INTERSPEECH 2009 BRIGHTON

Cross-Cultural Perception of Discourse Phenomena

Rolf Carlson [1], *Julia Hirschberg* [1,2*]

[1] CTT, KTH, Sweden
[2] Columbia University, USA
rolf@speech.kth.se, julia@cs.columbia.edu
*names in alphabetic order

Abstract

We discuss perception studies of two low level indicators of discourse phenomena by Swedish, Japanese, and Chinese native speakers. Subjects were asked to identify upcoming prosodic boundaries and disfluencies in Swedish spontaneous speech. We hypothesize that speakers of prosodically unrelated languages should be less able to predict upcoming phrase boundaries but potentially better able to identify disfluencies, since indicators of disfluency are more likely to depend upon lexical, as well as acoustic information. However, surprisingly, we found that both phenomena were fairly well recognized by native and non-native speakers, with, however, some possible interference from word tones for the Chinese subjects.

Index Terms: discourse, disfluency, phrase boundaries

10.21437/Interspeech.2009-519

1. Introduction

The cross-cultural study of discourse phenomena has attracted increasing attention both from the point of view of perception and production. How are phenomena such as information structure, turn-taking cues, and speaker state conveyed in different cultures? How are such phenomena perceived when the listener is from a different culture? Both of these have implications for cross-cultural communication as well as second language learning. In this paper we examine two low level phenomena associated with information structuring, turn-taking behavior and speaker state – the perception of prosodic boundaries and of disfluencies.

In earlier studies we found that listeners could detect the presence or absence of fluent prosodic boundaries in their native language and in a language they did not speak but which is phonologically close to their native language with considerable accuracy [1,2]. These studies tested the hypothesis that speakers not only encode prosodic breaks locally at the places where they occur (e.g. in the form of silent pauses), but that they also signal these breaks in advance. We found that native listeners of Swedish could perceive upcoming boundaries even in the absence of pause information, and furthermore that native speakers of a related language (English) could also perceive these, even in the absence of possible lexical and syntactic cues. Significant correlations for several f0 features were found for successful predictions for both groups.

Studies of speech disfluencies differ in their findings about native and non-native perception, with some studies finding little accuracy even for native speakers [3] and others finding considerable accuracy for native speakers asked to monitor for filled pauses and fragments, somewhat less accuracy for non-native monitoring of a related language, and still less for non-native monitoring of an unrelated languages [4].

In this paper, we will report results of studies of non-native identification of upcoming phrase boundaries and of disfluencies, where the non-native subjects' first language (Japanese and Chinese) is prosodically quite different from the language material being judged (Swedish) and compare these to our prior findings for Swedish and American native speakers. We also compare Swedish and Chinese native speakers' ability to identify disfluencies in Swedish material as a further test of native vs. non-native perception of discourse phenomena. We hypothesize that speakers of prosodically unrelated languages should be less able to predict upcoming phrase boundaries but potentially better able to identify disfluencies, since indicators of disfluency are more likely to depend upon lexical, as well as acoustic information.

2. Upcoming Boundary Identification

For our boundary identification study we presented spontaneous Swedish utterance fragments to listeners and asked them to judge whether or not each fragment was followed by a prosodic boundary; for hypothesized boundaries, we asked them to rate boundary strength on a scale from 1 to 5.

2.1. Speech Stimuli

The stimuli we used for these experiments were the same as those used in our previous studies, chosen from a Swedish Radio interview given by a female politician, which was prosodically labeledby three independent labelers [5,6] for presence and strength of boundary from listening alone and disagreements were resolved by majority vote. From this corpus, 58 utterance fragments (each about 2 seconds long) were selected, approximately one third from locations where our labelers found a strong prosodic boundary at the end of the fragment, about one third from contexts where the labelers found a weak boundary ending the fragment, and the rest from contexts with no boundary after the fragment. All fragments preceded the word "och" (and) in their original context, and were excised just before the silent interval (if any) preceding "och". The initial cut point was placed at the nearest word boundary occurring 2 seconds before the final cut point. From each of these 2-second fragments, we also constructed a short version, consisting of only the final word of the fragment. Thus, we used a total of 116 tokens for the perception experiments.

2.2. Subjects

Our Swedish subjects (SW) had included 13 students of logopedics from Umeå University and our American subjects (AM) had consisted of 29 staff and students at Columbia University, USA, all native speakers of standard American English with no knowledge of Swedish. Our Chinese subjects

(CH) were 8 students from the Chinese University of Hong Kong, all of whom we presume had some knowledge of English, half native speakers of Cantonese and half of Mandarin. Our Japanese speakers (JPN) were 12 subjects from University of Tokyo. We report on the Chinese and Japanese subjects' performance below and compare them to Swedish and American performance on the same data.

2.3. Method

Our method for this experiment is the same as that reported in (Carlson et al 2005). We randomized our 116 stimuli (long and short fragments, preceding a strong boundary, weak boundary or no boundary) and presented them sequentially to our listeners via a GUI interface, which allows us to run perception experiments over the internet using a standard web browser with audio facilities. To minimize potential learning effects, each subject was presented with a differently randomized list of stimuli. Subjects' task was to rate each stimulus on a 5-point scale from 'no boundary at all follows this fragment' (1) to 'a strong boundary follows this fragment' (5). Subjects were first given a short introduction briefly explaining concepts such as prosodic boundary as well as the task. Three examples tokens were presented in the introduction to the listeners, a 2-second fragment: när man tog avstånd naturligtvis .. (when you looked at it from a distance of course ..); a long word: paragrafen .. (the paragraph ..); a short word: den .. (it ..)

During the experiment subjects could listen as many times as they wished to a given stimulus before making a judgment, but they could not return to a previous stimulus after they had entered a response.

Introduction:
Thank you for participating in this test. We are studying how people make "breaks" between words. For example, speakers can put pauses in their utterances or can signal otherwise that there is some boundary between two consecutive words.
In this experiment, you will be presented with spoken utterance fragments (Swedish) that are either 2 seconds long or that consist of only one word.
These fragments could look like this:
2-second fragment: när man tog avstånd naturligtvis ..
long word: paragrafen ..
short word: den ..
In this experiment we would like you to judge on how strong a break will follow these fragments, for instance after the words "naturligtvis", "paragrafen" and "den" in the examples above. You will need to express your judgment on a 5-point scale. If you think there will be strong break after the last word, then you respond with 5. If you feel there will be no break after the last word, then you respond with 1. The rest of the scale you can use to mark the in between categories. We ask you to always give an answer, even if you are unsure about your answer.

2.4. Results of Perception Judgments

Results of the perception experiments with Chinese and Japanese speakers show that, while Japanese speakers could make decisions about upcoming boundaries from both the short and the long fragment stimuli that were not significantly different in accuracy from Swedish and American raters (Tukey HSD post hoc tests), Chinese speakers' judgments differed from these significantly (Tukey, $p<.025$). This difference is localized to the single word condition and the judgment of strong boundaries. In all other conditions the Chinese judgments are not significantly different from the Swedish, American, and Japanese. Japanese ratings show no significant differences from Swedish and American in any condition. Figures 1 and 2 compare Chinese and Japanese subject judgments for one word and 2sec fragments with our previously-reported Swedish and American judgments.

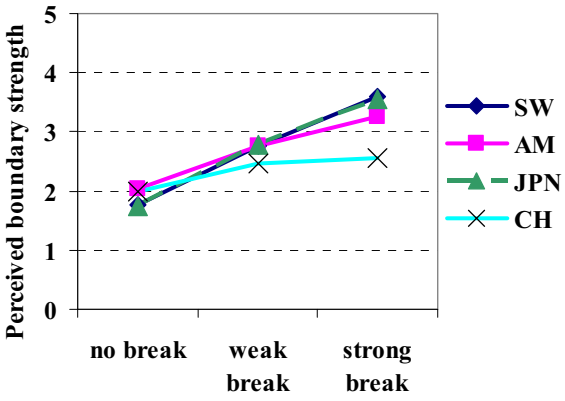

Figure 1: *Mean perceived upcoming boundary strength for one-word fragments for Swedish (SW), American (AM), Japanese (JPN) and Chinese (CH) subjects.*

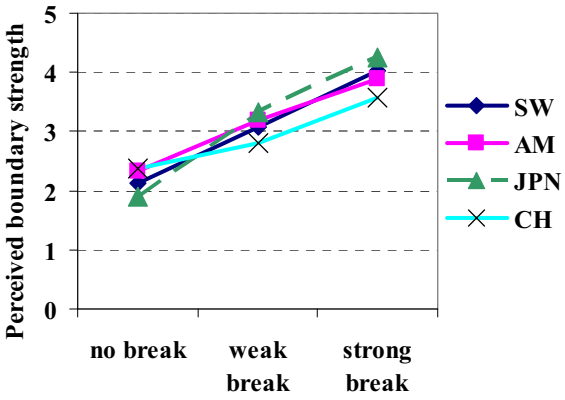

Figure 2: *Mean perceived upcoming boundary strength for 2sec fragments for Swedish (SW), American (AM), Japanese (JPN) and Chinese (CH) subjects.*

A repeated-measures ANOVA with between subjects factors of boundary type (none, weak, strong) and fragment size (1 word, 2s phrase) shows significant main effects for both ($F(2,120)=860$, $p<.001$; $F(1,120)=238$, $p<.001$). There is a significant interaction between boundary type and fragment size ($F(2,120)=33.70$, $p<.001$) and between boundary type and native language ($F(6,120)=33.80$, $p<.001$), although not between fragment size and native language. However, the interaction between type, size, and native language is due to a difference between Chinese speakers vs. other subjects in their ratings of strong, 1 word boundaries.

To isolate possible features of the stimuli which might account for those differences in subject judgments we found, we examine some possible acoustic and prosodic cues.

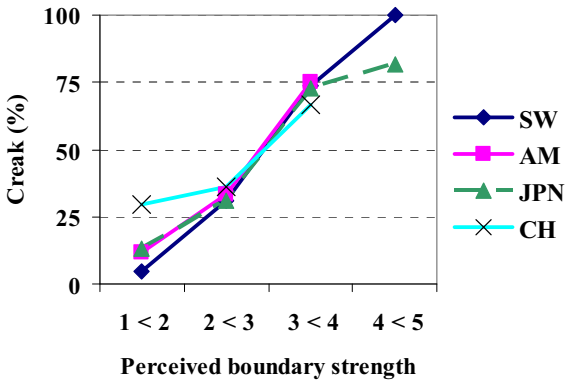

Figure 3: *Number of stimuli with creaky voice (in %) for different judged boundary strength intervals (one word). No American and Chinese data with a mean higher or equal 4 was found and thus the corresponding bar is missing.*

2.5. Acoustic and Prosodic Correlates

We had previous found that the presence of final creak (glottalization), median F0 for the last voice 50ms of the final word, phrase-final F0 slope during the same 50ms appeared to influence Swedish and American subjects' judgments about boundary strength, with final creak (Figure 3), lower median F0, and falling slope correlating with stronger boundaries. However, we did not find any strong correlation between final lengthening and boundary judgments. For the Japanese and Chinese raters only the Japanese subjects showed a significant dependency on the intonation cues, as shown in Table 1. We hypothesize that this may account for their greater similarity of judgments to the American and Swedish judges.

Table 1. *Regression analysis of subject judgments and F0 median and slope in the final 50 ms of the stimuli.*

	SW	AM	JPN	CH
F0 median	r =0.62 p < .01	r =0.43 p < .01	r =.45 p < .01	r =.17 p < .18
F0 slope	r =0,51 p < .01	r =0,49 p < .01	r =.47 p < .01	r =.18 p < .17

3. Disfluency Detection Experiment

For our disfluency detection studies we presented native speakers of Swedish and Chinese with fluent and disfluent utterances selected from a spontaneous conversation between two Swedish native speakers.

3.1. Speech Stimuli

Stimuli were chosen from conversations between male and female native speakers of Swedish. A trained labeler had labeled instances of disfluency in the conversations, including filled pauses, prolongations, and self-repairs. We chose 20 fluent and 20 disfluence (containing one or more disfluencies) from the male speaker's speech, subsequently recognizing that one of the 'fluent' utterances indeed might be labeled disfluent. So our final materials for the experiments included 21 disfluent and 19 fluent utterances.

3.2. Subjects

Subjects were 8 native speakers of Swedish including students at KTH and 12 Chinese speakers including faculty and students from the Chinese University of Hong Kong, 5 native speakers of Cantonese and 7 native speakers of Mandarin.

3.3. Method

Subjects were presented with the (mixed) fluent and disfluent phrases in a GUI interface with associated speech files that they could download on their own PCs. Subjects could see the speech waveform and play the speech as often as they wished. They were also given the ability to pause and start the playing at any point in the file. They were asked to mark the location of any disfluency they found, at the beginning of the disfluent speech. They then saved and sent their judgments to the experimenters in email. A short introduction briefly explained the concepts and the task.

Introduction
Sometimes, when people speak, some of their speech seems disfluent. They may hesitate or change their minds about what they are going to say. For this experiment, we are interested in finding out what people think is fluent vs. disfluent speech. We will ask you to listen to a small number of sentences which have been spoken by different people. Some sentences are fluent and some contain disfluencies. We ask you to mark each place in a sentence where you think the speaker is beginning a disfluent segment of speech. Here is what you should do to participate in the experiment:
.....
ii. Start at the top file in the left window. For each file, please do the following:
1. Play the file as often as you like. You may stop/start the output using the "space" bar.
2. When you believe you have heard a disfluency in the speech, mark the place where that disfluency begins, as best you can, by right-clicking on the ruler at the point where the disfluencies starts. A small triangle will appear on the ruler. Do this for each disfluency you hear in the speech. If you hear no disfluencies, don't mark anything in this file.
.....

3.4. Results of Perception Judgments

Figure 4 shows native and non-native perception of Swedish disfluencies. The first pair of bars on the left shows the percentage of fluent utterances identified by each language group. While the difference is not significant (Welch's t test, t=1.51, p=.15) the Chinese subjects nonetheless identify fewer of the fluent utterances as fluent, with a standard deviation of 4.45 compared to the Swedish subjects' 2.30. The second pair of bars **does** show a significant difference in Chinese subjects' perception of disfluent utterances as disfluent (t=3.30, p<.005). The third pair of bars compares how accurately each set of subjects could locate the disfluencies in disfluent utterances. Again, this difference is significant (t=2.89, p<.01), with non-native speakers having more difficulty than native speakers in identifying the position of disfluencies in the utterance. We conclude from these comparisons that Chinese subjects performed similarly to Swedish native speakers in identifying utterances as fluent, but differed from them in their ability to identify disfluent utterances as disfluent and in their ability to identify the location of disfluencies in such utterances correctly.

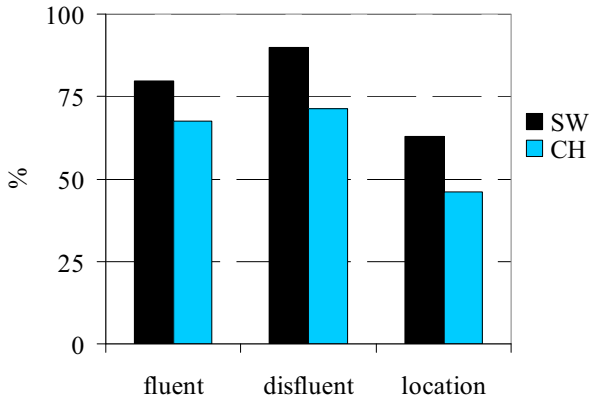

Figure 4: *Perceptual result of judging fluent, disfluent phrases overall and of locating disfluencies.*

We next examined the disfluencies identified by Swedish and Chinese speakers in regions which were **not** considered so by trained labelers (i.e. *insertions*). We noted that, over all, Chinese speakers hypothesized many more disfluencies than Swedish speakers. Figure 5 shows hypothesized vs. 'correct' disfluency labels for both sets of raters. The difference in mean number of disfluencies markings between the groups is in fact not significant (t=1.31, p=.21), although the Chinese speakers on average (mean=23.42) did hypothesize more disfluencies per utterance than the Swedish subjects (mean=15.13). However, the Swedish standard deviation is only 4.36 while the Chinese subjects' is 21.19, reflecting the presence of several outliers among the Chinese subjects. If we exclude these subjects, it would appear that, while the Chinese raters found fewer disfluencies than did the Swedish raters, they did not produce a larger number of false hypothesies.

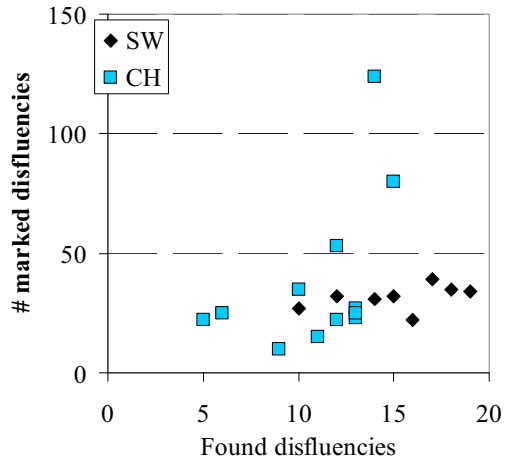

Figure 5: *Hypothesized disfluencies vs. 'correct' disfluencies.*

4. Conclusion

We have presented results of a series of perception studies of two phenomena important in the production and interpretation of discourse, prosodic boundaries and disfluencies, when observed by native speakers and by non-native speakers with different language backgrounds. We have found that, for the perception of upcoming boundaries, Swedish, American and Japanese subjects perform very similarly when judging Swedish data. However, Chinese speakers' performance differs, in particular in the perception of strong upcoming boundaries when given single words as input; their performance when given longer preceding context is comparable to the other language groups. We hypothesize that the difference observed in single word judgments may be due to an interference of word tone in Mandarin and Cantonese with the function of F0 slope as a cue to upcoming boundaries. For the other groups, F0 strongly correlates with boundary decisions. The overall success of non-native speakers in predicting upcoming boundaries indicates to us that there is a basic level of acoustic information available to all groups, given sufficient context and absent the interference of other functions of that information. Our study of Swedish and Chinese identification of disfluencies in Swedish data confirms this overall conclusion, since both groups perform similarly on the identification of utterances as fluent, while differing somewhat on the perception of disfluent utterances. This difference appears to be due to a few Chinese raters who tended to hear many more disfluencies than did their fellow Chinese or the Swedish raters. In future work we will examine in more detail the potential acoustic and lexical sources of this difference.

5. Acknowledgements

We thank the following researchers for their cooperation and support: Marc Swerts, University of Tilburg; Helen Meng and Tan Hee, Chinese University of Hong Kong; Keikichi Hirose, University of Tokyo; Anna Hjalmarsson and Samer Al Moubayed, KTH.

6. References

[1] Carlson, R., Hirschberg, J., & Swerts, M. (2004). Prediction of upcoming Swedish prosodic boundaries by Swedish and American listeners. In Bel, B., & Marlin, I. (Eds.), Proc of Intl Conference on Speech Prosody 2004 (pp. 329-332). Nara, Japan.

[2] Carlson, R., Hirschberg, J., & Swerts, M. (2005). Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. Speech Communication, 46, 326-333

[3] Lickley, R., and Bard, E. (1996): "On not recognizing disfluencies in dialogue", In ICSLP-1996, 1876-1879

[4] Lai, C., Gorman, K., Yuan, J. and Liberman, M. (2007): "Perception of disfluency: language differences and listener bias", Proc. INTERSPEECH-2007, 2345-2348.

[5] Carlson R., Granström B., Heldner M., House D., Megyesi B., Strangert E., Swerts M., 2002. Boundaries and groupings - the structuring of speech in different communicative situations: a description of the GROG project. Proc of Fonetik 2002, TMH-QPSR, 44.

[6] Heldner M., Megyesi B., 2003. Exploring the prosody-syntax interface in conversations, Proc. ICPhS 03.