Lexical Growth and Acquisition of Morphological Forms

Filip Smolík
University of Kansas

1. Introduction

When attempting to explain the order in which lexical and inflectional are acquired, the absolute or relative frequency of the forms in the input language is an obvious candidate for a causal factor. After all, children will not learn words or word forms that they don't hear. However, another obvious observation is that some highly frequent categories of words, such as articles, connectives and other functional items, are acquired rather late. The process of acquiring linguistic categories, both lexical (noun, verb) and inflectional (plural, past) is more complex.

Brown (1973) was the first to document a limited relationship between input frequency and the acquisition order of morphological forms. Using rank-order correlation coefficient, he did not find any detectable relationships between the input frequency of 14 grammatical morphemes and the order of acquisition. This led to the hypothesis that the order of acquisition is determined by syntactic *complexity* or semantic *generality*. However, empirical studies in this area did not provide unequivocal results because syntactic and semantic variables are confounded with frequency of the items and categories. For instance, Bloom, Merkin, and Wootten (1982) analyzed the order of acquisition of wh-words and concluded that the order depends on the complexity of their syntactic functions as well as on the semantic generality of the verbs that typically combine with each wh-word. However, Rowland, Pine, Lieven, and Theakston (2003) pointed out that the predictions made by semantic and syntactic factors are very similar to those made by the frequency of individual wh-forms. In their own analysis, Rowland et al. analyzed the acquisition order of wh-words and their combinations with verbs. Using input frequency and semantic/syntactic factors as concurrent predictors, they only found detectable relationship with input frequency of wh-words and their combinations, and no effects of semantic/syntactic variables. Similarly, Theakston, Lieven, Pine, and Rowland (2004) studied the acquisition order of verbs, and after accounting for input frequency, they did not find any effect of semantic generality on the acquisition sequence.

The last two studies are an example of a renewed interest in the effects of input frequency on the acquisition of lexical items and linguistic categories. This interest is related to the work in the lexicalist-constructivist framework (Tomasello, 2000; MacWhinney, 2002), which assumes that young children do not have gram-

matical rules represented in terms of abstract linguistic categories. The early language, according to this position, is represented in terms of individual forms and their combinations.

Some authors working in constructivist framework argued that input frequency may be responsible for the order of acquisition of morphosyntactic forms. Joseph, Serratrice, and Conti-Ramsden (2002) found similar distribution of copula and auxiliary forms in children's and mothers' language and suggested that frequency of the forms may be a factor responsible for the order of their acquisition. Serratrice, Joseph, and Conti-Ramsden (2003) found correlations between maternal frequency and the order of acquisition in past tense forms. They claim that "... the lexical statistics of maternal input can account for between approximately one-half and three-quarters of the variance seen in child productions of past tense forms." (p. 341) However, this claim is somewhat problematic as it is based on frequency in child data, not age of acquisition, and it is an aggregate figure for the 16 most frequent forms.

We see that a notable line of research today emphasizes the role of maternal frequency in language acquisition. This suggests a question if the input frequency influences the acquisition order of *categories* of word forms, as opposed to individual words. The present study attempts to test the this possibility. In particular, it asks the following questions:

- can the differences between acquisition timing of different morphological forms be explained by their different average frequency in the input language?
- how strong is the relationship of input frequency with the order of acquisition, compared to other predictors?
- are forms belonging to different morphosyntactic categories equally influenced by input frequency?

2. Data

The present study analyzed the data from Manchester corpus (Theakston, Lieven, Pine, & Rowland, 2001), which is available via CHILDES (MacWhinney, 2000). The data consists of ca. 34 transcribed sessions from each of 12 children, 6 boys and 6 girls. The age of the children is between 1;8 and 3;0 years, and the period of observation for each child is approximately 1 year, with the sessions evenly spaced. Each session lasted ca. 30 minutes. The mean length of utterance (MLU) in the transcripts ranges between 1.06 and 4.1. In general, there are over 10 000 utterances available for each child.

The analytic method described below required to extract detailed information about the first-time usage of individual noun and verb forms in each child. The first step was to find the unmarked (uninflected) word forms for all nouns and

Table 1: Numbers of nouns and regular and irregular verbs, and number of inflected target forms in each category.

	Nouns	Regular verbs			Irregular verbs
	plural	3s	reg. past	prog.	irreg. past
Targets	1417	258	393	1110	223
Total	5183	2486	2545	2677	520

verbs used by each child. For each such unmarked word, the marked target forms were searched. The target forms were 3sg present, progressive, and regular or irregular past for verbs, and plural for nouns. For both marked and unmarked forms, the age of first-time usage was searched. Therefore, the children's data set consisted of the data about first time occurrences of unmarked and five marked word forms, separately for each of the 12 children. In other words, the verb "walk" could be included up to 12 times in the analysis, if all children used the verb in the unmarked form.

All the analyses and counts were performed automatically using custom-written Perl programs. The programs used the part-of-speech and suffix marks on the `%mor` line of the CHILDES transcripts.

Because the total number of different verbs and nouns in all children would be extremely large, the total number of observational units was reduced by only including words that occurred in two different inflectional forms in at least one of the children. The total numbers of nouns and regular and irregular verbs across all children can be found in Table 1, along with the number of marked forms found for each target category. Note that the total numbers of observational units for progressives and 3 sg. present are slightly different even though the set of verbs that can be inflected for progressive and 3 sg. present is the same. These differences are due to the fact that some children did not use the target words neither in the marked nor in the unmarked form.

The analysis also included the input frequency count for each marked and unmarked form. The input frequency estimate was taken from mothers' utterances 7 early samples. Child data from these samples were not included in the analyses to avoid spurious findings due to common discourse situation or children's imitations. Maternal frequency of each marked target form was calculated separately for each mother, and used as a predictor of individual child's use of such marked form.

3. Method

The study used survival analysis and Cox regression to analyze the emergence of inflectional word forms. These methods are suited for the analysis of event occurrence, and the occurrences studied here are the first-time usages of marked inflectional forms of verbs and nouns. The challenge in studying event

occurrences is that within a set of observational units, the event of interest may not occur during the observational period. Methods like standard regression analysis are not appropriate because the dependent variable, time of occurrence (e. g. age of acquisition), is not available for all observational units (the phenomenon is called *censoring*).

Survival analysis estimates a nonparametric *survival curve*. The curve describes the relative number of observational units for which the event of interest did not occur yet. I. e., the initial survival ratio is 1 because no events occurred yet (100 % of observational units have no events observed). The value of the survival ratio at the end of observational period is the relative number of censored observations, i. e. units for which the time of the event is not available.

Survival curves are used to estimate the *hazard* of the event occurrence; in the present context, the term *hazard* is nothing else than the estimated rate of acquisition of the inflected form in a particular time of the development. The estimated hazard at a certain time point gives, in an arbitrary metrics, the rate of first-time usages of word forms in the category of interest (e. g. plural).

The estimated hazard curves are used in Cox regression to estimate the influence of categorical or continuous predictors on the event hazard (acquisition rate, in this study). This study examined the following categorical and continuous predictors:

- category membership: 3 sg. present, regular or irregular past, progressive, plural
- maternal frequency (logarithm) calculated from 7 early samples, as described above
- frequency of the unmarked form (logarithm) in a child's language
- age of first-time occurrence of the unmarked form (logaritgm)

Category membership and maternal frequency are the key predictors in question here. If there are no effects of category membership and significant effects of input frequency, the input frequency is likely to be responsible for the different acquisition timing of different morphological forms. On the other hand, if category membership is a significant predictor, the input frequency is not a sufficient explanation for the timing differences; such a finding would suggest that some categories of morphosyntactic forms are acquired earlier than others, regardless of the frequency with which they are used in maternal language.

For many word forms, the observed maternal frequency in the samples was zero. Therefore, the logarithm of maternal frequency was taken from observed frequency + 1. This transformation reflects the fact that expected input frequency of the forms not attested in maternal input (zero observed frequency) is, on the average, lower than the expected frequency of forms that occurred in the input at least one time. However, the number of forms not attested in the input was large

(about 85 % of the observational units). Therefore, two sets of analyses were performed, one involving all the words included in the study, and the other only involving words with nonzero observed maternal frequency.

The frequency of unmarked forms in children's language served as a partial correction for the sampling bias (for discussion of sampling bias in the study of longitudinally collected language samples, see Tomasello & Stahl, 2004).

The calculations used ages in days. For each child, the age was recentered according to the lowest MLU value available for all children (reference MLU = 1.71). In this transformation, all children's age was set equal at the time point when the MLU value of their language samples was closest to the reference value. This method preserves the time interval metrics in days but corrects for differences in the onset time of individual children's development (Smolík, 2004).

The analyses were performed using the statistical package R (R development core team, 2003), particularly the library `survival` (Lumley, 2004).

4. Results

The results of the analyses are summarized in Tables 2 and 3. The tables show hazard ratios and their 95 % confidence intervals; the hazard ratio is interpreted as the increase or decrease in hazard (i. e. acquisition rate, in our context) associated with a unit increase of a predictor. For categorical predictors, it is the relative acquisition rate for a category in comparison with a reference category, which was the 3 sg. present form. For the continuous predictors here, a unit increase actually means an increase by the factor of 2.72 because of the log-transformations.

The first analysis showed significant main effects of all predictors included in the analysis: maternal frequency ($z = 3.81$, $p < 0.001$), frequency of the unmarked form ($z = 18.31$, $p \ll 0.001$), first-time occurrence of the unmarked form ($z = 5.60$, $p < 0.001$), and each category's difference from 3sg forms ($p \ll 0.001$ for all four remaining categories). Besides the main effects, there were several significant interactions. Interaction of form level and maternal frequency suggested that plurals and regular past were influenced by input frequency more than the remaining categories ($p \ll 0.001$ for both). Also significant was the interaction of form level and frequency of unmarked form: the smallest effect showed in 3sg forms, strongest effect in regular past, with the remaining morphological forms in-between. Finally, the analysis also indicated significant but not particularly large interactions between maternal frequency and the frequency of unmarked form, and between the unmarked form frequency and age of its first occurrence.

The second analysis, which excluded observations with zero observed frequency in maternal corpus, used only about 15 % of the observations available. However, the overall pattern of results was similar. One important difference was that the main effect of maternal frequency was no more significant. This indicates that the first analysis, which used many observations with equal minimal value of observed maternal frequency, did *not* attenuate the effects of maternal frequency

Table 2: Hazard ratios in the analysis that included 12633 observations.

Predictor	Hazard ratio	.95 % lower	upper
Mat. freq. (log)	***1.51	1.22	1.88
Plural	***7.03	5.08	9.72
Progressive	***15.37	11.10	21.30
Irregular past	***5.87	3.37	10.22
Regular past	***6.13	4.33	8.68
Freq. of unmarked form (log)	***2.14	1.97	2.32
Acquisition age of the zero-form (log, 3 months)	***0.76	0.69	0.83
Mat. frq. $\times$ plural	***2.26	1.89	2.70
Mat. frq. $\times$ prog.	0.89	0.73	1.08
Mat. frq. $\times$ irreg. past	0.89	0.75	1.07
Mat. frq. $\times$ reg. past	***2.65	2.11	3.34
Unmarked frq. $\times$ plural	***0.72	0.65	0.79
Unmarked frq. $\times$ prog.	***0.73	0.66	0.80
Unmarked frq. $\times$ irreg. past	0.88	0.75	1.03
Unmarked frq. $\times$ reg. past	***0.52	0.47	0.57
Mat. frq. $\times$ base frq.	***0.91	0.88	0.95
Base frq. $\times$ zero age	*1.03	1.00	1.07

$^{*}p < 0.05$, *** $p < 0.001$

on the acquisition sequence. Also, it suggests that the effect of input frequency is mainly due to the differences between rare forms and the rest of the input. In other words, once the form is present in the input to the extent sufficient for detecting it at least once ca. 4-hour-long input sample, the differences between less and more frequent input forms do not show across-the-board influence on the acquisition order.

Unlike in the complete analysis, the second, reduced analysis did not indicate significant interactions between the continuous predictors. This may be due to the decreased sample size; the interactions were not particularly large in the first analysis. The pattern of interactions between inflectional category levels and continuous predictors (maternal frequency, frequency of unmarked form) is similar. The differences in absolute values of estimated hazard ratios can be attributed to the differences in sample size and sampling error.

The results of the the first analysis (using the complete data set) are represented graphically in Figure 1. For each category, the figure shows estimated rate of acquisition over time for two groups of forms in a given category: high-frequency (above median) and low-frequency. The figure illustrates clearly that the effects of input frequency are not uniform across categories. Note that the two curves are almost identical in progressives and irregular past.

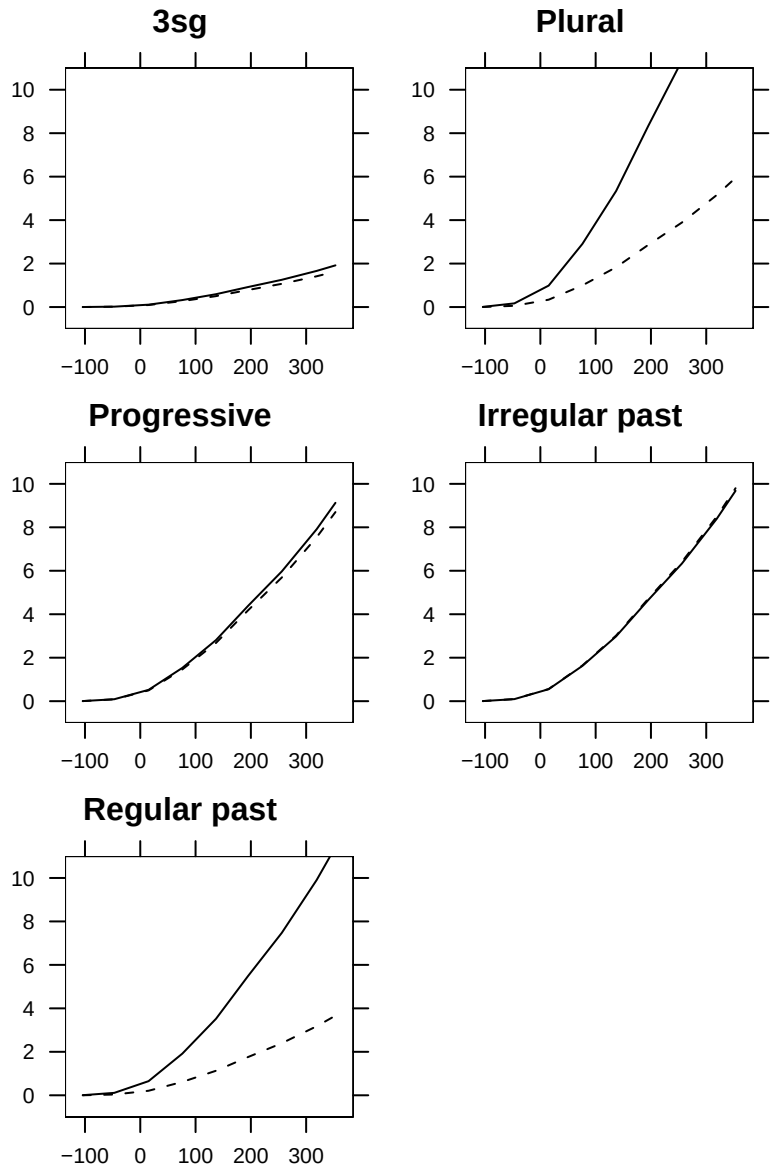

Figure 1: Estimated hazard curves for words with high (full line) or low (broken line) input frequency in each category. Remaining predictors are set to the mean value within each category. X-axis: hazard (common arbitrary scale); Y-axis: time in days (with 0 corresponding to the centering age with MLU=1.7).

Table 3: Hazard ratios in the analysis excluding the forms with zero maternal frequency (1873 observations).

Predictor	Hazard ratio	.95 % lower	upper
Mat. freq. (log)	1.036	0.862	1.246
Plural	***17.539	7.943	38.727
Progressive	***15.665	6.755	36.330
Irregular past	**4.128	1.649	10.332
Regular past	***15.534	6.831	35.323
Freq. of unmarked form (log)	***1.905	1.584	2.292
Acquisition age of the zero-form (log, 3 months)	***0.855	0.787	0.928
Mat. frq. × plural	***1.726	1.393	2.140
Mat. frq. × prog.	1.056	0.807	1.382
Mat. frq. × irreg. past	0.919	0.740	1.142
Mat. frq. × reg. past	**1.544	1.137	2.098
Unmarked frq. × plural	***0.654	0.534	0.801
Unmarked frq. × prog.	***0.679	0.548	0.840
Unmarked frq. × irreg. past	0.931	0.736	1.178
Unmarked frq. × reg. past	***0.530	0.430	0.653

$^{*}p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6. Conclusions and discussion

The results of the present study demonstrate that inflected forms in different morphosyntactic categories show different rate of acquisition, even if effects of input frequency and other variables are removed. Inter-categorial differences in input frequency *cannot* explain why some categories are acquired earlier than others. Specifically, the findings show that plurals and past forms (regular and irregular) have higher rate of acquisition than 3 sg. present forms, and progressive forms are acquired even faster than forms of plural and past. The effect of maternal frequency on the rate of acquisition *differs* between categories: higher input frequency increases rate of acquisition for plurals and regular past more than for the remaining forms studied here. Also, the frequency of unmarked form in children's language and acquisition timing of the unmarked forms influence the probability of acquisition of the marked form. This indicates different forms of a lexeme are not independent, i. e. that children have some control over lexical items regardless of their particular forms.

In general, the findings show that the input frequency by itself is not a crucial causal factor in the acquisition of categories. From early steps in the language development, there is something that makes progressives easier to acquire than e. g. 3 sg. present forms, and it is not just input frequency. The results indicate

that effects of linguistic categories are detectable before the age of 3. It does not mean that the inflectional categories are represented in children's grammars in the same form as in adults; however, the constructivist claim that early knowledge of language is purely item-based is not supported by these findings. Further research is needed to elucidate mechanisms responsible for the differences between categories, and for the sensitivity of developing language system to these categories.

Acknowledgments

The study was supported by NICHD core grant 5P30HD002528-37 to Kansas University Mental Retardation and Developmental Disabilities Research Center. The author is also affiliated with the Institute of Psychology of the Czech Academy of Sciences. The grant No. B-7025104 *Documentation and Analysis of the Czech Language Acquisition* awarded by Grant Agency of the Czech Academy of Sciences also supported this study. The author is grateful to Mabel Rice and Susan Kemper for helpful comments on the draft of this study, and to Jim Bovaird for mentorship in statistics. The contribution was typeset using LaTeX, graphics was produced in R.

References

Bloom, L., Merkin, S., & Wootten, J. (1982). Wh-questions: Linguistic factors that contribute to the sequence of acquisition. *Child Development, 53*, 1084–1092.

Brown, R. (1973). *A First Language: The Early Stages*. Cambridge, Mass.: Harvard University Press.

Joseph, K. L., Serratrice, L., & Conti-Ramsden, G. (2002). Development of copula and auxiliary BE in children with specific language impairment and younger unaffected controls. *First Language, 22*, 137–172.

Lumley, T. (2004). *Survival (a library for the statistical package r).* (Available from: http://cran.us.r-project.org/)

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk.* Mahwah, NJ: Lawrence Erlbaum.

MacWhinney, B. (2002). Language emergence. In P. Burmeister, T. Piske, & A. Rohde (Eds.), *An Integrated View of Language Development: Papers in Honor of Henning Wode* (pp. 17–42). Trier: Wissenshaftliche Verlag.

R development core team. (2003). *R (programmable environment for statistical computing)*. Vienna. (Available from: http://www.r-project.org)

Rowland, C. F., Pine, J. M., Lieven, E. V. M., & Theakston, A. L. (2003). Determinants of acquisition order in wh-questions: Re-evaluating the role of caregiver speech. *Journal of Child Language, 30*, 609–635.

Serratrice, L., Joseph, K. L., & Conti-Ramsden, G. (2003). The acquisition of past tense in preschool children with specific language impairment and unaffected controls: Regular and irregular forms. *Linguistics, 41*, 321–349.

Smolík, F. (2004). *MLU-based age recentering: a tool for studying developmental trajectories.* (Poster presented at The 25th Annual Symposium on Research in Child Language Disorders, Madison, WI, June 4.)

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: an alternative account. *Journal of Child Language,, 28*, 127–152.

Theakston, A. L., Lieven, E. V. M., Pine, J. M., & Rowland, C. F. (2004). Semantic generality, input frequency and the acquisition of syntax. *Journal of Child Language*, *31*, 91–99.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, *74*, 209–253.

Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language*, *31*, 101–121.