Syllable Onset Intervals as an Indicator of Discourse and Syntactic Boundaries in Taiwan Mandarin *

Janice Fon[1] **and Keith Johnson**[2]

[1] *National Taiwan Normal University, Taiwan*
[2] *The Ohio State University, U.S.A.*

Key words

boundary cues

hierarchical structure

spontaneous speech

syllable onset interval

Taiwan Mandarin

Abstract

This study looks at the syllable onset interval (SOI) patterning in Taiwan Mandarin spontaneous speech and its relationship to discourse and syntactic units. Monologs were elicited by asking readers to tell stories depicted in comic strips and were transcribed and segmented into Discourse Segment Units (Grosz & Sidner, 1986), clauses, and phrases. Results showed that the degree of final lengthening was modulated by boundary types. Lengthening before discourse boundaries was longer than that before clausal boundaries, which was in turn longer than that before phrasal boundaries. Final SOI lengthening also seemed to reflect cognitive load. At the discourse and clausal levels, the degree of lengthening is modulated by narration order. First narrations tended to have longer final SOIs than second narrations. In addition, there was also a mild lengthening effect in nonfinal SOIs, as was evidenced by the length differences in initial and medial SOIs and the differential lengthening effect regarding the positioning of phrases in a clause.

* *Acknowledgments:* This study is derived from a research project of the first author under the supervision of the second author in the Department of Linguistics, The Ohio State University. Parts of this article were presented at the Brown Bag Series at the Center for Cognitive Sciences, The Ohio State University, and at ICSLP 2000 in Beijing, China. Many thanks go to Dr. Mary Beckman, Dr. Chris Brew, Mr. Martin Jansche and all of our colleagues in the Phonetics Laboratory in the Department of Linguistics, The Ohio State University. Special thanks go to Ms. Laura Hsiu-min Liu for kindly sharing her recordings with us, Dr. Tammy Miao-Hsia Chang for helping out with the discourse analyses, and Mr. Hsin-fu Chiu for acting as a second labeler. We would also like to thank Mr. Pang-Hung Hsu for the technical support and Dr. Kuenhi Tsai for the statistical analyses. Many thanks also go to the Editor Dr. Marc Swerts and the two reviewers, Dr. Brigitte Zellner Keller and Dr. Anders Eriksson for the invaluable comments. Without them, this paper could not have been finished. Naturally, any faults are ours.

Address for correspondence: Janice Fon, Assistant Professor, Department of English, 162 Hoping E. Rd., Sec. 1, National Taiwan Normal University, Taipei 106, Taiwan; e-mail: <jfon@cc.ntnu.edu.tw>.

1 Introduction

Listeners are able to immediately and accurately slice an incoming stretch of speech signal into meaningful chunks so that on-line speech processing is possible. This does not seem to be a big feat until one looks at the actual speech signal. Unlike written texts, which contain punctuation marks that clearly demarcate syntactic and discourse units, spoken language often obscures its punctuation equivalents, if one believes they even exist in the first place.

Of all the potential acoustic punctuation cues, those in the time domain are frequently suggested to be likely candidates for discourse and syntactic boundary demarcation. In particular, silent pause and final lengthening are the most commonly proposed cues.

Goldman-Eisler (1972) was one of the first researchers to demonstrate the concurrence of silent pauses and syntactic boundaries. Compared to sentence-internal word boundaries, silent pauses are more likely to occur at clausal boundaries, and even more likely at sentential boundaries. A closer look at the pauses occurring at different syntactic boundaries showed a categorical distinction (Grosjean, 1980). Breathing pauses tend to co-occur with major syntactic boundaries while nonbreathing pauses tend to co-occur with minor boundaries. O'Malley, Kloker, and Dara-Abrams (1973) also showed that silent pause duration correlates positively with structural hierarchy. Phrases of higher attachments tend to be accompanied by longer pauses than those of lower attachments.

Final lengthening is another frequently proposed cue for syntactic boundary demarcation. Klatt (1975) found that sentence-final syllables are likely to be lengthened compared to sentence-internal ones, while Lehiste (1975b) and Scott (1982) claimed that it is the sentence-final foot that is lengthened, not the syllable. Regardless of the domain of lengthening, both silent pause and lengthening are useful in disambiguating sentences with multiple possible syntactic phrasings (Lehiste, 1973; Lehiste, Olive, & Streeter, 1976).

Ladd (1988) used "boundary duration" as a combined measure for final lengthening and silent pause to look at English hierarchical encoding. This is defined as the interval between the onset of the last stressed syllable preceding the boundary and the onset of phonation of the first syllable after the boundary. In other words, this includes the duration of the final stressed syllable, the following unstressed syllables if there are any, and the following pause. Ladd showed that boundary duration is a good indicator of hierarchical organization of clauses. Clausal boundaries of higher disjunctures have longer boundary duration than those of lower disjunctures.

Acoustic encoding of discourse boundaries is more equivocal. Lehiste and colleagues (1975a; 1979; Lehiste & Wang, 1977) found no consistent indication of phonetic paragraph boundary markers. On the other hand, more recent studies showed that pause duration can be a robust cue (Grosz & Hirschberg, 1992; Hirschberg & Nakatani, 1996). Discourse-final phrases are followed by longer pauses than discourse initial ones.

All of the above studies have focused their attention on Indo-European languages, mainly English. It is therefore unclear how and whether timing cues and syntactic/discourse boundaries interact with each other in languages that are as different as, say, Mandarin Chinese. In addition, except for a few (Hirschberg & Nakatani, 1996; Lehiste, 1979; Lehiste & Wang, 1977), most studies focused on read instead of spontaneous speech. Given the fact that spontaneous speech contains various phenomena not as available in read speech (e.g., hesitation, stutter, false start, etc.), it would be interesting to see whether timing cues said to be associated with discourse/syntactic boundaries in read speech still exist in spontaneous speech.

In this study, we examined whether timing plays a consistent role in demarcating discourse and syntactic boundaries in Taiwan Mandarin spontaneous narratives.[1] Specifically, we were interested in the patterning of lengthening and pause with regards to discourse, clausal, and phrasal boundaries. By looking for reliable timing cues at discourse/syntactic boundaries, we are not suggesting that there is necessarily a causal relationship between the two, as both are highly coterminous with prosodic boundaries (Butcher, 1980; Keller, Zellner, Werner, & Blanchoud, 1993; Zellner, 1994). Instead, we are mainly interested if there is a consistent correlation between the two in Taiwan Mandarin. How this correlation interacts with prosodic boundaries is beyond the scope of this paper.

2 Hypotheses

Two effects were tested in this study. The first was the boundary effect. That is, whether final lengthening exists in Taiwan Mandarin spontaneous speech. Realization of a boundary effect would be that signals at syntactic and discourse boundary positions are different (usually magnified) compared to those not located at boundary positions. The second effect examined is the hierarchy effect. That is, whether the degree of lengthening reflects the level of structural boundaries. Realization of such an effect would be a (usually positive) correlation between boundary strength and acoustic strength at the boundary position.

3 Methods

3.1 *Subjects*

Five female and five male native speakers of Taiwan Mandarin participated in this study. All of them were students from the National Taiwan University. Their ages ranged from 20 to 25 years at the time of recording.

1 Although the content of the narration was constrained (see Methods below), the term *spontaneous* is still used here because how subjects were to complete the task was not constrained. Describing a comic strip one has seen may not be an everyday task one usually performs. However, it does happen from time to time, and it is somewhat analogous to, say, telling a friend about a movie one has watched. Therefore, we believe that although the narratives collected in this study were fairly task-oriented, the style was still more spontaneous than controlled.

3.2
Materials

In order to elicit speech, two four-frame comic strips with no printed dialogs were chosen from *Shuangxiangpao*,[2] a very popular comic series in Taiwan by a famous cartoonist Te-Yong Chu. Please see Appendices A and B for the comics.

3.3
Procedure

Subjects were seated in a quiet room and were given the comic strips. They were told to study the comic strips and retell the stories as if telling them to a friend. Recordings were made individually with a SONY TCM-5000EV recorder and a SONY ECM-G3M superdirectional microphone. Half of the subjects received Comic 1 first and half received Comic 2 first. Appendix C shows the duration of the recording of each subject.

3.4
Measurements

Recordings were digitized at 22050 Hz and the onset of each syllable was determined using the waveform and the spectrogram on CSL KAY4300. Following Ladd (1988), the dependent measure was the syllable onset interval (SOI), which is the interval between the onset of one syllable and that of the next, which would include the duration of the syllable itself, and possible following silent pauses. This measure is a necessary modification of Ladd's "boundary duration" because Taiwan Mandarin is a syllable-timed language and stress does not play as an important role as in English. Moreover, perceptual studies have also shown that people are more sensitive to on-time intervals (i.e., the interval between onsets of 2 consecutive signals) rather than on-time duration (i.e., duration of a signal) with both speech (Huggins, 1972) and nonspeech stimuli (Gabrielsson, 1993; Handel, 1993). Since it would be more reasonable if the timing cues we chose to measure are also perceptually viable, SOI was used instead of syllable duration or silent pause. Figure 1 is an example of how SOI was measured. Only narration-initial and -medial SOIs were included for analyses, as narration-final SOIs by definition do not exist. In total, there were 1544 SOIs from Comic 1 and 1264 SOIs from Comic 2 for the analysis. The first author served as the main labeler. A second trained labeler was recruited and relabeled 10% of the data. The correlation of the two labels was high ($r = .99, p < .0001$). Therefore, labeling of the first author was used for subsequent analyses.

3.5
Discourse / syntactic analyses

Three levels of structural analyses were conducted. The first was the discourse level, at which Grosz and Sidner's (1986) model was used as a basis for discourse segmentation and analysis. According to their model, there are three separate but interrelated components to discourse structure — the linguistic structure, the intentional structure, and the attentional state. What concerns us in this study is the intentional structure.

---

[2] The Pinyin system is used for romanization of texts throughout the paper.

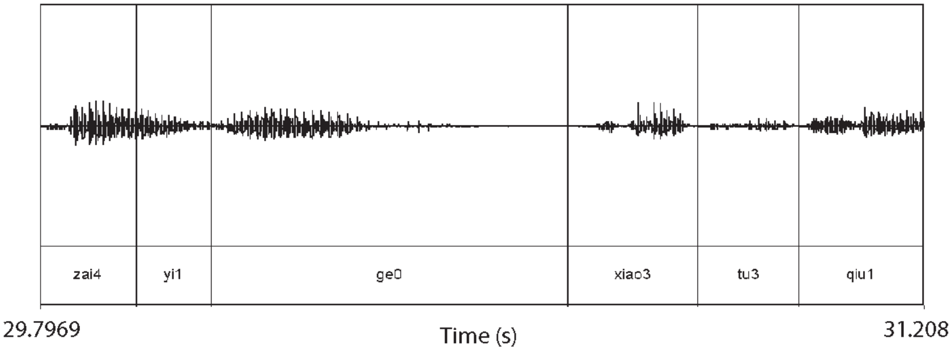


Figure 1

An illustration of SOIs. The vertical lines indicate SOI boundaries. In this example, *zai yi-ge xiao tuqiu* 'on a small mound', the SOIs of the first two and the last three syllables are equal to their syllable durations, while the third syllable *ge*, which is followed by a silent pause, has a longer SOI than its syllable duration

The intentional structure is what the talker intends to achieve when conducting a particular discourse. For example, telling a friend about a movie one has seen recently may achieve the goal of describing the story (in addition to entertaining the friend). This goal is fairly broad and has to be broken down into several subgoals, and maybe even subsubgoals (e.g., describing the plot, the cast, etc.). For each (sub-)subgoal, there will be a discourse segment to fulfill the intention. The intention of a discourse segment is called a discourse segment purpose (DSP).

For the purpose of this study, narratives were segmented into DSPs. As most subjects described the comic strips by overtly referring to each frame, most of the DSPs were conveniently divided accordingly. Comments on the comic strips were labeled as separate DSPs.

One of the authors, JF, who is a native speaker, served as the main DSP labeler. A second native speaker labeler who was familiar with discourse analysis was recruited to test interlabeler reliability. The labeler read the guidelines outlined in Grosz and Sidner (1986) and labeled all the data. As the agreement was very high (98.4% of the cases labeling were the same), segmentation by JF was used for the subsequent statistical analyses. Please see Appendices D and E for examples of DSP segmentation of Comics 1 and 2.

The second level is the clause. Only main clauses were considered, since there were not enough cases of subordinate clauses for analyses. The third level is the phrase. In order to examine possible position effects while at the same time control for different types of phrases, only NPs were included in the phrasal level analysis. This is because NPs are prevalent in clauses and can assume various positions. As Taiwan Mandarin is mostly an SVO language, NPs can be clause-initial, as in nominative NPs and some of the NPs functioning as ADVPs, clause-medial, as in oblique NPs, or clause-final, as in accusative and existential NPs. Since the positioning of SOIs is hard to define in monosyllabic and bisyllabic NPs, only those containing more than two syllables were included in the analyses. Also, units containing narration-final SOIs were excluded. In total, there were 124 DSPs, 211 clauses, and 187 NPs for analyses.

Examples of DSPs, clauses, and phrases are shown in (1)–(3) below. Examples (4)–(8) are NPs in different cases. The relevant units are bracketed.

1. DSP: segments of discourse intentions

[*ranhou zou dao shenfu mianqian shenfu shuo laibuji-le*]$_{DSP6}$
[then walk To priest front priest say too-late-PFV [3]]$_{DSP6}$
[Then (they) walked to the front of the priest, the priest said (it was) too late]$_{DSP6}$
(LZE: DSP0106)

2. Clause: main clauses only.

[*ranhou zou dao shenfu mianqian*]$_{CL11}$ [*shenfu shuo laibuji-le*]$_{CL12}$
[then walk To priest front]$_{CL11}$ [priest say too-late-PFV]$_{CL12}$
[Then (they) walked to the front of the priest]$_{CL11}$ [(the) priest said (it was) too late]$_{CL12}$
(LZE: CL0111 & CL0112)

3. Phrase: NPs only.

ranhou zou Dao [*shenfu mianqian*]$_{PH12}$ [*shenfu*]$_{PH13}$ *shuo laibuji-le*
then walk To [priest front]$_{PH12}$ [priest]$_{PH13}$ say too-late-PFV
Then (they) walked to [the front of the priest]$_{PH12}$ [the priest]$_{PH13}$ said (it was) too late
(LZE: PH0112 & PH0113)

4. Nominative:

[*ta-de zhuyao gongzuo*]$_{PH5}$ *shi ge shuangyanpi*
[3sg-GEN main job]$_{PH5}$ be cut double-eyelid
[His main job]$_{PH5}$ is double-eyelid surgery.
(ZZH: PH0205)

5. Accusative:

ranhou ju-chu [*kaolyu-de paizi*]$_{PH9}$ *a*
then hold-up [consider(n.)-ASSOC[4] sign]$_{PH9}$ PART:RF[5]
then (they) held up [(a) “to consider” sign]$_{PH9}$.
(ZZH: PH0109)

---

[3] PFV stands for the perfective aspect.

[4] ASSOC: associative phrase denoting the linkage of two noun phrases (Li & Thompson, 1981, p. 113).

[5] PART:RF is a particle used to reduce forcefulness (Li & Thompson, 1981, p. 313).

6. Oblique

… jieyou [*zhe-zhong manhua*]$_{PH23}$ *jiu shi fengci na-ge …*
… by-way-of [this-kind comic]$_{PH23}$ emphatic be satirize that-CL[6] …
… by (using) [this type of comics]$_{PH23}$ to satirize the …
(ZZH: 0123)

7. Existential

you [*yi-ge yisheng*]$_{PH1}$ *a*
have [one-CL doctor]$_{PH1}$ PART:RF
There was [a doctor]$_{PH1}$.
(LCX: PH0201)

8. ADVP

… [*ta-men zou dao zhenghunren qianmian de shihou*]$_{PH15}$ *na zhenghunren …*
… [3sg-pl. walk to officiant front NOM[7] time]$_{PH15}$ that officiant …
… (at) [the time when they walked to the front of the officiant]$_{PH15}$, that officiant …
(LZH: PH0115)

4 Results

In this section, we will first present some descriptive summary statistics. Then, we will discuss the results of each level in turn. Finally, we will look at possible hierarchical effects on SOIs.

4.1 *Summary statistics*

Figure 2 shows the mean and median non-narration-final SOIs of each subject. Filled circles indicate Comic 1 and empty circles Comic 2. Except for PSZ in Comic 2, there was not much difference between the two comic strips (between Comics 1 and 2: $r(10) = .80$, $p < .01$ for mean; $r(10) = .84$, $p < .01$ for median). PSZ mentioned comprehension difficulty prior to her narration of Comic 2, and as a result, had extremely long mean SOIs. The distributions of the two comic strips were both skewed to the right (Comic 1: skewness = 3.03; Comic 2: skewness = 3.00), which was expected for a duration measure. For numbers of SOIs each subject contributed, please see Appendix F.

Average speech rate (SR), calculated by dividing the total number of syllables by the total amount of time including pause (syll / sec), fluctuated between 2.18 syll / sec to 6.45 syll / sec, with an overall average of 4.08 syll / sec, which was roughly comparable to the reading rate of English-speaking college students, 5.4 syll / sec (Kowal, O'Connell, O'Brien, & Bryant, 1975). Although this seems to be quite a wide range,

[6] CL: classifier.

[7] NOM: nominalizer.

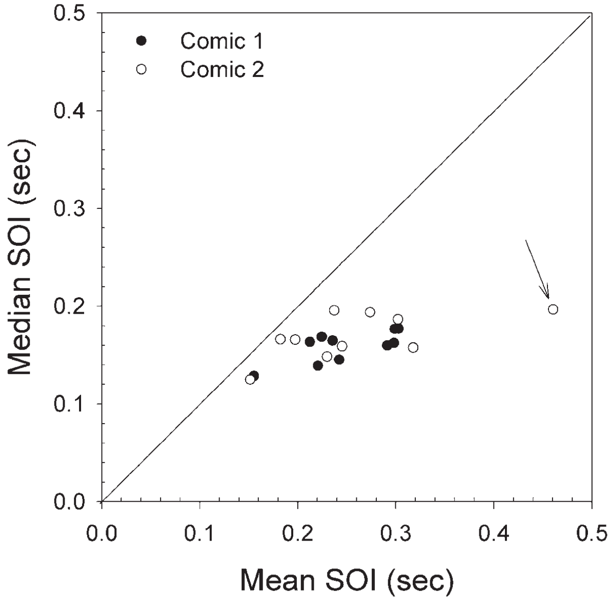


Figure 2

A scatter plot indicating the mean and median non-narration-final SOIs of each subject. Filled circles represent Comic 1 and empty circles Comic 2. The arrow points to PSZ in Comic 2

a closer look at the data showed that most of the subjects actually had an average SR at around 3–4 syll/sec for Comic 1 and 3–5 syll/sec for Comic 2, as shown in Table 1. The maximum and minimum mean values were contributed by Subjects XYL and PSZ, respectively. Subject XYL seemed to be a habitual fast speaker, as both of her narrations were conducted at an extremely fast yet consistent rate (≥6 syll/sec). On the other hand, PSZ showed abnormally slow SR only when she was narrating Comic 2. As mentioned above, this probably has to do with the fact that she had trouble understanding Comic 2.

Besides using SR as a measure for overall timing, articulation rate (AR) was also used. AR was calculated by dividing the total number of syllables by the total amount of articulating time; in other words, silent pause time was excluded.[8] Thus, by definition, AR should always be higher than SR. In general, most of the subjects had ARs at around 5–6 syll/sec, regardless of comic strips. Subjects XYL and PSZ were again exceptions. XYL had an average AR of over 7 syll/sec for both comics while PSZ had an AR of 4.6 syll/sec for Comic 2.

Table 2 shows the intercorrelations among SR, AR and pause time in percentages (%P). There was a high positive correlation ($r \geq .80$) between SR_1 and SR_2, AR_1 and AR_2, and $\%P_1$ and $\%P_2$, indicating consistency in rate and rhythm across the two comic strips. There was also a significant negative correlation between SR_1 and $\%P_1$, and SR_2 and $\%P_2$, which was expected, since the calculation of SRs included silent pause time. However, one also found a negative correlation between AR_1 and $\%P_1$, and AR_2 and $\%P_2$. In other words, speakers with a faster articulation rate also

[8] Thanks to one of the reviewers, Dr. Anders Eriksson, for pointing out the difference between SR and AR.

TABLE 1

Mean speech rate and articulation rate for the two comics for each subject

	Comic 1		Comic 2	
Subject	*SR*[a] *(syll/sec)*	*AR*[b] *(syll/sec)*	*SR (syll/sec)*	*AR (syll/sec)*
Male				
CCF	4.53	6.50	4.08	5.55
LZE	4.44	5.81	5.50	6.14
LZH	3.44	5.75	3.32	5.06
LZY	3.31	5.41	3.65	5.20
ZZH	4.14	6.62	4.37	6.63
Female				
LCX	4.66	5.67	5.04	5.69
LJY	4.26	5.79	4.22	5.08
LJS	3.37	5.25	3.16	5.09
XYL	6.45	7.64	6.60	7.48
PSZ	3.35	5.26	2.18	4.56

[a]*SR: speech rate*

[b]*AR: articulation rate*

TABLE 2

Intercorrelations among speech rate, articulation rate, and percentage of pause

	1	2	3	4	5	6
1. SR_1[a]	—	.90***	.86**	.84**	−.85**	−.72*
2. SR_2[b]		—	.73*	.88***	−.83**	−.89***
3. AR_1			—	.88***	−.49	−.44
4. AR_2				—	−.57(*)	−.58(*)
5. $\%P_1$[c]					—	.83**
6. $\%P_2$						—

[a] *SR: speech rate. Subscripts 1 and 2 refer to Comic 1 and Comic 2, respectively.*

[b] *AR: articulation rate*

[c] *%P: pause time percentage*

*$p < .05$. **$p < .01$. ***$p < .001$. (*)$p < .1$.*

paused less. Although the correlation was not significant for the former, it was near significant for the latter (p = .15 for Comic 1 and p = .08 for Comic 2), implying that there was a tendency for talkers who articulate faster to pause less.

4.2
DSP boundaries

There were in total 124 non-narrational-final DSPs in the corpus containing 2402 SOIs in the DSP analysis. The distribution of SOIs with regards to their DSP positions is shown in Table G-1 in Appendix G. Figure 3a shows the distribution of SOIs in various DSP positions with utterances lined up at the boundary.

A 2 (SEX) × 2 (ORDER) × 3 (DSP POSITION) three-way mixed ANOVA on non-narration-final SOIs was performed, with DSP POSITION being the within-subject variable.[9] There was no effect involving SEX; therefore, groups were collapsed and a 2 (ORDER) × 3 (DSP POSITION) two-way ANOVA was performed instead. Results showed that all three effects were significant, ORDER: $F(1, 122) = 4.79$, $p < .05$, $\hat{\eta}^2 = .04$; POSITION: $F(1.26, 153.36) = 120.90$, $p < .0001$, $\hat{\eta}^2 = .50$; ORDER ×POSITION: $F(1.26, 153.36) = 4.71$, $p < .05$, $\hat{\eta}^2 = .04$.[10]

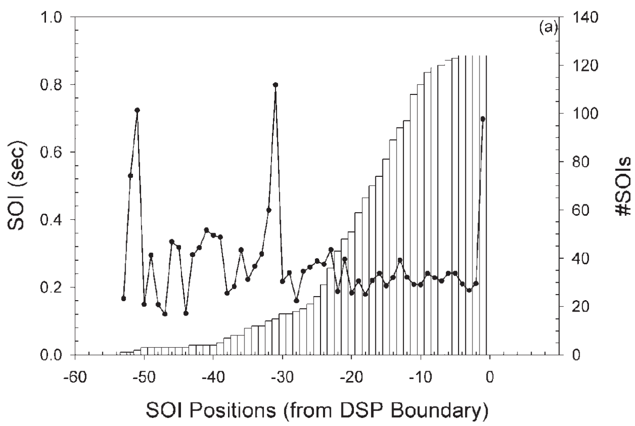


Figure 3a

SOIs in various DSP positions with utterances lined up at the boundary. The line graph (left *y*-axis) indicates the average duration of SOIs in each DSP position, with 0 denoting the DSP boundary. In other words, Position –1 implies one syllable before the boundary, Position –2 implies two syllables before the boundary, and so on. The histogram (right *y*-axis) reflects how many SOIs there were in each position

As shown in Figure 3b, there was a strong DSP POSITION effect, regardless of the recording order. SOIs were lengthened according to their positioning in a DSP. Post hoc tests using Bonferroni's adjustments showed that DSP-final SOIs were the longest while DSP-initial SOIs were the shortest for the first narration ($p < .01$). For the second narration, DSP-final SOIs were still the longest ($p < .0001$), while DSP-medial SOIs were marginally longer than DSP-initial ones ($p = .06$).

9 Although the distributions of SOIs violate the normality assumption, ANOVA was still applied because ANOVA tests are known to be fairly robust to non-normal distributions (Rietveld & van Hout, 1993).

10 Mauchly's Test showed that the sphericity assumption was violated. Therefore, Huynh-Feldt adjusted *df*'s were used accordingly. Adjusted *df*'s were used throughout the paper when the sphericity assumption was violated.

As to the effects involving ORDER, it is clear from Figure 3b that the main effect was actually a byproduct of the interaction. Post hoc independent *t*-tests showed that final SOIs were significantly longer in the first than the second narration, $t(112.23) = 2.36$, $p < .05$.[11]

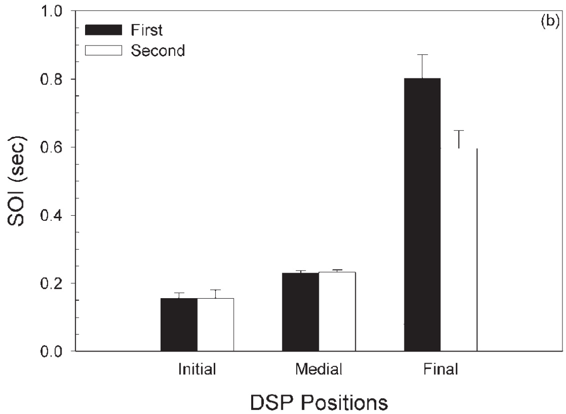


Figure 3b

A bar graph showing SOIs at different DSP positions. Black bars represent the first narration and white bars the second narration. The error bars represent SE. Only the top portions are shown to avoid clutter

4.3
Clausal boundaries

There were in total 193 non-narrational-final main clauses in the corpus containing 2355 SOIs in the clause analysis. The distribution of SOIs with regards to their clausal positions is shown in Table G-2 in Appendix G. Figure 4a shows the distribution of SOIs in various clausal positions with utterances lined up at the boundary.

A 2 (SEX) × 2 (ORDER) × 3 (CLAUSE POSITION) three-way mixed ANOVA on non-narration-final SOIs was performed. As in the DSP analysis, there was no effect involving SEX. Therefore, groups were collapsed and a 2 (ORDER) × 3 (CLAUSE POSITION) two-way ANOVA was performed instead. Results showed that both of the main effects and the interaction effect were significant, ORDER: $F(1, 315) = 9.71$, $p < .01$, $\hat{\eta}^2 = .003$; POSITION: $F(1.26, 396.296) = 255.78$, $p < .0001$, $\hat{\eta}^2 = .45$; ORDER × POSITION: $F(1.26, 396.26) = 9.31$, $p = .001$, $\hat{\eta}^2 = .03$.

As shown in Figure 4b, there was a strong CLAUSE POSITION effect. Post hoc tests using Bonferroni's adjustments showed that SOIs were lengthened according to their positioning in a clause. Clause-final SOIs were the longest and clause-initial SOIs the shortest ($p < .0001$).

As in the DSP analysis, the ORDER main effect seemed to be mainly due to the interaction effect. Post hoc independent *t*-tests showed that final SOIs were longer in the first narration than in the second, $t(281.28) = 3.33$, $p < .001$. Nonfinal SOIs did not show any difference in duration.

[11] Levene's Test for equality of variances showed that the assumption for equal variances was violated. Therefore, the *df* was adjusted accordingly. Adjusted *df*'s were used throughout the paper when the assumption was violated.

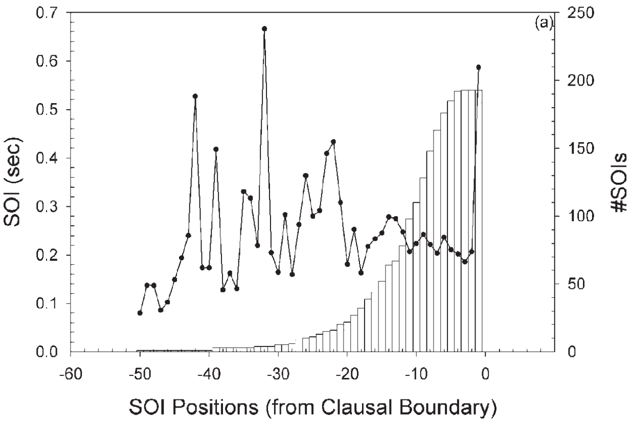


Figure 4a

SOIs in various clausal positions with utterances lined up at the boundary. The layout of the graph is similar to that of Figure 3

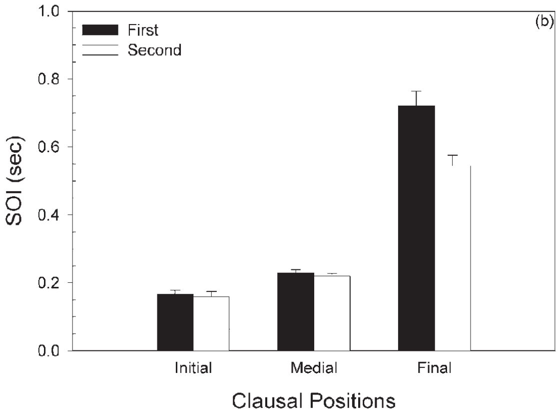


Figure 4b

A bar graph showing SOIs at different clausal positions. Black bars represent the first narration and white bars the second narration. The error bars represent SE. Only the top portions are shown to avoid clutter

4.4
Phrasal boundaries

There were in total 221 trisyllabic-or-longer non-narration-final NPs in the corpus. Of the 221 NPs, there were 52 nominatives, 78 accusatives, 15 obliques, 57 existentials and 19 NPs as ADVPs (see Table G-3 in Appendix G). Figure 5a shows the distribution of SOIs in various phrasal positions with utterances lined up at the boundary.

Since there were not enough cases of oblique NPs and NPs as ADVPs, only nominative, accusative, and existential NPs were included in the phrase analyses, resulting in 187 NPs and 1093 SOIs. Table G-4 in Appendix G shows the detailed distribution.

A 2 (SEX) × 2 (ORDER) × 3 (PHRASE POSITION) × 3 (CASE) mixed four-way ANOVA on SOIs of non-narration-final NPs that were trisyllabic or longer was performed. There was no significant effect involving ORDER or SEX, so groups were collapsed and a 3 (PHRASE POSITION) × 3 (CASE) two-way ANOVA was performed instead. Results showed that only the two main effects were significant, POSITION: $F(1.25, 230.15) = 62.20$, $p < .0001$, $\hat{\eta}^2 = .25$; CASE: $F(2, 184) = 3.32$, $p < .05$, $\hat{\eta}^2 = .03$.

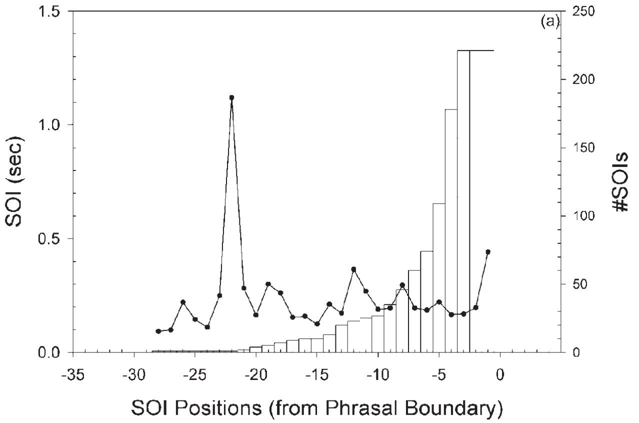


Figure 5a

SOIs in various phrasal positions with utterances lined up at the boundary. The layout of the graph is similar to that of Figure 3

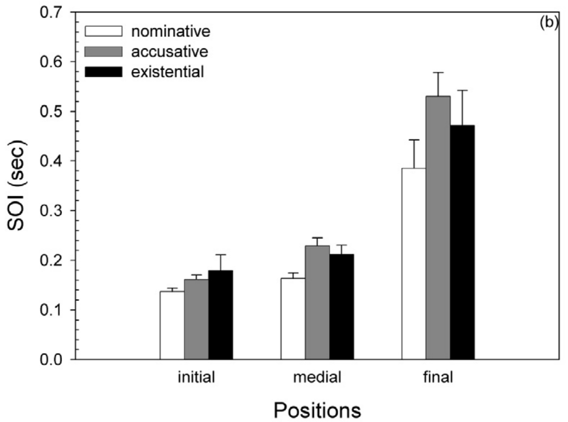


Figure 5b

A bar graph showing SOIs at different phrasal positions with regards to case. The error bars represent SE

As shown in Figure 5b, there was a strong POSITION effect. Post hoc tests using Bonferroni's adjustment showed that final SOIs were the longest and initial SOIs were the shortest ($p < .01$). As to the CASE main effect, post hoc Tukey's-*b* test showed that SOIs in accusative NPs were significantly longer than those in nominative NPs ($p < .05$).

4.5
Hierarchical encoding

Since the final SOI lengthening effect is fairly robust across all three levels of structural boundaries examined, one would like to know whether such an acoustic encoding is sensitive to any hierarchical organization components. However, it is virtually impossible to find final SOIs that only belong to one level but not the other, except for perhaps the very lowest phrasal level. Therefore, the hierarchy effect was examined by comparing final SOIs that are at all three boundaries (D + C + P), those that are at only clausal and phrasal boundaries (C + P), those that are at only phrasal boundaries (P), and those that are not at any of the three boundaries. In total, there were 57 instances of D + C + P SOIs, 31 instances of C + P SOIs, 121 instances of P SOIs, and 1023 instances of SOIs that were not at a boundary location of any level.[12]

A hierarchical regression analysis was conducted on duration (see Table 3). Results showed that all three levels of boundaries contributed to final SOI lengthening (Fig. 6). ΔR^2s were significant for all three models. However, DSP and phrasal boundaries had more unique contributions than clausal boundaries (Phrase: *pr* = .13; Clause: *pr* = .07; DSP: *pr* = .14).

TABLE 3

Summary of hierarchical regression analysis for variables predicting final SOI lengthening ($N = 1232$)

	Variable	*B*	*SE B*	*b*
Step 1	Phrase	.25	.02	.32****
Step 2	Phrase	.12	.03	.15****
	Clause	.31	.04	.28****
Step 3	Phrase	.12	.03	.15****
	Clause	.13	.05	.11*
	DSP	.28	.06	.21****

Note: $\Delta R^2 = .10$, $\Delta F = 141.94$, $p < .0001$ for Step 1; $\Delta R^2 = .05$, $\Delta F = 69.93$, $p < .0001$ for Step 2; $\Delta R^2 = .02$, $\Delta F = 23.19$, $p < .0001$ for Step 3.

*$p < .05$. ****$p < .0001$.

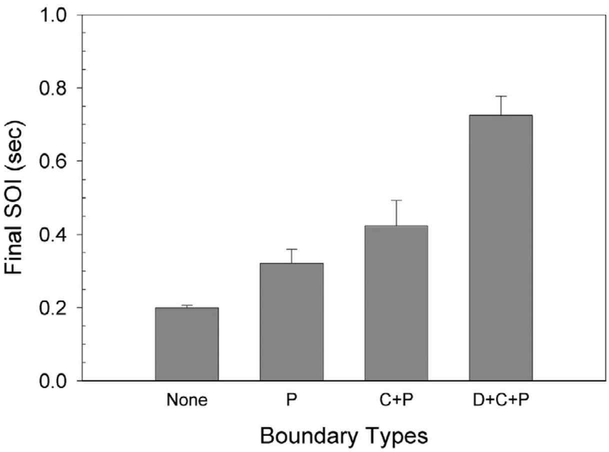


Figure 6

A bar graph showing final SOIs of different boundary types. The error bars represent SE. None: no boundary; P: phrasal boundary; C + P: clausal and phrasal boundary; D + C + P: DSP, clausal and phrasal boundary

[12] Since selection criteria used at each level were applied here, the number of SOIs included was smaller than that at each level.

5 Discussion

Although spontaneous speech is full of variability, timing regularities do seem to exist. Extremely fast or slow speaking rates can often be traced to idiosyncrasy (e.g., Talker XYL) or difficulty in cognitive processing (e.g., PSZ in Comic 2). In general, inter- and intra-talker speaking rates were fairly consistent (see SRs and ARs in Tables 1 and 2).

The negative relationship between AR and %P was interesting (Table 2), although the results were only near significant at best. Since the calculation of AR excluded pause, this implies that speakers who tend to utter syllables short will have fewer and shorter pauses. Further studies will be needed in order to determine the stability of this effect.

Strong POSITION effects were found as predicted by our first hypothesis (see the hypothesis section above). Compared to nonfinal SOIs, final SOIs underwent a significant amount of lengthening at all three levels. However, nonfinal SOIs also displayed a milder but steady lengthening effect, as was evidenced by the significant difference between initial and medial SOIs at all three levels, and among SOIs of different cases of NPs at the phrasal level, as nominative NPs, which always occur at or near the beginning of an utterance, tended to have shorter SOIs than accusative and existential NPs, which always occur at or near the end of an utterance. In other words, there are at least two kinds of lengthening at work in Taiwan Mandarin spontaneous speech. One is the steady but mild lengthening effect, the scope of which applies to all nonfinal SOIs; the other is the strong final lengthening effect, the scope of which is mostly limited to final SOIs. As the latter effect is more clearly documented in the literature of other languages, more studies would be needed in order to see if the former is stable across various languages and genres.

There was also a strong HIERARCHY effect on final SOIs. DSP-final SOIs were longer than clause-final SOIs, which were in turn longer than phrase-final SOIs. In contrast to the mild lengthening effect mentioned above, where the magnitude is determined mainly by the positioning of nonfinal SOIs within a structural unit, the magnitude of the final lengthening effect seems to be determined to a large extent by the level of structural hierarchy. Higher structural boundaries elicit greater SOI lengthening.

However, at a first glance, the three levels of structural hierarchy do not seem to impose the same degree of lengthening force on final SOIs, as shown in Table 3. Phrase and DSP contributed more to final SOI lengthening than clause. Nevertheless, this is likely due to an artifact of floor and ceiling effects of the study, as only three levels of structural hierarchy were examined. Influences from structural levels below the phrase (e.g., word) and above the DSP (e.g., discourse topic), if any, were necessarily combined with those from the phrase and the DSP, respectively, and thus boosted the nominal contributions of the two levels. Future studies will be needed to investigate the influences on SOIs from other levels of hierarchy.

The interaction effects concerning ORDER were surprising but interesting. At the DSP and the clausal levels, the final lengthening effect was modulated by the recording order. Final SOIs were longer in the first narration than the second. This might have something to do with a reduction in cognitive load (Grosjean, 1980). Although lengthening does not always imply a high cognitive demand, general processing difficulties are often reflected by an overall lengthening effect, as was shown in PSZ in Comic 2. In other words, when talkers became more used to the task, the cognitive load became lighter, and thus the degree of final lengthening became slighter. Interestingly, the phrasal level was rather impervious to such modulations, as no interaction involving ORDER was found at this level. Practice seemed to alleviate the cognitive load from mainly higher structural levels, probably because the processing demand required for lower structural levels was already at a necessary minimum in this study, as the stimuli used was especially selected to prevent difficult lexical selection problems. More research will be needed to better understand the practice effect.

6 Conclusion

This study promotes a hierarchical view of speech timing (Campbell, 1991; Cooper, Paccia, & Lapointe, 1978; Ferreira, 1993, Nakatani, O'Connor, & Aston, 1981; Zellner, 1998). Cues for segmenting structural units are readily available in the speech signal. The rhythmic regularities found can act as indicators for both cognition processing load and structural organization. From a speaker's point of view, the degree of SOI lengthening reflects the cognitive processing load the speaker is bearing. Speakers tend to produce longer SOIs when they are performing tasks they are not familiar with and demand more cognitive processing. As a result, the second comic strip in this study was always told with a smaller final SOI lengthening than the first, even though the two comic strips were different in content. Similarly, higher level structural units demand more cognitive planning than lower level units. Therefore, it is not surprising that the degree of final SOI lengthening was proportional to the level of hierarchy and final SOIs at the discourse level were longer than those at the clausal level, and those at the clausal level were in turn longer than those at the phrasal level.

From a listener's point of view, the degree of SOI lengthening can be used as an indicator for the structural organization of a discourse. Within a unit, the degree of lengthening reflects its positioning; the closer an SOI is to the end of a unit, the more likely it is to be lengthened. Across different units, the degree of lengthening can also reflect their hierarchical relationship, since final lengthening is proportional to the level of structural hierarchy. Since languages tend to have their preferred SOI duration in spontaneous speech (Fon, 2002), it should not be a difficult task for native speakers to recognize the existence of lengthening and the degree of such. By providing cues indicating whether a unit is coming to an end, and the level of hierarchy the unit is at, this might alleviate some of the burden listeners have in doing on-line segmentation and building discourse hierarchy. Further studies will be needed in order to confirm listeners' ability in and propensity for using such cues.

Received: June 26, 2001; first revision received: December 9, 2002; second revision received: July 2, 2003; accepted: August 28, 2003

References

BUTCHER, A. (1980). Pause and syntactic structure. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech: Studies in honor of Frieda Goldman-Eisler* (pp. 85–90). Berlin: Mouton.

CAMPBELL, W. N. (1991). Theories of prosodic structure: Evidence from syllable duration. *Proceedings of 12th International Congress of Phonetics Sciences*, 290–293.

COOPER, W. E., PACCIA, J. M., & LAPOINTE, S. G. (1978). Hierarchical coding in speech timing. *Cognitive Psychology*, **10**, 154–177.

FERREIRA, F. (1993). Creation of prosody during sentence production. *Psychological Review*, **100**(2), 233–253.

FON, Y.-J. J. (2002). *A cross-linguistic study on syntactic and discourse boundary cues in spontaneous speech*. Dissertation. The Ohio State University.

GABRIELSSON, A. (1993). The complexities of rhythm. In T. J. Tighe & W. J. Dowling (Eds.), *Psychology and music: The understanding of melody and rhythm* (pp. 93–120). Hillsdale, NJ: Lawrence Erlbaum.

GOLDMAN-EISLER, F. (1972). Pauses, clauses, sentences. *Language and Speech*, **15**, 103–113.

GROSJEAN, F. (1980). Linguistic structures and performance structures: Studies in pause distribution. In H. W. Dechert & M. Raupach (Eds.), *Temporal variables in speech: Studies in honor of Frieda Goldman-Eisler* (pp. 91–106). Berlin: Mouton.

GROSZ, B., & HIRSCHBERG, J. (1992). Some intonational characteristics of discourse structure. *Proceedings of the 2nd International Conference of Spoken Language Processing*, 429–432.

GROSZ, B. J., & SIDNER, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, **12**(3), 175–204.

HANDEL, S. (1993). The effect of tempo and tone duration on rhythm discrimination. *Perception & Psychophysics*, **54**(3), 370–382.

HIRSCHBERG, J., & NAKATANI, C. H. (1996). A prosodic analysis of discourse segments in direction-giving monologs. *Proceedings of International Congress of Speech and Language Processing*, 286–293.

HUGGINS, A. W. F. (1972). Just noticeable differences for segment duration in natural speech. *Journal of the Acoustical Society of America*, **51**, 1270–1278.

KELLER, E., ZELLNER, B., WERNER, S., & BLANCHOUD, N. (1993). The prediction of prosodic timing: Rules for final syllable lengthening in French. *Proceedings of the European Speech Communication Association (ESCA) Workshop on Prosody*, 212–215. Lund, Sweden.

KLATT, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal Phonetics*, **3**, 129–140.

KOWAL, S. O'CONNELL, D. C. O'BRIEN, E. A., & BRYANT, E. T. (1975). Temporal aspects of reading aloud and speaking: Three experiments. *American Journal of Psychology*, **88**(4), 549–569.

LADD, D. R. (1988). Declination "reset" and the hierarchical organization of utterances. *The Journal of the Acoustical Society of America*, **84**(2), 530–544.

LEHISTE, I. (1973). Phonetic disambiguation of syntactic ambiguity. *Glossa*, **7**(2), 107–121.

LEHISTE, I. (1975a). The phonetic structure of paragraphs. In A. Cohen & S. Nooteboom (Eds.), *Structure and process in speech perception* (pp. 195–206). Heidelberg, Germany: Springer-Verlag.

LEHISTE, I. (1975b). The role of temporal factors in the establishment of linguistic units and boundaries. In W. U. Dressler & F. V. Mares (Eds.), *Phonologica, 1972* (pp. 115–122). München-Salzburg, Germany: Wilhelm Fink Verlag.

LEHISTE, I. (1979). Perception of sentence and paragraph boundaries. In B. Lindblom & S. Oehman (Eds.), *Frontiers of speech communication research* (pp. 191–201). London: Academic Press.

LEHISTE, I., OLIVE, J. P., & STREETER, L. A. (1976). Role of duration in disambiguating syntactically ambiguous sentences. *Journal of the Acoustical Society of America*, **60**(5), 1199–1202.

LEHISTE, I., & WANG, W. S.-Y. (1977). Perception of sentence boundaries with and without semantic information. In W. U. Dressler & O. E., Pfeiffer (Eds.), *Phonologica, 1976: Innsbrucker Beiträge zur Sprachwissenschaft* (pp. 277–283). Innsbruck, Germany: Institut für Sprachwissenschaft der Universität Innsbruck.

LI, C., & THOMPSON, S. (1981). *Mandarin Chinese: A functional reference grammar*. University of California Press.

NAKATANI, L. H. O'CONNOR, K. D., & ASTON, C. H. (1981). Prosodic aspects of American English speech rhythm. *Phonetica*, **38**, 84–106.

O'MALLEY, M. H., KLOKER, O. R., & DARA-ABRAMS, D. (1973). Recovering parentheses from spoken algebraic expressions. *I.E.E.E. Transactions on Audio and Electroacoustics*, **AU–21**(3), 217–220.

RIETVELD, T., & VAN HOUT, R. 1993. *Statistical Techniques for the study of language and language behavior*. Mouton de Gruyter.

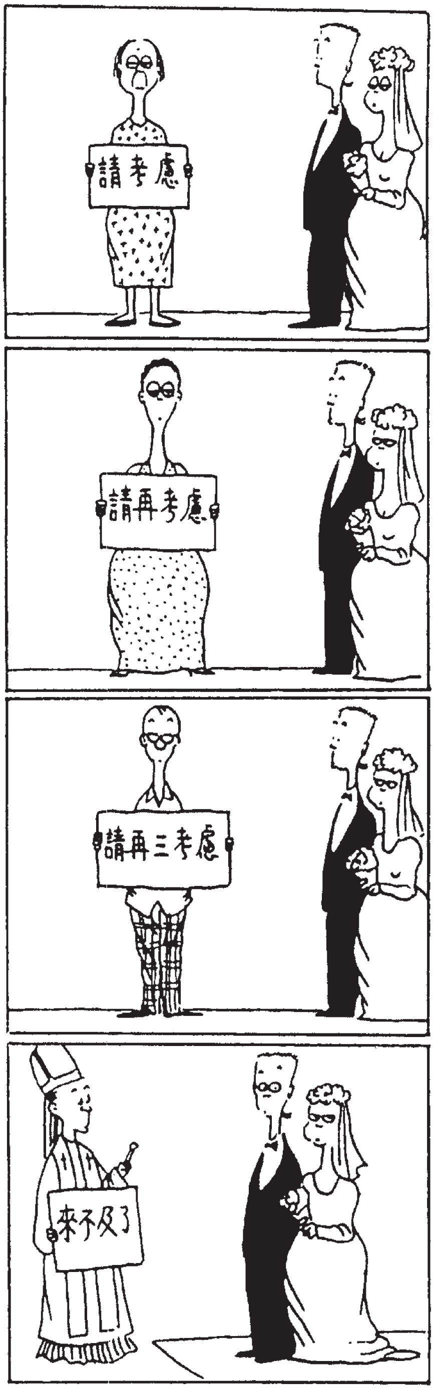
SCOTT, D. R. (1982). Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America*, **71**(4), 996–1007.

ZELLNER, B. (1994). Pauses and the temporal structure of speech. In E. Keller (Ed.), *Fundamentals of speech synthesis and speech recognition* (pp. 41–62). Chichester: John Wiley.

ZELLNER, B. (1998). Fast and slow speech rate: A characterization for French. *Proceedings of the 5th International Conference on Spoken Language Processing*, **7**, 3159–3163.

Appendix A

Comic 1



Please think about it.

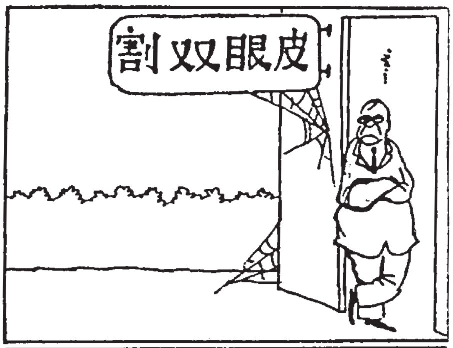
Please think about it again.

Please do think about it
once more.

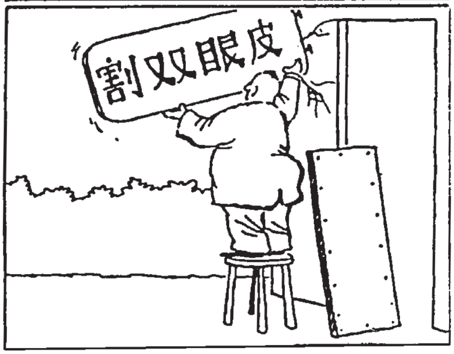
Too Late.

Appendix B

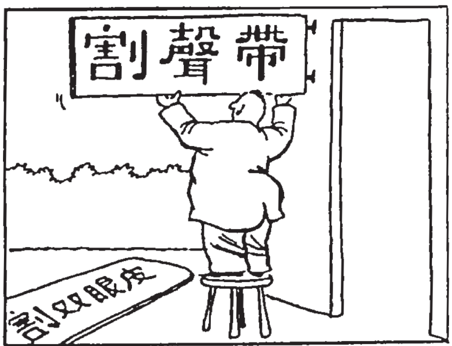
Comic 2



Double-eyelid surgery.



Double-eyelid surgery.



Vocal cord operation.

Vocal cord operation.

Appendix C

Duration of recording

Subject	Comic 1	Comic 2
Male		
CCF	00:22:933	00:47:986
LZE	00:31:967	00:17:263
LZH	00:36:630	00:24:994
LZY	01:07:608	00:38:344
ZZH	01:00:878	00:35:928
Female		
LCX	00:47:634	00:29:357
LJS	00:26:093	00:26:884
LJY	00:26:761	00:26:300
PSZ	00:38:756	00:56:844
XYL	00:23:563	00:20:444

Appendix D

Below is an example of DSP analysis in Comic 1

Subject: CCF

DSP1

wo	*xiang*	*ta-men*	*jiu*	*shi*	*haoxiang*	*zai*	*zou*	*na-ge*	*hong*	*tan*	*ma*
I	think	3sg-pl.	emphatic	be	seem	DUR[1]	walk	that-CL[2]	red	carpet	PART[3]

I think they seemed to be just walking (on) that red carpet.

DSP2

zhe-ge	*xinlang*	*gen*	*xinniang*	*yao*	*jiehun*	*le*
this-CL	groom	and	bride	want	marry	PART:CRS[4]

The bride and groom wanted to get married.

DSP3

na	*cong*	*diyi-ge*	*kaishi*	*jiu*	*chuxian*	*na-ge*	*qing*	*kaolyu*
then	from	first-CL	start	emphatic	appear	that-CL	please	consider

Then starting from the first (one), that “please consider” (sign) just appeared.

DSP4

ranhou	*zou*	*zou*	*dao*	*dier-ge*	*jiu*	*qing*	*zai*	*kaolyu*	*ma*
then	walk	walk	to	second-CL	emphatic	please	again	consider	PART

Then (he) walked to the second (one), and (it) just (said) “please reconsider.”

DSP5

na dao disan-ge jiu shi qing zaisande kaolyu

then to third-CL emphatic be please again-and-again consider

Then to the third (one) (it) was just “please reconsider again.”

DSP6

jiu shi haoxiang yijing kuai laibuji-le zhe-yangzi

emphatic be seem already almost too-late-PFV[5] this-way

(It) was as if (it was) already almost too late.

DSP7

ranhou dao disi-ge zhongyu laibuji-le zhe-yangzi

then to fourth-CL finally too-late-PFV this-way

Then to the fourth (one), finally (it was) too late.

DSP8

ta-men zou dao zheli le

3sg-pl. walk to here PART:CRS

They arrived here.

Appendix E

Below is an example of DSP analysis in Comic 2

Subject: LJY

DSP1

you yi-ge yisheng

have one-CL doctor

There was a doctor.

DSP2

ta shi zhuanmen zai ge shuang-yanpi de

3sg be specialize DUR cut double-eyelid NOM[6]

He was one who specialized in double-eyelid operation.

DSP3

danshi ne ta deng-le you deng zongshi deng-bu-dao
but PART 3sg wait-PFV again wait always wait-NEG[7]-arrive

keren shangmen
customer patronize

But, he waited and waited, and no customer patronized the clinic.

DSP4

shenzhi ta-de zhaopai dou zhang-chu zhizhu-wang le

even 3sg-GEN[8] shop-sign also grow-appear spider-web CRS

There were even spider webs on his shop sign.

DSP5

suoyi ta jueding lingji-yi-dong

so 3sg decide sudden-inspiration-one-move

So he decided (he) had a sudden inspiration,

ta gancui huan ge zhaopai gai cheng ge shengdai

he simply change CL shop-sign change as cut vocal-cords

He simply changed the shop sign to "vocal cord surgery."

DSP6

mashang jiu haha menting-ruo-shi a

immediately then haha courtyard-like-marketplace PART:RF[9]

immediately haha (his clinic) was like a crowded marketplace.

DSP7

xuduo-de xiansheng fenfen zhua-zhe ta-men-de laopo

many-NOM husband one-after-another grab-DUR 3sg-pl.-GEN wife

One after another, many husbands grabbed their wives

lai yao ge shengdai

come want cut vocal cords

to have (their wives') vocal cords removed.

DSP8

shenzhi ta lian zhaopai dou hai mei gua-hao ne

even 3sg even shop-sign even still NEG hang-ASP PART:Rex[10]

His sign had not even been properly hung yet.

---

[1] DUR: durative aspect (Li & Thompson, 1981).

[2] CL: classifiers.

[3] PART: particles.

[4] PART:CRS is a particle that implies the "currently relevant state." It refers to a state that "has special current relevance with respect to some particular situation" (Li & Thompson, 1981, p. 240).

[5] PFV: perfective aspect.

[6] NOM: nominalizer.

[7] NEG: negation.

[8] GEN: genitive.

[9] PART:RF is a particle that implies "reduced forcefulness'. It softens the query/statement (Li & Thompson, 1981).

[10] PART:Rex is a particle of "response to expectation." It points out "to the hearer that the information conveyed by the sentence is the speaker's response to some claim, expectation, or belief on the part of the hearer" (Li & Thompson, 1981, p. 300).

Appendix F

Number of SOIs each subject contributed

Subject	*Comic 1*	*Comic 2*	*Total*
Male			
CCF	103	195	298
LZE	141	94	235
LZH	125	82	207
LZY	223	139	362
ZZH	251	156	407
Female			
LCX	221	147	368
LJY	113	110	223
LJS	87	84	171
PSZ	129	123	252
XYL	151	134	285
Total	1544	1264	2808

Appendix G

Tables G-1, G-2, and G-4 show the number of units contributed by each speaker at different structural levels. Table G-3 shows the distribution of NPs with regards to case.

TABLE G-1

Distribution of 2402 SOIs in the DSP analysis. Only DSPs containing non-narration-final SOIs were included in the table. The numbers in the parentheses after each subject indicate the number of DSPs each subject contributed. There were in total 124 DSPs

		DSP Positions						
		Initial		***Medial***		***Final***		
Subject		*Comic 1*	*Comic 2*	*Comic 1*	*Comic 2*	*Comic 1*	*Comic 2*	***Total***
Male								
CCF	(14)	7	7	83	132	7	7	243
LZE	(12)	7	5	123	80	7	5	227
LZH	(9)	5	4	70	44	5	4	132
LZY	(13)	6	7	209	112	6	7	347
ZZH	(14)	7	7	205	107	7	7	340

Female								
LCX	(13)	6	7	157	97	6	7	280
LJS	(9)	3	6	71	56	3	6	145
LJY	(11)	5	6	93	85	5	6	200
PSZ	(14)	7	7	102	99	7	7	229
XYL	(15)	7	8	115	114	7	8	259
Total		60	64	1228	926	60	64	2402

TABLE G-2

Distribution of 2355 SOIs in the clause analysis. Only clauses containing non-narration-final SOIs were included in the table. The numbers in the parentheses after each subject indicate the number of clauses each subject contributed. There were in total 193 clauses

		Clausal Positions						
		Initial		***Medial***		***Final***		
Subject		*Comic 1*	*Comic 2*	*Comic 1*	*Comic 2*	*Comic 1*	*Comic 2*	***Total***
Male								
CCF	(22)	10	12	77	155	10	12	276
LZE	(21)	13	8	111	74	13	8	227
LZH	(13)	8	5	75	42	8	5	143
LZY	(24)	15	9	180	104	15	9	332
ZZH	(16)	7	9	152	124	7	9	308
Female								
LCX	(29)	18	11	147	99	18	11	304
LJS	(9)	3	6	40	56	3	6	114
LJY	(16)	6	10	77	77	6	10	186
PSZ	(22)	11	11	94	91	11	11	229
XYL	(21)	12	9	93	101	12	9	236
Total		103	90	1046	923	103	90	2355

TABLE G-3

Distribution of cases of 1327 SOIs in trisyllabic-or-longer, non-narration-final NPs in the corpus. There were in total 221 such NPs. The numbers in parentheses after each case indicate the number of instances

		Phrasal Positions						
		Initial		*Medial*		*Final*		
Case		*Comic 1*	*Comic 2*	*Comic 1*	*Comic 2*	*Comic 1*	*Comic 2*	***Total***
Nom.	(52)	24	28	83	75	24	28	262
Acc.	(78)	39	39	160	129	39	39	445
Obl.	(15)	12	3	32	9	12	3	71
Exist.	(57)	29	28	137	135	29	28	386
ADVP	(19)	12	7	100	25	12	7	163
Total		116	105	512	373	116	105	1327

TABLE G-4:

Distribution of 1093 SOIs in trisyllabic-or-longer nominative, accusative, and existential NPs in the phrase analysis. Only phrases containing non-narration-final SOIs were included in the table. The numbers in the parentheses after each subject indicate the number of NPs each subject contributed. There were in total 187 NPs

	Phrasal Positions						
	Initial		*Medial*		*Final*		
Subject	*Comic 1*	*Comic 2*	*Comic 1*	*Comic 2*	*Comic 1*	*Comic 2*	***Total***
Male							
CCF (19)	5	14	16	65	5	14	119
LZE (16)	9	7	38	28	9	7	98
LZH (16)	10	6	26	29	10	6	87
LZY (20)	11	9	47	42	11	9	129
ZZH (26)	11	15	84	49	11	15	185
Female							
LCX (26)	14	12	46	37	14	12	135
LJS (12)	6	6	21	16	6	6	61
LJY (18)	10	8	43	20	10	8	99
PSZ (18)	7	11	30	35	7	11	101
XYL (16)	9	7	29	18	9	7	79
Total	92	95	380	339	92	95	1093