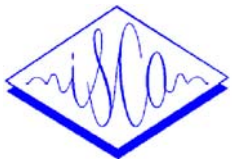

Speech Prosody 2004
Nara, Japan
March 23-26, 2004

Perception of Discourse Boundaries by Taiwan Mandarin Speakers

Janice Fon

Department of English
National Taiwan Normal University, Taiwan
jfon@cc.ntnu.edu.tw

Abstract

This study looks at whether Taiwan Mandarin speakers were able to detect discourse boundary cues in Mandarin (Guoyu and Putonghua), English, and Japanese. Results showed that there was a distinct language effect. Mandarin boundaries were harder to detect than English and Japanese boundaries for these listeners. This is thought to be due to the different boundary cue compositions in the four languages, as the magnitude of Mandarin boundary cues was not as strong as that of English and Japanese (Fon, 2002). However, the perceptibility difference disappeared once listeners became more familiar with the stimuli. Motor preparedness and subjects' expectation also played a role in determining RT.

1. Introduction

Segmentation is very important in language comprehension. Unlike written English, and many other written languages using some form of phonetic alphabets, where words are conveniently separated by spaces, sentences by periods, and paragraphs by paragraph breaks, spoken languages are more like ancient Chinese writings, where both word and sentential boundaries are left unmarked. Nonetheless, human beings seem to process auditory information with graceful ease. An average listener hardly has any problem identifying words, sentences, and even topic boundaries upon hearing a stretch of continuous speech. Although one might argue that through word identification and lexical access, the segmentation problem is virtually nonexistent, much is still left unexplained.

First of all, there is no one-to-one mapping between pronunciation of a word and the word itself. Using the Buckeye Speech Corpus, which consists of 300,000 words recorded from interviews of 40 speakers, Raymond et al. (2001) showed that a word as simple as *and* can have as many as more than 30 pronunciations. If more than two dozens of pronunciations can be associated with a simple *and*, the order of complexity would be astronomical for any sentence of average length. Even if all the tokens are stored in one's mental lexicon under the entry *and*, this segmentation-by-lookup method would still not be as efficient as it needs to be.

Secondly, people with little or no mental lexicon also seem to be capable of segmenting speech, albeit in perhaps a rudimentary fashion. Studies have shown that infants as young as 7½ months old can segment out at least some words in a sentential context (Jusczyk, Houston, & Newsome, 1999). In addition, anecdotal stories on non-speakers of a language being able to tell above chance level where a topic ends and another begins are not uncommon.

Previous studies showed that discourse boundary cues are fairly robust to a native ear. Swerts & Geluykens (1993), for example, asked Dutch speakers to listen to filtered Dutch spontaneous speech and indicate where a major discourse unit ended. They found that listeners were able to detect discourse boundaries at above chance level even when only prosodic or rhythmic information was available. A related study by Swerts, Collier, & Terken (1994) showed that listeners were especially accurate in distinguishing discourse-final, -prefinal, and -nonfinal clauses. It seems that they used both local (e.g., final lengthening) and global cues (e.g., overall pitch contour) to achieve this feat (Geluykens & Swerts, 1994).

However, subjects in the Dutch studies were all mature native speakers. Therefore, it is not very surprising that they were able to utilize boundary information in discourse segmentation even when segmental information was wiped out and some cues were altered. It is unclear from these studies whether such ability also exists in non-speakers and non-native speakers. Since all human infants begin their journey of language acquisition as non-speakers, and since many people acquire more than one language throughout their lives, it should be essential for human beings to possess some kind of ability to segment languages that are foreign or semi-foreign to them in addition to their native languages.

This study thus hopes to investigate whether and how well Taiwan Mandarin (hereafter Guoyu) speakers are able to detect discourse boundaries of Mandarin, English, and Japanese. Since English is taught in high schools in Taiwan, almost all Guoyu speakers know some English. On the other hand, although Japanese is also a commonly studied foreign language among students in Taiwan due to the popularity of Japanese TV dramas, none of the subjects in this study knows any Japanese. As to Mandarin, two major dialects, Guoyu and Putonghua (i.e., Mainland Mandarin), were included. Due to the gradual open-up of the Mainland market and government policies, people in Taiwan have more access to Putonghua than before via travel and mass media. Since the two dialects are mutually intelligible but distinct in prosody and rhythm, it would be interesting to see whether Guoyu speakers can also detect a variety of Mandarin that is somewhat different from their own.

There are three research questions that this study wishes to address. First of all, this study plans to investigate whether it is language familiarity or the richness of phonetic cues provided by the language that influences the perceptibility of discourse boundaries. If it is the former, then Guoyu speakers should discern discourse boundaries in their native language (Guoyu & Putonghua) faster than those in their nonnative language (English). Boundaries in their non-speaking language (Japanese) should be detected the slowest. Of the two dialects of Mandarin, boundaries in the native dialect (Guoyu) should be detected faster than those in the nonnative dialect (Putonghua). On the other hand, if richness of cues is the main determining factor, then English and Japanese boundaries should be detected faster than the two Mandarin dialects, as the former languages show more distinct boundary cues than the latter (Fon, 2002). Secondly, since Fon showed

10.21437/SpeechProsody.2004-164

that the four languages have different ways of coding discourse boundaries of different sizes, it would be interesting to also look into whether and how discourse boundary sizes may influence perceptibility. In other words, would bigger boundaries be perceived faster than smaller ones by default or would perceptibility depend mainly on cue composition regardless? Finally, this study would also like to see whether the degree of perceptibility can be improved by extended exposure. That is, if there is a difference in detecting boundaries of different languages due to either familiarity or differential richness in boundary cues, would this effect be mitigated once listeners become more accustomed to the stimuli? If so, how much exposure is needed?

2. Methods

2.1. Subjects

26 college students participated in the experiment. All of them were native Guoyu speakers and were born and raised in the Taipei Metropolitan area. The main language in their families was Guoyu and Guoyu was also their everyday language. All of the subjects had no experience with foreign languages before three, and on a scale of 1 (very disfluent) to 7 (very fluent), their ratings of their English skills were 4 or below. This is to control for their familiarity with English, as many people in Taipei have experiences of staying in an English-speaking country for an extended period of time. None of the subjects have studied Japanese as a foreign language.

2.2. Stimuli

Stimuli of four languages—English, Guoyu, Putonghua, and Japanese—were picked from a corpus from Fon (2002). Each stimulus was of 2.5 s long containing (parts of) two clauses with a discourse boundary in-between. The discourse boundary can be a major (Discourse Boundary Index 2, hereafter DBI2), minor (DBI1), or a potential but not realized boundary (DBI0). For detailed definition of the DBIs, please refer to Fon. In order to make the location of the discourse boundaries more variable so as not to create an expectation effect, boundaries were placed at 30% (0.75 ms), 40% (1 s), 50% (1.25 s), 60% (1.5 s), and 70% (1.75 s) of the total duration. To avoid subjects taking advantage of the semantic knowledge of the languages, the stimuli were low-pass filtered at 500 Hz, leaving only prosodic information available.

2.3. Equipment and software

E-prime 1.1 and its accompanying button box Model #200a were used to collect the reaction time (RT) data. Subjects wore Sony MDR-7502 headphones for the experiment.

2.4. Procedure

Subjects were seated in a quiet room and were told that they were to hear stretches of distorted speech over the headphones. Each stretch of speech contained parts of two sentences and they were required to press a designated button on the button box as quickly as possible when they heard the end of the first sentence. In the beginning of each stimulus, there was a warning beep. Each stimulus was repeated three times. The stimuli were blocked by language and each session contained four blocks. Subjects were allowed to rest between blocks. The order of presentation within each block and the order of the blocks were randomized for each session. Subjects were asked to come in four times within a week to repeat the process. However, they were not told that the four sessions were repetitions beforehand. On average, a session lasted about 40 min. Subjects were paid for their efforts.

2.5. Measurement

RTs were measured from the end of the first clause. To avoid outliers, RTs beyond 3 *SD* from the overall average were excluded from further analyses.

3. Results

Three planned analyses were done to test whether subjects can detect discourse boundaries and if so, whether this ability is influenced by language differences, size of the boundary, location of the boundary, and repetitions.

3.1. Session × Language × Trial

A Session (4) × Language (4) × Trial (3) three-way repeated ANOVA was done to test whether subjects showed (1) the ability of detecting a discourse boundary, (2) a learning effect throughout the four sessions, and (3) a language effect for the four languages. Results showed that all the main effects were significant [Session: $F(2.98, 1325.93) = 3.50$, $p < .05$, $\eta^2 = .01$; Language: $F(3, 1335) = 3.18$, $p < .05$, $\eta^2 = .01$; Trial: $F(1.05, 468.92) = 118.76$, $p < .0001$, $\eta^2 = .21$]. A two-way interaction was also significant [$F(3.57, 1588.30) = 11.75$, $p < .0001$, $\eta^2 = .03$].

Figure 1 shows the average RT over the four sessions. There was a significant session effect. RT became shorter as subjects repeated the process. Post-hoc analyses showed that RT on Day 1 was significantly longer than that on Days 2 ($p < .05$) and 3 ($p = .07$).

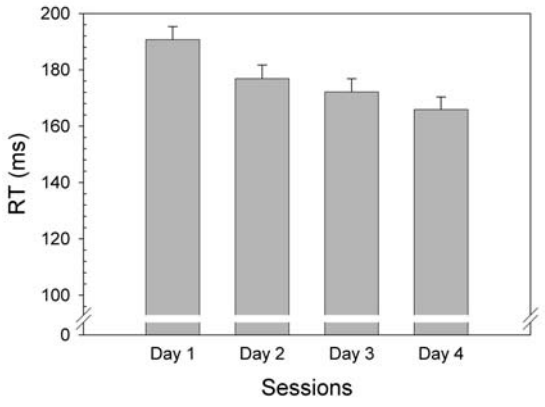

Figure 1: *Average RT for the four sessions.*

There was also a significant trial effect. As shown in Figure 2, RT was the longest for Trial 1 ($p < .0001$), but there was no significant difference between Trials 2 and 3.

An interaction effect of language and trial also existed. As shown in Figure 2, the language effect was only found for Trial 1. RTs for the two Mandarins were longer than those for English and Japanese ($p < .01$). Specifically, RT for Guoyu was the longest ($p < .01$) and that for Japanese was the shortest ($p < .0001$).

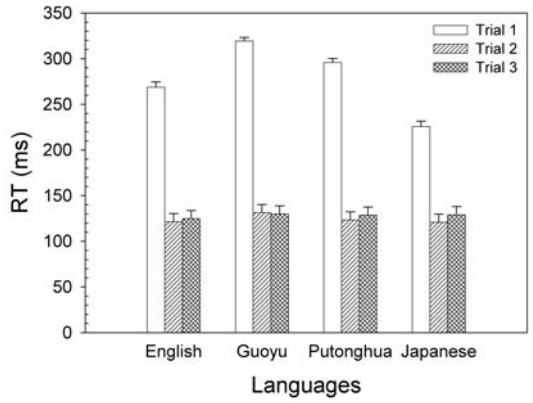

Figure 2: *Interaction effect of Language and Trial.*

3.2. Language × DBI × Trial

A Language (4) × DBI (3) × Trial (3) three-way repeated ANOVA was done to test whether DBIs of different sizes would influence the perceptibility. Results showed that two of the main effects and two of the interaction effects were (near-)significant [Language: $F(3, 2049) = 3.50$, $p < .001$, $\hat{\eta}^2 = .01$; Trial: $F(1.06, 722.44) = 147.02$, $p < .0001$, $\hat{\eta}^2 = .18$; Language × Trial: $F(3.72, 2540.59) = 14.41$, $p = .0001$, $\hat{\eta}^2 = .02$; Language × DBI: $F(6, 4098) = 1.96$, $p = .07$, $\hat{\eta}^2 = .003$]. The three-way interaction effect was also significant [$F(7.54, 5148.32) = 6.88$, $p < .0001$, $\hat{\eta}^2 = .01$].

Figure 3 shows the three-way interaction effect involving language, DBI, and trial. As shown in the figure, the size of DBI influences only RT in Trial 1. Post-hoc Tukey's-*b* test showed that for Japanese and Putonghua, DBI0 elicited the longest RT (Japanese: $p < .0001$; Putonghua: $p < .01$). In contrast, DBI0 elicited the shortest RT in Guoyu and English ($p < .05$). In addition, DBI1 in English elicited the longest RT ($p < .0001$).

3.3. Session × Position × Trial

A Session (4) × Position (5) × Trial (3) three-way repeated ANOVA was done to test whether positioning of discourse boundaries would influence perceptibility. Results showed that all of the main effects were significant [Session: $F(3, 822) = 4.17$, $p < .01$, $\hat{\eta}^2 = .01$; Position: $F(3.90, 1069.06) = 1038.56$, $p < .0001$, $\hat{\eta}^2 = .79$; Trial: $F(1.12, 305.92) = 261.70$, $p < .0001$, $\hat{\eta}^2 = .49$]. The two-way interaction effect between Position and Trial was also significant [$F(4.49, 1230.27) = 284.90$, $p < .0001$, $\hat{\eta}^2 = .51$].

As shown in Figure 4, RT became shorter as discourse boundaries came later ($p < .0001$). This was true regardless of trials. However, there was a difference between when discourse boundaries were placed at 30% of the total duration, and when they were placed at other places. At 30%, Trial 1 showed the shortest RT ($p < .05$), while at other places, Trial 1 showed the longest RT ($p < .05$).

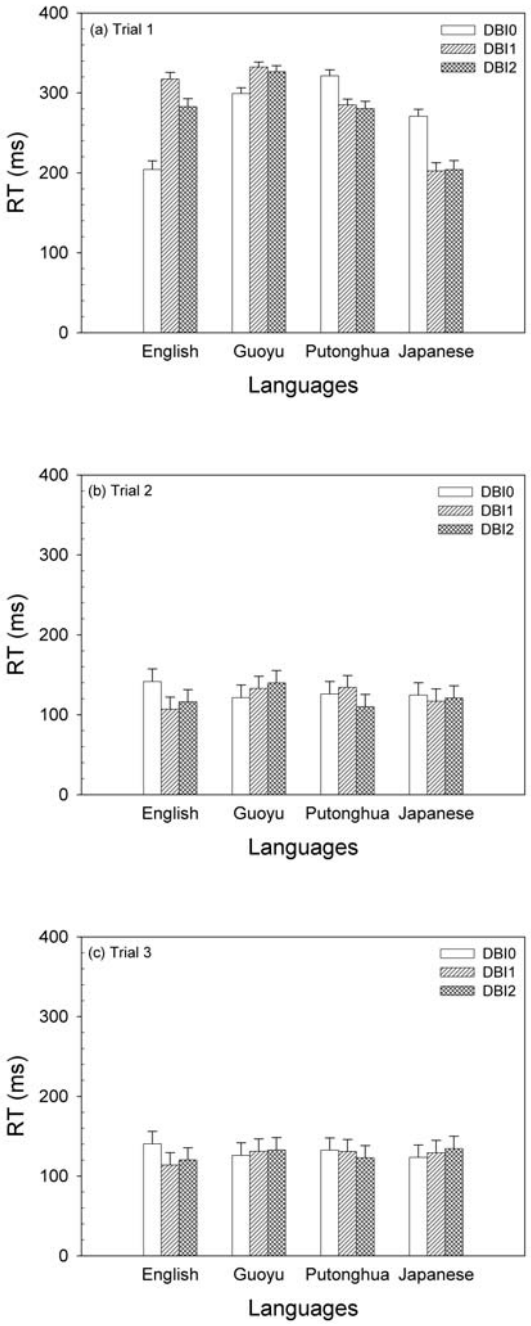

Figure 3: *Interaction effect of Language × DBI × Trial.*

4. Discussion

4.1. Language effect

Although one might expect that, as subjects were all native speakers of Guoyu, there might be a native language effect. That is, Guoyu (and thus Putonghua) boundaries might be the easiest for these subjects to detect and Japanese be the hardest. Detection of English boundaries should be somewhere in-between. However, this was not the case. In fact, the two Mandarins seemed to be the most difficult languages, Japanese was the most simple, and English was somewhere in

the middle, as shown in Figures 2 and 3a. Therefore, it seems that detection of discourse boundaries is not dependent on language familiarity, but instead on richness of boundary cues. According to Fon (2002), of the four languages, the two Mandarins show the least degree of final lengthening and the shortest pause duration. The degree of pitch reset is also small. On the other hand, Japanese shows the longest boundary pause duration and the biggest pitch reset, which might explain why listeners in this study showed faster RT with the language.

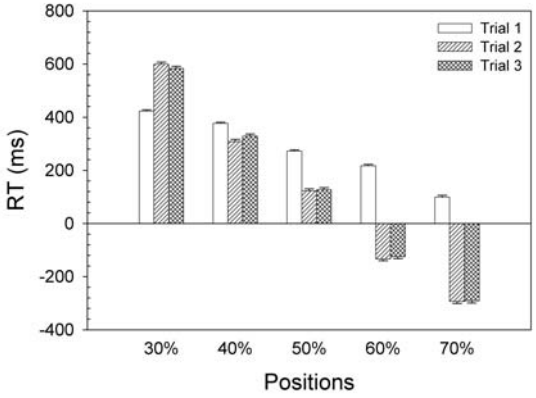

Figure 4: *Interaction effect of Position × Trial*.

4.2. Boundary size effect

As shown in Figure 3, boundary sizes only seemed to matter in Trial 1. In Putonghua and Japanese, the direction of RT went as expected. DBI0 showed the longest RT. However, this pattern was not observed in English and Guoyu. In both languages, DBI0 showed the shortest RT. It seems that the perceptibility of discourse boundaries does not necessarily correspond to the discourse hierarchy. Instead, it is fairly language-dependent. According to Fon (2002), English is peculiar in its DBIs of lower levels in that final lowering is seldom found due to a prevalent high-rising boundary tone. Guoyu also has its peculiarity in that boundary syllables with lower DBIs are actually *longer* than those with higher DBIs. Although Japanese also shows a similar pattern, having a long enough boundary pause probably mitigates this situation.

4.3. Learnability

In general, subjects showed a significant learning effect on perception of discourse boundaries. This can be demonstrated in two aspects—the session and the trial effects. As shown in Figure 1, subjects showed significant improvement on RT over the four sessions, especially between Day 1 and Day 2. In other words, the more one repeats the process, the faster one becomes in detecting discourse boundaries. Many of the subjects also mentioned that they had heard the stimuli before after the second session. The trial effect also demonstrates the learnability of boundary perception. As shown in Figures 2 and 3, there was a significant improvement on RT between Trial 1 and Trials 2 and 3, indicating that a single exposure of the stimuli was enough for subjects to estimate where the boundaries were, even for languages that have relatively weaker cues than others.

4.4. Motor preparedness & listeners' expectation

As shown in Figure 4, the positioning of DBIs made a difference on the RT. The earlier a DBI boundary was placed, the more likely its location was overestimated. A closer look at the data indicated that the cause for the patterning of Trial 1 and that of Trials 2 and 3 might be somewhat different. In Trial 1, RT was more influenced by subjects' motor preparedness. When boundaries were placed earlier in the stimuli, they were more likely to be overestimated since subjects were not prepared enough. As boundaries were placed closer to the end of the stimuli, RT became shorter since subjects were more prepared.

In Trials 2 and 3, however, subjects' motor control system should have already been well prepared since they had heard the stimuli once. Thus, the patterning could not be due to motor preparedness. Instead, it might have more to do with subjects' expectation of where the boundaries should be. In general, listeners expected the positioning of the boundary to be closer to the middle point of the stimuli than towards the beginning or the end.

That the two effects were at work can be evidenced by two peculiarities in Figure 4. First of all, underestimation of boundaries at 60% and 70% in Trials 2 and 3 could only have occurred if there had been an expectation effect. Motor preparedness alone could not have caused the underestimation. Secondly, RT patterning at 30% was quite different than that at other positions. Contrary to what one would usually expect (from the learning effect), RT for Trial 1 was the *shortest* at this position, compared to that for Trials 2 and 3. This could be explained if one assumes that during the first trial, subjects pressed the button when they thought they heard the boundary, but decided that the boundary was way too early to be possible during their second and third trials, and thus shifted the perceived boundaries to a later time. If only the motor-preparedness factor was at work, the pattern at 30% should have been no different from that at other positions. Trial 1 should always show a longer RT. Therefore, one concludes that both motor preparedness and subjects' expectation were at work during the experiment.

5. References

Fon, Y.-J. J., 2002. A Cross-linguistic Study on Syntactic and Discourse Boundary Cues in Spontaneous Speech. Dissertation. The Ohio State University.

Geluykens, R., & Swerts, M., 1994. Prosodic cues to discourse boundaries in experimental dialogues. *Speech Communication*, *15*(1-2), 69-77.

Jusczyk, P. W., Houston, D. M., & Newsome, M., 1999. The beginnings of word segmentation in English-learning infants. *Cognitive Psychology*, *39*, 159-207.

Raymond, W. D., Makashay, M. J., Dautricourt, R., Johnson, K., Hume, E., & Pitt, M., 2001, December. Variation in conversation: An introduction to the Buckeye Speech Corpus. Poster session presented at the annual meeting of the Acoustical Society of America, Ft. Lauderdale, FL.

Swerts, M., Collier, R., & Terken, J., 1994. Prosodic predictors of discourse finality in spontaneous monologues. *Speech Communication*, *15*(1 - 2), 79 - 90.

Swerts, M., & Geluykens, R., 1993. The prosody of information units in spontaneous monologue. *Phonetica: International Journal of Speech Science*, *50* (3), 189-196.