MANDARIN MULTIMEDIA CHILD SPEECH CORPUS: CASS_CHILD

Jun Gao, Aijun Li, Ziyu Xiong

Institute of Linguistics, Chinese Academy of Social Sciences, China

ABSTRACT

In order to deepen the understanding of the nature of first language acquisition, we carry a longitudinal recording of speech production of twenty-three Mandarin-speaking children from one year old to four years old at Chinese Academy of Social Sciences (CASS) in Beijing since 2009. The interactive communication between children and parents is recorded in both audio and video mode. Apart from longitudinal recordings, other three kinds of cross-sectional recording are also conducted. One of them is to invite parents to record their children's speech production at home through a web-based platform developed by the lab at CASS. Based on these recordings, a multimedia child speech corpus, i.e. CASS_CHILD, is constructed.

Index Terms— Mandarin-speaking children, longitudinal audio/video recording, cross-sectional recording, web-based platform, CASS_CHILD

1. INTRODUCTION

After birth, infants' language skill develops very fast. Only within a few years, without resorting to any formal instruction or conscious efforts, infants can master their mother tongue very well. In contrast to the complexity of language system, first language acquisition is characterized with naturalness and ease. This contrast arouses great interest of researchers in various fields. Over the past forty years, researchers have been devoted to probing the nature, developmental course and mechanism of first language acquisition.

Researchers are interested in finding out the milestones of language development. Although languages in the world are different from each other, their acquisitions by children in different languages feature similar developmental phases. Usually, in children's speech production, vowels are acquired earlier than consonants and for tonal languages, lexical tones are acquired earliest. In the development, the speech production of children across languages is characterized with errors and deviations. The errors and deviations in the speech of children in different languages are originated from similar strategies, such as substitution, deletion, assimilation, fronting, deaspiration and so on. But for a particular language, children might only use a subset of these strategies. What are the typological implications of these cross-linguistic similarities and dissimilarities? Does the parents' input play a crucial role in causing these dissimilarities? Does the native language ability or the general cognitive ability or the physiological ability underlie the similarities?

To answer these questions, first of all, it is needed to have a full and clear picture of language development. The accurate and detailed description of the characteristics of each phase in children's language development lays a foundation for the exploration of the nature and mechanism of first language acquisition. Secondly, to explore the role of language experience, the features of the caretakers' speech have to be referred to in the examination of children's language development. Both the depiction of children's language development and the depiction of the characteristics of the caretakers' speech have to rely on the analysis of children's and caretakers' speech. As a result, the recording of the two kinds of speech, especially the parallel longitudinal recording of the two kinds of speech, is very essential.

2. PAST RESEARCH

The past research on language acquisition mainly made longitudinal recordings of only one or a few children's speech by writing dairies or tape-recording. Even in cross-sectional studies, only a few children were under investigation. With the development of computer science and multimedia technology, the methodology in studying acquisition has been enhanced and digitized. Many new technologies have been applied. One historical advance in acquisition research mode is the development of CHILDES. In the mid of 1980s, Brain MacWhinney and Catherine Snow in America began to establish Child Language Data Exchange System (CHILDES, http://childes.psy.cmu.edu/). Researchers throughout the world can upload their data of all kinds of formats (such as text, audio and video files) to the platform and these resources can be accessed and shared by researchers in the other parts of the world. The format for transcription, CHAT, is standardized and can be searched with the tool, CLAN. The platform has been developing and enriched. So far, materials on CHILDES

cover twenty-five languages. However, most data are contributed by English-speaking children. Materials of Chinese-speaking children (including those from [1][2][3], data from the team led by Tardif and the team led by Chang) are relatively less.

3. MOTIVATIONS FOR BUILDING CASS_CHILD

Since there is a typological difference between Chinese and Indo-European languages, the investigation of the acquisition of Chinese will definitely promote the understanding of the nature of acquisition and the understanding of linguistic system. Reliable conclusions about acquisition should be based on the quantitive analysis of a large-scale corpus. In order to promote the research on language acquisition, our lab is dedicated to establishing a large-scale multimedia Mandarin Child Speech Corpus.

Apart from the speech data of Chinese children available on CHILDES, there are other data and materials collected by several research groups (like the team led by Zhu of Beijing Normal University, Zhang and Zou of Capital Normal University, Fan of Beijing Forestry University, [4], [5], [6], [7]). These speech data mainly focus on children's grammar development and collected from a few children. In the studies about children's phoneme development, except the studies in [8][9][10], researches of [11][12][13] are all case studies. So, a corpus with more child speakers is needed to make the quantitive investigation into phoneme development.

4. CASS_CHILD

The principle of building the corpus is to try to diversify the channels of raw material collecting and to integrate advanced technology to make data processing as automatical as possible.

CASS_CHILD is largely based on the longitudinal audio and video recordings of twenty-three Mandarin-speaking children (the recording is still continuing now). During the recording, caretakers and children play in a sound-proof and child-friendly room full of toys. They communicate naturally and do not need to fulfill any task. Their spontaneous dialogues are recorded with both wireless microphones and video cameras. Accompanying caretakers in the recording include parents, grandparents or babysitters. Only the high quality parts of these recordings are chosen for the corpus. The audio files are transcribed orthographically and annotated phonetically in conjunction with linguistic and paralinguistic information.

4.1. Participants

We record twenty-three Mandarin-speaking children, 13 boys and 10 girls. Originally, we planned to keep the record of the language development of 10 boys and 10 girls and to start recording around their first birthday and to record till their fourth birthday. But at the beginning in 2009, it was difficult to find the parents with one-year-old child willing to be involved in such a long longitudinal study, so four children started at around 2 years old of age in 2009 and some children joined in our study in 2010 or 2011. To balance the age, in 2011, we added four new children around 1 year old to our study. From 2009 to 2012, two children stopped recording due to personal reasons. One stopped after four sessions of recording and the other stopped after eighteen sessions.

All families in our study speak Mandarin. Of them, a few speak accented Mandarin. Two thirds of the families are Beijing locals.

All the parents participating in our study signed a consent form before they started recording.

4.2. Recording

Recordings are conducted in two scenarios. One is to record in a sound-proof room. The other is to record at home. The idea to record in a sound-proof room is to have good-quality recording which can be used for acoustic analysis. The sound-proof room is set up as a play room to attract parents to bring their child to our lab, and also in this way it is prone to elicit the natural spontaneous dialogues between the parents and the child. For some parents who can not bring their child to our lab regularly, they record at home with the equipments provided by our lab.

Recording in the sound-proof room. The sound-proof room (3.25m*3.5m*2.4m) is decorated as a play room full of various kinds of toys, books and flash cards. The caretakers and the child are playing freely in the room and communicate naturally. The whole scene is videotaped and at the same time their speech is recorded separately with wireless microphones. Two wireless pin microphones (AKG, SR400/SR40) are used. One is for one of the accompanying caretakers, the other for the child. The wireless microphones are clipped a bit lower below the collar, around 15 cm far from the mouth. The sound signals are recorded through Cool Edit Pro 2.0. The caretaker's speech is recorded in left channel and the child's speech in right channel.

The sound is sampled at 44 KHz and quantized to16 bit and saved in WAV format. To obtain the most natural and real data, the recording does not pause even when there is long silence between the caretaker and the child, or when the child is crying, drinking, eating or goes to the restroom outside the sound-proof room.

Besides audio modality, video is taken at the same time by two Panasonic cameras (TM700, with wide-angle lens) in 1920*1080 HG resolution. One camera is placed on a shelf at appropriate height at one corner of the room, the

other is held by the experimenter or by one of the accompanying caretakers to trace the child. The corner camera has the panoramic view of the whole room. The hand-held camera is to highlight the local scenes of the interaction between the child and the parents and the face of the child. But usually the child moves around, the video camera taker has to move to follow, so sometimes it is difficult to keep the pictures stable.

In order to diversify the content of the communication, the toys and books change regularly. Moreover, particular toys and books are prepared for a particular child to keep up his/her interest to come for a second time.

Recording at home. Recording at home has both audio and video recordings too. The audio is made with a digital recorder (Samson, ZOOM H4N). The digital audio recorder is placed near the child and the interlocutor. The audio recording is in mono and in WAV format, sampled at 44 KHz and quantized to 16 bit. The video is made with a Panasonic camera (TM700) held by one caretaker. The recording content is usually the daily life of the child or the natural interaction between the caretaker and the child when they are playing.

4.3. Recording Period

Of the twenty-three children, five were recorded once half a month before they were three years old in the sound-proof toy room. When they were three years old, they can only come once a month due to the reason that they have to go to kindergarten. Another fifteen children are recorded once a month in the sound-proof toy room. The other three children are recorded at home for 1-2 hours per month. Each time, for each child, the recording lasts one hour. So far, in total, we have around 570 hours of recording.

4.4. Transcription and Annotation

The corpus contains the basic information of the participants and their parents. The audio file of each recording session is named as: date-name-months (age). The video files of each session are named in the way: date-name-months(age)-A (for hand-held camera) and date-name-months(age)-B (for corner camera) respectively. In each session, the experimenter fills out both a paper table of experimental log and a .DOC version. In the table, the relationship of the caretaker accompanying the child, the kinds of toys and books, the mood, behavior and performance of the child (including counting, singing, dancing, reciting poems, reading Chinese characters, playing toys), the way of interaction with the caretaker, some phonetic and grammatical characteristics of the child's speech production (for example, the deviations of phoneme production and the usages of sentence patterns) and the situations whether the equipments go wrong or not are specified. All the files are sorted out based on participants. Under each participant, their related files are categorized according to session.

4.4.1. Orthographic Transcription

For the convenience of transcription, the audio files are first converted from stereo into mono through Cool Edit Pro 2.0. The audio file of one session of a child is transcribed as a whole. The trained transcribers from iFLYTEK make the sound-to-text transcription with PRAAT (http://www.fon.hum.uva.nl/praat/). The text transcription is first checked among transcribers. Then the transcription is rechecked by the experienced transcriber in charge. If the errors are more than 10%, the transcription is returned to the original transcriber for a second check and correction. If the errors are less than 10%, the transcriber in charge corrects the errors by herself. The word with a retroflex final is transcribed as the word plus 儿 (*er*, the retroflex final) with a underline in between, i.e. 花_儿(flower_*er*), to be distinguished from the non-retroflexed pronunciation transcribed without a underline, i.e. 花儿 (flower).

4.4.2. Phonetic annotation

After the transcription is finalized, G2P and POS are automatically processed. Afterward, speech sound is automatically segmented into syllables, initials and finals (with lexical tones).

After the automatic annotations, the transcribers check carefully the result of each step. When checking the transformation into pinyin, polyphones, words with two versions of pronunciation, and words with neutral tone are paid special attention to. When checking the boundary of syllables and the boundary of syllabic initials and finals, the well-trained annotators in our lab realign the boundaries automatically identified to the correct and precise positions. At the same time, the annotators label the real pronunciation of the child with the machine-readable phonetic alphabet labeling system of SAMPA-C. SAMPA-C, Chinese SAMPA conventions formulated for labeling continuous speech, include symbols for consonants and vowels, initials and finals, retroflex finals, and lexical tones, and can label sound variations (such as insertion, deletion, pharyngrealization, voiced, voiceless, nasalization, rounding, aspiration, centralization and phoneme change) and non-speech phenomena in spontaneous speech [14][15][16].

Apart from the annotations mentioned above, prosodic annotations are made including break index and stress information. Linguistic information of sentence type and paralinguistic information such as expressive emotions are annotated as well. Finally, fourteen tiers of information in annotation are as follows:

1) Tier 1, INFO. Brief information of the child and the caretaker.

2) Tier 2, BH (BianHao). The numbers of utterance turns.

3) Tier 3, HL (HuaLun). The speech is segmented according to speakers. Each segment is marked with the abbreviation of the speaker's identity, i.e. MOT for mother, FAT for father, GRM for grandmother, GRF for grandfather, CHI for child, OTH for other persons accompanying the child in the recording. If the speech in a turn produced by one speaker is too long, the turn is further segmented into sentences.

4) Tier 4, HZ (HanZi). In this tier, there is the text transcribed from the sounds with the marking of the identity of the speaker. If the speech is overlapped, it is separated and marked with OV. For example, [OV MOT：这是什么啊？(What's this?) CHI：这是球。(This is a ball.)].

5) Tier 5, FC (FenCi). The information in this tier is the result of word segmentation. An example of word segmentation and parts-of-speech tagging is shown in Figure 1 indicated by the red box. The word boundary is marked with slashes. The letters following the slash mean the parts-of-speech tagging.

The first five tiers are shown in Figure 1.

Figure 1: *An example of the annotations of the first five tiers (the utterance in the example means to put on a hat.).*

6) Tier 6, PY (PinYin). This tier shows the orthography of Chinese alphabets of the text on Tier 4 (See Figure 2).

7) Tier 7, YJ (YinJie). This tier shows the result of automatic segmentation. The Chinese alphabets shown are also the orthography of the syllables (See Figure 2).

8) Tier 8, SY (ShengYun). This tier shows syllable initials and finals. The sound variations produced by the child are labeled and marked in the parentheses as indicated in the red circles (See Figure 2).

The first eight tiers are shown in Figure 2.

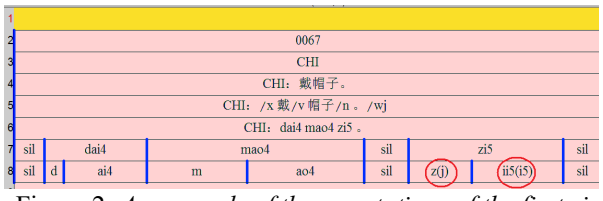

Figure 2: *An example of the annotations of the first eight tiers (the utterance in the example means to put on a hat.).*

9) Tier 9, BI (Break Index). The prosodic structures are annotated in this tier with reference to the labeling system of C-ToBI developed by our lab (http://ling.cass.cn/yuyin/product/product_10.htm).

10) Tier 10, SI (Stress Index). Sentential stress is specified in this tier.

11) Tier 11, VQ (Voice Quality). The tier is to specify the voice quality of the speech. The phenomena include creaky (CR), breathy (BT), tremulous (TR), ingressive (IN), glottal-attack (GL), falsetto (FA), whisper (WH), rough (RO), nasal (NA).

12) Tier 12, MIS (Miscellaneous information). It is about phenomena special in spontaneous speech. The phenomena are shown in Table 1 with reference to SAMPA-C.

Table 1: *Phenomena in spontaneous speech and the labels.*

phenomena	symbols	phenomena	symbols
repairs	[RP]	crying	[CR]
disfluencies	[DS]	noise	[NS]
silences	[SIL]	lengthening	[LE]
laughing	[LA]	modal	[MO]
coughing	[CO]	murmur	[MUR]
breathing	[BR]	smack	[SM]

13) Tier 13, JX (JuXing), sentence type. Sentence types include statement (s), imperative sentence (is), exclamatory sentence (es), yes-no question (qyn) echo question (eq), positive and negative question (qpn), wh-question (qw), question-tags (qt), alternative question (qa), open-ended question (qo), rhetorical question (qr).

14) Tier 14, CX (CiXing), parts of speech. The information of parts-of-speech is extracted from Tier 5 to be shown separately on this tier.

The full annotations are shown in Figure 3.

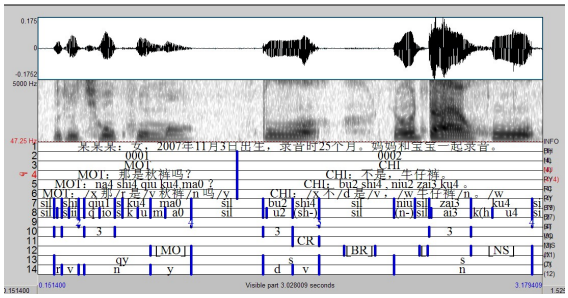

Figure 3: *An example of the annotations of the fourteen tiers.*

4.5 'Retrieval' tool

The corpus has the key word retrieval function. The audio file, the two video files and the corresponding textgrid are aligned in the database by the time information in the tiers

of the textgrid. With the *retrieval* tool, by typing a key word, all the textgrids where the key word is present can be searched out and form a list. By clicking the item in the list, the corresponding audio and video file can be played. The information contained in all the tiers of the textgrid can be retrieved.

4.6 'Word frequency' tool

The words in all the textgrids can be listed out and counted with the tool, *word frequency*, developed in our lab for the corpus. The tool can list all the words in the textgrids in a .csv file. For each word, the information of parts-of-speech, frequency, the total number of the words in the corresponding textgrid will be listed in parallel.

4.7. Elicited recordings

Although longitudinal data is useful, it has limitations for the lack of enough data produced by different children and parents. So, besides the longitudinal recordings, we record child-directed mother speech based on the prepared prompts, the Question-Answer (Q-A) dialogues between parents and children, and children's production of words (web-based recording platform). These three kinds of recordings are made to focus on one particular aspect of language acquisition.

The aim of the recording of child-directed mother speech is to examine the effect of input on language acquisition, especially acquisition of word classes. During the recording, the mother read the preset prompts to their infants. The mother's speech is recorded in the lab in the same way as the longitudinal recording mentioned above. The analyses of the data of twenty mothers show that homophone nouns and verbs can be distinguished to a large extent by phonetic differences [17]. Thus when infants learn nouns and verbs, they might make use of phonetic cues to make word class categorization.

In the Q-A dialogue, pictures are presented in front of parents and children. For each picture, there are a couple of questions. The parents ask the preset questions and try to elicit the child's answers. The answers of the child can be of any kind. If the child doesn't know the object in the picture, the parents utter the word and make the child repeat. This recording aims to investigate the phoneme production by children. The possible answers cover the most important aspects of Mandarin phonology, such as all consonants, neutral tones, retroflex finals, tone sandhi, one-, two- and three-syllable words. We record the production of the children who start speaking and that of children younger than three and half years old. The dialogues are recorded in the lab in the same way as the longitudinal recording mentioned above. The preliminary results based on sixteen children reveal that children speak more slowly than adults. Their production of syllables has longer duration. Accordingly, longer VOTs are seen in children's stop production [18].

4.8 Web-based recording platform

To investigate the development of children's phoneme production, a large number of samples are needed owing to the fact that there are great individual differences among children. Recording in the lab or in the kindergartens needs a lot of labor and is with many limitations. The effective and labor-saving way is to invite the parents to record their children's speech production at home through a network platform and then the recordings can be uploaded to our database. Therefore, besides recording in the lab, the third way of cross-sectional recording for our corpus is web-based home recording. We have a network platform developed by our lab with the technical support of Anhui USTC iFLYTEK Co., Ltd. On this platform, there is a list of words familiar to children which are embodied with pictures. After the parents log in the site, they can click on a picture, then a recording tool based on tcl/tk [19] and snack [20] is activated (See Figure 4). The parents can ask their children to say the name of the picture with headphones and are suggested to record in a relatively quiet room. The parents can listen to their children's speech at any time as long as they log in. Furthermore, they can listen to the production of the word by other children. Also they can have a longitudinal recording of their children's speech. In this way, the parents can have a digital diary of their child's language development. The privacy of the parents and the child is strictly protected. By this way, for researchers, there will be a very rich corpus of child language development.

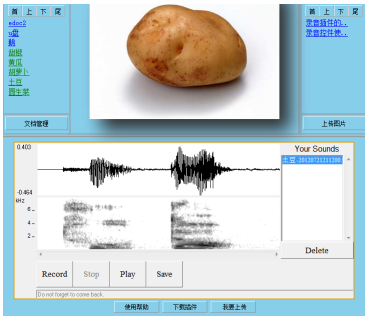

Figure 4: *An example of the webpage for recording*.

5. SUMMARY

The CASS-CHILD is still under construction. The longitudinal recordings are mainly made in the sound-proof room, which makes acoustic analysis possible. The longitudinal materials can be used not only for the study of phonology development but also for the study of grammar

development. So far, we have about 570 hours of longitudinal recordings; 74 pieces of recordings from mothers lasting around 37 hours; 105 pieces of the Q-A dialogues, totaling around 24 hours. 80 hours of recordings have been transcribed into Chinese characters.

The establishment of child speech corpus is a huge work. The deviations of the children's production make transcription and phonetic labeling very difficult and time-consuming. For one-hour-long recording, it takes around 40 hours to transcribe into Chinese characters and about 400 hours to annotate only the phoneme variations. In consequence, the mode of studying child language development from the perspective of speech production, regardless of longitudinal or cross-sectional, is not the most effective way. To overcome the disadvantage, other research modes are needed, such as experimental study on children's perception or comprehension of language. Another reason for perception and comprehension studies is that children's language perception and comprehension come before production. Study on perception and comprehension can really help understand the beginning of acquisition. Therefore, the best way to look into the nature of first language acquisition is to combine the examination of children's speech production with the investigation of children's perception and comprehension with experiments. In our lab, perceptual and comprehension experiments are conducted as well for better understanding of the nature of acquisition [21][22][23][24][25].

6. ACKNOWLEDGEMENT

The work is supported by the fund of CASS Innovation Project (2011-2015) to the lab, the Social Sciences foundation (2008-2011) and the CASS Key project (2009-2011) funds to the second author. Thanks to the lab members: Zhigang Yin, Zhao Zhang, Hongli Liang, Yao Yue. Thanks to all the parents and children in our study.

7. REFERENCES

[1] T. H.-t. Lee and C. Wong, "Cancorp-The Hong Kong Cantonese Child Language Corpus," Cahiers de Linguistique Asie Orientale, vol. 27, pp. 211-228, 1998.

[2] Virginia Yip, "Early grammar development of Cantonese-English bilingual children," Contemporary Linguistics, 6, pp. 1-18, 2004.

[3] J. Zhou and J. Zhang. Studies of language development of Chinese-speaking children: application and development of the method of international child speech corpus. Beijing: Educcation Science publisher, 2009.

[4] L. Kong, et al., Acquisition of content words by Chinese-speaking children. Anhui: Anhui University Press, 2004.

[5] Y. Li, Children language development. Wuhan: Central China Normal University press, 1995.

[6] Y. Zhang, "Lexical research based on a 3-6-years-old Mandarin-speaking children Spoken Corpus" Doctoral dissertation. Jinan: Shandong University, 2010.

[7] G. Zhou and B. Wang, Study on children's sentence pattern development and acquisition theory. Beijing: Beijing Language and Culture University Press, 2001.

[8] H. Zhu, Phonological development in specific contexts: Studies of Chinese-speaking children: Multilingual Matters Ltd., 2002.

[9] W. Li and H. Zhu, "Phoneme acquisition by Chinese-speaking children," Journal of Psychology, 2000.

[10] T. Wu and Z. Y, " Preliminary Analysis of record of children's language development from birth to 3-years-old," Journal of Psychology, 1979.

[11] Y. Li, "Study of articulation of an infant from birth to 120-days-old " Psychological Science, 1991.

[12] X. Deng, "Strategies used by children in acquisition of phonemes," Journal of Tsinghua University (Philosophy and Social Sciences), S1, 2004.

[13] Y. Si, "Case study of phoneme acquisition of a Mandarin-speaking child, " Contemporary Linguistics, 2006.

[14] X. Chen, A. Li, G. Sun, W. Hua, and Z. Yin, "AN APPLICATION OF SAMPA-C IN STANDARD CHINESE," in Sixth International Conference on Spoken Language Processing, Beijing, 2000.

[15] A. Li, X. Chen, G. Sun, W. Hua, Z. Yin, and Y. Zu, "SPEECH CORPUS COLLECTION AND ANNOTATION," in Sixth International Conference on Spoken Language Processing, Beijing, 2000.

[16] A. Li, F. Zheng, W. Byrne, P. Fung, T. Kamm, Y. Liu, Z. Song, U. Ruhi, V. Venkataramani, and X. Chen, "CASS: A PHONETICALLY TRANSCRIBED CORPUS OF MANDARIN SPONTANEOUS SPEECH," in Sixth International Conference on Spoken Language Processing, Beijing, 2000.

[17] A. Li, R. Shi, and Z. Zhang, "The prosodic features of verbs and nouns in infant-directed Mandarin Chinese," Zhongguo Yuwen, 2011.

[18] Y. Zhang, J. Gao, and A. Li, "Stop production by Mandarin-speaking children," 2012, in preparation.

[19] "http://www.speech.kth.se/snack/."

[20] "http://www.tcl.tk/software/plugin/."

[21] J. Hu, R. Li and P. Lee. "Scope acquisition at the interfaces". The 19th IACL. Nankai University, Tianjin, June 11-13, 2011.

[22] R. Li and J. Hu. "Null object construction vs. VP ellipsis construction: An experimental study, " Accepted for oral presentation, International workshop 'ELLIPSIS2012: Crosslinguistic, formal, semantic, discoursive and processing perspectives. November 9-10, 2012, Vigo, Spain

[23] J. Gao, R. Shi, and A. Lli, "Phonological neutralization and the representation of lexical tones in toddlers," in the 12th International Congress for the Study of Child Language, 2011.

[24] R. Shi, J. Gao, and A. Li, "Perception of Lexical Tones in Infants," in the 36th ANNUAL BOSTON UNIVERSITY CONFERENCE ON LANGUAGE DEVELOPMENT, 2011.

[25] Z. Zhang, R. Shi, and A. Li, "Grammatical Categorization of Nouns versus Verbs in Mandarin-Chinese-Learning Infants," Journal of Psycholinguistic Research, under review.

[This paper was published on COCOSDA 2012]