Occurrence Frequency and Transition Probability of the Chinese Four Tones

Shizuo Hiki[1], Kazuko Sunaoka[2], Liming Yang[3] and Yasuyo Tokuhiro[4]

[1] School of Human Sciences, [2] School of Political Science and Economics,
[3] Institute of Language Teaching, Waseda University, Japan
{hiki; ksunaoka; youl}@waseda.jp
[4] Graduate School of Japanese Applied Linguistics, Waseda University, Japan
tokuhiro@toki.waseda.jp

Abstract

Occurrence frequency and transition probability of the Chinese four tones were analyzed statistically. The database used for the analysis was *the Grammatical Knowledge-base of Contemporary Chinese* (S-W. Yu, editor, Tsinghua University Press, China, 1998). One, two, three and four syllable words, about 55,000 words in total, were categorized into more than twenty kinds of parts of speech in this database.

The occurrence frequency of the four tones deviated from 25%, depending on number of syllables in a word or position of the syllables in the word. In nouns, Tone-2 was always around 25% but Tone-3 was 5% to 10% less. In the 1st syllables of the words with any number of syllables, Tone-1 were more than 25%, but it became less in the back position of syllables. Tone-4 was 5% more even in the 1st syllables and became as large as 40% in the last syllables.

In the total 38,968 syllables of the nouns, occurrence frequencies of Tone-1, Tone-2, Tone-3 and Tone-4 were 24%, 23%, 18% and 33%, respectively. The deviations were much larger in the cases of some of the other parts of speech such as verb, adjective, adverb and onomatopoeia.

The entropy of the four tones, which is 2 bits when the ratios are equal, decreased by less than 0.05 bit when averaged over all positions of syllables. This indicates that the four tones are used fairly effectively in Chinese language.

In the nouns, transition probability between Tone-2 and Tone-3 were less than average. Transitions beginning at Tone-1 and ending at Tone-4, or beginning at Tone-4 and ending at Tone-4 in the two, three and four syllable words were much more. Some other kind of parts of speech showed characteristic transition such as repetition of the same tone in successive syllables.

1. Database used for the statistical analysis

The database used for extracting occurrence frequency and transition probability of the Chinese four tones was *the Grammatical Knowledge-base of Contemporary Chinese* edited by Shiwen Yu of Institute of Computational Linguistics, Beijing University, and published by Tsinghua University Press, China, 1998. The database was converted into Japanese computer code by Ming Yin and Kazuko Sunaoka, under a project of the Chinese division, Cross-Cultural Distance Learning, Waseda University, Japan.

Tone symbols of the four tones of Standard Chinese were furnished to each of the syllables involved in one, two, three and four syllable words, about 55,000 words in total. About 124,000 syllables were involved in those words.

Parts of speech of the words were categorized in this database into more than twenty kinds, but following nine parts of speech were used for the present analysis:

Number of words (Items less than 20 are not shown.)

	Number of syllables in a word				
	One	Two	Three	Four	Total
Part of Speech					
Noun	325	10,039	5,511	508	16,383
Verb	1,660	12,400	415		14,479
Morpheme	7,129				7,142
Set phrase				5,254	5,252
Idiom			450	2,584	3,045
Adjective	284	2,531	39		2,856
Adverb	183	908	78		1,172
Word of state		341	420	225	986
Onomatopoeia	68	63			145

2. Occurrence frequency of the four tones

In order to make it easier to grasp the ratio among the occurrence frequency of the four kinds of tones, namely, Tone-1: high flat, Tone-2: rising, Tone-3: low flat and rising, and Tone-4: falling, the value of occurrence frequency is converted into percentage, then, plotted on a two dimensional plane along the upward, rightward, downward and leftward axes from the origin, respectively, and connected by solid lines in this order in Figure 1. As most of the values stayed in the range from 10% to 40%, 10% is taken as the origin and up to 40% was plotted on each axis, so that their differences are magnified. The case of all 25% is indicated by gray bold lines for reference.

2.1. Nouns

The number of nouns was most among the parts of speech, and they occupied about 30% of the total of words accommodated in the database. Total of the syllables involved was 38,968 syllables. By taking the noun as a typical example of the part of speech, firstly, occurrence frequency of the Chinese four tones was examined for each number of syllables in a word, and for each position of syllables in the word.

2.1.1. Position of syllable in the word

Ratio of the occurrence frequency for one, two, three and four syllable words are shown from the top in the left column of Figure 1. In each figure, the data is subdivided by their position of syllables such as the 1st, 2nd, 3rd and 4th syllable, and the plots are superposed in each figure.

The occurrence frequency of the four tones deviated from the average of 25%, depending on number of syllables of the word or position of syllable in the word.

Tone-2 was always around 25% but Tone-3 was 5% to 10% less. In the first syllable of the word with any number of syllables, Tone-1 were more than 25%, but it became less than 25% in the syllables in the back position in the word. Tone-4 was 5% more even in the first syllables and became as large as 40% in the last syllables, regardless of the number of syllables in the words.

2.1.2. Number of syllable in a word

The bottom of the left column in Figure 1 is average of all positions of syllables for each of the one, two, three and four syllable words of nouns. Although the ratios were characteristic of the position of syllables, the frontal and back syllables cancelled each other characteristics and the average became quit similar among the words with different number of syllables.

2.1.3. Average occurrence frequency

In the total 38,968 syllables of the nouns, occurrence frequencies of Tone-1, Tone-2, Tone-3 and Tone-4 were 24%, 23%, 18% and 33%, respectively. These ratios are very close to the traditionally referred ratios; “3, 3, 2 and 4,” as they become 2.88, 2.76, 2.16 and 3.96, if 25% is converted to 3.

2.2. Other parts of speech

In other kinds of parts of speech, the characteristic deviations of occurrence frequency of kind of tones were observed.

1) In the average of all syllable positions of one, two and three syllable words of verbs, Tone-2 was about half of that of noun, as shown in the top of the right column of Figure 1.

2) In the adverbs, Tone-1 and Tone-3 were less than that of the nouns, while Tone-4 was more than twice of any other tones, as shown in the second from the top.

In the one and two syllable words of adjectives, the rate of occurrence frequency was similar to that of nouns, but, Tone-1 was few in any syllable positions of the three syllable words. In the three syllable word adjectives, Tone-4 was frequent but Tone-1 was few in any position of syllables.

3) In onomatopoeias, most of the syllables were Tone-1, as shown in the third from the top.

4) In the four syllable words of idioms and set phrases, Tone-2 was more than that of the nouns, as shown in the bottom of the right column.

2.3. Entropy of the four tones

In order to assess the efficiency of use of the four kind of tones, entropy of the tone in each number of syllables in the word or each position of syllable in the word was calculated from their occurrence frequency.

The entropy is derived by summation over the four kind of tones of - pi log(2) pi, where pi is occurrence probability of Tone-i in each number of syllable in a word, or each position of syllable in the word. The value is 2 bits when the occurrence frequencies are even among the four tones, but, decreases from 2 bits according to the degree of deviation from the even.

Amount of the decrease from 2 bits are shown for each plot in Figure 1. The entropy decreased by more than 0.1 bit only in the last syllables of some of the words. But, the decrease was less than 0.05 bit when averaged over all positions of syllables, and this indicates that the four kind of tones are used fairly effectively in the Chinese language.

3. Transitions probability of the four tones

As the transitions among the four tones have 16 combinations (4 kinds in preceding syllable x 4 kinds in following syllable), 5% which corresponds to about their average, is set as base plane in this figure, and the combination of tones from the preceding to following syllables are coordinated as array composed by columns of tones of preceding syllables and rows of tones of following syllables on this base plane. The transition probabilitiy above the base plane is shown three-dimensionally by the top of upward cones. The vertical axes are limited up to 10% in order to magnify the difference, as the value for most combinations stayed within this range. When the value exceeds 10%, the top of the cone becomes a flat disk.

3.1. Nouns

In the three syllable words of nouns, as shown in the upper half of the left column in Figure 2, Tone-1 was frequent in the first syllables, but, Tone-4 was frequent in the following syllables. Similar tendency was observed in the two or four syllable words of nouns.

Transition probability between Tone-2 and Tone-3 were less than average. Transitions beginning at Tone-1 and ending at Tone-4, or beginning at Tone-4 and ending at Tone-4 in the two, three and four syllable words were much more.

3.2. Other parts of speech

Kind of part of speech also showed characteristic transition of kind of tone.

1) In the three syllable words of the words of state, as shown in the lower half of the left column of Figure 2, most of the 2nd syllables were Tone-1. From 2nd to 3rd syllables, repeating transitions such as Tone-1+Tone-1, Tone-2+Tone-2, Ton-3+Tone-3 and Tone-4+Tone-4 were frequent.

These transitions along diagonal line on the base plane were also observed in the onomatopoeias.

2) In the adjectives, as shown in the right column, transitions in the two syllable words were similar as those of the first two syllables in the noun. But, from 1st to 2nd syllables in the three syllable words, Tone-4+Tone-2, -3 and -4 but not +Tone-1 were frequent. The transitions along the anti-diagonal line on the base plane, such as Tone-2+Tone-4 and Tone-4+Tone-2, from the 2nd to 3rd syllables were noticeable.

Similar tendency was observed in the adjectives.

Tone-5 (light tone or weakened tone) were more than 5% in the 2nd and 3rd syllables of the three syllable words of adjectives. Tone-5 was not shown in other figures, because the percentage was very small.

4. Influence of syllable structure

Preliminary analysis on this database shows that the syllabic structures and categories of consonants and vowels are associated with particular kind of tones. This suggests their relevance to the properties of occurrence frequency and transition probability of the Chinese four tones obtained in this statistical analysis.

.

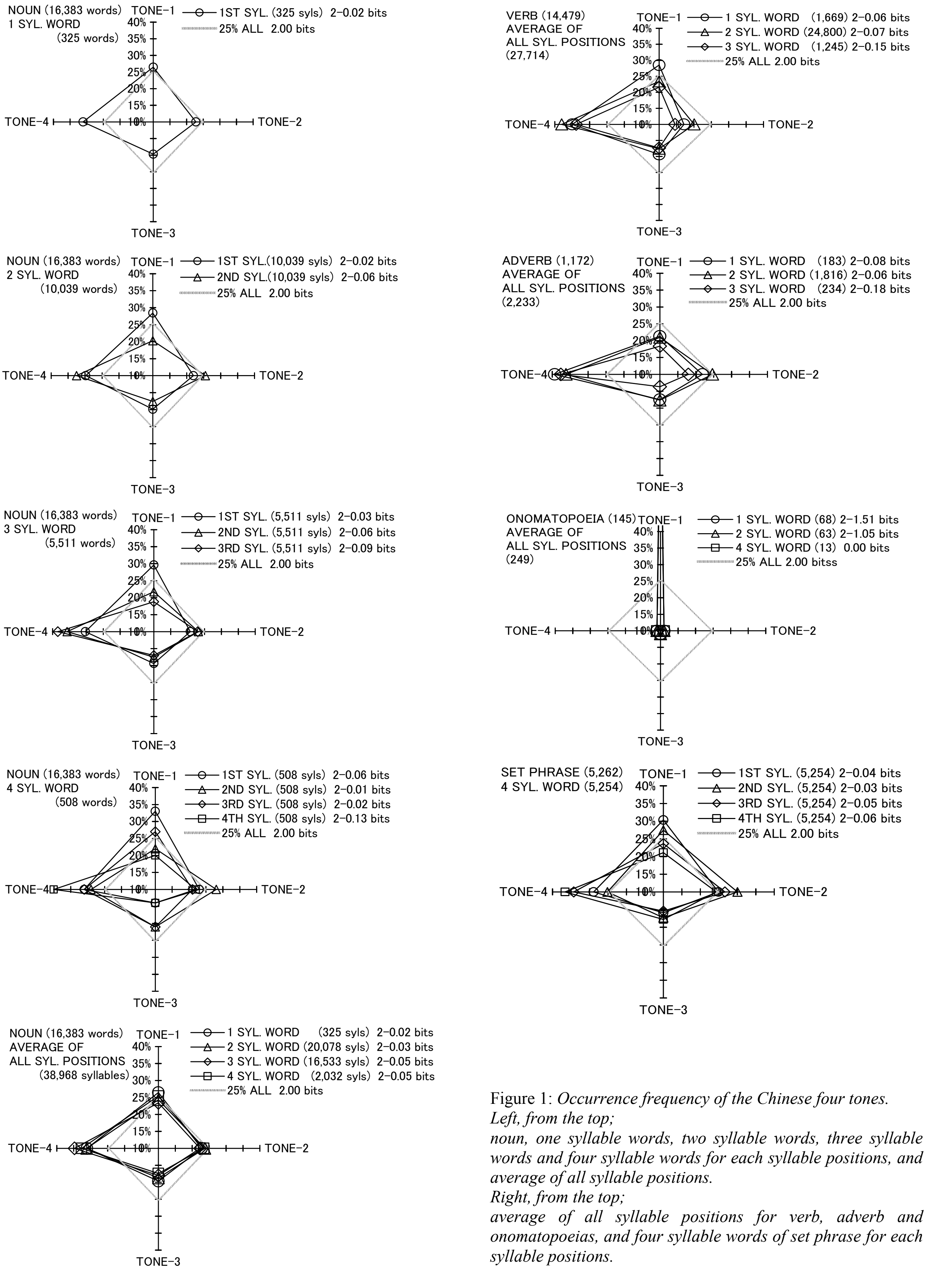

Figure 1: *Occurrence frequency of the Chinese four tones. Left, from the top;*

noun, one syllable words, two syllable words, three syllable words and four syllable words for each syllable positions, and average of all syllable positions.

Right, from the top;

average of all syllable positions for verb, adverb and onomatopoeias, and four syllable words of set phrase for each syllable positions.

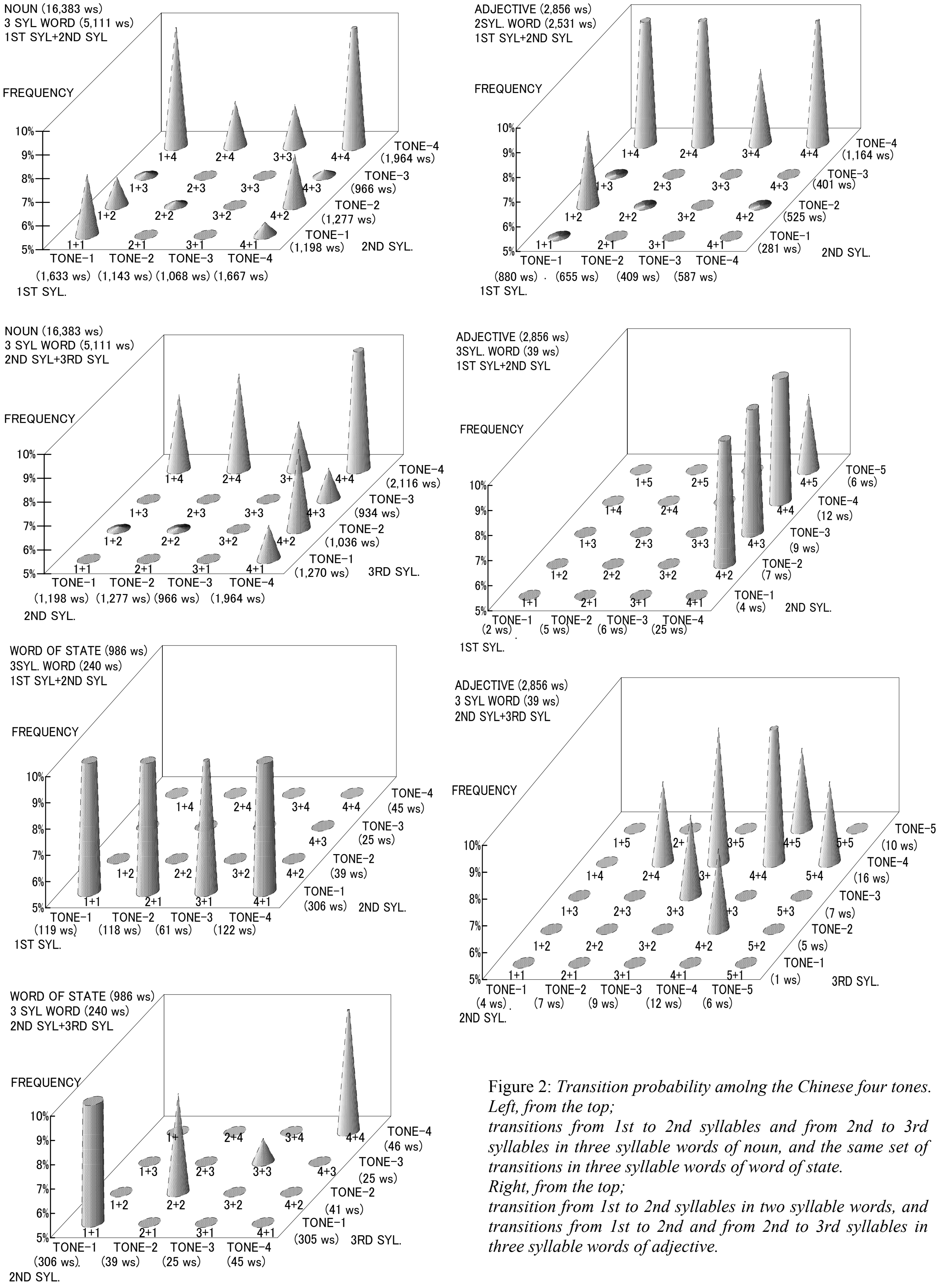

Figure 2: *Transition probability amolng the Chinese four tones. Left, from the top;*
transitions from 1st to 2nd syllables and from 2nd to 3rd syllables in three syllable words of noun, and the same set of transitions in three syllable words of word of state.
Right, from the top;
transition from 1st to 2nd syllables in two syllable words, and transitions from 1st to 2nd and from 2nd to 3rd syllables in three syllable words of adjective.