Speech Prosody and Eye Movements in Processing Discourse Information: A Preliminary Study in Mandarin Chinese

Ying Chen*, Wentao Xiao, Jie Cui and Hanyu Xu
Nanjing University of Science and Technology, Nanjing, China
*E-mail: ychen@njust.edu.cn Tel: +86-17714341101
E-mail: wtxiao@njust.edu.cn Tel: +86-18801595912
Email: jiecui@njust.edu.cn Tel: +86-13140745799
Email: xhyso61819@163.com Tel: +86-18951973566

***Abstract*—This study investigates variations in speech prosody and eye movements and their potential correlations in processing discourse information of map direction in Mandarin Chinese. A production experiment was conducted to collect mean duration, F0, intensity of target words in speech prosody and fixation counts and fixation duration of target areas of interest in eye movements for statistical analyses. The results show fixation counts, fixation duration, and syllable duration of the target words decreased, syllable intensity increased, but syllable pitch remained intact as the information became old to the speaker in the discourse. Prosodic reduction of duration, F0, and intensity was found in speech repetition and in the processing of old information.**

I. INTRODUCTION

Repeated words in discourse have been found often reduced in prosodic prominence [1]. This finding is in line with Smooth Signal Redundancy Hypothesis (SSRH, [2]), which holds the viewpoint that duration, pitch, and intensity are all supposed to have a negative correlation with informational redundancy to maintain the communication process. Among the three prosodic parameters, duration has been found as the most robust predictor for utterance planning, recent auditory experience, phonological encoding of target words and the retrieval difficulty of mental lexicon [1, 3, 4, 5]. Then repetition and predictability were discovered to significantly predict not only duration but also intensity [6, 7, 8]. The overall effects of repetition on pitch and intensity have been found less consistent than duration. Nonetheless, an earlier and steeper rising pitch movement in new discourse was found as well as that the rise became smaller and starting later in old discourse in Malaysian English [9]. In contrast, mean pitch in repeated words was found higher than that in their first mentions in Papuan Malay [10]. These studies suggest that speech prosody is associated with information status and their alternation thereof.

Eye-tracking technology has been widely used to study spoken word recognition, speech comprehension, sentence processing, and also the online processing of information structure. Scenario pictures and printed words were used to track listeners' eye movements during online processing for the function of lexical stress or intonation in spoken word recognition and discourse status prediction [11, 12]. One of the most influential eye-tracking paradigms taking both speech processing and cognition into account is the Visual World Paradigm (VWP), proposed by Tanenhaus and colleagues [13]. The earliest work using VWP was designed to monitor participants' eye movements as they followed experimenter-generated spoken instructions to pick up and move objects arranged on a table to investigate word recognition, reference resolution, and syntactic processing. VWP has been adopted to investigate the role of pitch accent in online processing of information structure and reference resolution in spoken English comprehension [14, 15, 16]. On the other hand, VWP has been used to examine how Mandarin tone perception was affected by the pitch height of tones at onset, turning point and offset in Mandarin Chinese. The processing of fine-grained pitch information prior to lexical access was then discovered in this study [17].

Previous studies have employed the speech perception approach in the eye-tracking experiments, but none of them have combined eye-tracking methods with an acoustic analysis of speech production. Also, studies on repetition reduction in Mandarin Chinese remain scarce. The current study was designed to investigate online processing of information status in Mandarin via eye movements and prosodic encoding taking advantage of the remote function of the eye-tracking equipment. Two research questions are explored: (1) Is there a reduction in speech prosody as well as eye movements when processing old information vs. new information? (2) Are the patterns found in eye movements in accordance with those observed in the speech prosody during online processing of discourse information?

II. METHODS

A. Participants

Twenty-four native Mandarin speakers—12 males and 12 females, participated in the current study. All participants were born and raised in northern China and reported to have normal hearing with normal vision or corrected visual acuity.

The participants were not informed of the research purpose of the experiment.

B. Stimuli

The stimuli were presented in the form of a map-based direction task designed with a pseudo VWP. A map display consisted of four images in the four quadrants of the computer screen. Each quadrant (i.e., area of interest) contained two target words—a destination word and a distance word (see Figure 1, note that only the Chinese characters of orientation words, destination words and distance words and their corresponding images were shown in the map). In order to track a continuant F0 trajectory, syllables in these target words were designed with all sonorants except for the unavoidable voiceless unaspirated /p/ in the syllable-initial positions of *bai* 'hundred' and *ba* 'eight.'

During the experiment, participants were requested to provide oral responses to a serial of pre-recorded direction inquiries. The stimulus questions are shown in following order:

(1) *Wo jiao Zhang Wei/Wang Li, xianzai zai youeryuan, xiang qu youleyuan gei xiaopengyou mai jitipiao, gai zenme zou?*

'My name is Zhang Wei/Wang Li, now in the kindergarten. I want to go to the amusement park to buy group tickets for the kids. How do I get there?'

(2) *Cong youleyuan chulai, wo xiang qu yanglaoyuan lianxi kanwang laoren de shiyi, gai zenme zou?*

'From the amusement park, I want to go to the nursing home to consult the plan of future visits. How do I get there?'

(3) *Cong yanglaoyuan chulai, wo xiang qu meirongyuan zuo ge xinfaxing, gai zenme zou?*

'From the nursing home, I want to go to the salon for a new hairstyle. How do I get there?'

(4) *Cong meirongyuan chulai, wo xiang jinkuai huidao youeryuan, zuikuai de luxian gai zenme zou?*

'From the salon, I want to get back to the kindergarten as soon as possible. What is the fastest route?'

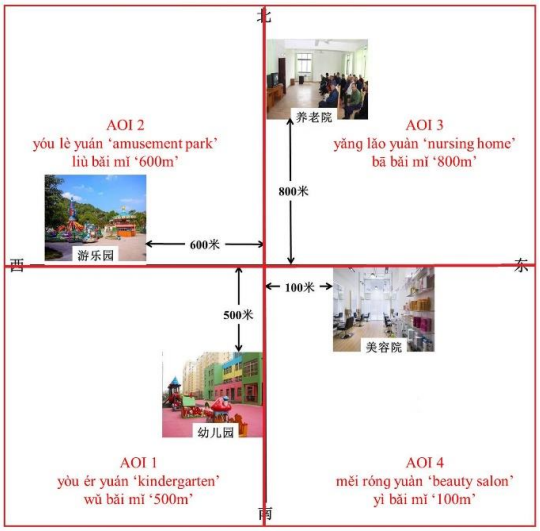

Fig. 1 The map, target words, and areas of interest.

C. Procedures

Recordings took place in the sound-attenuated booth of Language Cognition and Speech Science Lab at Nanjing University of Science and Technology. All participants were given a brief introduction of the experiment procedure and prepared for tracking their eye movements. The speech was recorded with a Shure professional head-worn dynamic microphone linked to a Marantz professional solid state recorder. All of the recordings were digitized and saved on an SD card into a personal computer. The map was displayed only after each stimulus question was played aloud. The experiment consisted of three trials. Before each trial, a nine-point calibration was conducted to ensure the reliability and stability of each participant's eye movements.

After the first trial, participants were told that the recording was not properly recorded and were requested to answer the same questions again to generate the second trial. When the second trial was finished, the participants were requested to answer the same questions but raised by a new inquirer via playing the pre-recorded stimuli of the other gender. The experiment was designed as such in order to elicit speech realizations of three types of information statuses: (1) speaker-new and hearer-new; (2) speaker-old and hearer-old; (3) speaker-old but hearer-new. As a result, three repetitions in total were recorded from each participant.

D. Analyses

In this paper, two types of data were analyzed. The eye-tracking data, including fixation duration and fixation counts of the four interest areas (see Figure 1), were extracted by the Data Viewer software of Eye-link 1000 Plus, but only the data of the two related interest areas which contain the related target words in each sentence were analyzed. The prosodic data, including mean duration, pitch (F0), and intensity of the syllables of the target words, were extracted using ProsodyPro [18]— a Praat script for prosody analysis.

All dependent variables (fixation duration, syllable duration, syllable pitch, and syllable intensity) were analyzed with a generalized linear mixed model using the lme4 and lmerTest package of R (version 3.5.0). For the fixation counts per sentence, we fitted the linear mixed effect model with a Poisson distribution. Sentence (4 levels) and Trial (3 levels) were set as fixed effects while word and participant ID were treated as random factors. Sentence 1 and Trial 1 were set as the reference levels respectively for each measured parameter. If both Trial 2 and Trial 3 differed from Trial 1, another set of regression analysis was conducted, taking Trial 2 as the reference level. Each trial includes four sentences, and each sentence involves two destination words and two distance words and thus two interest areas of eye movements.

III. Results and Discussion

A. Fixation Counts

The statistic results show that both trial and sentence had significant effects on participants' fixation counts (see Table I). A follow-up regression analysis taking Trial 2 as reference

level shows another difference between Trial 2 and Trial 3 (z = -2.664, p = 0.008). Figure 2 indicates a fairly consistent pattern across trials with more fixation counts in Sentence 1, fewer in Sentences 2 and 3, and slightly recurrent in Sentence 4.

Table I
STATISTIC RESULTS OF FIXATION COUNTS.

Predictor	Est. SD	SE	*z*	*p*
(Intercept)	2.796	0.079	35.283	0.000
Trial 2	-0.430	0.054	-7.956	0.000
Trial 3	-0.595	0.057	-10.462	0.000
Sentence 2	-0.354	0.061	-5.762	0.000
Sentence 3	-0.447	0.068	-6.578	0.000
Sentence 4	-0.235	0.056	-4.202	0.000
Trial 2: Sentence 2	0.205	0.085	2.423	0.015
Trial 2: Sentence 3	0.223	0.087	2.565	0.010
Trial 3: Sentence 2	0.320	0.087	3.667	0.000
Trial 3: Sentence 3	0.292	0.090	3.236	0.001

The decrease of fixation counts over trials reflects the familiarity of the speaker to the map information regardless of the hearer. The downward trend of fixation counts within trials can be referred back to the stimulus questions and the map design. Within each trial, Sentence 1 involved two sets of new target words (a destination word and a distance word in each) in two areas of interest (both new information); however, the target words in the 2nd interest area became second-mentioned and old information in Sentence 2 and the 3rd interest area entered into Sentence 2 as new information and so as Sentence 3 with the 3rd interest area (old) and the 4th interest area (new). On the other hand, the 4th and the 1st interest areas that are involved in Sentence 4 are both old information within trial; however, the fixation times in Sentence 4 were counted more than that in Sentences 2 and 3, which may be attributed to the "unpredictable" words *zuikuai de luxian* 'the fastest route' [19]. This key phrase in Question 4 could have attracted participants' attention and consequently costed them more efforts to work out the fastest route and thus increased the fixation counts.

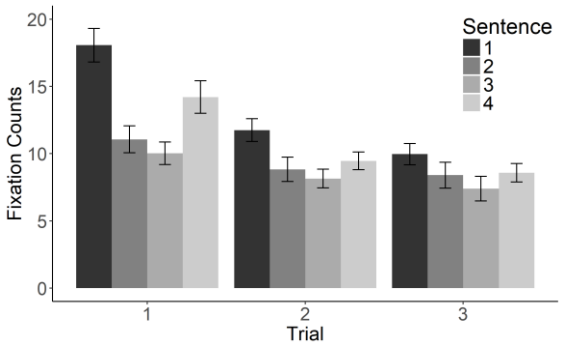

Fig. 2 Mean fixation counts by sentence and trial.

B. *Fixation Duration*

The regression results of fixation duration are listed in Table II. Sentence, as well as the interaction between sentence and trial, emerged as significant predictors. Specifically, the fixation duration for Sentence 3 was significantly longer than Sentence 1 but interacted with Trial 2 and Trial 3. The mean fixation duration of all sentences in the three trials is illustrated in Figure 3, indicating that there was no remarkable variation among the three trials. However, in Trial 1, participants tended to gaze the interest areas of Sentence 3 longer than those in Sentence 1.

Table II
STATISTIC RESULTS OF FIXATION DURATION.

Predictor	Est. SD	SE	*df*	*t*	*p*
(Intercept)	205.963	33.270	28.719	6.191	0.000
Sentence 3	115.115	43.716	192.529	2.633	0.009
Trial 2: Sentence 3	-20.500	56.503	537.959	-2.133	0.033
Trial 3: Sentence 3	-70.625	56.503	537.959	-3.020	0.003

Unlike fixation counts, the effect of trial on fixation duration was not significant, which is contrary to the previous findings in the literature that fixation duration on the target word is shorter when it becomes more predictable [20]. However, Sentence 3 stood out with longer fixation than Sentence 1 in Trial 1. One possible explanation is that the distance from the center of the map to *yanglaoyuan* 'nursing home' was longer than those to the other three destinations. Therefore, it could have taken participants more time to find and fixate on the corresponding image. This pattern was not replicated in Trial 2 and Trial 3 as participants became more familiar with the map layout. Specifically, the location of *yanglaoyuan* 'nursing home' on the map became old information in Trial 2 and Trial 3.

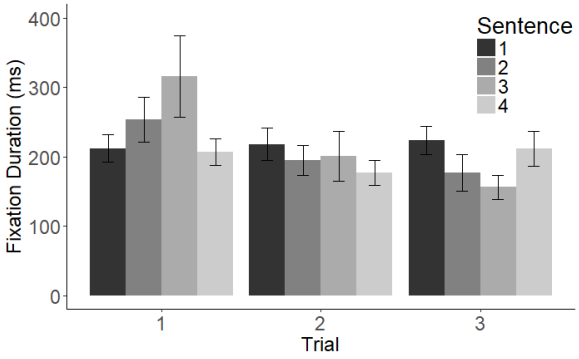

Fig. 3 Mean fixation duration by sentence and trial.

C. *Syllable Duration*

Regression analyses of syllable duration (see Table III) show significant differences between trials. A notable decrease across trials can be seen in Figure 4. Table III and Figure 4 indicate that Sentence 4 was shorter than Sentence 1 and interacted with Trial 2 and Trial 3.

Table III
STATISTIC RESULTS OF SYLLABLE DURATION.

Predictor	Est. SD	SE	*df*	*t*	*p*
(Intercept)	175.124	7.928	17.861	22.09	0.000
Trial 2	-12.276	3.733	1110.046	-3.288	0.001
Trial 3	-15.255	3.733	1110.046	-4.086	0.000
Sentence 4	-13.702	4.024	1116.291	-3.405	0.001
Trial 2: Sentence 4	11.739	5.280	1110.046	2.223	0.026
Trial 3: Sentence 4	15.393	5.280	1110.046	2.915	0.004

The reduction trend in the duration of target words across trials implies an effect of information status—old information results in shorter duration of the speech. In particular, as the information status gets old in Trial 2 and Trial 3, it was less challenging for the participants to plan the answers and thus resulted in a much faster speech rate. Within trials, the duration of the target words in Sentence 4 was shorter than that in Sentence 1 but interacted with Trial 2 and Trial 3. This can be attributed to the fact that the two sets of target words (*meirongyuan* 'beauty salon' and *yibaimi* '100m' vs. *youeryuan* 'kindergarten' and *wubaimi* '500m') were mentioned respectively in Sentence 1 and Sentence 3, and they are all repeated words in Sentence 4 even in Trial 1. These within-trial and cross-trial findings reconfirm the viewpoint in the literature that word duration decreases as talkers become familiar with the word or able to anticipate the upcoming words, or in speech repetition [2, 3, 6].

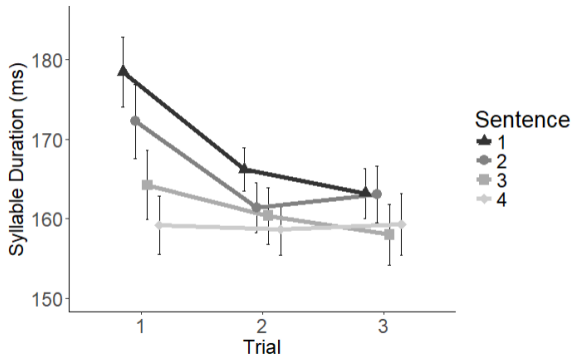

Fig. 4 Mean syllable duration by sentence and trial.

D. Syllable Pitch

The statistical analyses suggest that neither trial nor sentence has a significant effect on the pitch of the target words (see Table IV and Figure 5). Only Sentence 4 shows a marginal difference from Sentence 1 (p = 0.067).

Table IV
STATISTIC RESULTS OF SYLLABLE PITCH.

Predictor	Est. SD	SE	*df*	*t*	*p*
(Intercept)	151.242	9.989	25.515	15.141	0.000
Sentence 4	-5.641	3.069	774.772	-1.838	0.067

No significant fluctuation was found in the mean F0 of target words across sentences and trials. The slightly lower pitch in target words in Sentence 4 detected within trials may be attributed to the same reason as in duration that both sets of target words in Sentence 4 were second-mentioned within trials. Although the cross-trial finding is inconsistent with the results in previous studies [10, 12, 14, 21] and SSRH [2], the marginal difference of Sentence 4 compared to Sentence 1 within trials indicates that pitch functions as a predictor in repeated speech.

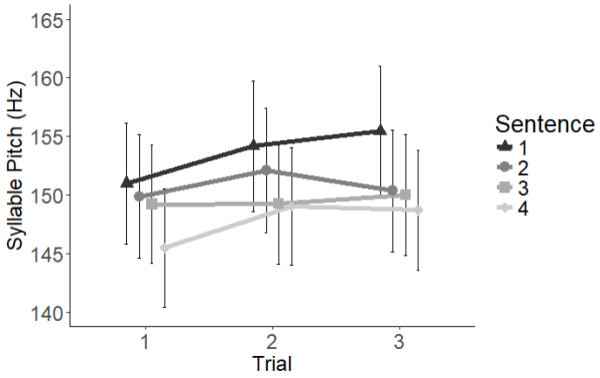

Fig. 5 Mean syllable pitch by sentence and trial.

E. Syllable Intensity

The results of syllable intensity show a main effect of trial. Sentence 4 differed from Sentence 1 and interacted with Trial 2 (See Table V). Figure 6 indicates that mean intensity of the target words increased in Trials 2 and 3 compared to Trial 1. However, it was found lower in Sentence 4 than Sentence 1 in Trial 2.

Table V
STATISTIC RESULTS OF SYLLABLE INTENSITY.

Predictor	Est. SD	SE	*df*	*t*	*p*
(Intercept)	66.475	1.070	27.566	62.101	0.000
Trial 2	1.257	0.454	1109.985	2.772	0.006
Trial 3	1.46	0.454	1109.985	3.22	0.001
Sentence 4	1.578	0.489	1116.678	3.23	0.001
Trial 2: Sentence 4	-1.474	0.641	1109.985	-2.299	0.022

The increase of intensity in Trial 2 and Trial 3 from Trial 1 was unexpected and differed from previous findings [6, 7, 8]. This may be plausibly due to the experiment design of the current study, which led to a fact that the participants unconsciously spoke louder in Trial 2 after they were told that Trial 1 had not been recorded and so as to a new hearer in Trial 3. Again, similar to the results of duration and F0, Sentence 4 showed lower intensity than Sentence 1 in Trial 2 and Trial 3, however, not in Trial 1, which may be due to the same reason as in fixation counts that *zuikuai de luxian* 'the fastest route' aroused more attention of the participants when they suddenly perceived it.

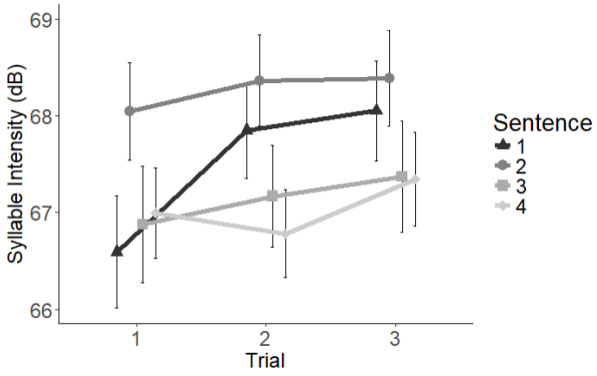

Fig. 6 Mean syllable intensity by sentence and trial.

IV. CONCLUSIONS

The current study investigates eye movements and prosodic variations in processing new and old information in Mandarin. The results indicate that fixation counts, fixation duration, and syllable duration of the target words decrease while syllable intensity increases as the information gets old to the speaker in the discourse. Syllable pitch shows no change across information status. However, within-trial repetition and old information resulted in the reduction of duration, F0 and intensity of the target words.

Future work will take pitch range as a dependent variable to further examine F0 variations across information status and take target word as an independent variable to examine the effects of departure vs. arrival status as well as within-trial information status on prosodic variations. Correspondingly, eye-movement parameters will involve regression-in and regression-out counts of the target areas of interest for insight into the effect of discourse information.

ACKNOWLEDGMENT

This work was supported by the National Social Science Foundation of China, approval number 19BYY043.

REFERENCES

[1] S. H. Fraundorf, D. G. Watson, and A. S. Benjamin, "Reduction in prosodic prominence predicts speakers' recall: implications for theories of prosody," *Language, Cognition and Neuroscience*, vol. 30, pp. 606-619, 2015.

[2] A. Bell, J. Brenier, M. Gregory, C. Girand, and D. Jurafsky, "Predictability effects on durations of content and function words in conversational English," *Journal of Memory and Language*, vol. 60, pp. 92-111, 2009.

[3] A. Christodoulou, "Variation in word duration and planning," Ph.D. dissertation, University of North Carolina, 2012.

[4] C. L. Jacobs, L. K. Yiu, D. G. Watson, and G. S. Dell, "Why are repeated words produced with reduced durations? evidence from inner speech and homophone production," *Journal of Memory & Language*, vol. 84, pp. 37-48, 2015.

[5] L. K. Yiu, and D. G. Watson, "When overlap leads to competition: effects of phonological encoding on word duration," *Psychonomic Bulletin & Review*, vol. 22, pp. 1701-1708, 2015.

[6] T. Q. Lam, and D. G. Watson, "Repetition is easy: why repeated referents have reduced prominence," *Memory and Cognition*, vol. 38, pp. 1137-1146, 2010.

[7] L. W. Shields, and D. A. Balota, "Repetition and associative context effects in speech production," *Language & Speech*, vol. 34, pp. 47, 1991.

[8] T. Q. Lam, and D. G. Watson, "Repetition reduction: lexical repetition in the absence of referent repetition," *Journal of Experimental Psychology Learning Memory & Cognition*, vol. 40, pp. 829, 2014.

[9] U. Gut, S. Pillai, and Z. M. Don, "The prosodic marking of information status in Malaysian English," *World Englishes*, vol. 32, pp. 185-197, 2013.

[10] C. Kaland, C. Bracks, and N. P. Himmelmann, "Repetition reduction in Papuan Malay prosody," in 9th International Conference on Speech Prosody, Poznań, Poland, 2018.

[11] C. Féry, E. Kaiser, R. Hornig, T. Weskott, and R. Kliegl, "Perception of intonational contours on given and new referents: a completion study and an eye-movement experiment," *Phonology in Perception*, vol. 3, pp. 267-292, 2009.

[12] E. Reinisch, A. Jesse, and J. M. McQueen, "Early use of phonetic information in spoken word recognition: Lexical stress drives eye movements immediately," *The Quarterly Journal of Experimental Psychology*, vol. 64, pp. 772-783, 2010.

[13] M. K. Tanenhaus, M. J. Spivery-Knowlton, K. M. Eberhard, and J. C. Sedivy, "Integration of visual and linguistic information in spoken language comprehension," *Science*, vol. 268, pp. 1632-1634, 1995.

[14] D. Danhan, M. K. Tanenhaus, and C. G. Chambers, "Accent and reference resolution in spoken language comprehension," *Journal of Memory and Language*, vol. 47, pp. 292-314, 2002.

[15] A. Chen, E. D. Os, and J. P. Ruiter, "Pitch accent type matters for online processing of information status: evidence from natural and synthetic speech," *The Linguistic Review*, vol. 24, pp. 317-344, 2007.

[16] D. G. Watson, M. K. Tanenhaus, and C. A. Gunlogsun, "Interpreting pitch accents in online comprehension: H* vs. L+H*," *Cognitive Science*, vol. 32, pp. 1232-1244, 2008.

[17] J. Shen, D. Deutsch, and K. Rayner, "On-line perception of Mandarin Tones 2 and 3: Evidence from eye movements," *Acoustical Society of America*, vol. 135, pp. 3016-3029, 2013.

[18] Y. Xu, ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. *Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, Aix-en-Provence, France. pp.7-10, 2013.

[19] S. F. Ehrlich, and K. Rayner, "Contextual effects on work perception and eye movements during reading," *Reading Research Quarterly*, vol. 16, pp. 227-235, 1981.

[20] P. E. Dussias, "Using the eye-tracking data in second language sentence processing research," *Annual review of applied linguistics*, vol. 30, pp. 149-166, 2010.

[21] Pierrehumbert, J. "The phonetics and phonology of English intonation." Ph.D. dissertation, Massachusetts Institute of Technology, 1980.