



PROJECT MUSE®

A Multimedia Corpus of Child Mandarin: The Tong Corpus

Deng Xiangjun, Virginia Yip

Journal of Chinese Linguistics, Volume 46, Number 1, January 2018, pp. 69-92
(Article)

Published by Chinese University Press

DOI: <https://doi.org/10.1353/jcl.2018.0002>



➔ *For additional information about this article*

<https://muse.jhu.edu/article/684589>

A MULTIMEDIA CORPUS OF CHILD MANDARIN: THE TONG CORPUS

Deng Xiangjun Virginia Yip


The Chinese University of Hong Kong

ABSTRACT

This article features a new multimedia corpusⁱ with 22 hours of recordings of a Mandarin-speaking child from the age of 1;7 to 3;4. We review the state of the art in the use of corpora for first language acquisition of Mandarin, and highlight the importance of corpus studies in evaluating children's language developmental patterns vis-a-vis adult input. The transcripts in our new corpus are annotated with a morphological tier indicating parts of speech, and linked to audio or video files. This corpus goes beyond existing published corpora of child Mandarin in having more data for a single child, as well as media linking. It contributes to a number of fields including language acquisition, Chinese linguistics, corpus linguistics, developmental psycholinguistics, education, and speech and language therapy.

Acknowledgment We are grateful to the reviewers of this paper, our subject Tong and his family members, Brian MacWhinney, Stephen Matthews, Mai Ziyin, Zhong Jing, Hannah Lam and participants at the International Symposium on Psycholinguistics of Second Language Acquisition and Bilingualism held at the Chinese University of Hong Kong (CUHK) in 2015. The research was supported by a start-up grant to the Bilingualism and Language Disorders Laboratory at Shenzhen Research Institute of CUHK, funding for the University of Cambridge-CUHK Joint Laboratory for Bilingualism, and CUHK-Peking University-University System of Taiwan Joint Research Center for Language and Human Complexity, and a General Research Fund from the Hong Kong Research Grants Council (Project no. 14413514).

Authors claim no conflict of interests to publish this paper in *Journal of Chinese Linguistics*.

Deng Xiangjun [dengxj@szu.edu.cn]; The Research Centre for Language and Cognition, School of Foreign Languages, Shenzhen University, 3688 Nanshan Ave., Nanshan District, Shenzhen, Guangdong Province 518060, P. R. China.  <https://orcid.org/0000-0002-9693-5593>.

Virginia Yip [vcymatthews@cuhk.edu.hk], Department of Linguistics and Modern Languages, G/F, Leung Kau Kui Building, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong.

i. TalkBank/CHILDES [Index to Corpora: Chinese/Tong Corpus (DOI: 10.21415/T5PC7Q)]

KEYWORDS

Child language corpus Mandarin Chinese Language input
Media linking Morphological tier

1. INTRODUCTION

Corpus-based studies have played an important role in many subfields of linguistics, including child language acquisition. Since Brown (1973), corpora have been used extensively in child language research. A number of child Mandarin corpora have also been constructed and important publications have been based on corpus data. This article aims to introduce a new corpus, review the state of the art in child Mandarin corpora, and give future researchers some ideas as to how to construct a child language corpus to address theoretical and methodological issues. We start with a discussion of the importance of constructing child language corpora, and compare our new multimedia corpus with existing child Mandarin corpora. This new corpus will be the first publicly available multimedia longitudinal corpus documenting a Mandarin-speaking child's language. Given the importance of Mandarin as a language spoken by more than one billion people in China and overseas communities, this corpus has far-reaching implications for the field.¹

2. SIGNIFICANCE OF CHILD LANGUAGE CORPORA

Before the advent of audio- and video-recording technology, the diary method was exclusively used by child language researchers (see Yip and Matthews 2007 for a review). Along with advances in the technology of audio- and video-recording, researchers are able to systematically document children's language development, as it has become the norm in the new millennium. The main argument of this section is that the advancement of child language research depends on good, publicly available multimedia child language corpora. A multimedia corpus enables researchers to capture a child's linguistic and non-linguistic production as well as contextual information, which in turn facilitates accurate transcription of the speech data. A well-constructed corpus lends itself to quantitative and qualitative analyses: quantitatively, the distribution and properties of different structures can be revealed with relative ease by calculating their types, tokens, and percentages; qualitatively, both target

and non-target forms produced by children can be analyzed systematically.

Unlike experimental studies, a naturalistic corpus enables researchers to observe young children unobtrusively, without imposing severe psychological demands on them. It also provides invaluable information about adult input, which is completely missing in experimental studies. The role of input in language acquisition has long been an important theoretical issue. In particular, some usage-based researchers argue that input plays a determinant role in shaping language development (Tomasello 2000a, 2000b, 2009) whereas nativists contend that innate Universal Grammar, rather than input, accounts for children's rapid and uniform acquisition of a first language (Crain 1991, Gleitman and Newport 1995); still others emphasize the combined role of nature and nurture (Pinker 1984; Yang 2002, 2004). To evaluate the role of language input, one has to sample and analyze adult input and compare it with the child's production. Previously, researchers have resorted to corpora from formal written English sources such as the *Wall Street Journal* to evaluate the poverty of stimulus argument of nativism (Pullum and Scholz 2002). Such corpora, however, tell us little or nothing about how adults address children (MacWhinney 2004). To address issues related to input in language acquisition, it is important to document the actual adult speech directed to the child.

Moreover, longitudinal child language corpora provide a window into individual children's developmental pathways. First, age of first emergence of a certain structure or feature can easily be ascertained, provided that recording starts early enough to capture the very beginning of child language development. First emergence is an important index in examining the acquisition of a structure, as it is the most sensitive measure of grammatical competence that is least affected by production constraints (Snyder and Stromswold 1997, Campbell and Tomasello 2001). Second, the ensuing developmental pattern of a structure can be revealed by investigating the frequencies of the structure over stages defined by the child's Mean Length of Utterance (Brown 1973, Theakston et al. 2001) or arbitrarily defined intervals (Demuth 1989, Wilson 2003, Theakston et al. 2005).

3. EXISTING MANDARIN-SPEAKING CHILD CORPORA

Researchers in the 1980s and early 1990s began to conduct

longitudinal studies of Mandarin-speaking children by collecting naturalistic speech data from one or two children (e.g. Li and Tang 1990). Erbaugh's (1982, 1992) classic study marks the beginning of the modern study of child Mandarin, using longitudinal naturalistic speech data. She conducted biweekly audio recordings for two children in Taiwan for twelve months from 1;9 and 2;10 respectively. Based on the data, she examined the development of a wide range of structures in Mandarin in a systematic and comprehensive way, and made the observation that learning Mandarin is comparable in difficulty to learning other languages including English, French and so on. However, early works were published without making their corpora available to the research community. Since the 1990s, a number of child Mandarin corpora have been made accessible in the Child Language Data Exchange System (CHILDES),ⁱⁱ including the Beijing corpusⁱⁱⁱ and the Context corpus^{iv} (Tardif 1993, 1996; Tardif, Gelman and Xu 1999), the Zhou1^v and Zhou2^{vi} corpora (Zhou 2001, 2009), the Xu/Min/Chen corpus^{vii} (Min 1994, Chen 2008), and the Chang1 corpus^{viii} (Chang 1998). Table 1 summarizes the information on published corpora, which will facilitate the comparison across corpora and our new corpus. We will introduce our new corpus – the Tong corpus^{ix} – in Section 4.

Table 1 Child Mandarin corpora in CHILDES

Corpus	no. of children	region	age range	no. of child utterances	no. of adult utterances	total
Longitudinal						
Tong	1	Shenzhen	1;7-3;4	9111	21247	30358
Beijing	10	Beijing	1;9-2;2	23983	52225	76208
Xu/Min/Chen	5	Beijing	0;11-3;5	9018	13992	23010
Cross-sectional						
Context	24	Beijing	2;0	3340	16144	19484
Zhou1	46	Nanjing	1;2-2;8 ²	2948	8609	11557
Zhou2	140	Nanjing	3;0-6;0	15757	37473	53230
Chang1	24	Hsinchu	3;6-6;5	2585	2601	5186

Source: Data related to each individual corpus are from CHILDES (TalkBank/CHILDES [Index to Corpora: Chinese]. See footnotes ii-ix).

ii. TalkBank/CHILDES [Index to Corpora: Chinese]

iii. TalkBank/CHILDES [Index to Corpora: Chinese/Beijing Corpus (DOI:10.21415/T5MK5D)]

iv. TalkBank/CHILDES [Index to Corpora: Chinese/Context Corpus (DOI:10.21415/T52C8H)]

v. TalkBank/CHILDES [Index to Corpora: Chinese/Zhou1 Corpus (DOI:10.21415/T5BS37)]

vi. TalkBank/CHILDES [Index to Corpora: Chinese/Zhou2 Corpus (DOI:10.21415/T5M59S)]

vii. TalkBank/CHILDES [Index to Corpora: Chinese/Xu/Min/Chen Corpus (DOI:10.21415/T59W3Z)]

viii. TalkBank/CHILDES [Index to Corpora: Chinese/Chang1 Corpus (DOI:10.21415/T52K5F)]

ix. See footnote i.

The existing corpora in CHILDES contain both longitudinal and cross-sectional data: the Beijing corpus and the Xu/Min/Chen corpus are longitudinal whereas the rest are cross-sectional, containing data from a large group of children at one specific point in time. While cross-sectional data have a large sample size to control for individual variation, the longitudinal data of a child allow us to make in-depth observation of the language development of the child over time. The Beijing corpus is considered longitudinal, as it contains 4-6 sessions of data from 10 children, starting from around 1;9 and ending around 2;2, with a one-month interval between each one-hour session for each child (Tardif 1993, 1996). Based on this corpus, Tardif (1993, 1996) challenged the universal noun bias, showing that nouns are not learned before verbs by Mandarin-speaking children, unlike English-speaking children. The Beijing corpus has also been used in Chen (2008), Deng and Yip (2015a, 2016). The Xu/Min/Chen corpus documented 5 children's longitudinal development from around 0;11 to 3;5 in Beijing (Min 1994; Chen 2008; Chen and Shirai 2010, 2015). Each child was recorded for half an hour each week or fortnightly in his/her home from 1983 to 1986. Chen (2008: 73) compared children sampled in the Beijing corpus with those in the Xu/Min/Chen corpus, and found the two groups of children to be comparable in terms of language development as measured by Mean Length of Utterance. On the other hand, the Context corpus contains cross-sectional data based on naturalistic conversation between 25 two-year-olds and their mothers in three activities: regular toys, mechanical toys and book reading (Tardif, Gelman and Xu 1999). The Zhou1 corpus contains cross-sectional data from 45 children aged between 1;2 to 2;8 and 5 sessions of longitudinal data from one child from 1;2 to 4;0, and the Zhou2 Corpus includes cross-sectional data from 140 children aged between 3;0 and 6;0 (Zhou 2001, 2009). The longitudinal data in the Zhou1 corpus were recorded in a naturalistic home setting, and the cross-sectional data in Zhou1 and Zhou2 were semi-structured, where the child draws, reads or plays with the mother. The Chang1 corpus includes narratives from 24 3- to 6-year-olds from Hsinchu, Taiwan (Chang 1998).³ The children were presented with a set of toy props and asked to set up a jungle, as well as continue the story after being given narrative prompts. One thing to note is that the Beijing, Context and Xu/Min/Chen corpora were transcribed with *pinyin* without Chinese characters.

Besides the Beijing and Xu/Min/Chen corpora, there are at least three more longitudinal multimedia child Mandarin corpora. The Chinese Early Language Acquisition (CELA) corpus^x developed by Lee (2010, 2012) contains longitudinal data of four Mandarin-speaking children in Beijing, and three children in Changsha who grew up acquiring Mandarin in a dialect environment, over a period of six months to 1.5 years.⁴ Using the Beijing CELA data, Fan (2007), Yang and Xiao (2008) and Yang and Yang (2015) investigated Mandarin-speaking children's acquisition of negation, the *ba* construction and control. The CASS_CHILD corpus^{xi} is an ongoing project that records 23 children biweekly or monthly from around 1;0 to 4;0, which has yielded 570 hours of recording in the initial stage (Gao, Li and Xiong 2012). The recordings are mainly made in a soundproof lab, which makes acoustic analysis possible, and phonetic annotation is provided after orthographic transcription, making CASS_CHILD an ideal corpus to study phonological development. The corpus developed by Zhang^{xii} (2014: 2) contains longitudinal data of six children from around 1;0 to around 5;0 who were audio- and video-taped for one hour each week. However, none of these three longitudinal corpora have been released with transcripts or audios/videos. There are also some more cross-sectional corpora (Zhou and Wang 2001, Kong et al. 2004) which have not been published.⁵

Child Mandarin is also documented in bilingual and multilingual acquisition research. Chang-Smith (2010) compares a monolingual child in Taiwan recorded fortnightly between 1;8 and 3;1, yielding 31 1.5-hour-long videos,^{xiii} and a Mandarin-English bilingual child in Australia, whose utterances were recorded in the diary^{xiv} on a daily basis from 0;10 to 3;2. Qi (2010, 2011) compared the development of Mandarin and English personal pronouns in a heritage Mandarin-speaking child in Australia, based on 82 20- to 30-minute-long audio recordings^{xv} from 1;7

x. Language Acquisition Laboratory, Chinese University of Hong Kong/Corpora. <http://www.arts.cuhk.edu.hk/~lal/corpora.html> (for Chinese Early Language Acquisition (CELA); accessed June 28, 2017). This longitudinal child Mandarin corpus has not been made available in the public domain. The above link only contains a description of the corpus.

xi. This corpus has not been made available in the public domain.

xii. This corpus has not been made available in the public domain.

xiii. The videos have not been made available in the public domain.

xiv. The diary has not been made available in the public domain.

xv. The audio recordings have not been made available in the public domain.

to 4;6. Yang and Hua (2010) investigated the phonological development of a trilingual child who acquired Spanish, Mandarin and Taiwanese simultaneously in Paraguay and the child was audiotaped^{xvi} at one-month intervals and occasionally videotaped from 1;3 to 2;0. While these works involve Mandarin in bilingual and multilingual contexts, there is no published corpus on bilingual first language acquisition of Mandarin Chinese so far.

This section has reviewed the state of art of corpora for child Mandarin. So far, we have a rich body of corpus data, but there is still room for improvement. First, to date, none of the published corpora provides multimedia data such as audio or video. In particular, no transcripts are linked with videos. Video can capture the entire context, and paralinguistic phenomenon such as gesture, eye gaze, and joint attention. As contextual information is crucial in interpreting young children's utterances, a video-linked corpus will provide very important data for a comprehensive study of language development. For instance, the video-linked multimedia corpus built by Yip and Matthews^{xvii} (2007: ch. 3) based on 6 Cantonese-English bilingual children has enabled researchers to work on phenomena such as object omission for which tracking contextual information such as joint attention and physical presence of an object is important (Zhou, Mai and Yip 2015).

Second, most published corpora lack sufficient density and duration to allow researchers to have a close look at a child's developmental trajectory. If a child's speech is regularly recorded for one hour every one to two weeks, the data will only represent 1-1.5% of the language a child hears and produces during the sampling period (Tomasello and Stahl 2004). For high-frequency phenomena such as the *ba* construction in Mandarin, such a density of recording may be adequate, but for some low-frequency phenomena such as the passive *bei* construction, the sampling rate is not dense enough to support valid and reliable analyses (see Deng and Yip 2015b). The majority of existing corpora in CHILDES are cross-sectional, as table 1 shows. Although the Beijing corpus and the Xu/Min/Chen corpus contain longitudinal data, the former contains only 4 to 6 one-hour sessions

xvi. The audiotapes have not been made available in the public domain.

xvii. TalkBank/CHILDES [Index to Corpora: Bilingual/YipMatthews Corpus (DOI:10.21415/T5MS3Q)]

of data for each child and the latter contains 9 30-minute sessions for each child, with the exception of one child having 17 sessions of data. The two corpora documented 5 to 10 children's longitudinal development, so the number of recordings per child was limited by constraints of resources. As a result, the sample size for each child may not be adequate to capture the acquisition process of a wide range of structures in this child. The duration of recording is another issue. Most of the existing corpora do not cover a period long enough to capture the full developmental trajectory if the focus is on syntactic and semantic development. Some corpora begin after age 2, and may be unable to capture the emergence of syntax as children generally start to combine words around 1;6. Some corpora finish before age 3 and may miss the development of some complex constructions. For example, Deng (2014) shows that 3- to 6-year-old Mandarin-speaking children are not sensitive to the aspectual properties of locative subject sentences (comparable to locative inversion sentences in English) in the experimental setting. It would be interesting to see if children in this age range produce locative subject sentences with the target aspectual properties in a naturalistic setting. However, there is no longitudinal corpus with naturalistic data from children aged between 3 and 6 to address the question. Moreover, there are considerable differences between speech directed to 1- and 2-year-olds and that directed to older children (as reviewed in Lee 1996). Input to 1- and 2-year-olds lacks sufficient cues for complex constructions such as the locative subject construction: the verb types and the frequency of aspect markers in locative subject sentences are quite limited in the adult input in the Beijing corpus which covers the age range 1;9-2;3 (Deng 2014). However, it is not clear whether such cues are frequent in adult speech addressed to 3- to 6-year-olds. If one wants to find out whether the late acquisition of the aspectual properties of locative subject sentences is due to insufficient input or due to the inherent difficulty of this construction, it is necessary to investigate corpora that contain adults' naturalistic speech directed to older children.

Third, except for the Beijing, Context, and Xu/Min/Chen corpora, previously published child Mandarin corpora cannot be considered to contain completely naturalistic adult input. In the Chang1 corpus, it was mainly the investigators that interacted with children and elicited narratives from them. Parents did not participate in the recording sessions.

The investigators' utterances may not represent the adult input that the child receives on a daily basis. The cross-sectional data in Zhou1 and Zhou2 were semi-structured, with the child and mother engaged in a limited number of activities, whereas a child's daily communication includes a broader range of activities. Longobardi et al. (2015) show that the contexts in which spontaneous speech is recorded may influence the proportion of early nouns vs. verbs: book reading gives rise to more nouns than verbs in child speech, whereas children produce fewer nouns when playing with toys. As different activities may yield conflicting findings, one needs to record spontaneous adult-child interactions in contexts that are representative of the child's daily life.

Lastly, except for the Chang1 corpus and the Zhou1 corpus, previously published corpora do not provide the morphological tier. These corpora only allow the key word search function, but are not searchable in terms of grammatical categories such as sentence final particles or classifiers. The search by part of speech can only be achieved if the part-of-speech information of each word in an utterance is provided in the morphological tier. Therefore, the morphological tier is indispensable in researching morpho-syntactic development.

To summarize, it is important to include information about the adult input and context in constructing child Mandarin corpora. In an effort to improve on existing corpora, we constructed an audio- or video-linked longitudinal corpus with denser sampling, and documented naturalistic adult-to-child input over the entire period of study. Moreover, the corpus provides a morphological tier which facilitates grammatical analysis of the data.

4. THE TONG CORPUS

Since Brown (1973), the common practice in the field of child language acquisition has been to sample the child's speech regularly for an hour once every week or every two weeks. Tomasello and Stahl (2004) discussed the issue of density of sampling and argued for denser samples especially when the phenomena under investigation are infrequent and rare. However, corpus builders are constrained by the time needed to transcribe the recordings. In order to build a general-purpose corpus to be used to examine various structures in Mandarin, frequent or infrequent,

we used the sampling rate of one hour per week. In addition, the first author, the mother of the child Tong, kept a diary of the child's language development. Informed by the Cantonese-English bilingual child language corpus – the YipMatthews corpus (Yip and Matthews 2007), this corpus was constructed as an audio- or video-linked longitudinal corpus with morphological tagging.

4.1 Background of the Child

Our subject Tong was raised in Shenzhen, China where Mandarin is the language of the community. Members of the family who interacted with Tong include his mother, father, and grandmother, all of whom speak Mandarin to the child on a daily basis. The child is also exposed to some input in Rudong dialect, a variety of Jianghuai Mandarin spoken in Jiangsu province, China, as the father and grandmother sometimes speak the dialect between them. From 2;5 on, the child attended a kindergarten in Shenzhen where he also received some Cantonese input. However, the influence of Cantonese is minimal, and the child only produced Mandarin in his speech. From 3;3, for three hours a day he attended a kindergarten in Hong Kong which used Cantonese as the medium of instruction. Previous corpora documenting child Mandarin were mostly based in Beijing. However, Mandarin, or Putonghua, as lingua franca and the official language of China, is also spoken by people from outside Beijing. Shenzhen, for instance, is a metropolitan city in Southern China with Mandarin as the dominant language in the society. It is representative of big cities in contemporary China where Mandarin and other Chinese dialects are in contact. Beijing is no exception in this regard: the children of migrants use Mandarin at schools in the urban centers, but are exposed to other dialects spoken by their parents and grandparents at home. The findings on Mandarin-speaking children in Shenzhen are suggestive of situations with millions of children from big cities in China.

4.2 Overview of the Corpus

Our corpus contains naturalistic interactions between a Mandarin-speaking child Tong and his caregivers mainly in the home setting, from age 1;0 to 4;6.⁶ The spontaneous interaction in play situation was audiotaped for one hour each week since 1;0, and video recording started at

The Journal of Chinese Linguistics vol.46, no.1 (January 2018): 69-92

©2018 by The Journal of Chinese Linguistics. ISSN 0091-3723/2018/46-0003\$10: A multimedia corpus of child Mandarin: The Tong Corpus. By Deng Xiangjun and Virginia Yip. <https://doi.org/10.1353/jcl.2017.0025>. All rights reserved.

2;3, producing more than 250 hours of audio recording and 90 hours of video recording.⁷ As we can only record one child at this stage, we try to have as many recordings for this child as possible to build a balanced corpus with fair sample size. In the initial phase, 22 one-hour recordings with one-month intervals from 1;7 to 3;4 have been released. They were transcribed and checked by native speakers of Mandarin with linguistic training.⁸ Table 1 in Section 3 compares the Tong corpus with other available existing corpora in CHILDES. The table shows that our corpus contains a larger number of child and adult utterances for one child than other corpora, as it has more hours of data for a single child. A special feature of the Tong corpus is that it provides video and audio files which are not available in the other child Mandarin corpora. The audio- or video-linked transcripts with morphological tiers will be deposited in CHILDES by installment.

4.3 Coding

The CHAT format of CHILDES was used for data coding (MacWhinney 2000). A sample transcript of Tong's speech data is shown in figure 1. The transcript begins with some header lines introducing the child's language, participant information, and so on. The lines marked with *MOT were produced by the mother, and the lines marked with *CHI by the child. After each utterance, a dependent tier %mor, namely the morphological tier, provides word-by-word glosses to the main tier. Each line in the transcripts is linked to the video to enable the researchers to see the context in which the child interacts with the adult. For instance, in figure 1, the video frame shows the context of the highlighted utterance.

Figure 1 Sample transcript of Tong's speech data at 2;5;30



@Begin

@Languages: zho⁹

@Participants: CHI Target_Child, MOT Mother, GRA Grandmother, FAT Father

The Journal of Chinese Linguistics vol.46, no.1 (January 2018): 69-92

©2018 by The Journal of Chinese Linguistics. ISSN 0091-3723/2018/46-0003\$10: A multimedia corpus of child Mandarin: The Tong Corpus. By Deng Xiangjun and Virginia Yip. <https://doi.org/10.1353/jcl.2017.0025>. All rights reserved.

@Birth of CHI: 17-JUN-2011
 @Media: 20131215, video
 @Date: 15-DEC-2013
 @Transcriber: Zhong Jing, Deng Xiangjun
 @Location: Shenzhen, Guangdong Province, China
 @Situation: playing with mother at home

...
***MOT: 同同，你先头拿那几个桶子来跟我说什么啦？**

%mor: n:name|tong2tong pro|ni3=you n|xian1tou2=earlier_time v|na2=hold pro|na4=that num|ji3=several class|ge4 n|tong3-NOM=pail v:dirc|lai2=come prep|gen1=with pro|wo3=I v|shuo1=say pro:wh|shen2me=what sfp|la1 ?

*CHI: <你 你> [] <你 拿> [] 你 拿 一个 我 拿 一个 .

%mor: pro|ni3=you v|na2=hold num|yi1=one class|ge4 pro|wo3=I v|na2=hold num|yi1=one class|ge4 .

*MOT: 好啊！

%mor: co|hao3=good sfp|la1 !

*MOT: 你把那个桶子放到哪里去了？

%mor: pro|ni3=you prep|ba3=object_marker pro|na4=that class|ge4 n|tong3-NOM=pail v|fang4=put v:resc|dao4=get pro:wh|na3li3=where v:dirc|qu4=gosf|le ?

*CHI: 放在你的电脑后面 .

%mor: v|fang4=put prep|zai4=at pro|ni3=you poss|de n|dian4nao3=computer post|hou4mian4=behind .

...
 @End

In processing the data, our first step was transcribing utterances into Chinese characters and segmenting a continuous string of syllables into words. In Chinese, the concept of word is by no means clear and intuitive; now the most common view is that words are minimal units of syntax (Packard 2000: 13, 14). The transcriber used spaces to show word boundaries. Mandarin Chinese is notorious for the indeterminate word boundaries: Mandarin words may consist of only one morpheme such as *ku* ‘cry’ and *putao* ‘grape’ or a combination of two or more morphemes. The combination of morphemes is of three types: compounding such as *feng-che* ‘wind-vehicle=windmill’, affixation such as *hai-zi* ‘child-NOM=child’, and reduplication such as *jiao-jiao* ‘teach-teach=teach a little’ (Li and Thompson 1981). Among the compounds, the resultative verb compound (RVC) such as *da-po* ‘hit-break’, and the V(erb)-O(bject) compound such as *guan-xin* ‘close-heart=care for’ are worth mentioning. There is a long-standing debate as to whether the RVC is a word or a syntactic phrase. Scholars taking the syntactic approach maintain that the RVC is derived in syntax via verb incorporation (Zou 1994, Tang 1997, Sybesma 1999, Huang 2006). The alternative lexicalist approach holds that RVCs are lexical items rather than products of syntactic operation (Y. Li 1990,

Gu 1992, Huang et al. 2009). Even though some acquisition studies treated the RVCs as one word (see Chang-Smith 2010), in our corpus, the two verbs in a RVC such as *da* ‘hit’ and *kai* ‘open’ in *da-kai* ‘hit-open=open’ are treated as two separate words, as we recognize that the child’s ability to combine the two morphemes suggests syntactic productivity. Mandarin V-O forms such as *chi-fan* ‘eat-rice=eat’ seem to have dual status as both words and phrases (Packard 2000: 115, Guo 2002: 190). We treat idiomatic V-O forms such as *shui-jiao* ‘sleep-sleep=sleep’ as one word, but those that are easily separated and combine with other elements such as *pa shan* ‘climb mountain’ as two words. In this study, we stick to this working definition of word: a word cannot be decomposed into parts that can stand alone as full free words with the same meaning. Functional or grammatical words such as sentence-final particles, classifiers, aspect markers, and localizers, which cannot stand alone as full free words are also treated as words (see Packard 2000: 299). In indeterminate cases, we consulted *Xiandai Hanyu Cidian* (Modern Chinese dictionary), (ZSKYYCB 2008) and the *Zho (Mandarin) dictionary* (TalkBank/Mor Grammars/Chinese (zho)). The intuition of wordhood is captured and represented in the form of dictionaries (Guo 2002: 32): if a string of morphemes is listed as a word in the dictionary, we consider it a word rather than a phrase. In parsing utterances into words, we try to be as consistent as possible. As the main purpose of the corpus is to evaluate language development within a single child, the indexes of grammatical competence such as Mean Length of Utterance, namely average number of words per utterance, will be precisely captured if transcription is consistent.

After transcription, the first author double-checked all the transcription. We then used the MOR command in the CLAN software provided by CHILDES to add morphological tagging to every utterance. In the morphological tier, CHILDES converts the Chinese characters into *pinyin* with four lexical tones marked with numbers 1 to 4, and provides part of speech information and English translation, based on its own ‘zho (Mandarin)’ dictionary. The current Chinese ‘zho’ dictionary (TalkBank/Mor Grammars/Chinese (zho)) is based on a simplification of a larger lexicon provided by the Academia Sinica in Taiwan. Words that are not in the ‘zho’ dictionary were added manually. The major parts of speech and their codings used in the ‘zho’ dictionary are shown in table 2. There are some controversial categories in this table. For instance, the

category of postposition is also treated as the localizer that bears the property of nouns (A. Li 1990, Huang et al. 2009). We remain non-committal to any theoretical stance, but since previous child Mandarin corpora abide by the CHILDES convention, we follow the same convention to gloss words such as *hou-mian* ‘behind/back’ in figure 1 as postpositions. CHILDES actually makes finer distinctions within some categories in this table. For instance, verbs also include directional verb complements (coded as v:dirc) such as *shang* ‘go up’ in *la-shang* ‘pull-ascend=pull up’, resultative verb complements (v:resc) such as *kai* ‘open’ in *da-kai* ‘hit-open=open’, and auxiliary verbs (v:aux) such as *hui* ‘will’. The category of noun is further divided into 7 subcategories: common nouns (n), proper names (n:prop), family names (n:fam), names (n:name), geographical names (n:geo), kinship terms (n:relat) such as *ba* ‘father’, and time expressions (n:tm) such as *baitian* ‘daytime’.

Table 2 Major parts of speech used in Chinese (Zho) dictionary

No.	Category	Code	Example
1	Adjective	adj	小 <i>xiao3</i> ‘small’
2	Adverb	adv	老 <i>lao3</i> ‘always’
3	Aspect marker	asp	了 <i>le</i> ‘perfective’
4	Classifier	class	分钟 <i>fen1zhong1</i> ‘minute’
5	Communicator ¹⁰	co	哎呀 <i>ai1ya1</i> ‘jeez’
6	Conjunction	conj	不但 <i>bu2dan4</i> ‘not only’
7	Interjection	int	对不起 <i>dui4bu4qi3</i> ‘sorry’
8	Noun	n	鱼 <i>yu2</i> ‘fish’
9	Negation	neg	不 <i>bu4</i> ‘not’
10	Number	num	八 <i>ba1</i> ‘eight’
11	Onomatopoeia	on	轰隆 <i>hong1long2</i> ‘rumble’
12	Postposition	post	后面 <i>hou4mian4</i> ‘behind’
13	Preposition	prep	从 <i>cong2</i> ‘from’
14	Pronoun	pro	我 <i>wo3</i> ‘I’
15	Quantifier	quant	各 <i>ge4</i> ‘each’
16	Sentence final particle	sfp	吗 <i>ma</i> ‘question’
17	Small (functional) words	small ¹¹	的 <i>de</i>
18	Verb	v	逛 <i>guang4</i> ‘hang out’

Source: Chinese (zho) (TalkBank/Mor Grammars/Chinese (zho))

After the morphological tier is added by the MOR command, we then use the POST command to automatically disambiguate morphemes that have multiple parts of speech or different meanings, as multiple syntactic functions and polysemy are abundant in Mandarin. For instance, the morpheme 给 *gei* can either be a verb meaning ‘give’ or a preposition meaning ‘for’. Moreover, homophones may also give rise to mistaggings. For instance, the character 还 can be pronounced as *hai* which is an adverb meaning ‘still’, or *huan* where it functions as a verb meaning ‘return’. To disambiguate the ambiguous morpheme, the CHILDES program computes the frequencies of various usages of the morpheme and tracks local contexts in some training transcripts. Based on the training results, the program assigns a tag to every token of the morpheme in our corpus. After the automatic disambiguation, we checked each line of our transcripts to manually correct mistaggings, based on the classification in Li and Thompson (1981) and *Xiandai Hanyu Cidian* (Modern Chinese dictionary) (ZSKYYCB 2008). The automatically generated tagging still requires manual checking, but the program is actively being developed by CHILDES to improve the accuracy of tagging.

5. SUMMARY

In this article we have reviewed corpora available to the research community on first language acquisition of Mandarin, with a focus on child Mandarin corpora deposited in CHILDES and introduced a new multimedia corpus featuring Tong’s development of Mandarin from 1;0 to 4;6. The first instalment includes 22 one-hour transcripts linked to audio/video files between 1;7 and 3;4. The new data will benefit researchers interested in the acquisition of Mandarin, including bilingual and multilingual acquisition researchers who need monolingual child data as the baseline. It will also contribute to related fields including Chinese linguistics, corpus linguistics, developmental psycholinguistics, education, and speech and language therapy. To provide the core longitudinal data needed for fundamental theoretical advances, more children’s data are needed, but we hope that this case study can inform other Chinese researchers about the method of data collection, data coding, and media linkage.

NOTES

1. China's population was 1.37 billion in 2014 according to National Bureau of Statistics of China (Zhonghua Renmin Gongheguo Tongjiju 2015). As the one-child family planning policy was officially replaced by two-child policy in China in 2016, we expect to see the new policy give rise to a larger number of children acquiring Mandarin in monolingual and bilingual contexts.

2. In the Zhou1 corpus, only one child was recorded 5 times from 1;2 to 4;0.

3. The Chang1 corpus documented Taiwan Mandarin that may have some grammatical differences from Mandarin used in mainland China (see Lin 2001: 15). Besides the Chang1 corpus, there are two narrative corpora called Chinese-Guo (TalkBank/CHILDES [Index to Corpora: Frogs/Chinese-Guo Corpus (DOI:10.21415/T5PC83)]) and Chinese-Tardif (TalkBank/CHILDES [Index to Corpora: Frogs/Chinese-Tardid Corpus (DOI:10.21415/T5PK7S)]) under the 'Frogs' folder in CHILDES. The two corpora used Mercer Mayer's wordless picture book "Frog, where are you?" as a tool for eliciting narrative descriptions. They contain "frog stories" elicited from subjects aged 3, 4, 5, 7, 9, and 20 (12 subjects in each age group) in Guo and Chen's (2009) and Chen and Guo's (2010) studies, and 604 children aged between 2;11 and 5;0 in Tardif's corpus. Following the paradigm set in Berman and Slobin (1994), they allow cross-linguistic comparison of narratives. As their contents are limited to narratives based on a picture book, we do not elaborate on them here.

4. The children were observed at weekly to biweekly intervals in hourly sessions. There are approximately 198 hours of recording for the Beijing children and 94 hours of recordings for the Hunan children. More details of this corpus can be found at <http://www.arts.cuhk.edu.hk/~lal/corpora.html#CELA>.

5. Zhou and Wang (2001) and Kong et al. (2004) collected 3 sessions of data from more than 90 children aged between 1;0 and 5;0. Kong et al. (2004) also includes a longitudinal observation of two children from around 1;0 to around 2;8. As the two studies did not track the development of each child but presented all the data as a whole, we consider their corpora as cross-sectional.

6. The recording was still going on at the time of writing.

7. The regularity of recording was sometimes disrupted by some

uncontrollable factors, such as the child's sickness.

8. The transcribers included Zhong Jing, Lam Ho Ching, Xie Shanrong, Zhou Jiangling, Lu Yaqiao, Lyu Lu, Yao Yao, and Zhishu Yu. The first author, Tong's mother, checked all the transcripts. Zhong Jing and Xie Shanrong helped with the morphosyntactic tagging, and Au Chui Yee, Lyu Lu and Yao Yao with the media-linking.

9. *Zho* stands for Mandarin Chinese (TalkBank/Mor Grammars/Chinese (zho)).

10. The boundary between communicator and interjection is tricky. According to CHILDES online manual (MacWhinney 2000), communicators are used for interactive and communicative purpose. Interjections are similar to communicators, but they typically can stand alone as complete utterances, rather than being integrated as parts of the utterances. They include forms such as *wow*, *hello*, *good-bye*, *please*, and *thank-you*. The definitions are different from those of some Chinese linguists (e.g. Guo 2002: 236-8) whose word categories do not include communicators, and the communicators in the CHILDES system are treated as interjection words. To be consistent with previous corpora, we stick to the CHILDES convention.

11. The term 'small' covers all the functions served by a few function words. For instance, the function word 的 has multiple usages: it is coded as 'poss' if it links a possessor and a possessee, as 'nom' if it is a nominalizer, 'cleft' if it is part of the cleft construction (*shi*)...*de*.

REFERENCES

- BERMAN, Ruth, and Dan Slobin. 1994. *Relating Events in Narrative: A Crosslinguistic Developmental Study*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- BROWN, Roger W. 1973. *A First Language: The Early Stages*. Cambridge, Mass: Harvard University Press.
- CAMPBELL, Aimee L., and Michael Tomasello. 2001. The acquisition of English dative constructions. *Applied Psycholinguistics* 22:253-67.
- CHANG, Chien-ju. 1998. The development of autonomy in preschool Mandarin Chinese-speaking children's play narratives. *Narrative Inquiry* 8(1):77-111.

- CHANG-SMITH, Meiyun. 2010. Developmental pathways for first language acquisition of mandarin nominal expressions: Comparing monolingual with simultaneous Mandarin-English bilingual children. *The International Journal of Bilingualism* 14(1):11-35.
- CHEN, Jidong. 2008. The acquisition of verb compounding in Mandarin Chinese. PhD diss., Vrije Universiteit Amsterdam.
- CHEN, Jidong, and Yasuhiro Shirai. 2010. The development of aspectual marking in child Mandarin Chinese. *Applied Psycholinguistics* 31:1-28.
- _____. 2015. The acquisition of relative clauses in spontaneous child speech in Mandarin Chinese. *Journal of Child Language* 42:394-422.
- CHEN, Liang, and Jiansheng Guo. 2010. From language structures to language use: A case from Mandarin motion expression classification. *Chinese Language and Discourse* 1(1):31-65.
- CRAIN, Stephen. 1991. Language acquisition in the absence of experience. *The Behavioral and Brain Sciences* 14:597-650.
- DENG, Xiangjun. 2014. Space, events, and language acquisition in Mandarin. PhD diss., The Chinese University of Hong Kong.
- DENG, Xiangjun and Virginia Yip. 2015a. The linguistic encoding of space in child Mandarin: A corpus-based study. *Linguistics* 53(5):1079-112.
- _____. 2015b. A corpus study of the acquisition of *ba* and *bei* constructions in Mandarin. Paper presented at The International Symposium on Psycholinguistics of Second Language Acquisition and Bilingualism, Chinese University of Hong Kong.
- _____. 2016. Cognition and perception in the linguistic encoding of space in child Mandarin. *Journal of Chinese Linguistics* 44(2):284-325.
- DEMUTH, Katherine. 1989. Maturation and acquisition of Sesotho passive. *Language* 65:56-81.
- ERBAUGH, Mary S. 1982. Coming to order: Natural selection and the origin of syntax in the Mandarin speaking child. PhD diss., University of California, Berkeley, CA.
- _____. 1992. The acquisition of Mandarin. In *The Crosslinguistic Study of Language Acquisition*, vol. 3, ed. D. I. Slobin, 373-455. Hillsdale, NJ: Lawrence Erlbaum Associates.

- FAN, Li 范莉. 2007. Ertong dui Putonghua zhong foudingci de zaoqi huode 儿童对普通话中否定词的早期获得 (Early child language acquisition of negative words in Mandarin Chinese). *Xiandai Waiyu* 现代外语 30(2):144-154.
- GAO, Jun (高军), Aijun Li (李爱军), and Ziyu Xiong (熊子瑜). 2012. Mandarin multimedia child speech corpus: CASS_CHILD. Paper presented at 2012 Oriental-COCOSDA, Shanghai.
- GLEITMAN, Lila, and Elissa Newport. 1995. The invention of language by children: Environmental and biological influences on the acquisition of language. In *Invitation to Cognitive Science*, vol. 1, eds. L. Gleitman and M. Liberman, 1-24. Cambridge, MA: The MIT Press.
- GU, Yang. 1992. The syntax of resultative and causative compounds in Chinese. PhD diss., Cornell University, Ithaca, NY.
- GUO, Jiansheng, and Liang Chen. 2009. Learning to express motion in narratives by Mandarin-speaking children. In *Crosslinguistic Approaches to the Psychology of Language: Research in the Tradition of Dan Isaac Slobin*, eds. J. Guo, E. Lieven, S. Ervin-Tripp, N. Budwig, S. Ozcaliskan, and K. Nakamura, 193-208. Mahwah, NJ: Lawrence Erlbaum Associates.
- GUO, Rui 郭锐. 2002. *Xiandai Hanyu Cilei Yanjiu* 现代汉语词类研究 (Research on the word categories in modern Chinese). Beijing: Shangwu Yinshu Guan.
- HUANG, C.-T. James. 2006. Resultatives and unaccusatives: A parametric view. *Bulletin of the Chinese Linguistic Society of Japan* 253:1-43.
- HUANG, C.-T. James, Y-H. Audrey Li, and Yafei Li. 2009. *The Syntax of Chinese*. Cambridge: Cambridge University Press.
- KONG, Lingda 孔令达, Deming Hu 胡德明, Junlin Ouyang 欧阳俊林, Changhui Chen 陈长辉, Lingyun Ding 丁凌云, Xiangrong, Wang 王祥荣, Wanxi Zhu 朱万喜, Manyi Fu 傅满义, and Wenbing Yao 姚文兵. 2004. *Hanzu Ertong Shici Xide Yanjiu* 汉语儿童实词习得研究 (Research on the acquisition of content words by Chinese children). Hefei: Anhui Daxue Chubanshe.
- LEE, Thomas H.-T. 1996. Theoretical issues in language development and Chinese child language. In *New Horizons in Chinese Linguistics*, eds. C.-T.

- J. Huang and Y. A. Li, 293-356. Dordrecht: Kluwer Academic Publishers.
- _____. 2010. Nominal structure in early child Mandarin. In *Chinese Matters: From Grammar to First and Second Language Acquisition*, eds. C. Wilder and T. A. Afarli, 75-109. Trondheim, Norway: Tapir Academic Press.
- _____. 2012. Quantificational structures in three-year-old Chinese-speaking children. In *Plurality and Classifiers across Languages of China*, ed. D. Xu, 243-82. Berlin: Mouton de Gruyter.
- LI, Charles N., and Sandra A. Thompson. 1981. Word structure. In *Mandarin Chinese: A Functional Reference Grammar*, 28-81 (Chap. 3). Berkeley, CA: University of California Press.
- LI, Y.-H. Audrey. 1990. *Order and Constituency in Mandarin Chinese*. Dordrecht: Kluwer Academic Publishers.
- LI, Yafei. 1990. On V-V compounds in Chinese. *Natural Language and Linguistic Theory* 8:177-207.
- LI, Yuming 李宇明, and Zhidong Tang 唐志东. 1990. San sui qian ertong fanfu wenju de fazhan 三岁前儿童反复问句的发展 (The development of A-not-A questions in children before three years of age). *Zhongguo Yuwen* 中国语文 no.2 (1990):91-6.
- LIN, T.-H. Jonah. 2001. Light verb syntax and the theory of phrase structure. Phd diss., University of California, Irvine, CA.
- LONGOBARDI, Emiddia, Clelia Rossi-Arnaud, Pietro Spataro, Diane L. Putnick, and Marc H. Bornstein. 2015. Children's acquisition of nouns and verbs in Italian: Contrasting the roles of frequency and positional salience in maternal language. *Journal of Child Language* 42:95-121.
- MACWHINNEY, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk*, 3rd ed. Mahwah, NJ: Lawrence Erlbaum.
- _____. 2004. A multiple process solution to the logical problem of language acquisition. *Journal of Child Language* 31:883-914.
- MIN, Rui-Fang. 1994. The acquisition of referring expressions by young Chinese children: A longitudinal study of the forms and functions of early noun phrases. PhD diss., Catholic University of Nijmegen (Netherlands).
- PACKARD, Jerome. 2000. *The Morphology of Chinese: A Linguistic and*

- Cognitive Approach*. Cambridge: Cambridge University Press.
- PINKER, Steven. 1984. *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.
- PULLUM, Geoffrey K., and Barbara C. Scholz. 2002. Empirical assessment of stimulus poverty arguments. *The Linguistic Review* 19:9-50.
- QI, Ruying. 2010. Pronoun acquisition in a Mandarin-English bilingual child. *The International Journal of Bilingualism* 14(1):37-64.
- _____. 2011. *The Bilingual Acquisition of English and Mandarin: Chinese Children in Australia*. Amherst, NY: Cambria Press.
- SNYDER, William, and Karin Stromswold. 1997. The structure and acquisition of English dative constructions. *Linguistic Inquiry* 28:281-317.
- SYBESMA, Rint. 1999. *The Mandarin VP*. Dordrecht, Netherlands: Kluwer Academic Publishers.
- TalkBank/CHILDES Child Language Data Exchange System. <http://childes.talkbank.org/> (Index to Corpora: Bilingual/YipMatthews Corpus[doi:10.21415/T5MS3Q]; accessed June 28, 2017).
- _____. (Index to Corpora: Chinese; accessed June 28, 2017).
- _____. (Index to Corpora: Chinese/Beijing Corpus [doi:10.21415/T5MK5D]; accessed June 28, 2017).
- _____. (Index to Corpora: Chinese/Chang1 Corpus; accessed June 28, 2017).
- _____. (Index to Corpora: Chinese/Context Corpus [doi:10.21415/T52C8H]; accessed June 28, 2017).
- _____. (Index to Corpora: Chinese/Tong Corpus [doi:10.21415/T5PC7Q]; accessed June 28, 2017).
- _____. (Index to Corpora: Chinese/Xu/Min/Chen Corpus [doi:10.21415/T59W3Z]; accessed June 28, 2017).
- _____. (Index to Corpora: Chinese/Zhou1 Corpus [doi:10.21415/T5BS37]; accessed June 28, 2017).
- _____. (Index to Corpora: Chinese/Zhou2 Corpus [doi:10.21415/T5M59S]; accessed June 28, 2017).
- _____. (Index to Corpora: Frogs/Chinese-Guo Corpus [doi:10.21415/T5PC83]; accessed June 28, 2017).
- _____. (Index to Corpora: Frogs/Chinese-Tardif Corpus [doi:10.21415/T5PK7S]; accessed June 28, 2017).

- TalkBank/Mor Grammars. <http://talkbank.org/morgrams/>(For Chinese (zho)); accessed June 28, 2017). These include word lists with part of speech information and English translation.
- TANG, Sze-Wing. 1997. The parametric approach to the resultative construction in Chinese and English. In *UCI Working Papers in Linguistics* 3, eds. L. C.-S. Liu and K. Takeda, 203-26. Irvine, CA: Irvine Linguistics Students Association.
- TARDIF, Twila. 1993. Adult-to-child speech and language acquisition in Mandarin Chinese. PhD diss., Yale University.
- _____. 1996. Nouns are not always learned before verbs: Evidence from mandarin speakers' early vocabularies. *Developmental Psychology* 32:492-504.
- TARDIF, Twila, Susan A. Gelman, and Fan Xu. 1999. Putting the "noun bias" in context: A comparison of Mandarin and English. *Child Development* 70(3):620-35.
- THEAKSTON, Anna L., Elena V. M. Lieven, Julian. M. Pine, and Caroline Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language* 28:127-52.
- _____. 2005. The acquisition of auxiliary syntax: *be* and *have*. *Cognitive Linguistics* 16(1): 247-77.
- TOMASELLO, Michael. 2000a. Do young children have adult syntactic competence? *Cognition* 74:209-53.
- _____. 2000b. The item-based nature of children's early syntactic development. *Trends in Cognitive Sciences* 4(4):156-63.
- _____. 2009. The usage-based theory of language acquisition. In *The Cambridge Handbook of Child Language*, ed. E. Bavin, 69-88. New York: Cambridge University Press.
- TOMASELLO, Michael, and Daniel Stahl. 2004. Sampling children's spontaneous speech: How much is enough? *Journal of Child Language* 31:101-21.
- WILSON, Stephen. 2003. Lexically specific construction in the acquisition of inflection in English. *Journal of Child Language* 30:75-115.
- YANG, Charles D. 2002. *Knowledge and Learning in Natural Language*. Oxford: Oxford University Press.

- _____. 2004. Universal Grammar, statistics or both. *Trends in Cognitive Sciences* 8(10):451-6.
- YANG, Hsueh-Yin, and Zhu Hua. 2010. The phonological development of a trilingual child: Facts and factors. *The International Journal of Bilingualism* 14(1):105-26.
- YANG, Xiaolu 杨小璐, and Dan Xiao 肖丹. 2008. Xiandai hanyu ba zi ju xide de ge'an yanjiu 现代汉语把字句习得的个案研究 (A case study of the acquisition of the *ba* construction in modern Chinese). *Dangdai Yuyanxue* 当代语言学 10 (3):200-10.
- YANG, Xiaolu, and Cheng Yang. 2015. Control in Mandarin-speaking children's early naturalistic production. *Lingua* 163:1-22.
- YIP, Virginia, and Stephen Matthews. 2007. *The Bilingual Child: Early Development and Language Contact*. Cambridge: Cambridge University Press.
- ZHANG, Yunqiu 张云秋. 2014. *Hanyu Ertong Zaoqi Yuyan de Fazhan* 汉语儿童早期语言的发展 (Chinese-speaking children's early language development). Beijing: Shangwu Yinshuguan.
- Zhongguo Shehui Kexueyuan Yuyan Yanjiusuo Cidian Bianjishi 中国社会科学院语言研究所词典编辑室. 2008. *Xiandai Hanyu Cidian* 现代汉语词典 (Modern Chinese dictionary), 5th ed. Beijing: Shangwu Yinshuguan.
- Zhonghua Renmin Gongheguo Tongjiju 中华人民共和国国家统计局. 2015. 2014 Nian Guomin Jingji he Shehui Fazhan Tongji Gongbao 2014 年国民经济和社会发展统计公报. (Statistical communique of the people's Republic of China on the 2014 national economic and social development) http://www.stats.gov.cn/tjsj/zxfb/201502/t20150226_685799.html (accessed June 28, 2017).
- ZHOU, Guoguang 周国光, and Wang Baohua 王葆华. 2001. *Ertong Jushi Fazhan Yanjiu he Yuyan Xide Lilun* 儿童句式发展研究和语言习得理论 (The study of construction development in Chinese children's speech and the theory of language acquisition). Beijing: Beijing Yuyan Wenhua Daxue Chubanshe.
- ZHOU, Jing. 2001. Pragmatic development of Mandarin-speaking children: from 14 months to 32 months. PhD diss., The University of Hong Kong.
- ZHOU, Jing 周兢 and Chien-ju Chang 张鑑如. eds. 2009. *Hanyu Ertong*

Yuyan Fazhan Yanjiu: Guoji Ertong Yuliaoku Yanjiu Fangfa de Yingyong yu Fazhan 汉语儿童语言发展研究：国际儿童语料库研究方法的应用与发展 (Research on the language development of Chinese children: The application and development of the research method of international child language corpora). Beijing: Jiaoyu Kexue Chubanshe.

ZHOU, Jiangling, Ziyin Mai, and Virginia Yip. 2015. Null objects in Cantonese-English bilingual children. Paper presented at the 23rd Annual Conference of the International Association of Chinese Linguistics (IACL-23), Seoul, Korea.

ZOU, Ke. 1994. Resultative V-V compounds in Chinese. *MIT Working Papers in Linguistics* 22:271-90.

ZSKYYCB. See Zhongguo Shehui Kexueyuan Yuyan Yanjiusuo Cidian Bianjishi.

一个多媒体汉语普通话儿童语料库：
同语料库 (TONG CORPUS)

邓湘君 叶彩燕
香港中文大学

提要

本文发布一个新的多媒体语料库的首阶段成果。这部分内容记录了一名普通话儿童从1岁7个月到3岁4个月期间的语言发展，共录得22个小时的语料。借此机会，我们回顾了汉语普通话一语习得研究中语料库使用的最新情况，强调语料库研究在考察儿童语言发展和成人语言输入时的重要作用。在我们这个新的语料库中，文字转写材料添加了词类注释层，并已实现与多媒体材料的链接。这个语料库在单个普通话儿童数据量和音频视频链接上超越了现有已发表的语料库。它将为语言习得、汉语语言学、语料库语言学、发展心理语言学、教育以及言语治疗等诸领域做出贡献。

关键词

儿童语料库 汉语普通话 语言输入 多媒体链接 词类注释层