the present analysis I adopt the latter analysis. (16) summarizes the syllabic difference between Mandarin and Shanghai.

(16) Syllabic difference between Mandarin and Shanghai
  a. In Mandarin all full syllables are heavy (bimoraic).
  b. Shanghai syllables are unspecified for weight underlyingly.

In addition, I assume the constraints in (17), to be further specified shortly. (17) is similar to (5). The major difference is that the minimal word is bimoraic in (5e) but disyllabic in (17e).

(17) Constraints on Mandarin and Shanghai (preliminary)
  a. The lowest metrical unit is moraic trochee.
  b. The tone-bearing unit is the moraic segment.
  c. A tonal domain is a stress domain.
  d. Unstressed syllables can lose their underlying tones.
  e. A minimal word must be disyllabic.

The fact that a language can be metrically sensitive to both moras and syllables was noted by Prince (1980) and maintained through later works, such as McCarthy and Prince (1986), Prince and Smolensky (1993), and Kenstowicz (1995). But the metrical structure of a minimal word has been vague. I suggest that the minimal word must satisfy the two constraints

in (18). Following a convention in Optimality Theory, constraint names are given in small caps.

(18) a. TROCHEE (at both moraic and syllabic levels):
The head of a foot is on the left.

b. BINARITY (at both moraic and syllabic levels):
A foot must be (at least) binary.

Following Kager (1993), I assume that moraic trochee is the only metrical structure for a heavy syllable. Metrical binarity was first proposed by Prince (1980) and is now well known in metrical phonology. Since a minimal word has two levels of metrical constituents, I use two terms to distinguish them: I use M-foot to refer to a moraic trochee and S-foot to refer to a syllabic trochee. A minimal word, therefore, is an S-foot. There are four possible S-foot structures, but only two are good, shown in (19).

(19) S-foot (minimal word) structures

a. Good

x	x	
(x x)	(x .)	S-foot
(mm) (mm)	(mm) m	M-foot
heavy-heavy	heavy-light	

b. Bad

x	x	
(x .)	(x x)	S-foot
(m) m	(m) (mm)	M-foot
* light-light	* light-heavy	(violating BINARITY)

In (19a), a heavy syllable always forms a binary M-foot. A light syllable does not form an M-foot but is still visible at the S-foot level, shown by a dot. The light-light pattern in (19b) is bad because the stress on the first syllable requires it to be an M-foot, yet because it is monomoraic, it cannot form a binary M-foot, and so it violates BINARITY. Similarly, the light-heavy pattern is bad because of the violation of BINARITY.

There are several apparent problems for (19). First, one can think of American English words like *city* and *happy*, which have been thought to be light-heavy (the second syllable is heavy because it has a tense vowel [i]). Second, one can think of monosyllabic words, such as *John* or *no* in English. Third, one can think of words with iambic stress, such as *today*. For the first problem, I follow Hammond (1997) that the stressed syllable in such words is heavy. The latter two problems will be discussed in section 4.2. Finally, (19) differs from a popular assumption that Mandarin

compound stress is final (e.g., Chao (1968)). However, the assumption is mainly based on tradition rather than on facts. As Chao (1968, p. 32) points out, native intuition on this matter is rather vague. The only clear argument for final stress is that, when an expression is read in isolation, the final full syllable is the longest (Lin et al. (1984), Yan and Lin (1988)). However, there are two problems with this argument. First, although the final syllable is the longest, it does not have the greatest amplitude or pitch range, but the initial syllable does (Yan and Lin (1988), Yang (1992)). Second, when an expression is read in a carrier sentence, which is a more appropriate phonetic method, the final syllable is not the longest, but the initial syllable is (Wang and Wang (1993)). Thus, when an expression is read in isolation, the longer duration of the final syllable is likely due to final lengthening, a well-known phonetic effect (Klatt (1975)), instead of final stress. Phonetic facts from expressions read in a carrier sentence clearly show that the initial syllable is the most prominent overall. Phonologically, there is no evidence for final stress either, but there is evidence for initial stress. For further discussion, see Duanmu (1997).

(19) correctly predicts disyllabic expressions in Mandarin, which can be heavy-heavy or heavy-light, but not light-light or light-heavy, a point made by Lin (1994). (19) also offers a better analysis of Shanghai. Since no Shanghai syllable is inherently heavy, only the initial syllable gets stress. In addition, since a stressed syllable must be heavy, the Shanghai S-foot must be heavy-light, instead of light-light as proposed by Duanmu (1993). This analysis agrees with the phonetic fact that the initial Shanghai rime is as long as a full Mandarin rime, and about twice as long as a non-initial Shanghai rime or a weak Mandarin rime (cf. (15)).[6]

Next consider tone. The analysis of Mandarin is similar to that of Duanmu (1993), where each full syllable forms a bimoraic M-foot, carries stress, and has two tone-bearing units. Thus, every full syllable can keep its underlying tones. The problem now is to explain why the initial syllable in Shanghai, which is also bimoraic and has two tone-bearing units, takes just one tone. I propose the following analysis. First, the syllable tones in Shanghai are H and L, instead of HL and LH. A tonal polarity constraint requires an initial tone to be followed by an opposite tone. Thus, H leads to HL, and L leads to LH. A further constraint requires a syllable to carry a simple tone. This forces the polarity tone to occur on the second syllable, even though the initial syllable has two tone-bearing units. Finally, both the polarity constraint and the simple tone constraint can be overridden by the requirement for tonal categories to remain distinct, to be seen below. My analysis will be cast in Optimality Theory (Prince and Smolensky (1993)) in that I assume a set of ranked constraints. For illustration, I list the relevant metrical con-

straints in (20) and tonal constraints in (21), where "A ≫ B" means "A is ranked higher than B." The constraints apply to both Mandarin and Shanghai.

(20) Metrical constraints
   a. S-FOOT: heavy-heavy or heavy-light (see (19)).
   b. KEEP WEIGHT: Syllables must preserve their underlying weight.
   S-FOOT ≫ KEEP WEIGHT

(21) Tonal constraints
   a. T-DISTINCTION: A stressed syllable must maintain its tonal distinction.
   b. POLARITY: An initial tone is followed by an opposite tone at surface in an S-foot.
   c. SIMPLE TONE: Avoid contour tones.
   T-DISTINCTION ≫ POLARITY ≫ SIMPLE TONE

It is possible to decompose S-FOOT into other constraints, such as TROCHEE and BINARITY (along with mora counting and syllable counting). For simplicity I forego the decomposition. T-DISTINCTION and KEEP WEIGHT can be related to FAITHFULNESS of McCarthy and Prince (1995) or ANTI-ALLOMORPH of Burzio (1996). It is important to note that KEEP WEIGHT has different effects on Mandarin and Shanghai. It makes Mandarin syllables keep their underlying weight, yet it allows Shanghai syllables to surface as either light or heavy because Shanghai syllables are unspecified for weight underlyingly. T-DISTINCTION is sensitive to the number of tonal categories in a language. For example, if a language has two categories, H and L, H can change to HL without violating T-DISTINCTION since HL is still distinct from L. However, if a language has three categories, H, HL, and L, H cannot change to HL, otherwise their distinction will be lost. POLARITY has been reported in several African tone languages (cf. Newman (1997)),[7] and can probably be related to the obligatory contour effect (cf. Leben (1971)). The condition "at surface" means that an unrealized tone cannot satisfy POLARITY (see below). SIMPLE TONE can be related to the genereal principle of economy. I illustrate the analysis of disyllables in section 4.1, monosyllables in section 4.2, and trisyllables and longer domains in section 4.3.

4.1. *Disyllables*

Let us start with the metrical structure in Shanghai, illustrated with (2a) in (22). Following the convention in Optimality Theory, constraints are shown on the top row, and output candidates are shown in the first column. An asterisk * indicates a violation of a constraint, and an exclamation mark ! indicates a violation that rules out a candidate. Shaded cells mean that they are no longer relevant for the evaluation. The best candidate is indicated by ✓.

(22) Metrical analysis of Shanghai (2a) ((22b) will be ruled out by tonal constraints)

	/se pe/	S-FOOT	KEEP WEIGHT
a. ✓	x (x .) (mm) m se e pe		
b. ✓	x (x x) (mm) (mm) se e pe e		
c.	x (x .) (m) m se pe	*!	

The imput is shown in the top left cell. Since Shanghai syllables are unspecified for weight underlyingly, all candidates satisfy KEEP WEIGHT. (22a) and (22b) satisfy S-FOOT, but (22c) violates it. So we get two best candidates for metrical structure, (22a) and (22b). Next consider tone, analyzed in (23). (23a–c) show three tone patterns for the metrical structure (22a), and (23d) shows a tone pattern for the metrical structure (22b).

(23) Tonal analysis of Shanghai (2a)

	H H /se pe/	T-DISTINCTION	POLARITY	SIMPLE TONE
a. ✓	H L see pe x			
b.	H H see pe x		*!	
c.	HL se e pe x			*!
d.	HL H se e pee x			*!

It is immaterial whether the underlying tones are prelinked or not. In (23a), the second syllable has changed to L, which satisfies POLARITY, and since it is unstressed, there is no violatiion of T-DISTINCTION. In (23b) there is a violation of POLARITY. In (23c), the first syllable has changed to HL. Since HL is still distinct from L, it does not violate T-DISTINCTION, but it does violate SIMPLE TONE. In (23d), both syllables are heavy and stressed. To satisfy T-DISTINCTION, both must keep H. To satisfy POLARITY, L must follow the first H, so that the first syllable becomes HL. However, the HL on the first syllable violates SIMPLE TONE. Thus, the best candidate is predicted to be (23a), correctly.

Next consider (2b). The metrical analysis is same as that for (2a). The tonal analysis is shown in (24). Since the heavy-heavy structure will violate SIMPLE TONE, only the heavy-light structure is shown.

(24) Tonal analysis of Shanghai (2b)

	H L /se bø/	T-DISTINCTION	POLARITY	SIMPLE TONE
a. ✓	H L \ | see bø x			
b.	H H \ | see bø x		*!	
c.	HL || see bø x			*!

In (24a) there is no violation. (24b) violates POLARITY. (24c) violates SIMPLE TONE. The best candidate is again correctly predicted to be (24a).

Let us now consider disyllables in Mandarin. First, consider the metrical structure, examplified with (1c) in (25).

(25) Metrical analysis of Mandarin (1c)

	/sii bei/	S-FOOT	KEEP WEIGHT
a. ✓	x (x x) (mm) (mm) s i i be i		
b.	x (x .) (mm) m s i i bei		*!
c.	x (x x) (m) (mm) s i be i	*!	*

It will be noted that [sii] 'four' is a full syllable, whose weight was not reflected in (1) because the Pinyin transcription does not mark length. Since both syllables are underlyingly heavy, they must surface as such, or KEEP WEIGHT will be violated. Thus, the best output is (25a). Next consider tone. Mandarin has four tonal categories on full syllable, given in (26). I have omitted some tonal variations. For example, Tone 4 is often a shorter fall 53 before another Tone 4 (cf. Chao (1968)). In addition, the Third Tone Sandhi is postponed till section 5.1.

(26) Mandarin tones

Categories	Isolation	Nonfinal	Underlying
First	55	55	H
Second	35	35	LH
Third	214	21	L
Fourth	51	51	HL

Isolation and non-final tones are given in Chao digits (Chao (1930)). The marking 21 for the Third tone may suggest a fall, but it is largely a low tone. A more accurate marking is either 211 (Chao (1968, p. 27)) or 11 (Chao (1931, p. 323), Wang (1979)). Ignoring tonal registers, the First, Second, and Fourth tones can be represented as H, LH, and HL respectively. For the Third tone, some take the isolation form 214 to be basic (e.g., Chao (1968), Milliken (1989)), and some take the nonfinal form 21 to be basic (e.g., Yip (1980), Wang (1997)). I adopt the latter, to be justified below. Besides full syllables, Mandarin also has light syllables, which are underlyingly toneless, or are said to belong to the Fifth tone category.[8] (27) shows the analysis of (1c).

(27) Tonal analysis of Mandarin (1c)

	HL H /si i bei/	T-DISTINCTION	POLARITY	SIMPLE TONE
a. ✓	HL H | | | si i bei x x			*
b.	HL || si i bei x x	*!		*
c.	H L \ \ si i bei x	*!*		

In (27a) each syllable keeps its underlying tones, and there is a violation of SIMPLE TONE. In (27b) the second syllable loses its underlying tones, which causes a violation of T-DISTINCTION. In addition, there is a violation of SIMPLE TONE. In (27c), the second syllable loses its underlying tones, and the L of the first syllable is moved to the second. This causes two violations of T-DISTINCTION since both syllables changed their categories: the first changed from the Fourth tone to the First tone, and the second syllable changed from the First tone to the Third. Thus, the best candidate is (27a), which has no tone change. Next consider (1a), analyzed in (28).

(28) Tonal analysis of Mandarin (1a)

	H H /san bei/	T-DISTINCTION	POLARITY	SIMPLE TONE
a.	H L \ \ san bei x x	*!		
b. ✓	H H \ \ san bei x x		*	

In (28a), the second syllable has changed to L. While this satisfies POLARITY, it violates T-DISTINCTION. In (28b) both syllables keep their underlying tones. While this violates POLARITY, it satisfies the higher ranked T-DISTINCTION. Thus, (28b) is the better output.

It is interesting to compare (28) = (1a) with its Shanghai counter-part (23) = (2a). They have the same underlying tones [H H] but different surface tones. In Mandarin, because both syllables are stressed, they both keep their underlying tones, giving surface [H H]. In Shanghai, because the second syllable is unstressed, it can change to L to satisfy POLARITY without violating T-DISTINCTION, giving surface [H L]. In other words, Mandarin shows tonal stability and an apparent lack of the POLARITY effect because KEEP WEIGHT requires all full syllables to remain heavy, hence stressed, and T-DISTINCTION requires them to keep their underlying tones. In contrast, Shanghai lacks tonal stability because its syllables – being underlyingly unspecified for weight – are not constrained by KEEP WEIGHT. The tonal constraints then choose the heavy-light structure, in which the non-initial syllable is light and need not keep its underlying tone. However, the POLARITY effect can be observed in Mandarin when a full syllable is followed by a light syllable. According to Chao (1968, p. 36), a light syllable has the highest pitch after the Third tone and lower values after other tones. Cheng (1973, p. 56) suggests that the speaker intends to produce H in the formeer case and L in the latter. Tha data in (29) follows Cheng's interpretation.

(29) POLARITY in Mandarin

a.	H			H	L	
	fei	le	→	fei	le	'fly Asp (flied)'
b.	LH			LH	L	
	lai	le	→	lai	le	'come Asp (came)'
c.	L			L	H	
	mai	le	→	mai	le	'buy Asp (bought)'
d.	HL			HL	L	
	mai	le	→	mai	le	'sell Asp (sold)'

The aspect marker (Asp) [le] is underlyingly light. It surfaces as H after L but as L after other tones. Assuming that the default value of an unstressed syllable is L, one must explain why it is H after L. In the present analysis, it is due to POLARITY, by which the light syllable should be L after H, as in (29a), and H after L, as in (29c). After LH and HL, which already satisfy POLARITY, the light syllable should take the default value L, as in (29b) and (29d). The analysis of (29a) and (29c) are shown in (30) and (31).

(30) Tonal analysis of Mandarin (29a)

	H /fei le/	T-DISTINCTION	POLARITY	SIMPLE TONE
a. ✓	H L \ | fei le x			
b.	H H \ | fei le x		*!	
c.	HL || fe i le x	*!		*!

(31) Tonal analysis of Mandarin (29c)

	L /mai le/	T-DISTINCTION	POLARITY	SIMPLE TONE
a.	L L \ | mai le x		*!	
b. ✓	L H \ | mai le x			
c.	LH || ma i le x	*!		*!

Because of T-DISTINCTION, the heavy initial syllable cannot change its tone. Because of POLARITY, the light second syllable must take an opposite tone.

I mentioned above that some studies consider the Mandarin Third tone to be underlyingly 214. For example, Milliken (1989) porposes that the

Third tone is underlyingly L(H), i.e., an L followed by a floating H. The floating H helps to explain why a light syllable is H after the Third tone but L after other tones. In the present analysis there is no need to posit the floating H. Its appearance after the Third tone is due to POLARITY, which is independently motivated in Shanghai and African languages. I return to the analysis of the monosyllabic 214 below.

4.2. *Monosyllables*

I proposed in (19) that a minimal word, or S-FOOT, must be disyllabic. Let us now consider the analysis of stressed monosyllables, such as *yes* or *John*. There are two approaches. The first, such as Halle and Vergnaud (1987), considers monosyllables to be exceptions, where the disyllabic requirement is relaxed. The other, such as Hogg and McCully (1987) and Burzio (1994), considers there to be no exception to the disyllabic requirement. Specifically, when a stressed monosyllable is followed by an unstressed syllable, they form a disyllabic S-foot, such as *(ˈeat an) (ˈapple)* or *(ˈthis po)(ˈtato)*. When a stressed monosyllable is not followed by an unstressed syllable, it is accompanied by a "zero" or "silent" syllable, and they together form a disyllabic S-FOOT. Although the choice is not consequential for the present analysis, I will adopt the second approach since there seems to be considerable evidence for it.

Many people have made arguments for the zero syllable (e.g., Abercrombie (1967), Liberman (1975), Selkirk (1984), Giegerich (1985), Hogg and McCully (1987), Burzio (1994)). Abercrombie (1967, pp. 35–36) gives an excellent example which shows that a syllable need not be audible. He points out that the English expression *thank you* is often pronounced as ['kju], where the first syllable *than(k)* is acoustically silent. But there still is muscular action during the silent syllable, which resembles the production of a stressed syllable. Similarly, Hogg and McCully (1987) argue that a zero syllable always follows a stressed monosyllable (unless a weak syllable follows) so that they form a disyllabic trochee, which, rhythmically, occupies the same amount of time as an overt disyllabic trochee. This can be seen in the nursery rime in (32).

(32) (Ding ø) (dong ø) (bell ø)
(Kit-ty's) (in the) (well ø)

In (32) each line forms three trochees. On line one each trochee consists of a monosyllable and a zero syllable. On line two the first two trochees are filled by overt syllables, and the third consists of a monosyllable and a zero syllable. The zero syllable need not always be realized as silence

but can be filled by lengthening the preceding monosyllable. For example, the first two trochees on line one can be read as a prolonged *ding* and a prolonged *dong*, without pauses. What is important is that the zero syllable adds an extra unit of time, or beat, as the second part of a binary S-FOOT. The same argument applies to words like *today* and *again*, which appear to be iambic. In the present analysis *today* is *to(day ø)*, where the stressed second syllable forms an S-FOOT with a zero syllable.[9]

Since a heavy syllable can form a binary S-FOOT with a zero syllable, a reviewer wonders whether the heavy syllable still needs to form a bimoraic foot itself. The answer is yes. A heavy syllable always forms a bimoraic foot whether it is also the head of an S-FOOT or not. The moraic foot is needed to explain why in a heavy-heavy trochee, the second heavy syllable also has stress and can carry two tones. Its stress comes from its being a binary M-FOOT. Another reviewer asks why Chinese still uses an overt syllable to make S-FEET, as seen in (10)–(14), if a monosyllable can form an S-FOOT with a zero syllable. The reason, apparently, is that an overt syllable is preferred to a zero syllable. The zero syllable is used only when an overt syllable is not available.

Let us now consider the metrical structure of monosyllables. Assuming that the zero syllable has one mora, a Shanghai monosyllable is analyzed in (33).

(33) Metrical analysis of the Shanghai monosyllable

	/se/ 'three'	S-FOOT	KEEP WEIGHT
a. ✓	x (x .) (mm) m s e e ø		
b.	x (x .) (m) m s e ø	*!	

Since Shanghai syllables are underlyingly unspecified for weight, KEEP WEIGHT is always satisfied. Since (33a) satisfies S-FOOT, and (33b) does not, (33a) is a better form. Next consider a full Mandarin monosyllable, analyzed in (34).

(34) Metrical analysis of the full Mandarin monosyllable

	/san/ 'three'	S-FOOT	KEEP WEIGHT
a. ✓	x (x .) (mm) m s a n ø		
b.	x (x .) (m) m san ø	*!	*

Since the syllable is underlyingly heavy, (34a) satisfies both S-FOOT and KEEP WEIGHT. Changing the syllable to light will violate both S-FOOT and KEEP WEIGHT, as shown in (33b). Thus, a Shanghai syllable and a full Mandarin syllable are both heavy when occurring alone.

Let us now consider tone. First consider Shanghai. (35) shows the syllable [se] 'three', which has an underlying H tone.

(35) Tonal analysis of an H syllable in Shanghai

	H /se/ 'three'	T-DISTINCTION	POLARITY	SIMPLE TONE
a.	H \ see ø x		*!	
b. ✓	HL | | se e ø x			*
c.	H L \ | see ø x		*!	

In (35a) there is no tone change, and POLARITY is violated. In (35b) H changes to HL, and SIMPLE TONE is violated. In (35c) the polarity tone is linked to the zero syllable and so it is not realized, and POLARITY is still violated. Since POLARITY ranks higher than SIMPLE TONE, (35b) is better. The fact that (35b) is better than (35c) shows that a tone linked to a zero syllable cannot satisfy POLARITY. Next consider [sz] 'four', which is underlyingly L. Since the polarity tone cannot be linked to a zero syllable, I omit that candidate in the illustration below.

(36) Tonal analysis of an L syllable in Shanghai

	L /sz/ 'four	T-DISTINCTION	POLARITY	SIMPLE TONE
a.	L | szz ø x		*!	
b. ✓	LH | | sz z ø x			*

Again, if there is no tone change, POLARITY is violated. If L changes to LH, SIMPLE TONE is violated. Thus, (36b) is predicted to be better, correctly.

We have seen that in Shanghai, the polarity tone can be realized on the monosyllable. This is because Shanghai has just two tonal categories (H and L), so linking a polarity tone to the monosyllable does not violate T-DISTINCTION. In particular, H plus a polarity L makes HL, which is still distinct from L, and L plus a polarity H makes LH, which is still distinct from H. In Mandarin, however, because there are more tonal categories, linking a polarity tone to a monosyllable will violate T-DISTINCTION, to be seen immediately.

Mandarin has four tonal categories on full syllables: H, LH, L, and HL. The analysis of LH and HL are shown in (37) and (38).

(37) Tonal analysis of an LH syllable in Mandarin

	LH /pan/ ‘dish’	T-DISTINCTION	POLARITY	SIMPLE TONE
a. ✓	LH | | pan ø x			*
b.	L | pan ø x	*!	*	

(38) Tonal analysis of an HL syllable in Mandarin

	HL /sii/ ‘four’	T-DISTINCTION	POLARITY	SIMPLE TONE
a. ✓	HL | | si i ø x			*
b.	H | si i ø x	*!	*	

In each case, the input tone satisfies POLARITY but violates SIMPLE TONE. To satisfy SIMPLE TONE, the higher ranked T-DISTINCTION and POLARITY will be violated. Thus, the best solution is to make no change. Next consider a syllable with H, shown in (39).

(39) Tonal analysis of an H syllable in Mandarin

	H /san/ ‘three’	T-DISTINCTION	POLARITY	SIMPLE TONE
a. ✓	H | san ø x		*	
b.	HL | | sa n ø x	*!		*

If there is no tone change, POLARITY is violated. If H changes to HL, T-DISTINCTION and SIMPLE TONE are violated. It is interesting to note that changing H to HL does not violate T-DISTINCTION in Shanghai, as shown in (35), but does in Mandarin. The reason is that Shanghai has only two tone categories, H and L, whereby HL is still distinct from L. In contrast, Mandarin has four tone categories, H, LH, L, and HL, and for H to change to HL will neutralize the two and violate T-DISTINCTION. Finally, let us consider the Mandarin category L, shown in (40).

(40) Tonal analysis of an L syllable in Mandarin

	L /wuu/ ‘five’	T-DISTINCTION	POLARITY	SIMPLE TONE
a. ✓	L | wuu ø x		*	
b.	LH | | wuu ø x	*!		*

Again, L cannot change to LH since LH is another category in Mandarin. Thus, L is predicted to be a better form. However, as mentioned in (26), a monosyllabic Third tone in Mandarin usually surfaces as 214, which ends in a H. Therefore, the analysis in (40) is not quite correct.

It is well-known, however, that 214 is longer than other full syllables. For this reason, some researchers consider 214 to have three moras (e.g., Woo (1969), Shih (1997)). In addition, as Chao (1933, p. 132) points out, a monosyllabic Third tone "often breaks into two syllables," with a glottal stop appearing in between, as shown in (41), transcribed in phonetic symbols. (This fact has been reconfirmed by several linguists in a recent discussion on the e-mail list *Chinese*, April–May, 1998.)

(41) Mandarin Third tone as two syllables (Chao 1933, 132)

L		L H		L		L H	
xɑʊ	→	xɑɑ-ʔʊ		nii	→	nii-ʔi	
'good'				'you'			

If the 214 form of the Third tone is indeed a heavy-light disyllable, then it satisfies all the constraints in the present analysis and is predicted to be good. This is shown in (42).

(42) Tonal analysis of a 214 syllable in Mandarin

	L /wuu/ 'five'	T-DISTINCTION	POLARITY	SIMPLE TONE
a.	L H \ | wuu ø x		*!	
b. ✓	L H \ | wuu ʔu x			
c.	LH || wuu ʔu x	*!		*

The occurrence of the extra syllable has good motivations. It can help satisfy S-FOOT and POLARITY at the same time. It is less clear though why the Mandarin H monosyllable is not realized in a similar way, i.e., as a heavy-light disyllable H-L. A possible answer is as follows. First, there is a constraint against a surface syllable that does not come from an underlying one. Let us call it FAITH-SYLLABLE (a Faithfulness constraint, cf. McCarthy and Prince (1995)). Second, POLARITY is made of two parts; let

us call it H-POLARITY (H → HL) and L-POLARITY (L → LH). Third, FAITH-SYLLABLE is ranked below L-POLARITY and above H-POLARITY. This is shown in (43).

(43) a. L-POLARITY: An initial L must be followed by H at surface.
b. H-POLARITY: An initial H must be followed by L at surface.
c. FAITH-SYLLABLE: Surface syllables must correspond to underlying syllables.
L-POLARITY ≫ FAITH-SYLLABLE ≫ H-POLARITY[10]

Since the POLARITY constraints are still ranked below T-DISTINCTION and above SIMPLE TONE, the revision does not affect other examples discussed earlier. The reanalysis of the H and L syllables in Mandarin are shown in (44) and (45), where s indicates an overt syllable, ø indicates a zero syllable, and a hyphen indicates a syllable boundary.

(44) Tonal analysis of a H syllable in Mandarin

	/H/ s	T-DISTINCT.	L-POLAR.	FAITH-SYLL.	H-POLAR.
a.	HL s-ø	*!			
b. ✓	H s-ø				*
c.	H-L s-s			*!	

(45) Tonal analysis of a L syllable in Mandarin

	/L/ s	T-DISTINCT.	L-POLAR.	FAITH-SYLL.	H-POLAR.
a.	LH s-ø	*!			
b.	L s-ø		*!		
c. ✓	L-H s-s			*	

In both cases, S-FOOT requires the output to be disyllabic. T-DISTINCTION prevents the input syllable from changing its tone. Because FAITH-SYLLABLE

is ranked above H-POLARITY, the H syllable cannot be followed by an added overt syllable. However, because FAITH-SYLLABLE is ranked below L-POLARITY, the L syllable can be followed by an added overt syllable. Once again, the 214 form of the Third tone can be accounted for with the same constraints that apply to both Mandarin and Shanghai. There is no need to posit a floating H for the Third tone.

A reviewer points out that in expressions like [Lao35 Li214] 'Old Li', where [Li214] is already disyllabic in the present analysis, there should be no metrical or tonal reason to add [Lao]. Why then is it still added? The answer is as follows. [Li] is [214] only in final position. In non-final positions it is [21] or [11], which is monosyllabic, to which [Lao] must be added, as in [Lao35 Li11 he35 wo214] 'Old Li and I'. The fact that [Lao] is added in all positions, including final position, can be attributed to a general requirement to avoid variation of the same expression (cf. the ANTI-ALLMORPH constraint of Burzio (1996) and the UNIFORM EXPONENCE constraint of Kenstowicz (1996)).

4.3. *Trisyllabic and Longer Expressions*

Mandarin and Shanghai also differ in trisyllabic and longer expressions. In Mandarin, all full sllables basically keep their underlying tones, as shown in (46). In Shanghai, however, syllables beyond the second can become toneless (or get default L), as shown in (47). (For the formation of metrical and tonal domains in multisyllabic expressions in Shanghai, see Duanmu (1995).) The underlying tones of a foreign name are the tones of the syllables used to translate it. In Mandarin, they are often the same as the surface tones. In Shanghai, the underlying tones can surface in hyper-articulated speech, where each syllable is stressed and forms a tone domain.

(46) Mandarin (transcribed in Pinyin)

H - H - H → H-H-H
zhi-jia-ge
'Chicago'

(47) Shanghai (transcribed in phonetic symbols)

H - H - H → H-L-L
tsz-ka-ku
'Chicago'

In (46) there are three full syllables, which have two possible metrical structures, shown in (48).

(48) a.
```
  x
 (x   x)  x
(mm) (mm) (mm)
```
b.
```
  x          x
 (x   x)  (x   .)
(mm) (mm) (mm) ø
```

In (48a) there is one S-FOOT, and in (48b) there are two. In either case the three full syllables are stressed, so they must all keep their tonal categories. In (47), the three syllables form one S-FOOT, as shown in (49).

(49)
```
  x
 (x   .) .
(mm) m m
```

The initial syllable is stressed, but the latter two are not. The tonal analysis is given in (50), where S indicates a stressed syllable, s indicates an unstressed syllable, and ø represents tonelessness or default L. Irrelevant constraints are omitted.

(50) Tonal analysis of (47) in Shanghai

	/H-H-H/	T-DISTINCTION	H-POLARITY
a.	H-H-H S - s - s		*
b. ✓	H-L-H S - s - s		
c. ✓	H-L-ø S - s - s		

The analysis predicts that H-L-H and H-L-ø are both good. However, only H-L-ø is in fact good. Intuitively, H-L-H violates a well-known generalization in Chinese that unstressed syllables should become toneless. To rule out H-L-H, one can postulate the constraint in (51), ranked below H-POLARITY and SIMPLE TONE.

(51) TONELESS: Unstressed syllables should be toneless.
H-POLARITY ≫ SIMPLE TONE ≫ TONELESS

Since TONELESS is ranked at the bottom of the constraints discussed so far, it will not affect previous examples. The revised tonal analysis of (47) is given in (52).

(52) Tonal analysis of (47) in Shanghai

	/H-H-H/	T-DISTINCT.	H-POLAR.	SIMPLE TONE	TONELESS
a.	H-H-H S-s-s		*!		**
b.	H-L-H S-s-s				**!
c. ✓	H-L-ø S-s-s				*
d.	HL-ø-ø S - s-s			*!	
e.	H-ø-ø S-s-s		*!		

To satisfy TONELESS, unstressed syllables must be toneless. To satisfy H-POLARITY, the second syllable must be L. To satisfy SIMPLE TONE, the initial syllable cannot be HL. (52c) and (52d) show that SIMPLE TONE is more important than TONELESS. (52c) and (52e) show that H-POLARITY is more important than TONELESS. The best form is predicted to be (52c), correctly.

4.4. *Summary of Constraints*

I have shown that stress and tone in Mandarin and Shanghai can be analyzed in the same way. The ranked constraints discussed so far are given in (53).

(53) a. Metrical constraints
S-FOOT ≫ KEEP WEIGHT

b. Tonal constraints
T-DISTINCTION ≫ L-POLARITY ≫ FAITH-SYLLABLE ≫ H-POLARITY ≫ SIMPLE TONE ≫ TONELESS

The difference between the two dialects comes from the fact that full Mandarin syllables are underlyingly heavy but all Shanghai syllables are underlyingly unspecified for weight, and that Shanghai has just two tonal categories but Mandarin has four.

5. Further Issues

5.1. *The Third Tone Sandhi (T3S) in Mandarin*

In the preceding discussion, T-DISTINCTION is ranked high. This does not mean that it cannot be violated. As a reviewer points out, Mandarin has a well-known rule, Third Tone Sandhi (T3S), by which a Third tone (T3) changes to a Second tone (T2) before another T3, as shown in (54).

(54) Mandarin T3S: T3 → T2/_T3

Some researchers have suggested that the changed T3 is not entirely the same as a T2, but the perception study of Wang and Li (1967) shows that they are identical for the listener. Since T3S changes T3 to T2, it violates T-DISTINCTION.

In Optimality Theory, T3S can be translated into a number of smaller constraints, but for simplicity I will present it as one constraint. T3S must rank above T-DISTINCTION, as shown in (55) and illustrated in (56).

(55) T3S ≫ T-DISTINCTION

(56) Analysis of T3S

	/L-L/	T3S	T-DISTINCTION	H-POLARITY
a.	L-L	*		*!
b. ✓	LH-L		*	

T3S can be seen as a dissimilation between two low tones, but it also shows a peculiar behavior of Mandarin. For example, there is no reason why T3-T3 must change to T2-T3, instead of T3-T2, since both T2-T3 and T3-T2 are good. Second Tone Sandhi, by which T2 becomes T1 when it is after T1 or T2 and before any full tone (see Chao (1968)), is another special property of Mandarin. The fact that Mandarin (or Shanghai) has some peculiar behavior of its own should not be a problem. Any analysis must make such a provision. What the present paper claims is that all languages ar fundamentally the same and that there are more similarities than previously thought.

5.2. *Tone Sandhi in Old Shanghai*

Besides Mainstream Shanghai, there is another major variety, which has been called Old Shanghai (Shen (1981a, 1982), Xu et al. (1988)). According

to Xu et al. (1988), speakers of Old Shanghai were mostly people over 60 years old who used to live in the old town of Shanghai (presently the Nashi District of Shanghai City). By now the surviving speakers should be over 70 years old.

A main difference between Old Shanghai and Mainstream Shanghai lies in tone sandhi. First, Old Shanghai has more tone patterns, including contour tones on the initial and final syllables. Second, whereas in Mainstream Shanghai the tone pattern is basically determined by the initial syllable alone, in Old Shanghai it is sensitive to both initial and non-initial syllables. Third, tone patterns in Old Shanghai are not very stable (see below). To understand Old Shanghai, it is necessary to discuss historical tonal categories. Chinese has four historical tones, Ping, Shang, Qu, and Ru. Each is further divided into two according to the voicing of the onset consonant (sonorant onsets can go either way). This is shown in (57), where the categories are numbered from I though VIII.

(57) Historical tones in Chinese

	Ping	Shang	Qu	Ru
Voiceless onset	I	III	V	VII
Voiced onset	II	IV	VI	VIII

In Old Shanghai, historical categories II and IV have merged with others. For some speakers III is still a separate category, but it is unstable and often merges with V (Shen (1981b, p. 280) and Xu et al. (1988, pp. 56–57)). Thus, Shen (1981a) gives five tonal categories on isolated syllables, shown in (58). The phonological interpretation is based on the pitch values of both isolation tones and tone patterns in multisyllabic expressions.

(58) Tonal categories in Old Shanghai

Categories	I	V	VI	VII	VIII
Interpretation	HL	LH	LH	LH	LH

Although four of the categories are LH, they differ in other aspects. V and VI differ in tonal register, whereby V is a high rise (with clear voice) and VI a low rise (with murmured voice). VII and VIII also differ in tonal register. In addition, VII and VIII occur on glottalized rimes, whereas V and VI occur on non-glottalized rimes. The five categories in (58) are similar to those in Mainstream Shanghai and can be reduced to two underlyingly, H for I and L for V–VIII.

For simplicity, let us ignore the glottalized categories VII and VIII. In addition, let us focus on trisyllabic domains. Shen (1982) gives the patterns in (59), where S1 means the first syllable, S2 means the second, and S3

means the third. The tones of the initial syllable when it is read in isolation are given in parentheses.

(59)	Initial category	Trisyllabic	Condition
	I (HL)	a. H-H-HL	S3 is I or II
		b. H-H-L	S3 is II, III, IV, or V
		c. H-L-L	S2 is II, III, IV, or V
	V (LH)	a. LH-H-HL	S3 is I or II
		b. LH-H-L	S3 is III, IV, V, or VI
	VI(LH)	a. L-H-H	S1 is II
		b. LH-H-HL	S1 is IV or VI
		c. LH-H-L	S1 is IV or VI, S3 is V or VI

The condition for a given pattern is based on Shen (1982, p. 101) and Xu et al. (1988, pp. 60–61), which do not agree in all details. I will return to this point shortly.

The patterns in (59) are variable. According to Shen (1981b, p. 280, 1982, p. 108), an initial LH can be read as L (although an initial L cannot be read as LH). In addition, Shen (1981a, p. 143) points out that final HL can be read as H – the HL pattern carries some emphasis. Under these considerations, (59) can be simplified to (60).

(60)	Initial syllable	Trisyllabic	
	I (HL)	a. H-H-H	
		b. H-H-L	
		c. H-L-L	(same as Mainstream Shanghai)
	V (LH)	a. L-H-H	
		b. L-H-L	(same as Mainstream Shanghai)
	VI (LH)	a. L-H-H	
		b. L-H-H	
		c. L-H-L	(same as Mainstream Shanghai)

The patterns now look rather similar to those in Mainstream Shanghai. The only big difference is that in Old Shanghai H can spread to other syllables, whereas in Mainstream Shanghai H is linked to just one syllable. The alternation in H-spreading is not unusual but has been observed in African languages (cf. Pulleyblank (1986)). The spreading of H (and its optional linking to the initial syllable to create a LH contour there) in Old Shanghai requires some modification of the analysis. In particular, H-polarity, SIMPLE TONE and TONELESS must be made sensitive to the categories of non-initial tones in Old Shanghai. However, to account for

the patterns where H does not spread (i.e., those similar to Mainstream Shanghai), these constraints are still needed.

Let us now ask a moe fundamental question. Why are tone patterns more variable in Old Shanghai? And why should tone patterns be sensitive to historical categories, especially those that do not otherwise contrast with each other? For example, consider trisyllabic patterns starting with VI, repeated in (61).

(61)	VI (LH)	a. L-H-H	S1 is II
		b. LH-H-HL	S1 is IV or VI
		c. LH-H-L	S1 is IV or VI, S3 is V or VI

Since II has already merged with VI in isolation, why should it give rise to a different trisyllabic pattern in (61a)? In addition, since S3 loses its underlying tone anyway, why should it affect the tone pattern in (61c)? Following Duanmu (1993), I suggest that Old Shanghai is in a state of transition from a dialect whose full syllables are heavy (like Mandarin) to a dialect whose syllables have unspecified weight (like Mainstream Shanghai). The rimes of Old Shanghai have already been simplified (no diphthongs, no contrastive codas; see Duanmu (1993)), so none needs to be specified as heavy. However, like a dialect with heavy syllables, where most syllable tones have an effect on the surface pattern, non-initial syllables in Old Shanghai still show some remnant influence on the surface pattern. It is natural that even within Old Shanghai there is much variation, as reflected in the optional patterns noted above and in the discrepancies between the reports of Shen (1981a, 1982) and Xu et al. (1988). It is also natural that in a short course of a few decades, Mainstream Shanghai has become the dominant variety.

6. Conclusions

I have offered an analysis of the contrast in tonal behavior between two types of Chinese dialects, represented by Mandarin and Shanghai. I have argued that the two dialects differ mainly in whether their syllables are specified for weight underlyingly and in the number of tonal categories they have; otherwise they obey similar constraints, summarized in (53). The present analysis solves several problems in Duanmu (1993) and offers a better prediction of rime duration in Mandarin and Shanghai.

The central proposal is the notion of S-FOOT, introduced in (19), by which Chinese is both mora-counting and syllable-counting. In particular, a heavy syllable always forms a moraic trochee (M-FOOT), but a minimal word must be a syllabic trochee (S-FOOT). Since a stressed syllable must be heavy,

an S-FOOT always contains at least one M-FOOT. The mora-counting part explains why a heavy syllable always has stress and can carry two tones (LH or HL), even in the second position of a syllabic trochee. The syllable-counting part explains why a minimal Chinese expression must be disyllabic, which cannot be satisfied by a heavy syllable alone. This article also proposes the notion of tonal polarity, independently found in African languages (cf. Newman (1997)), which simplifies underlying tones in Shanghai and Mandarin and avoids a problem in tone deletion.

Theoretically, the notion S-FOOT clarifies an ambiguity in phonology. McCawley (1978) proposes a parameter by which languages are either mora-counting or syllable-counting. However, languages like Estonian, in which feet seem to count syllables, yet monosyllables must be heavy, pose a problem. Prince (1980) suggests that Estonian can count either syllables or moras. The idea is later reformulated as the generalized trochee (Kager (1992b), Hayes (1995)), by which the default foot is a syllabic trochee, yet a monosyllable can form a moraic trochee. The present proposal differs from the notion of the generalized trochee in that a minimal word (S-FOOT) has two levels of structure, a moraic level and a syllabic level, whereas in the generalized trochee a minimal word has only one level, either a moraic trochee or a syllabic trochee but not both. For example, in the generalized trochee a heavy monosyllable is a bimoraic foot, which is also a minimal word. In the present analysis a bimoraic foot is insufficient to be a minimal word but must be followed by a zero syllable. Similarly, in the generalized trochee a heavy-heavy disyllable forms one foot, in which only the first syllable has stress. In the present analysis, a heavy-heavy disyllable forms an S-FOOT that is made of two M-FEET, whereby both syllables are stressed.

The present analysis also has other implications. First, it shows that languages (or dialects) may appear quite different when they in fact obey similar linguistic principles. Second, it shows that the tone-bearing unit may appear to be the syllable when it is actually the mora. In particular, the constraint SIMPLE TONE can prevent a heavy syllable from taking two tones (as with the initial syllable in Shanghai), even though the syllable has two tone-bearing units. But when T-DISTINCTION or POLARITY overrides SIMPLE TONE, contour tones show up on heavy syllables. Third, the present analysis shows that a language may appear to be quantity-insensitive when it is in fact quantity-sensitive. For example, Shanghai does not have a phonemic contrast between heavy and light syllables, but its surface syllables do show a difference in quantity, which is predicted by metrical structure. This supports the idea that all languages are quantity-sensitive (Kager (1992a)). Fourth, the present analysis offers a more specific definition of stress, by

which there are three degrees in a minimal word: S-FOOT stress (associated with a disyllabic trochee), M-FOOT stress (associated with a heavy syllable), and lack of stress (associated with a light syllable). This differs from a popular view, such as Kenyon and Knott (1944) and Gimson (1979), according to which stress (or accent) is a relative notion, in that a syllable has stress if it is stronger than another in the same word. Thus, a monosyllable like *yes* has no stress since there is no other syllable to compare with. The final syllable in *bungalow* has secondary stress since it is stronger than the second syllable but weaker than the first. And the final syllable in *yellow* has no stress since it is the weakest in the word. In the present analysis, *yes* has as much stress as the first syllable in *yellow* or *bungalow* (S-FOOT stress), and the final syllables in *bungalow* and *yellow* have the same amount of stress (M-FOOT stress).[11] Fifth, the present analysis supports a simple theory of the relation between weight and stress, namely, all heavy syllables are stressed (Prince (1990)), and all stressed syllables are heavy (Hammond (1997)). Finally, the present analysis offers an intuitive metrical theory. Since all languages use segments (mora-sized units) and syllables, it is only natural that they count both moras and syllables in metrical structure.

NOTES

* Some ideas proposed here were presented in 1997 to audiences at University of Michigan and at NACCL9, University of Victoria, Canada. I thank the participants for their comments. In addition, I thank three anonymous reviewers for their comments, which have improved the quality and clarity of this article.

[1] The underlying tone of a syllable is usually (but not always) the tone that occurs on a syllable when it is read in isolation. It is also called the syllable tone.

[2] A reviewer points out that markedness requires that if a language has heavy syllables, it must also have light syllables. Mandarin does have light syllables, which are mostly grammatical particles, whereas heavy syllables are mostly content words. This is also true in English, where content words are either a heavy syllable or longer. In Shanghai, a syllable will also be heavy when it occurs alone, to be discussed below.

[3] The tonal domain here refers to what Duanmu (1993) calls the association domain, which is the domain of tone deletion and spreading. It does not include the domain of dissimilation, which can take place across stress domains.

[4] Zhu measured the duration of tones. Since all syllables in Zhu's word list began with oral stops, the tone duration was close to rime duration.

[5] Although Burzio (1994) and Hammond (1997) agree that the stressed syllable in disyllabic words is always heavy, they differ with regard to longer words. For example, in *metrify*, Burzio considers the stressed initial syllable to be light, but Hammond considers it to be heavy.

A reviewer points out that trisyllabic shortening in English is a problem for the claim that all stressed syllables are heavy. According to trisyllabic shortening, *metrify* is derived from *meter*, where the first vowel is underlyingly long and is shortened in a trisyllabic word. If the initial syllable is heavy in both words, the reason for the shortening requires

an explanation. Hammond's answer is that the Great Vowel Shift and trisyllabic shortening are independent processes, which take place before consonant gemination after a stressed lax vowel.

[6] A reviewer rightly points out that interpreting phonetic data can be complicated. While the present analysis points to a relation between weight and duration, namely, in read speech heavy rimes are about 200 ms, and light rimes are about 100 ms, it does not exclude possible adjustment under other conditions, such as very slow or ver fast speech, or the influence of vowel height on duration. The reviewer also rightly points out that duration is not the only cue for stress. However, duration is a major cue. According to Fry (1958), the most important cue for stress is F0 contour, followed by duration. According to Lin (1985), the main cue for stress in Chinese is duration, probably because F0 is taken up by tone.

[7] As Newman (1997) points out, some polarity cases are probably due to tonal dissimilation, and some are "true polarity." The Chinese cases appear to be the latter. However, since the surface result is the same, the distinction does not affect the present analysis. In the languages Newman cited, polarity can occur at the beginning and the end of an expression. Since no metrical structure was discussed, it is not clear whether polarity was related to S-foot in those languages.

[8] A reviewer points out that some surface light syllables in Mandarin may still have underlying tones. For example, [jie] in [xiao-jie] 'young lady' must have the Third tone underlyingly since it triggers tone sandhi in [xiao]. There are two solutions. First, as the reviewer suggests, [jie] is underlyingly stressed and carries the Third tone, but after tone sandhi it loses its stress and tone. This is the standard traditional analysis. Second, [jie] remains stressed at surface and carries the Third tone L. The question for the second solution is why [jie] cannot be 214 since a final Third tone is supposed to be 214. I do not have a good answer at this point, except pointing out that not all speakers use 214 for a final Third tone, a fact confirmed by a survey on the e-mail list *Chinese*.

[9] It will be noted that (19) requires a minimal word to be a disyllabic foot, but it does not require all syllables to be in a foot. In *to(day ø)* the first syllable is not in a foot. When *today* follows a stressed syllable, *to-* can form an S-FOOT with it, such as *(go to)(day ø)*.

[10] If we consider tonelessness to be non-H, then (H ø) in a sense satisfies H-POLARITY. In contrast, (L ø) does not satisfy L-POLARITY. Under this light, the ranking of L-POLARITY above H-POLARITY is not accidental.

[11] Three degrees of stress seem to be the necessary minimum for languages like English and Mandarin. It is not clear whether three is also the maximal limit or whether there is a maximal limit. See Gussenhoven (1991) for arguments for a maximal limit.

REFERENCES

Abercrombie, David (1967) *Elements of General Phonetics*, Aldine, Chicago.

Burzio, Luigi (1994) *Principles of English Stress*, Cambridge University Press, Cambridge.

Burzio, Luigi (1996) "Surface Constraints versus Underlying Representation," in J. Durand and B. Laks (eds.), *Current Trends in Phonology: Models and Methods*, European Studies Research Institute, University of Salford Publications, pp. 123–141.

Chao, Yuen-Ren (1930) "A System of Tone Letters," *Le Maître Phonétique* 45, 24–27.

Chao, Yuen-Ren (1931) "Fanqie yu ba zhong" [Eight types of Fanqie languages], *Bulletin of the Institute of History and Philology*, Academia Sinica 2.3, 312–354.

Chao, Yuen-Ren (1933) "Tone and Intonation in Chinese," *Bulletin of the Institute of History and Philology*, Academia Sinica 4.2, 121–134.

Chao, Yuen-Ren (1968) *A Grammar of Spoken Chinese*, University of California Press, Berkeley.

Cheng, Chin-Chuan (1973) *A Synchronic Phonology of Mandarin Chinese*, Monographs on Linguistic Analysis no. 4, Mouton, The Hague.

Duanmu, San (1993) "Rime Length, Stress, and Association Domains," *Journal of East Asian Linguistics* 2, 1–44.
Duanmu, San (1995) "Metrical and Tonal Phonology of Compounds in Two Chinese Dialects," *Language* 71, 225–259.
Duanmu, San (in press) "Word Stress in Mandarin Chinese," in D. Xu (ed.), *Studies in Chinese Phonology in the 90's*, China Social Sciences Publishing House, Beijing.
Fry, Dennis B. (1958) "Experiments in the Perception of Stress," *Language and Speech* 1, 126–152.
Giegerich, Heinz J. (1985) *Metrical Phonology and Phonological Structure: German and English*, Cambridge University Press, Cambridge.
Gimson, A.C (1979) "The Pronunciation of English," in P. Hanks (ed.), *Collins Dictionary of the English Language*, Collins, London, pp. xix–xxi.
Gussenhoven, Carlos (1991) "The English Rhythm Rule as an Accent Deletion Rule," *Phonology* 8, 1–35.
Halle, Morris, and Jean-Roger Vergnaud (1987) *An Essay on Stress*, MIT Press, Cambridge.
Hammond, Michael (1997) "Vowel Quantity and Syllabification in English," *Language* 73, 1–17.
Hayes, Bruce (1995) *Metrical Stress Theory: Principles and Case Studies*, University of Chicago Press, Chicago.
Hogg, Richard, and C.B. McCully (1987) *Metrical Phonology: A Coursebook*, Cambridge University Press, Cambridge.
Kager, René (1992a) "Are There Any Truly Quantity-insensitive Systems?" *BLS 18*, Proceedings of the 18th Annual Meeting of the berkeley Linguistics Society: General Session and Parasession on the Place of Morphology in a Grammar, Berkeley, 1991, pp. 123–132.
Kager, René (1992b) "Shapes of the Generalized Trochee," *WCCFL* 11, 298–312.
Kager, René (1993) "Alternatives to the Iambic-Trochee Law," *Natural Language and Linguistic Theory* 11.3, 381–432.
Kenstowicz, Michael (1995) "Cyclic vs. Noncyclic Constraint Evaluation," *Phonology* 12, 397–436.
Kenstowicz, Michael (1996) "Base Identity and Uniform Exponence: Alternatives to Cyclicity," in J. Durand and B. Laks (eds.), *Current Trends in Phonology: Models and Methods*, vol. 1, European Studies Research Institute, University of Salford Publications, pp. 363–393.
Kenyon, John Samuel, and Thomas Albert Knott (1944) *A Pronouncing Dictionary of American English*, Merriam, Springfield, Massachusetts.
Klatt, Dennis H. (1975) "Vowel Lengthening is Syntactically Determined in a Connected Discourse," *Journal of Phonetics* 3.3, 129–140.
Leben, William (1971) "Suprasegmental and Segmental Representation of Tone," *Studies in African Linguistics*, Supp. 2, 183–200.
Liberman, Mark (1975) *The Intonational System of English*, PhD dissertation, MIT.
Lin, Hua (1994) "Against Domain-final Stress in Mandarin," paper presented at the Sixth North American Conference on Chinese Linguistics, University of Southern California, Los Angeles.
Lin, Maochan, and Jingzhu Yan (1988) "The Characteristic Features of the Final Reduction in the Neutral-tone Syllable of Beijing Mandarin," *Phonetic Laboratory Annual Report of Phonetic Research*, 37–51, Phonetic Laboratory, Institute of Linguistics, Chinese Academy of Social Sciences, Beijing.
Lin, Maochan, Jing-zhu Yan, and Guo-hua Sun (1984) "Beijinghua liangzizu zhengchang zhongyin de chubu shiyan [Preliminary Experiments on the Normal Stress in Beijing Disyllables]," *Fangyan* 1984.1, 57–73.
Lin, Tao (1985) "Tantao Beijing hua qingsheng xingzhi de chubu shiyan [Preliminary Experiments in the Exploration of the Nature of Mandarin Neutral Tone]," in Tao Lin

and Lijia Wang (eds.), *Beijing yuyin shiyan lu* [Peking University Working Papers in Phonetics], Beijing University Press, Beijing, pp. 1–26.
Lu, Zhiji (1987) "Shanghai Tones: A Nonlinear Analysis," *Studies in the Linguistic Sciences* 17, 93–113.
McCarthy, John, and Alan Prince (1986) *Prosodic Morphology*, ms., University of Massachusetts, Amherst and Brandeis University.
McCarthy, John, and Alan Prince (1995) "Faithfulness and Reduplicative Identity," in J. Beckman, L. Walsh Dickey and S. Urbanczyk (eds.), *Papers in Optimality Theory*, University of Massachusetts Occasional Papers in Linguistics 18, 249–384. (Available as ROA 60.)
McCawley, James D. (1978) "What Is a Tone Language?" in V. A. Fromkin (ed.), *Tone: A Linguistic Survey*, Academic Press, New York, pp. 113–131.
Milliken, Stuart (1989) "Why There is No Third Tone Sandhi Rule in Standard Mandarin," paper presented at the Tianjin International Conference on Phonetics and Phonology, Tianjin Normal University.
Newman, Paul (1997) "Hausa Tonology: Complexities in an 'Easy' Tone Language," in J. Goldsmith (ed.), *The Handbook of Phonological Theory*, Blackwell, Cambridge, Massachusetts, pp. 762–781.
Prince, Alan (1980) "A Metrical Theory for Estonian Quantity," *Linguistic Inquiry* 11, 511–562.
Prince, Alan (1990) "Quantitative Consequences of Rhythmic Organization," *CLS 26*, Papers from the 26th Regional Meeting of the Chicago Linguistic Society, Volume 2: Parasession on the Syllable in Phonetics and Phonology, Chicago Linguistic Society, 1992, pp. 355–398.
Prince, Alan, and Paul Smolensky (1993) *Optimality Theory: Constraint Interaction in Generative Grammar*, ms., Rutgers University and University of Colorado.
Pulleyblank, Douglas (1986) *Tone in Lexical Phonology*, Reidel, Dordrecht.
Selkirk, Elisabeth (1980) "The Role of Prosodic Categories in English Word Stress," *Linguistic Inquiry* 11, 563–605.
Selkirk, Elisabeth (1984) *Phonology and Syntax: The Relation between Sound and Structure*, MIT Press, Cambridge.
Selkirk, Elisabeth, and Tong Shen (1990) "Prosodic Domains in Shanghai Chinese," in S. Inkelas and D. Zec (eds.), *The Phonology-Syntax Connection*, CSLI, Stanford University, pp. 313–337. Distributed by University of Chicago Press.
Shen, Tong (1981a) "Lao pai Shanghai fangyan de liandu biandiao [Tone Sandhi in Old Shanghai]," *Fangyan* 1981.2, 131–144.
Shen, Tong (1981b) "Shanghai hua laopai xinpai de chabie [Differences between Old and New Shanghai Dialects]," *Fangyan* 1981.4, 275–283.
Shen, Tong (1982) "Lao pai Shanghai fangyan de liandu biandiao [2] [Tone Sandhi in Old Shanghai, Part 2]," *Fangyan* 1982.2, 100–114.
Shih, Chi-lin (1997) "Mandarin Third Tone sandhi and Prosodic Structure," in Wang Jialing and N. Smith (eds.), *Studies in Chinese Phonology*, Mouton de Gruyter, Berlin, pp. 81–123.
Wang, Jialing (1997) "The Representation of the Neutral Tone in Chinese Putonghua," in Wang Jialing and N. Smith (eds.), *Studies in Chinese Phonology*, Mouton de Gruyter, Berlin, pp. 157–183.
Wang, Jing, and Lijia Wang (1993) "Putonghua duo yinjie ci yinjie shi chang fenbu moshi [The Types of Relative Lengths of Syllables in Polysyllabic Words in Putonghua]," *Zhongguo Yuwen* 1993.2 (233), 112–116.
Wang, Li (1979) "Xiandai Hanyu yuyin fenxi zhong de ji ge wenti [Some Problems in the Phonetic Analysis of Modern Chinese]," *Zhongguo Yuwen* 1979.4, 281–286.
Wang, William S-Y, and Kung-pu Li (1967) "Tone 3 in Pekinese," *Journal of the Speech and Hearing Research* 10.3, 629–36.
Woo, Nancy (1969) *Prosody and Phonology*, PhD dissertation, MIT.

Xu, Baohua, Zhenzhu Tang, and Nairong Qian (1981) "Xin pai Shanghai fangyan de liandu biandiao [Tone Sandhi in New Shanghai]," *Fangyan* 1981.2, 145–155.
Xu, Baohua, Zhenzhu Tang, Rujie You, Nairong Qian, Rujie Shi, and Yaming Shen (1988) *Shanghai Shiqü Fangyan Zhi* [Urban Shanghai Dialects], Shanghai Jiaoyu Chubanshe, Shanghai.
Yan, Jingzhu, and Maocan Lin (1988) "Beijinghua sanzizu zhongyin de shengxue biaoixian [Acoustic Characteristics of the Stress in Beijing Trisyllables]," *Fanyan* 1988.3, 227–237.
Yang, Shun-an (1992) "Beijinghua duoyinjie zuhe yunlu tezheng de shiyan jianjiu [An Experiment on the Prosody of Polysyllables in the Beijing Dialect]," *Fangyan* 1992.2, 128–137.
Yip, Moira (1980) *The Tonal Phonology of Chinese*, PhD dissertation, MIT.
Zee, Eric, and Ian Maddieson (1979) "Tones and Tone Sandhi in Shanghai: Phonetic Evidence and Phonological Analysis," *UCLA Working Papers in Phonetics* 45, 93–129.
Zhu, Xiaonong (1995) *Shanghai Tonetics*. PhD dissertation, Australian National University, Canberra.

Received: September 2, 1997
Revised: May 2, 1998

Program in Linguistics
University of Michigan
Ann Arbor, MI 48109-1285
U.S.A.
E-mail: duanmu@umich.edu