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

# A STUDY OF INFLECTIONAL MORPHEME DEVELOPMENT IN ENGLISH-SPEAKING CHILDREN USING *CHILDES* CORPUS[1]

**Abstract.** In this paper we focus on the inflectional morpheme development on the CHILDES corpus from 1 to 7 years of age which account for 97% of the entire CHILDES corpus. We discovered (i) that the correlations between D value and the children's age were different depending on the morphemes, (ii) that overregularization errors were found in all the inflectional morphemes, most frequently with past-tense *–ed*, and mainly in 2- to 3-year-old children showing a *U*-shaped development pattern and (iii) that some differences between American and English children were also found.

## 1. Introduction

This paper aims to investigate the inflectional morpheme development in child language using the entire CHILDES corpus[2]. Our eventual purpose of the study is to verify the earlier studies on inflectional morpheme development in child language. Brown (1973)[3], one of the prominent studies on child language, provided 14 grammatical morphemes seen in the longitudinal study on 3 children. In his study, the present progressive is acquired first and followed by plural, irregular past tense, possessive, regular past tense, and 3rd person present tense, in that order. Based on that study, Marcus et al. (1992)[4] assessed overregularization on CHILDES corpus of 83

[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/)**

[3] *Brown R. A first Language-The early Stages*. 1973. P. 274, Harvard University Press.

[4] *Marcus G.F., Pinker S., Ullman M., Hollander M., Rosen T.J., Xu F.* Overregularization in Language Acquisition, Monographs of the society for research in child development, 1992, 57 (4, Serial No. 228).

children. He stated that the overregularization errors are relatively rare but U-shaped development was confirmed. Earlier studies on inflectional morphemes were executed on a limited number of subjects insufficient for them to generalize child language development with their results.

In this paper, we focus on the inflectional morpheme development on the CHILDES corpus from 1 to 7 years of age, which accounts for 97% of the entire CHILDES corpus. First, we reorganized the existing files saved only with the researchers' names into files with the names of the researchers, child age and country. If there was more than one child in a file, we made each child a separate file by changing their name tier to CHI. Table 1 shows the files we examined for the study.

*Table 1*. Number of Files of CHILDES CORPUS (Ages of 1~7)

| Country / Age | Number of files | |
|---|---|---|
| | **UK** | **USA** |
| 1 | 203 | 960 |
| 2 | 1,742 | 1,652 |
| 3 | 194 | 1,124 |
| 4 | 51 | 834 |
| 5 | 57 | 665 |
| 6 | 6 | 124 |
| 7 | 19 | 210 |
| **sub-total** | **2,272** | **5,569** |
| **Total** | | **7,841** |

## 2. Inflectional morphemes in English

Individual transcript files of the same age and country were combined into a single folder. The FREQ program in the CLAN software package (**http://childes.talkbank.org/clan/**) counts the number of times that every word is used. The total number of types was 45,435 and the number of tokens was 4,708,936. We extracted the inflectional morphemes from the types of 45,435 by hand and to

verify the accuracy of the extraction we checked on it more than 3 times by cross validation. We grouped the inflectional morphemes by the same morphological features: *–d/–ed* for Regular/Irregular past tense, *–es/–s* for the 3rd person present tense and plural, *–ing* for the present progressive, *–er/–est* for the comparative and superlative, *–s'/'s* for the possessive. With these 5 types of inflectional morphemes, we added one category more, the Pronoun. Since irregular past tense forms like *wore,* superlative forms like *worst* and plural forms like *children* are not filtered by the typical morphemes, we checked all 13,528[5] types one by one and indicated them with *irr* and then classified *irr* by 3 categories; Verb (V), Noun (N) and Adjective (A). As seen below, pronoun is the lowest in type but the highest in tokens resulting in the lowest type per token ratio. Table 2 shows the results.

*Table 2.* TTR & D value of Inflectional morphemes by countries

| TTR | UK | | | | USA | | | |
|---|---|---|---|---|---|---|---|---|
| | **Type** | **Token** | **TTR** | **D** | **Type** | **Token** | **TTR** | **D** |
| -ing | 1,229 | 39,759 | 0.031 | 33.26 | 1,084 | 47,458 | 0.023 | 27.38 |
| -d_ed_irr(V) | 1,006 | 82,474 | 0.012 | 10.26 | 1,472 | 132,405 | 0.011 | 19.23 |
| -er_-est_irr(A) | 217 | 11,499 | 0.008 | 0.77 | 198 | 13,978 | 0.014 | 1.57 |
| -es_-s_irr(N) | 4,245 | 114,524 | 0.037 | 18.18 | 3,905 | 172,631 | 0.023 | 30.78 |
| pronoun | 52 | 165,778 | 0.000 | 2.64 | 51 | 321,019 | 0.000 | 1.82 |
| -s'_-'s | 1,359 | 49,219 | 0.028 | 4.48 | 1,904 | 71,172 | 0.027 | 3.77 |
| **Total** | **8,108** | **463,253** | **0.109** | **11.69** | **8,614** | **758,663** | **0.016** | **14.09** |

### *2.1. Present Progressive*

On the list of Brown's (1973) 14 grammatical morphemes, present progressive is on the top of the list. The correlations between D value and the children's age were not significant, which seems to indicate that children already apply the present progressive morpheme

[5] There is a difference between total number of types (13,628) and the total number of types on table 2. It is because there are the same types of inflectional morpheme in each country.

to diverse verbs from the age of 1. The words with the highest frequencies are *going, doing,* and *coming* in the UK and *going, doing,* and *making* in the USA. As the present progressive form shares the *-ing* form with the gerund, it requires further examination to extract the present progressive words.

*Table 3.* TTR and D value of Present Progressive

| TTR / Age | UK | | | | USA | | | |
|---|---|---|---|---|---|---|---|---|
| | **Type** | **Token** | **TTR** | **D** | **Type** | **Token** | **TTR** | **D** |
| 1 | 97 | 719 | 0.135 | 13.83 | 290 | 3,378 | 0.086 | 27.60 |
| 2 | 1,158 | 33,264 | 0.035 | 33.27 | 637 | 15,904 | 0.040 | 28.14 |
| 3 | 256 | 3,071 | 0.083 | 20.18 | 558 | 10,924 | 0.051 | 23.93 |
| 4 | 131 | 642 | 0.204 | 21.84 | 569 | 11,343 | 0.050 | 26.67 |
| 5 | 154 | 1,009 | 0.153 | 27.07 | 383 | 3,740 | 0.102 | 28.95 |
| 6 | 83 | 264 | 0.314 | 27.07 | 244 | 1,210 | 0.202 | 36.05 |
| 7 | 118 | 790 | 0.149 | 22.48 | 200 | 959 | 0.209 | 27.23 |
| **Total** | **1,997** | **39,759** | **0.153** | **23.68** | **2,881** | **47,458** | **0.106** | **28.37** |

### *2.2. Regular/Irregular Past Tense*

We extracted regular and irregular past tense types and tokens from UK and USA transcripts, resulting in 1,886 types with 82,474 tokens and 3,746 types with 132,405 tokens, respectively.

*Table 4.* TTR & D value of Regular and Irregular Past tense

| TTR / Age | UK | | | | USA | | | |
|---|---|---|---|---|---|---|---|---|
| | **Type** | **Token** | **TTR** | **D** | **Type** | **Token** | **TTR** | **D** |
| 1 | 94 | 1,800 | 0.052 | 5.24 | 223 | 5,145 | 0.043 | 16.34 |
| 2 | 913 | 68,022 | 0.013 | 11.45 | 726 | 33,190 | 0.022 | 19.92 |
| 3 | 262 | 6,474 | 0.040 | 12.63 | 757 | 33,020 | 0.023 | 21.24 |
| 4 | 166 | 1,600 | 0.104 | 15.68 | 820 | 38,482 | 0.021 | 21.00 |
| 5 | 181 | 2,327 | 0.078 | 14.82 | 547 | 13,106 | 0.042 | 22.60 |
| 6 | 108 | 610 | 0.177 | 12.34 | 352 | 4,755 | 0.074 | 24.52 |
| 7 | 162 | 1,641 | 0.099 | 13.39 | 321 | 4,707 | 0.068 | 20.05 |
| Total | 1,886 | 82,474 | 0.080 | 12.22 | 3,746 | 132,405 | 0.042 | 20.81 |

As children get older, the D value of past-tense increased by the age of 5 or 6 and decreased at the age of 7 in both the UK and the USA. The D value of the USA is higher than that of the UK, which means that the USA children applied past-tense morphemes to more diverse verbal words than the UK children did. A marginal correlation was found between the D value and the children's age (English children: $r = 0.643$ $p > .05$; American children: $r = 0.66$, $p > .05$; the critical value of significant correlation coefficient was 0.68). The words with the highest frequencies are occupied mostly by irregular verbs. They were found four times more than regular verbs in both countries.

### *2.3. Comparative and Superlative*

As children get older, the D values of comparative *-er* and superlative *-est* increased by the age of 6 and decreased at the age of 7 in both the UK and the USA. The strong correlations were found between D value and the children's age (English children: $r = 0.779$, $p > .05$; American children: $r = 0.776$, $p > .05$). The words with the highest frequencies are *more* and *better* in the UK and the USA. Since *more* can be used as a noun in a sentence like '*give me more'*, this is also subject to further examination.

*Table 5.* TTR and D value of Comparative and Superlative

| TTR / Age | UK | | | | USA | | | |
|---|---|---|---|---|---|---|---|---|
| | **Type** | **Token** | **TTR** | **D** | **Type** | **Token** | **TTR** | **D** |
| 1 | 11 | 661 | 0.017 | 0.10 | 19 | 1,660 | 0.011 | 0.26 |
| 2 | 76 | 9,940 | 0.008 | 0.74 | 70 | 4,242 | 0.017 | 0.92 |
| 3 | 26 | 420 | 0.062 | 1.24 | 99 | 2,758 | 0.036 | 2.03 |
| 4 | 21 | 126 | 0.167 | 2.89 | 122 | 3,349 | 0.036 | 2.70 |
| 5 | 28 | 194 | 0.144 | 2.92 | 84 | 1,211 | 0.069 | 3.44 |
| 6 | 20 | 57 | 0.351 | 5.82 | 47 | 376 | 0.125 | 4.06 |
| 7 | 17 | 101 | 0.168 | 2.54 | 41 | 382 | 0.107 | 2.17 |
| **Total** | **199** | **11,499** | **0.131** | **2.32** | **482** | **13,978** | **0.057** | **2.23** |

### *2.4. Third Person Singular Present Tense and Plural*

The third person singular present tense is the one of the morphemes acquired latest, and the plural takes the fourth place on the list of Brown's (1973) 14 grammatical morphemes. However, since the inflectional form *–es* and *–s* are shared by the two, it requires further examination.

*Table 6.* TTR and D value of Third Person Present tense and Plural

| TTR / Age | UK | | | | USA | | | |
|---|---|---|---|---|---|---|---|---|
| | **Type** | **Token** | **TTR** | **D** | **Type** | **Token** | **TTR** | **D** |
| 1 | 315 | 2,577 | 0.122 | 43.72 | 712 | 8,228 | 0.087 | 55.98 |
| 2 | 3,865 | 97,058 | 0.040 | 48.95 | 1,882 | 49,940 | 0.038 | 24.88 |
| 3 | 588 | 8,163 | 0.072 | 19.17 | 1,850 | 44,227 | 0.042 | 22.53 |
| 4 | 298 | 1,881 | 0.158 | 19.74 | 1,990 | 44,945 | 0.044 | 21.10 |
| 5 | 356 | 2,466 | 0.144 | 28.67 | 1,157 | 16,499 | 0.070 | 23.64 |
| 6 | 200 | 630 | 0.317 | 42.40 | 743 | 4,943 | 0.150 | 41.86 |
| 7 | 323 | 1,749 | 0.185 | 44.65 | 537 | 3,849 | 0.140 | 27.12 |
| **Total** | **5,945** | **114,524** | **0.148** | **35.33** | **8871** | **172,631** | **0.081** | **31.01** |

Regarding the frequency of the third person present tense, the word with the highest frequency is *is*, followed by *does* and *goes* in

both UK and USA. As for plural, the words used to indicate things that a child plays with, like *toys*, and body-related words, like *eyes* and *feet*, occupied the high position of frequency. Also, the plural forms of demonstrative pronouns, like *these* and *those*, are frequently used by children in the UK and the USA.

### *2.5. Possessive*

Since the possessive form shares *–'s* with the shortened form of *be*'s third person present tense and the shortened form of *have*'s third person present tense, it should be revised by examination of their usages. *Mummy's* and *Daddy's* are highly frequently used words. Children tend to apply the form *–'s* to the names of people around them.

*Table 7.* TTR and D value of Possessive

| TTR / Age | UK | | | | USA | | | |
|---|---|---|---|---|---|---|---|---|
| | **Type** | **Token** | **TTR** | **D** | **Type** | **Token** | **TTR** | **D** |
| 1 | 97 | 1,515 | 0.064 | 3.61 | 175 | 3,810 | 0.046 | 3.62 |
| 2 | 1,170 | 40,061 | 0.029 | 4.41 | 843 | 27,934 | 0.030 | 3.69 |
| 3 | 235 | 4,163 | 0.056 | 4.42 | 733 | 16,322 | 0.045 | 3.82 |
| 4 | 90 | 873 | 0.103 | 4.11 | 790 | 15,462 | 0.051 | 4.03 |
| 5 | 110 | 1,475 | 0.075 | 3.65 | 349 | 5,117 | 0.068 | 3.75 |
| 6 | 40 | 272 | 0.147 | 4.20 | 136 | 1,307 | 0.104 | 3.96 |
| 7 | 70 | 860 | 0.081 | 3.93 | 128 | 1,220 | 0.105 | 3.58 |
| **Total** | **1,812** | **49219** | **0.079** | **4.04** | **3154** | **71,172** | **0.064** | **3.79** |

### *2.6. Pronoun*

The pronoun has the lowest TTR among the 6 groups. The most frequently used word in the pronoun group is *I*, followed by *you, my, me, he*.

*Table 8.* TTR and D value of Pronoun

| TTR / Age | UK | | | | USA | | | |
|---|---|---|---|---|---|---|---|---|
| | **Type** | **Token** | **TTR** | **D** | **Type** | **Token** | **TTR** | **D** |
| 1 | 27 | 1,674 | 0.016 | 1.63 | 36 | 9,054 | 0.004 | 1.52 |
| 2 | 52 | 135,933 | 0.000 | 3.73 | 45 | 87,927 | 0.001 | 1.87 |
| 3 | 41 | 16,480 | 0.002 | 2.74 | 48 | 82,137 | 0.001 | 2.75 |
| 4 | 34 | 3,268 | 0.010 | 3.31 | 48 | 94,036 | 0.001 | 3.32 |
| 5 | 40 | 4,400 | 0.009 | 3.40 | 46 | 29,666 | 0.002 | 2.76 |
| 6 | 28 | 1,163 | 0.024 | 3.10 | 42 | 9,581 | 0.004 | 3.03 |
| 7 | 34 | 2,860 | 0.012 | 3.17 | 41 | 8,618 | 0.005 | 2.47 |
| **Total** | **256** | **165,778** | **0.011** | **3.01** | **306** | **321,019** | **0.003** | **2.53** |

## 3. Conclusions

In general, the results showed that the correlations between D value and the children's age were different depending on the morphemes: no correlation with *–ing*, marginal correlation with *–ed*, and strong correlation with *–er/-est*, which confirms to the developmental order in Brown (1973). Overgeneralization errors were found most frequently with past-tense *–ed*, but rarely found with present progressive *–ing*, which also closely related to the earlier acquisition of present progressive than past tense.

As mentioned above, each group of inflectional morphemes is to be examined in further detail to understand the inflectional morpheme development in child language thoroughly. However, our results can be applied to verify earlier studies on child language development. In addition, we introduced a useful method for evaluating child language acquisition by reorganizing the entire CHILDES corpus into an appropriate form for research by ages and countries.