EXPLORING INDIVIDUAL DIFFERENCES AND CONTEXTUAL VARIATIONS IN CHILD LANGUAGE CORPORA

Hintat Cheung[*] **Jing-chen Yang**
Hong Kong Institute of Education *National Taiwan University*

ABSTRACT

Corpora are usually built to serve specific purposes. Child language corpora are constructed mainly for examining the course of development of a target group of children. As in other developmental studies, individual differences are commonly found. Individual differences can lead to growth curves of different slopes and unexpected plateaus. These inherent variations raise the question of how representativeness of a child language corpus can be determined. To this end, the present study examined the range of variations that are inherent to contextual variations. Child language samples archived in Taiwan Corpus of Child Mandarin (TCCM, http://taiccm.org/) were analyzed. Two types of language samples were compared: spontaneous conversational samples and narratives elicited in experimental settings. D, an index of lexical diversity in child language samples, as well as several other indices on language development were computed. Our findings suggested that conversational samples and narrative samples are quite different in their capacities in gauging linguistic development. D showed sensitivity to the early stages of language development in typically developing children while Verb Type showed age effect in children with Specific Language Impairment (SLI).

[*] Corresponding author: hintat.htc@gmail.com

Linguistic Corpus and Corpus Linguistics in the Chinese Context
Journal of Chinese Linguistics Monograph Series 25 (2015): 30-38

2409-2878/2015/25-0002

KEYWORDS
Child language corpora **I**ndividual differences **C**ontextual variations **L**anguage development

1. INTRODUCTION

The use of corpus has quite a long tradition in the study of child language development. Most reviews considered Brown's seminal study in 1973 the standard model in examining the development pathway of child language by using rigorous language sample analyses of a small number of children. In 1984, CHILDES was started for the purpose of corpora-sharing. It has become the world's biggest archive of child language samples. CHILDES has provided notational guidelines on the format of language samples, such as the use of morphological tier (MOR) for part of speech, phonological tier (PHO) for phonological transcription, and computational tools for analyzing mean length of utterance (MLU), type/token ratio (TTR) in Computerized Language Analysis (CLAN). With the aid of the popularity of internet, CHILDES has successfully provided a platform for the circulation of language samples.

Corpora of Chinese-speaking children have been available for more than 20 years. CANCORP (Lee & Wong 1998), probably the first one, holds with more than 170 one-hour language samples from eight Cantonese-speaking children aged between one and a half and three. Subsequent efforts in enriching the Chinese child language corpora include those now archived at CHILDES (Chang 1998; Fletcher et al. 1999; Tardif 1996; Zhou & Li 2007).

Recently another child language corpus is available, TCCM, the Taiwan Corpus of Child Mandarin (Cheung et al. 2011), which is a web-based corpus modeled after CHILDES. One special feature of this corpus is its holding of language samples from children with specific language impairment. Children with specific language impairment suffer from unexplained language difficulties and often are considered test-cases of possible dissociations of cognition and language. Because of the isolating nature of Chinese language system, theoretically linguistic difficulties associated with Chinese SLI children should have no direct links with the morph-syntactic complexity of verbs, which have been observed in English as well as other European languages and became one of the core issues. Because of its possible role in theoretical and clinical applications, measuring the representativeness of Chinese child language samples, and indexing linguistic properties of child language samples should not be overlooked.

Our concern of the representativeness of TCCM is basically similar to that of most researchers in corpus linguistics – do we have enough data to even out sampling errors so as to acquire a normative status so that researchers who are working on individual differences of children of special population can make reference to it? The notion of a balanced corpus has yet been thoroughly explored in the field of child language acquisition. Technology in transcribing oral language sample is the first concern. Collecting child language samples is labor intensive and most samples are mostly manually processed. For a one hour child Mandarin sample, it will take at least twelve hours of processing time. With this technical barrier, the current issue in building child corpus is: how large is large enough, rather than how massive samples can be processed efficiently. Sampling in child language corpora is often not as dense as we wish and researchers can only safeguard against sampling errors from inferential procedures.

Contextual variation is the second factor. Child language samples can roughly be divided into two types if categorized according to the nature of the discourse: (1) spontaneous conversation, and (2) elicited narrative. It has been reported that children can produce longer and more complex utterances when conversing with adults, a phenomenon often referred as scaffolding. Narratives are more demanding since the child has to make both global and local plans alone. In clinical practice, narrative elicitations are often preferred because many possible confounding factors, such as social-cognitive development can be better controlled. It is therefore questionable if indices such as lexical diversity, mean length of utterances (MLU) which are extracted from spontaneous conversations comparable to those extracted from narratives. Variations are expected but it is our concern if such differences are negligible.

Individual difference in child development is difficult to tackle, be it in cognitive, social or language development. It is well-documented that the pace of language development in the first few years varies drastically among children of the same age range. Some children took four months to reach a production display of 50 words while some needed seven months. It is for sure that the developmental trajectory is an upward move but the slope of the growth curve can take many shapes. The growth curve of MLU in child Mandarin was examined but indices on lexical diversity remained to be explored. Since the study of the grammatical impairment often made use of MLU in matching subjects, a general index on lexical diversity may lead to a new option in matching subjects for studies that need to control over lexical strength of children.

1.1 Measuring Individual Differences and Contextual Variations

Grammatical errors produced by children formed one source of evidence for research that probed into the development of grammar. Percentage of correct use in obligatory context is an important index of children's mastery of the grammatical system. A failure in doing so of course can also reveal the status of development. For example, ill-formed verbal morphology such as *goed* for *went* or omission of subject-verb agreement such as *he sing everyday* are targets for the computation of error rate in English verbal morphology, which in turn inform us the stage of grammatical acquisition the language sample stands for. Probability of hit can be computed with different sampling rates, which is measured in terms of number of hours per week as independent variable (Tomasello & Stahl 2004; Rowland & Fletcher 2006).

Although error rate has been found to be a good index in examining the reliability of language samples, its extension to gauge individual differences in Chinese child language is not quite plausible. Chinese children did make grammatical errors but it is quite difficult to compute an error rate for there are no bases for making a decision of an omission error and for defining the obligatory use of a particular grammatical form. For example, if mama chi 'mommy eat' is produced by a child after his mother has finished eating lunch, the child form *chi* have several possible targets, such as *chi le* 'eat ASP', *chi wan fan* 'eat finish', *chi guang le* 'eat gone ASP'.

Grammatical productivity is another issue in grammatical acquisition that begs the question of sampling reliability. Rowland and her colleague (Rowland et al. 2008) found that sampling error in grammatical productivity can be caused by sample size, frequency statistics of the language and the vocabulary size of the child.

The present study strategically selected the measurement of lexical diversity as the target for measuring contextual variations. Lexical diversity of a language sample measures the range of word types used, which is often considered a measurement of vocabulary. Besides, an understanding of lexical diversity will allow us to further explore issues such as grammatical productivity.

2. METHOD

2.1 Data

Language samples from TCCM (Cheung et al. 2011; http://taiccm.org/) were analyzed. A total of 239 files, ranging from 1;6 to 8;5. Of the 139 files from typically developing children, 104 of them are samples of spontaneous conversation and 35 of them are narratives elicited experimentally. Another 110 language samples from Mandarin-speaking children with SLI: 56 of them are spontaneous conversations and 54 are narratives.

2.2 Measurement of Lexical Diversity: D

D is basically an adjusted Type-token Ratio (TTR), now offered via the software VOCD. TTR, a traditional index of lexical diversity, has been found to be too sensitive to sample size such that it will decline with increasing size of token samples. *D* is the parameter for estimating of type-token curve, a simplification of Sichel's (1986) model, in which N is the number token extracted from language samples:

$$\mathrm{TTR} = \frac{\mathrm{D}}{\mathrm{N}}\left[\left(1+2\frac{\mathrm{N}}{\mathrm{D}}\right)^{\frac{1}{2}} - 1\right]$$

In VOCD, D rendered as the average of TTRs, computed by restricting N to a random sampling of 35 to 50 tokens, reiterated for 100 times, with the whole process repeated three times.

2.3 Other Indices

Two other indices on lexical development are measured: Noun Type and Verb Type. Noun Type and Verb Type are tabulated directly from language samples, based on the POS tags.

3. RESULTS AND DISCUSSION

3.1 D with Age

As shown in Figure 1, there is a general growth of D from age two (mean = 33.16) to age 4 (mean = 45.60). Oneway ANOVA confirmed the general age effect. Post hoc analyses showed that D at age two is significantly lower than scores at other age but no other significant differences between age were found.

D-measure in adults' speech is quite stable across age-groups and no significant differences were found.

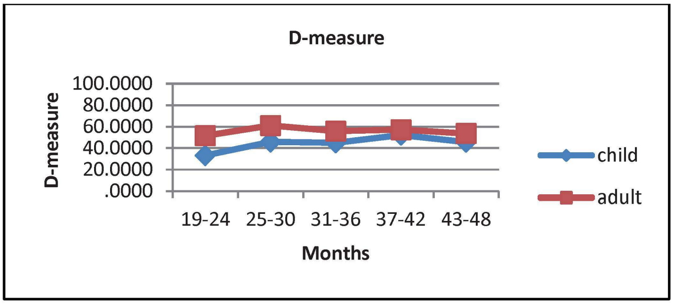

Figure 1 Variation of D with Age

As it has been reported that Chinese children displayed a verb-bias in early stage of development (Tardif 1996), noun types and verb types are also computed, which are summarized in Table 1. Both indices also showed age group effect. Of these three indices, noun type showed the biggest eta square (0.212), which suggests a stronger association with age than D and Verb Type.

Table 1 Variation of D, Noun Type and Verb Type with Age

Age group	N	D	SD	Noun Type	SD	Verb Type	SD
24	15	33.16	(13.85)	36.73	9.55	35.33	15.62
30	27	45.81	(13.53)	59.44	23.11	66.63	35.18
36	26	45.05	13.77	50.35	14.25	64.65	19.92
42	29	52.21	11.01	67.76	22.30	76.21	25.58
48	17	45.60	9.13	57.88	19.37	66.65	13.40
Total	114						

3.2 D for Narrative

D computed from narratives of four groups (from age four to age seven) of typically developing children were compared. Although a mild growth was observed, no significant age effect was found, even though the age four group showed a smaller D. Same results were found with Noun Type and Verb Type. Compared with the results in spontaneous speech, Ds are lower in narratives across all age groups. It is probably a reflection of the shorter length of the language samples and the limited range of thematic topics in the narrative elicitation task.

Table 2 Variation of Indices with Age (Narratives)

Age group	N	D	SD	Noun Type	SD	Verb Type	SD
4	9	32.62	14.62	26.67	11.30	43.00	18.32
5	12	37.86	19.59	24.67	6.98	35.50	11.41
6	6	36.42	9.20	28.00	8.39	46.50	8.41
7	8	35.07	11.02	25.63	5.50	44.88	9.63
Total	35						

3.3 D for Children with SLI

Language samples from Mandarin-speaking children with SLI, in three age groups were examined. A growth trend in the means of D, Noun Type, Verb Type were observed but results of three two-way ANOVAs (3 age group X 2 contextual type) conducted with D, Noun Type and Verb Type as the dependent variables only confirmed an age effect with Verb Type. Besides, all three ANOVAs showed significant contextual type effects. All narrative samples showed significantly smaller scores in D, Noun Type and Verb Type.

Table 3 Contextual Variations for Children with SLI

Contextual-type	Age Group	N	D	SD	Noun Type	SD	Verb Type	SD
Conversation	6	19	43.91	14.82	44.11	21.25	67.79	33.01
	7	21	51.15	10.34	56.29	25.22	94.33	37.92
	8	16	53.23	9.76	55.31	23.62	87.06	31.80
	Sub-total	56	49.29	12.35	51.88	23.73	83.25	35.91
Narrative	6	18	34.73	17.70	21.50	7.45	33.22	9.44
	7	21	36.20	9.42	24.86	6.17	40.05	7.91
	8	15	34.33	8.99	25.67	8.07	40.80	7.01
	Sub-total	54	35.19	12.49	23.96	7.25	37.98	8.76
	Total	110						

4. CONCLUSION

The present study analyzed language samples from typically developing children and those with language impairment, with the purpose of tapping the range of inherent variations in child Mandarin corpus. Three lexical indices, D, Noun Type and Verb Type were computed and were compared across age groups and contextual types. It is found that language samples from spontaneous conversation showed a wider range of vocabulary coverage than those of narratives, for both typically developing children and children with SLI. D, the measure for lexical diversity, is capable of detecting the differences between samples from young and older children (i.e. 24 months vs. 36 months) but not with samples from older children. In other words, individual difference in lexical diversity is getting more prominent after age three. For children with SLI, Verb Type differences are found between age 6 and age 7/8 groups. Verb learning difficulties in children with SLI has been reported and as the typically developing children did not show such a general age effect, it is very likely that the development of verbs in Mandarin-speaking children with SLI follows a very unique growth curve.

REFERENCES

CHANG, C. 1998. The development of autonomy in preschool Mandarin Chinese-speaking children's play narratives. *Narrative Inquiry* 8:77-111.

CHEUNG, H. 張顯達, C. Chang 張鑑如, H. Ko 柯華葳, and S. Tsai 蔡素娟. 2011. *Taiwan ertong yuyan yuliaoku zhi jianzhi* 台灣兒童語言語料庫之建置 (The Taiwan Corpus of Child Mandarin). Zhuanti yanjiu jihua chengguo baogao 專題研究計劃成果報告 Taiwan: Xingzhengyuan Guojia Kexue Weiyuanhui. https://drive.google.com/file/d/0B-fRmGHdeRXzekRNS0dVNmlQcXM/view?usp=sharing&pli=1

FLETCHER, P., T. H-T. Lee, S. Leung, and S. Stokes. 1999. Milestones in the learning of spoken Cantonese by pre-school children. Hong Kong: Language Fund.

LEE, Thomas H.T., and Colleen Wong. 1998. CANCORP: The Hong Kong Cantonese Child Language Corpus. *Cahiers de Linguistique Asie Orientale* 27: 211-228.

ROWLAND, C.R., and S.L. Fletcher. 2006. The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language* 33:859-877.

ROWLAND, C.R., S.L. Fletcher, and D. Freudenthal. 2008. How big is enough? Assessing the reliability of data from naturalistic samples. In *Corpora in Language Acquisition Research: History, Methods, Perspectives*, ed. by Heike Behrens. Amsterdam: John Benjamins Publishing Company.

SICHEL, H. S. 1986. Word frequency distributions and type-token characteristics. *Mathematical Scientist* 11:45-72.

TARDIF, T. 1996. Nouns are not always learned before verbs: Evidence from mandarin speakers' early vocabularies. *Developmental Psychology* 32:492-504.

TOMASELLO, M., and D. Stahl. 2004. Sampling children's spontaneous speech: How much is enough? *Journal of Child Language* 31:101-121.

ZHOU, Jing 周兢, and Xiaoyan Li 李曉燕. 2007. 0-6sui Hanyu ertong yuyong jiaoliu xingwei fazhan yu fenhua yanjiu 0-6 歲漢語兒童語用交流行為發展與分化研究 (Developmental differentiation of communicative acts of Chinese children in preschool years). *Zhongguo Wenzi Yanjiu* 中國文字研究 2007(4).

探索兒童語料庫中的個別差異與語境變異

張顯達 **楊靜琛**
香港教育學院 國立台灣大學

提要

兒童語料庫是以探究兒童語言發展為建置目的。兒童發展研究中經常報告個別差異的現象，如不同的成長曲線和預期之外的發展停頓。這一類的個別差異涉及如何判斷兒童語料樣本是否具備代表性。本文嘗試透過語料庫語料分析去探索這個方法學上的議題。語料是來自台灣兒童語料庫 (TCCM, http://taiccm.org/)，包括自發對話和誘發敘事說話兩種樣本。分析聚焦在詞彙的量化指標。結果顯示自發對話和誘發敘事說話這兩種不同語境中取的樣本在詞彙量化指標上表現不盡相同。詞彙多樣性指標較能反映初期的正常兒童語言發展，而動詞類別指標能夠呈現出語言障礙組的年齡差異。

關鍵詞

兒童語料庫 **個**別差異 **語**境變異 **語**言發展