S. Y. Lee, J. S. Jun, M. S. Min, J. H. Suh

A CORPUS STUDY ON THE ENGLISH-SPEAKING CHILDREN'S DEVELOPMENT OF LEXICAL DIVERSITY AND *MLU* USING *CHILDES* DATABASE[1]

Abstract. This study investigates English-speaking children's development of lexical diversity and MLU using the entire CHILDES data base. Frequency of types and tokens of vocabulary and type-per-token ratios (TTRs), D values in children's and their caregivers' utterances were counted for lexical diversity and mean length of utterances (MLUs) for early syntactic development. The results showed children's developmental trajectory with increase of Ds and MLUs by age and some influence of input from their caregivers.

1. Introduction

This study investigates English-speaking children's lexical and morphological development using the entire corpus of CHILDES data base (MacWhinney, 2000)[2]. CHILDES data base is a corpus of collection of transcripts of mainly conversation between a child and his or her caregivers for a certain period of time during the very early stage of language development. The corpus provides us with longitudinal and spontaneous speech data of children who are involved in a conversation with his or her mother in everyday life. This corpus has been used by linguists to investigate children's language development and influence of input from their caregivers. However, most of the previous studies used only a few children's data samples of CHILDES in the study due to its big size of the corpus. One of the main goals of this study is to reorganize the data base and create a useful way of using the entire set of data for investigation of

[1] This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012-2256-001).

[2] *MacWhinney B., Snow C. E.* The Child Language Data Exchange System: An Update. *Journal of Child Language 17.* 2000. **http://childes.psy.cmu.edu/**

children's language development and reexamine the findings of the previous studies with increased statistical power.

In this paper, we examined children's development of vocabulary and syntax by using the concepts of lexical diversity (D) and mean length of utterances (MLU) using the whole set of CHILDES data base. Previously, type-per-token ratio (Templin, 1957)[3] in children's utterances has been used to indicate productivity of children's use of vocabulary. High ratio of TTR means that children use various types of vocabulary in their conversation whereas low ratio of TTR means that children use limited size of vocabulary. However, a major problem of TTR has been pointed out, that is, TTR is affected by the length of the text. As the text gets longer, the TTR tends to become lower (e.g., Johansson, 2004)[4]. Therefore, D value was developed as an alternative in order to avoid the influence of the length of the text on the TTR to measure lexical diversity (Richards & Malvern, 1997[5]; Malvern et al., 2004).[6] More specifically, with the program CLAN to use the database in CHILDES, commend *Vocd* was developed in order to measure D value (MacWhinney, 2000). The higher value of D means higher level of lexical diversity. Therefore, in most cases, we expect increase of the D value by children's age at the early stage of language development.

On the other hand, MLU has been used to measure children's development of morphology in their production of utterance. Mean length of utterance indicates the morphological length of the utterance

[3] *Templin M.C.* Certain language skills in children. Minneapolis: University of Minnesota Press, 1957.

[4] *Johansson V.* Lexical diversity and lexical density in speech and writing: a developmental perspective, Lund University, Dept. of Linguistics and Phonetics, Working Papers 53. P. 61–79, 2008.

[5] *Richards B. J., Malvern D.* Quantifying lexical diversity in the study of language development. Reading: Faculty of Education and Community Studies, 1997.

[6] *Malvern D., Richards B., Chipere N., Duran P.* Lexical diversity and language development: quantification and assessment. New York: Palgrave Macmillan, 2004.

a speaker produces. It is calculated by counting the mean number of morphemes per 100 utterances. For example, children use only one word at the age of one, which means MLU of 1. As they get older, the length of their utterances gets longer with higher MLU (e.g., about 5 words or morphemes in a sentence). Therefore we expect to find the increase of MLU in the early period of children's language development. In this study, the D values and MLUs of the children's and caregivers' utterances were counted and compared to investigate children's development of lexical diversity and early syntax and any influence of input frequency from their caregivers on their development using the whole set of CHILDES data base. Our specific research questions are as follows:

1. Does D value increase as the children's ages increase?
2. Does MUL increase as the children's ages increase?
3. Is there any influence of parents' input on the children's development of lexical diversity (measured with D) and syntax (measured with MLU)?

2. CHILDES Data base

The data base of CHILDES investigated in this study contained 7,839 files of transcripts including 2,272 UK transcripts and 5,569 USA transcripts from a total of 1,630 English-speaking children. The number of files in each age is shown in Table 1.

Table 1. Number of Files in CHILDES Corpus analyzed

Age	UK		USA	
	No of children	No of files	No of children	No of files
1	62	203	203	960
2	67	1,742	339	1,652
3	69	194	302	1,124
4	24	51	204	834
5	28	57	134	665
6	6	6	89	124
7	19	19	84	210
Total	**275**	**2,272**	**1,355**	**5,569**

3. Analysis and Results

The entire CHILDES database was reorganized by age in order to analyze the whole corpus by age and country (UK, USA). Each of the files was renamed according to the name of the researcher of the corpus, age and the country of the target child. Frequencies of lexical types and tokens were counted using *freq* with the program CLAN. D values and MLUs were obtained using commends *vocd* and *mlu* respectively. The results of the D and MLU are provided in Table 2 for children and in Table 3 for caregivers.

Table 2. D and MLU in Children

Age	Types		Tokens		D		MLU	
	UK	USA	UK	USA	UK	USA	UK	USA
1	3,332	6,554	63,694	176,337	99.32	165.52	1.802	1.921
2	19,892	14,529	1,650,007	857,550	59.58	96.19	2.615	3.366
3	4,412	12,340	127,244	662,640	58.92	113.12	3.565	4.557
4	2,108	13,878	26,568	698,742	115.28	120.82	4.171	4.841
5	2,318	7,436	36,654	234,408	110.50	133.11	4.695	5.472
6	1,182	4,625	8,969	71,635	110.43	133.18	5.230	5.936
7	1,886	4,343	24,488	70,000	0.077	128.77	5.148	4.483

Table 3. D and MLU in Caregivers

Age	Types		Tokens		D		MLU	
	UK	USA	UK	USA	UK	USA	UK	USA
1	4,817	13,796	186,706	889,075	112.96	107.05	4.370	4.421
2	35,093	18,717	5,034,430	1,397,425	112.53	118.18	4.720	5.216
3	4,605	13,174	123,619	662,347	113.95	109.46	4.704	5.411
4	1,493	12,652	12,446	626,081	116.85	121.99	5.062	5.862
5	590	6,354	2,682	231,407	105.10	122.90	4.582	5.814

The results can be summarized as follows:

D in Children. There was a tendency that D values were higher with the older children than with younger children. However, there was a drop of D values at age 2–3 in the UK and at age 2 in the USA. The results seem to indicate a U-shaped developmental pattern of lexical diversity in children. The D values remained rather stable from age 4 in the UK and from age 5 in the USA. Comparing the US and the UK children, the D values were higher in the US children than in the UK children.

D in caregivers. The D values of caregivers are not different from those of children from age 4. The D values of caregivers remained rather stable across children's age. Comparing the US and the UK caregivers, the D values of caregivers were very similar across country.

MLU in children. MLUs increased as the children's age increased and they remained rather stable from age 6 in children from both countries. On the other hand, MLUs were higher in the US children than they were in the UK children.

MLU in caregivers. MLUs of caregivers remained rather stable across children's age. There was a slight tendency that MLUs of caregivers were lower at children's age 1. Comparing the UK and the US caregivers, MLUs of the USA caregivers were higher than those of the UK caregivers across age.

4. Discussion

The research questions of this study can be answered based on the results of our corpus analysis. First, regarding our research about lexical diversity, there was a big increase of lexical diversity between age 3 and age 4. Children's lexical diversity remained stable afterwards. A *U*-shaped developmental pattern was found in children's lexical development. There was difference between the children from the UK and the children from the USA. D remained rather stable from age 4 in the UK and from age 5 in the USA. D was higher in the USA children than in the UK children.

Next, regarding our research question about children's development of MLU, our study found that children's MLU increased as children got older. Children's MLU reached the level of adults at age 4. In general, MLUs of the USA children were higher than those of the UK children across age.

Finally, regarding our research question about the influence of parents' input, it was found that there was a tendency that children's D values were somewhat higher in the USA than those in the UK and that caregivers' D values were somewhat higher in the USA than those in the UK. There was a tendency that children's MLUs were higher in the USA than those in the UK and that caregivers' MLUs were higher in the USA than those in the UK. These differences between the UK and the US seem to indicate a possible influence of the parents' input on the development of children's lexical diversity measured with D and syntax measured with MLU.

5. Conclusion

This study investigated children's lexical development using measure D and syntactic development using MLU. Children's lexical and syntactic developmental patterns were detected using D measure and MLU. There was an indication of parents' input on the children's development of lexical and syntactic development.

In general, this study found that using the whole set of CHILDES to investigate children's development of vocabulary and morphology resulted in the supporting evidence for previous studies with bigger statistical power. This proves that our method can be used to reexamine the findings of earlier studies on child language development using CHILDES data base with more confidence.