Prosodic Encoding of Information Structure in Mandarin Chinese: Evidence from Picture Description Task

Yifei Bi [1, 2], *Lesya Y. Ganushchak* [1, 2, 3], *Agnieszka E. Konopka* [4], *Guiqin Ren* [5], *Xue Sui* [5] & *Yiya Chen* [1, 2]

[1] Leiden University Center for Linguistics, Leiden, Netherlands
[2] Leiden Institute for Brain and Cognition, Leiden, Netherlands
[3] Education and Child Studies, Faculty of Social and Behavioral Sciences, Leiden, Netherlands
[4] School of Psychology, University of Aberdeen, Aberdeen, UK
[5] College of Psychology, Liaoning Normal University, Dalian, China
y.bi@hum.leidenuniv.nl, lganushchak@gmail.com, agnieszka.e.konopka@gmail.com, renguiqin@126.com, suixue88@163.com, yiya.chen@hum.leidenuniv.nl

Abstract

This study investigates the extent to which prosodic cues are employed during online sentence production to distinguish three different notions of information structure (informational focus, corrective focus, and givenness) at two sentential focus locations (i.e. the subject and object positions). Participants were asked to describe pictures. The information status of the subject and object characters was manipulated in the discourse preceding the presentation of each picture. Acoustic data (including duration, F_0, and intensity) from 65 participants were analysed. Results showed consistent acoustic differences between corrective focus and givenness, confirming the findings from reading and semi-controlled production tasks. Contrary to previous studies, our results showed no durational and F_0 differences between informational focus and givenness, and no differences in intensity range between informational focus and corrective focus. Moreover, our results showed that the sentential focus locations of the target word also had an impact on the prosodic encoding of information structure in natural utterance production.

Index Terms: information structure, prosodic encoding, natural utterance production, Mandarin Chinese

1. Introduction

Various notions of information structure such as Focus and Givenness have been studied extensively [1]. Three basic notions are illustrated using the following example shown in Figure 1.

Figure 1: *Example of a target picture.*

(1) Q: 谁在抓瓢虫? (*Who* is catching the ladybugs?)
A: 男孩在抓瓢虫。(*The boy* is catching the ladybugs.)
(2) Q: 男孩在抓什么? (The boy is catching *what*?)
A: 男孩在抓瓢虫。(The boy is catching *ladybugs*.)
(3) S: 老太太在抓瓢虫。(*The old woman* is catching ladybugs.)
D: 男孩在抓瓢虫。(*The boy* is catching ladybugs.)
(4) S: 男孩在抓青蛙。(The boy is catching *frog*.)
D: 男孩在抓瓢虫。(The boy is catching *ladybugs*.)

Examples (1) and (2) illustrate *informational focus* and (3) and (4) *corrective focus*, in contrast to *given* information. In (1), *The boy* in the subject position bears *informational focus* since it has not been mentioned in the preceding context; *the ladybugs* in the object position is known to have the status of *givenness*, as it has already been mentioned previously. In (2), *the ladybugs* denotes *informational focus* and *the boy* denotes *givenness*. In (3), *The boy* is realized with *corrective focus* as the speaker has corrected the previous subject in the statement. Similarly, in (4), *the ladybug* is corrected and bears *corrective focus*. In summary, without the consideration of focus locations, there are three notions of information structure involved in the examples: *informational focus, corrective focus* and *givenness*. *Informational focus* denotes information that cannot be inferred from the preceding context, *corrective focus* signals the contrast with other information provided in the previous utterances and *givenness* marks the information that has already been mentioned in previous context.

These different notions of information structure are known to be encoded via prosodic cues such as duration, F_0 and intensity in many languages (e.g. [2-5] for Chinese, [6] for English, and [7] for Spanish). Studies on the acoustic cues used to signal different notions of information structure, however, suggest that their results were not always consistent. For example, studies on Chinese have reported conflicting results regarding the different prosodic encodings of informational focus vs. corrective focus. [8] showed that words in corrective focus had longer duration than those in informational focus and F_0 range did not differ between the two types of focus. However, [3] found contrastive results and showed that F_0 range differed, while no difference was found in duration between informational focus and corrective focus. Moreover, [3] examined the acoustic characteristics of given information. The results showed that words in informational focus had longer duration and larger F_0 range than those for

doi: 10.21437/SpeechProsody.2016-149

given words. In addition, the effect of focus location on the prosodic encoding of focus has also been considered [2, 9-11], in relation to which, inconsistency has also been reported on the F_0 range realization of focused utterances. For example, while F_0 was generally raised on the focused word and lowered on the post-focus words in the utterances, such an effect has been shown to be dependent on the focus location.

It is important to note that most existing studies have investigated the prosodic realization of information structure only in reading tasks [2, 3]. Participants were asked to read, often with several repetitions, the designed utterances with required notions of information structure (with Chinese characters shown on the screen or papers). Little has been investigated on how different notions of information structure have been encoded prosodically in tasks of relatively more natural utterance production. To our knowledge, only [4, 8] conducted semi-controlled production experiments. [4] examined the differences between informational focus and givenness. The results showed that words in informational focus had longer duration and larger F_0 range than those in givenness. No difference was found in intensity range between corrective focus and givenness. [8] studied two types of contrastive focus (corrective focus and counter-presupposition focus) and the results showed durational increase in corrective focus. Note that neither of the two semi-controlled studies considered different focus locations in the prosodic encoding of different notions of information structure.

Compared to the reading tasks, in the semi-controlled tasks adopted to tap into the focus realization of information structure in Chinese, the designed utterances typically were not shown to the participants. Instead, the Chinese characters of target words were shown on the screen and participants were requested to produce target words in a required sentence structure multiple times. Similar to the reading tasks, the semi-controlled tasks still fall short of providing a context where participants plan and produce an utterance online similar to natural speech communication setting.

Taking into consideration the design of both reading and semi-controlled production tasks, we have improved the task employed in this study in the following ways:

1. An online picture description task was adopted to elicit relatively more natural utterance production;
2. No explicit requirement on the syntactic structure of utterances in picture description;
3. No prompt in Chinese characters shown on the screen or papers;
4. No repeated production of utterances for each target picture to simulate more natural discourse settings.

With this setting, we aimed to address the following specific questions: 1) how do speakers of Mandarin Chinese prosodically encode different notions of information structure in natural utterances produced for online picture description? 2) What are the differences between the results in such relatively natural sentence production and those in reading and semi-controlled production? We quantitatively analysed multiple acoustic cues including duration, F_0 (both F_0 contour and range), and intensity range with functional data analysis and linear mixed-effects modelling to answer the research questions.

2. Methodology

2.1. Participants

65 native speakers (49 females; mean age: 22, SD: 3.5) of Mandarin Chinese (from Northern China) participated in the experiment. 30 participants were studying at Leiden University (the Netherlands) and 35 were students from Liaoning Normal University (China) at the time of the experiment [1]. All have normal vision and no history of speech disorders according to their self-report. Participants received small financial reward for their participation and were given written informed consent prior to participating in the experiment.

2.2. Experiment design and procedure

The experiment was designed to elicit picture description given a preceding discourse context, which aimed to elicit natural utterance productions with three different notions of information structure. In each trial, participants were first given a brief discourse context, which can be a WH-question or a statement (through the headphone), pre-recorded by a native Chinese female speaker. They were then presented with a colored picture on the computer screen. Participants were then asked to describe the colored picture, while providing information requested in the preceding WH-question (as illustrated in examples 1-2 of Figure 1) or to make corrections over the mistaken information in the preceding statement (as illustrated in examples 3-4 of Figure 1). Participants did not see the pictures beforehand. Each participant was asked to describe the picture according to one of the information-status conditions that were manipulated (see Section 2.3 for further details). The participants were recorded in a quiet room either at Leiden University or at Liaoning Normal University, but they all followed the same procedure.

2.3. Stimuli

The total corpus contained stimuli of 131 colored pictures displaying simple events [12] on the pictures. There were 43 target pictures, 84 filler pictures and 4 practice pictures. The stimuli were divided into five lists, which were counterbalanced on the notions of information structure expected for the utterances in picture description. Each target picture only occurred once in different conditions and focus locations on the five lists. Therefore, each participant saw each picture only once. There were three types of proceeding discourse contexts and two focus locations: informational focus and corrective focus in both the subject and object positions, in comparison to their counterparts as given information. All the characters shown in the target pictures could be described via disyllabic words. Different from [4], the Chinese characters were not shown on the colored pictures.

In order to have a better control of the tonal combinations for the target disyllabic words, only a subset of the data were analysed in this study. The subset included two tonal combinations for the target disyllabic words (T1T2 and T2T2), as shown in Table 1. In total, there were five disyllabic words elicited in the subject position and four in the object position, respectively.

[1] The preliminary results showed that there were no differences between the two participant populations in responses.

Table 1. *Disyllabic words for subject and object positions of focus locations.*

Subject Position	Object Position
sha1 yu2 *'shark'* ying1 er2 *'baby'* liu2 mang2 *'hooligan'* nan2 hai2 *'boy'* nan2 ren2 *'man'*	xin1 niang2 *'bride'* ying1 er2 *'baby'* nan2 hai2 *'boy'* piao2 chong2 *'ladybug'*

2.4. Data analysis

2.4.1 Perceptual verification

Before the acoustic analyses, a perceptual verification test was performed to separate the utterances into two groups (i.e. with-perceivable focus and without-perceivable focus). 28 native Chinese speakers (13 females; mean age: 28.3, SD: 1.1) were recruited to first listen to the utterances and then judged focus location of the utterances. Participants were given three choices to choose for the utterance: focus on subject position, focus on object position, and not sure about focus location. They were allowed to listen to the utterances repeatedly without time restriction, to avoid potential confusions caused by factors such as utterances produced with low voice. Note that none of the participants participated in the original production study.

Each target utterance was judged by five to seven listeners. Utterances with correct identification by more than 80% of the listeners were categorized as with-perceivable focus, while the rest were as without-perceivable focus. In the following, we will report results of analysis performed over the utterances that have been judged with correct focus locations.

2.4.2 Data preparation

The analysis of all acoustic data was conducted in Praat [13]. All the sound files were manually segmented. The onset and offset of the target disyllabic words determined the relevant time intervals for the extraction of duration. The onset and offset of target vowels for each syllable determined the time intervals for the F_0 and intensity values. F_0 and intensity values were sampled at 20 equidistant measurement points using a Praat script. The F_0 range and intensity range were determined by extracting the maximal and minimal F_0 and intensity values of each syllable, respectively.

2.4.3 Functional data analysis

Functional data analysis (FDA) [14] was used to investigate the F_0 contours. FDA provides a method to analyse the dataset consisting of entire curves with different durations. Two main procedures were taken: smoothing with a linear time registration and functional principal component analysis (FPCA). We analysed the Principle Component scores (PC scores), s_1 and s_2, which represent the slope and turning point of the F_0 contours, respectively. F_0 contours of the first syllable and the second syllable of the disyllabic words were analysed separately in order to have detailed comparison. All FDA analysis was carried out by the R package of '*fda*' [15]. Details of the implementation were presented in [16].

2.4.4 Linear mixed-effects modelling

The durational data (the whole disyllabic words), F_0 range (the first syllable and second syllable, separately) and intensity range (the first syllable and second syllable, separately) were modelled by linear mixed-effects modelling using *R* [17] with packages of *lme4* [18], and *lmerTest* [19]. Focus location, notions of information structure, tonal combinations, as well as their interactions were considered as fixed effects. Likelihood ratio tests were performed to decide the random terms of the model. By-subject slopes for the effect of tonal combinations and by-item slopes for the effect of information-structure notions and tonal combinations were kept as random effects (after model comparisons). We also performed the same modelling for the PC scores to investigate the differences in F_0 contours as a function of information-structure notions and focus location over the target words while taking into account the variation due to individual speakers and stimuli items.

3. Results

3.1. Duration

Results showed a main effect of focus location (df=1, F=12.53, p<0.001) and there was a significant interaction between notions of information structure and focus location (df=2, F=5.46, p<0.001). Therefore, we analysed how each notion of information structure and focus location contributed to the durational differences. The results showed that there were no durational differences in object nouns of the utterances across any of the three notions of information structure. Durational differences only existed in the subject nouns of the utterances between informational focus and corrective focus (df=1, F=4.35, p<0.01) and between corrective focus and givenness (df=1, F=6.07, p<0.01). The duration of corrective focus was 20ms longer than that of informational focus and 18ms longer than that of givenness.

3.2. F_0 contours

We have performed FDA for the F_0 contours among three notions of information structure for first and second syllable in the disyllabic words in different focus locations, separately.

The results of linear mixed-effects modelling for PC scores showed that there were significant differences of s_1 between corrective focus and givenness on both subject (df=1, F=65.43, p<0.001) and object positions (df=1, F=5.95, p<0.01) in the utterances of Tone 2 for the first and second syllable, respectively. No significant differences among other notions of information structure and focus locations were found on s_1 or s_2 of the F_0 contours.

3.3. F_0 range

Significant differences of F_0 range only existed in the second syllable of subject position in the utterances between corrective focus and givenness (df=1, F=8.87, p<0.001). F_0 range in corrective focus was larger than that in givenness. No significant differences were found in the object position across the three notions of the information structure.

3.4. Intensity range

Results showed that there were significant differences between informational focus and givenness and between corrective focus and givenness for both focus location and syllables (all with p<0.001). Regardless of focus location in the utterances, the intensity range of corrective focus was larger than givenness and there were no significant differences between informational focus and corrective focus. Table 2 summarised the significant results of these prosodic cues among the three notions of information structure.

Table 2. *Summary of the results from duration, F_0 and intensity (only significant results were shown).*

Prosodic cue	Focus location	Information-structure notions
Duration	Subject	Informational vs. Corrective
Duration	Subject	Corrective vs. Givenness
F_0 contour	Subject	Corrective vs. Givenness
F_0 contour	Object	Corrective vs. Givenness
F_0 range	Subject	Corrective vs. Givenness
Intensity range	Subject	Informational vs. Givenness
Intensity range	Subject	Corrective vs. Givenness
Intensity range	Object	Informational vs. Givenness
Intensity range	Object	Corrective vs. Givenness

4. Discussion and conclusions

We have examined the acoustic cues including duration, F_0 (both F_0 contour and F_0 range) and intensity in relatively natural online utterance production. Our results from 65 speakers showed that duration, F_0, and intensity range are employed, with varying degrees of reliability, in the prosodic encoding for different notions of information structure. To our knowledge, none of the related studies have compared directly the time-varying F_0 contours. With functional data analysis, our results have showed that, among the different notions of information structure, F_0 contour is also one of the important prosodic cues in differentiating information-structure notions probably more reliable than F_0 range. This suggests that the F_0 encoding of information structure such as corrective focus is not merely via pitch range expansion, but rather is through the distinctive realization of the F_0 contours' characteristics of the lexical tones, as argued in [20]. In the following, we will compare our results with those reported in reading or semi-controlled production studies in the literature.

In our experiment, there are prosodic differences between corrective focus and givenness in the subject position: words in corrective focus have longer duration, more exaggerated F_0 contour, larger F_0 range and intensity range than those in givenness. Furthermore, the results of duration showed that the duration of corrective focus was longer than informational focus and givenness. These results are consistent with those in [4].

There are also some discrepancies between our results and previous reports. For example, we found no durational difference between informational focus and givenness. There was no F_0 difference between informational focus and corrective focus or between informational focus and givenness. Our results then indicate that F_0 may not always be employed to convey information focus in relatively natural utterance production.

Focus location also affected the prosodic encoding of information structure in our study. Different from the results reported in reading tasks [2, 9-11], we did not find differences in F_0 range on object position for any of the information structure notions, which indicates the unreliable role of F_0 range in the prosodic encoding to differentiate information-structure notions in the object position. One might interpret the different results of focus location as due to the effect of first attention. 67% of the target pictures in the subset data have the subject images on the left of the whole pictures, which means that speakers have first looked at the image of the subject character before they start to describe the picture. This has also been confirmed with eye-movement data for the planning of the utterances [21]. This may have led to the differences in the prosodic encoding of subject and object nouns when they are produced with different information status, lending some preliminary support to the possibility that the prosodic encoding of information status may be, to some extent, due to speech planning processes rather than the mere encoding of specific linguistic representations for different information-structure notions. Further experiments are certainly needed to examine exactly how prosody has been planned and what mechanisms could have led to the different prosodic encodings of target words in subject and object positions.

For intensity range, different from [4], our results showed significant differences between corrective focus and givenness regardless of focus positions. A cautionary note is that intensity can be greatly affected by external factors such as the distance of the speaker's mouth to the microphone. Further studies on the role of intensity should have better control over such external factors by for example having speakers in fixed positions with head-mounted microphones.

To conclude, results of the acoustic analysis of words with different information-structure notions during online sentence production showed that between corrective focus and givenness, there are more consistent and reliable prosodic cues to encode the two different notions of information structure. The prosodic encoding of informational focus, however, is less stable, indicating the lack of specific and steady prosodic cues that can encode all the contrasts among different notions of information structure. In further studies, we are interested in the possible interactions between utterance planning and the prosodic representation of information structure notions that gives rise to the prosodic shape of an utterance.

5. Acknowledgements

The research was supported by the China Scholarship Council for the fellowship to YB, the European Research Council for the ERC-Starting Grant (206198) to YC and the National Natural Science Foundation of China (31471075; 31100732) to GR. We would like to thank Qian Li for comments, Qi Sun, Shuang Li and Weiwei Wang for collecting data as well as the support from Leiden University Center for Linguistics, the College of Psychology of Liaoning Normal University, Hongli Liu and Bo Liu from SR Research, Jos Pacilly for technical support in Leiden University Center for Linguistics and all the participants.

6. References

[1] M. Krifka, "Basic notions of information structure," *Acta Linguistica Hungarica,* vol. 55, pp. 243-276, 2008.

[2] Y. Xu, "Effects of tone and focus on the formation and alignment of f< sub> 0</sub> contours," *Journal of phonetics,* vol. 27, pp. 55-105, 1999.

[3] Y. Chen and B. Braun, *Prosodic realization of information structure categories in Standard Chinese*: Bibliothek der Universität Konstanz, 2006.

[4] I. C. Ouyang and E. Kaiser, "Prosody and information structure in a tone language: an investigation of Mandarin Chinese," *Language and Cognitive Processes,* pp. 1-16, 2013.

[5] Y. Chen, "Durational adjustment under corrective focus in Standard Chinese," *Journal of Phonetics,* vol. 34, pp. 176-201, 2006.

[6] M. Breen, E. Fedorenko, M. Wagner, and E. Gibson, "Acoustic correlates of information structure," *Language and Cognitive Processes,* vol. 25, pp. 1044-1098, 2010.

[7] R. van Rijswijk and A. Muntendam, "The prosody of focus in the Spanish of Quechua-Spanish bilinguals: A case study on noun phrases," *International Journal of Bilingualism,* August 28, 2012 2012.

[8] M. Greif, "Contrastive Focus in Mandarin Chinese," presented at the Speech Prosody, Chicago, IL, USA, 2010.

[9] S. Jin, "An acoustic study of sentence stress in Mandarin Chinese," The Ohio State University, 1996.

[10] W. E. Cooper, S. J. Eady, and P. R. Mueller, "Acoustical aspects of contrastive stress in question–answer contexts," *The Journal of the Acoustical Society of America,* vol. 77, pp. 2142-2156, 1985.

[11] E. Gårding, "Speech act and tonal pattern in Standard Chinese: constancy and variation," *Phonetica,* vol. 44, pp. 13-29, 1987.

[12] A. E. Konopka, "Speaking in context: Discourse influences formulation of simple sentences," in *the 27th CUNY Conference on Human Sentence Processing [CUNY 2014]*, 2014.

[13] P. Boersma and D. Weenik, "Praat: Doing phonetics by computer [Computer program]. Version 5.4.04, retrieved 28 December 2014. Online: http://www.praat.org/retrieved.," ed.

[14] J. O. Ramsay, *Functional data analysis*: Wiley Online Library, 2006.

[15] J. Ramsay, H. Wickham, S. Graves, and G. Hooker, "fda: Functional data analysis," *R package version,* vol. 2, 2011.

[16] Y. Bi, Y. Chen, and N. Schiller, "The effect of word frequency and neighbourhood density on tone merge," in *proceeding of International Congress of Phonetic Science (IcPhS)*, Glasgow, UK, 2015.

[17] R Core Team, "R: a language and environment for statistical computing. Vienna, Austria. R Foundation for Statistical Computing," Online: http:// www.R-project.org/ ed, 2013.

[18] D. Bates, M. Maechler, and B. Bolker, "lme4: Linear mixed-effects models using S4 classes," 2012.

[19] A. Kuznetsova, P. B. Brockhoff, and R. Christensen, "lmerTest: tests for random and fixed effects for linear mixed effect models (lmer objects of lme4 package)," ed, 2013.

[20] Y. Chen and C. Gussenhoven, "Emphasis and tonal implementation in Standard Chinese," *Journal of Phonetics,* vol. 36, pp. 724-746, 2008.

[21] L. Y. Ganushchak, A. E. Konopka, and Y. Chen, "What the eyes say about planning of focused referents during sentence formulation: a cross-linguistic investigation," *Frontiers in psychology,* vol. 5, 2014.