Psychology of Language and Communication 2001, Vol. 5, No. 2

ULLA VERES
Göteborg University

A CONTRASTIVE STUDY OF THREE DIFFERENT SWEDISH LANGUAGE CORPORA APPROXIMATED AS INPUT FOR CHILDREN'S ACQUISITION OF PAST TENSE[1]

This paper reports results from a contrastive study of three Swedish language corpora approximated as input to children. Most research addressing the role of input derives its input data from corpora of written language. This study has been done in order to shed light on whether corpora of written language are comparable to corpora of spoken language or of child-directed adult speech (CDS). The corpora were analyzed in terms of type and token frequencies of verbs marked for past tense. The results of the present study show that the frequency distributions are rather similar in the corpora. In the corpus of CDS, however, a small set of verbs is used more frequently in past tense than in the other corpora. These results are discussed in relation to some first results from a study on input and production in a longitudinal case study which show that speech addressed to children, i.e the input over time, is dynamic. The corpora based on communication between adults are static in the sense that they do not change over time. The dynamic interplay between the child's own production and the input to the child will be neutralized in a model that uses static corpora as input.

Introduction

Recent research on past tense morphology in children's first language acquisition has focussed on the role of input factors such as type and token frequencies. (Plunkett & Marchman, 1991; Bleses, 1998; Ragnarsdottir, Simonsen & Plunkett, 1999). It has been shown that this kind of input has an impact on children's acquisition of inflectional morphology. Past tense inflection varies between strong/irregular and weak/regular inflection and therefore provides information about how children gradually learn the morphology of their language. Children's acquisition of past tense has for this reason been studied within different research paradigms during the last decades. With few exceptions, previous research addressing the role of the input has derived its input data from corpora of written language. One reason for using other corpora to approximate children's input is that the sizes of corpora of child-directed adult speech often have been small and the datapoints

[1] I wish to thank Sally Boyd for valuable discussions on earlier drafts of this paper. Address for correspondence: Ulla Veres, Department of Linguistics, Göteborg University, Box 200, 405 30 Göteborg, Sweden, e-mail: ullav@ling.gu.se

few. It is, however, an open empirical question whether the frequency distributions in corpora of written language are comparable to the frequency distributions in corpora of spoken language or of child-directed adult speech, and which of them predicts children's production data the best. The input to preschool children consists mainly of spoken language and the question is whether written language corpora approximate the input that 2 – 8 year olds are modelling their acquisition of past tense on. Research on input to children has shown that child-directed adult speech is different from adult-directed adult speech in several aspects, for example prosodically and syntactically (Snow 1977; Snow, 1995). This is an important reason for trying to shed light on whether the frequency distributions for past tense in a corpus of adult-directed spoken language are the same as those in a corpus of child-directed spoken language.

In this study I have looked at the frequency distributions of past tense in one written language corpus, one corpus of adult-directed adult speech (henceforth ADS) and one corpus of child-directed adult speech (henceforth CDS). The corpora were analyzed in terms of type and token frequencies of verbs marked for past tense.

The results from this study are discussed in relation to some first results from ongoing analyses of the input and the production of past tense in a longitudinal corpus.

The research questions posed for thepresent study were the following:

– are the type/token frequency distributions in the corpus of written language comparable to the frequency distributions in the corpus of adult-directed adult speech and in the corpus of child-directed adult speech ?
– are the same verbs used in the different corpora or is a small set of verbs used more frequently in past tense in the corpus of child-directed speech than in the other corpora?
– is the input in the corpora comparable to the input to children in two longitudinal case studies?

Swedish verb past tense morphology

Swedish verbs are divided into inflection classes depending on which lexical or phonotactical varieties of the inflectional suffixes they are combined with. The two main types of verbs are lexically determined, the weak class that forms past tense/preterite with one of the dental suffixes -*de* or -*te* and the strong class that forms past tense usually with vowel change in the stem but without suffixation. The weak verbs are additionally divided into three subclasses depending on the phonological form of the end of the stem. The imperative is considered as the stem due to the fact that it is formed without a suffix. (Teleman et al, 1999; Jörgensen & Svensson, 1987; Andersson, 1993; Thorell, 1987).

The first weak subclass contains 4 189 verbs (listed in SAOL, the Swedish reference dictionary)[2]. This class contains verbs with stems ending on an unstressed -a (Teleman, Hellberg & Andersson, 1999), (*kasta-de* "throw", *mlĺa-de* "paint"). This class is by far the largest one and it is generally productive, most novel verbs and loan verbs in the language are going into this class. Additionally, there is a large subgroup in this class containing loan verbs with the derivation suffix -*era*; *fund-era* "reflect", *fung-era* "funktion".

[2] Thanks to Mats Gellerstam and Yvonne Cederholm who provided me with the Swedish verbs listed in SAOL, Svenska Akademiens ordlista, in machine readable form.

Table 1. An overview of the Swedish verbs

		Weak verbs N= 4 515 (97% of the Swedish verbs)				Strong verbs N=129 (3% of the Swedish verbs)
WL N= 4 189		WS N = 322			HW N= 4	Strong N= 129
-de	-de (-era verbs)	-de	-te	-dde	*	vowel change
N=2 682	N=1 507	N=162	N=123	N=37	N=4	N=129

* suffixes similar to the weak verbs, vowel change in the stem (3 of 4)

This subgroup is partly responsible for the fact that this verb class is so much larger than the other verb classes (see Table 1).

In spoken language there is a tendency to omit the preterite suffix *-de* on verbs from this weak subclass. (In the spoken corpora in the study 44% of adult directed speech and 62% of child-directed speech of the past tense tokens lose the suffix) (Veres, 2000).

The second weak subclass contains verbs with a stem ending in a consonant (voiced with the suffix *-te* (*blĺs-te* “blew”) or voiceless with the suffix *-de* (*hör-de* “hear-d”). The third weak subclass is small and unproductive. The verbs in this subclass have a monosyllabic stem (always ending with a stressed long vowel) in past tense and in supine the stem vowel is shortened. Past tense examples of verbs from this class are: *bo-dde* “live-d”, *trodde* “believe-d”. The second and the third weak subclasses contain together a total of 322 verbs. The strong class contains 129 verbs (*gĺ gick gatt* – “go went gone”, *sova sov sovit* “sleep slept slept”).

For the purpose of this study and an ongoing comparison with children’s acquisition of past tense in the Icelandic, Norwegian and Danish languages (Ragnarsdottir, Simonsen & Plunkett, 1999; Bleses, 1998) the verbs are further divided into three classes in this study; the larger weak class (WL) which contains the first weak subclass, the smaller weak class (WS) that contains the second and third weak subclasses, and finally the strong class (S). A small number of verbs can be considered as “halfweak” (HW) (Teleman et al, 1999). The reason is that they have past tense suffixes similar to the weak verbs, but they (three of four) have a vowel change. Compounds and derivations are not included in the above figures.

Data and method

The data for this study were obtained from three language corpora at the Department of Linguistics at Göteborg University. One written corpus includes material from newspapers and novels, comprising 1 million words (Allwood et al, 1999). One corpus of adult-directed adult speech which also comprises 1 million words (Allwood et al, 1999) was collected in order to cover as much variation as possible in the use of spoken language.The type of variation aimed at was in terms of activity rather than dialect and social class. The spoken language material is transcribed in MSO6 (Modified Standard Orthography), (Nivre, 1999). MSO-transcriptions are somewhat more like the spoken language than Swedish standard orthography but not as detailed as a phonetic transcription.

The corpus of child-directed adult speech comprises 455000 words (Strömqvist & Richthoff, forthcoming). In that corpus five Swedish-speaking children were recorded between the ages of 18 months and 47 months in interaction with one or more adults: mother, father and/or grandparents. The children and adults participated in different activities like conversation, playing, looking at picture books and meal times. All datapoints are transcribed in CHILDES/ CHAT format (MacWhinney, 1991). (For this contrastive study only the input data, i.e the adult verb forms, were used).

Emacs, a standard program for UNIX[3], was used to excerpt the verbs from the corpora and for the coding of the past tense forms. All verbs were excerpted from each corpus and put together in a separate file/corpus. The verb forms were coded and the past tense forms were analyzed in terms of type and token frequencies.

Solely the preterite form of the verb is counted as past tense, since the perfect tense refers to a time closer to speech time and does not refer to a certain point of time in the past.

Results and discussion

Type and token frequencies

Frequency factors such as type and token frequencies have been shown in several studies to have an impact on acquisition (Ragnarsdottir, Simonsen & Plunkett, 1999 and Bleses, 1998). Ragnarsdottir, Simonsen & Plunkett conclude in their study that "type frequency is most clearly manifested in order of acquisition effects" while they also found an overall effect of token frequency of verbs; the past tense forms of more frequent verbs were acquired earlier than the past tense forms of less frequent verbs in their study of Icelandic and Norwegian children.

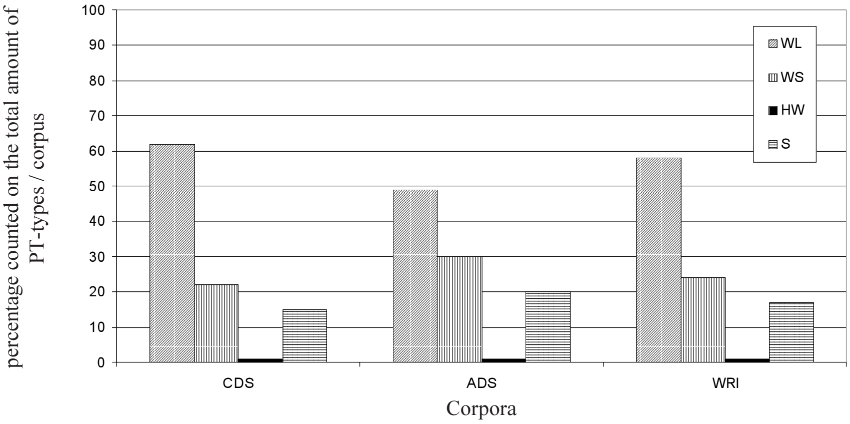
Type frequency refers to the number of verbs that undergo the same kind of change from stem to past tense form (Plunkett & Marchman, 1991), i.e the number of different verbs from each verb class in each corpus.

Figure 1 below shows the distribution of past tense types by verb class and corpus and compares the different corpora.

As the figure shows, there are only small differences between the corpora in this respect. The proportion of past tense types from the larger weak subclass makes up a somewhat larger part of the total number of different verb types in past tense, while the proportion of past tense types from the smaller weak subclass and the strong class thus make up a smaller part in the corpus of child-directed speech compared with the other two corpora. Of all verbs in past tense in the corpus of child-directed speech, verbs from the larger weak class constitute 62%, verbs from the smaller weak class 22%, verbs from the "half weak" class 1% and verbs from the strong class 15%. Of all verbs marked for past tense in the corpus of adult-directed speech, verbs from the larger weak class constitute 49%, verbs from the smaller weak class 30%, verbs from the "half weak" class 1% and verbs from the strong class 20%. Of all verbs in past tense in the written language, verbs from the larger weak class constitute 58%, verbs

[3] Thanks go to Leif Grönkvist who helped me a lot with the UNIX-program

Figure 1. Past tense types / verb class and corpus



from the smaller weak class 24%, verbs from the "half weak" class 1%, and verbs from the strong class 17%. The corpus of child-directed speech was more similar to the corpus of written language than to the other spoken corpus in this respect.

In all three corpora the largest proportion of past tense types was made up of verbs from the larger weak class.

Token frequency is based on how commonly the past tense form from each verb class is used in each corpus.

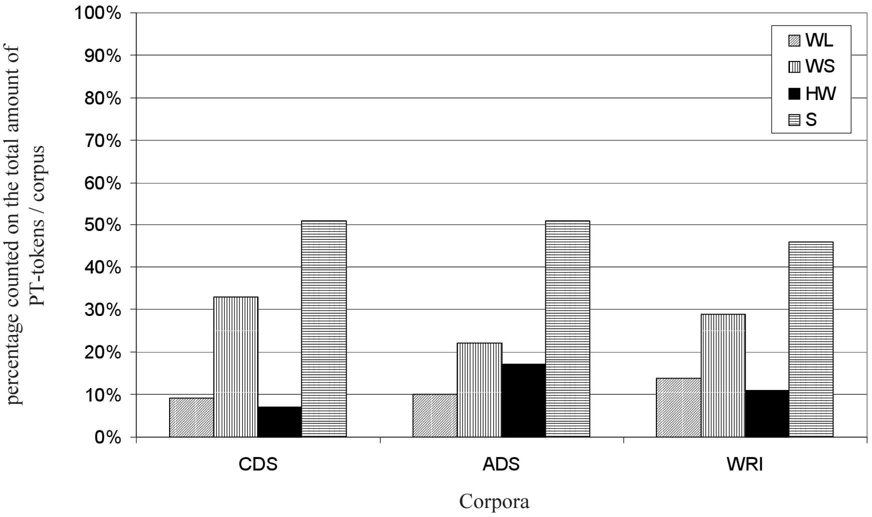
Past tense tokens in relation to the size of each corpus represent 2% in the spoken corpora and 7% in the written corpus. One explanation for the larger number of past tense tokens in the written corpus is that in written language, in particular in newspaper texts, events are often reported that took place in the past. The use of past tense is also a stylistic feature in written texts. It is interesting, however, that there is a correspondence between the spoken corpora in this sense. Child-directed speech is regarded as more constrained to here and now, as was pointed out by Snow (1995), but in the input to the children in the corpus the adults used nonpresent referents to the same extent as in the corpus of adult-directed speech.

Figure 2 shows how the past tense tokens are distributed over the three corpora respectively and in comparison between the corpora.

On the token level as well the results show only small differences between the corpora. In all three corpora the largest proportion of past tense tokens were represented by verbs from the strong verb class.

Of all past tense tokens in the corpus of child-directed speech, verbs from the larger weak class constitute 9%, those from the smaller weak class 33%, those from the "half weak" class 7% and those from the strong class 51%. Of all past tense tokens in the corpus of adult-directed speech, verbs from the larger weak class constitute 10%, those from the smaller weak class 22%, those from the "half weak" class 17% and those from the strong class 51%. Of all past tense tokens in the written language corpus, verbs from the larger weak class constitute 14%, verbs from the smaller weak class 29%, those from the "half weak" class 11% and those from the strong class 46%.

Figure 2. Past tense tokens / verb class and corpus



The corpus of child-directed speech used somewhat larger proportions of past tense tokens from the smaller weak class (WS) and from the strong class (S) in relation to the other corpora. This partly depends on the fact that one type from the strong verbclass is represented by a large number of past tense tokens, namely, the past tense form of the copula *var* "was" and two types from the smaller weak class, namely *gjorde* "did" *Vad* ***gjorde*** *du när du* ***var*** *hos mormor?* "What did you do when you were at Grandma's?" and *sa(-de)* "said" *Vad sa du?* "What did you say?". (See discussion about this small set of verbs below). Another study (Veres, 1999) shows that the past tense forms of these verbs are highly represented in questions directed to children.

Another factor responsible for the high proportion of the copula *var* "was" is the Swedish use of past tense in fixed expressions as *Det var gott* "It tastes good", which is also often used in questions *Var det gott?* "Do you like the food?"

A comparison between Figure 1 and Figure 2 shows that in all three corpora the relation between the distributions of past tense types and past tense tokens is such that the type of verb which had the greatest number of types (WL) had the fewest tokens, while the strong category had few types but many tokens. The smaller weak class shows more even numbers in all three corpora.

In summary, and in answer to the first research question, I conclude that from looking at the frequency distributions in the corpora it is obvious that they are very similar in this respect with only small differences. The type/token frequency distribution in the three corpora seems to be comparable at this level. But is this a general similarity at all levels or is it limited to frequency distributions by verb class?

The second research question for this study was whether the same verbs from the different verb classes are used in the corpora or if a small set of verbs are used more frequently in past tense in the corpus of child-directed speech than in the other corpora? With regard to the results presented above it could be expected that the verbs *var* "was", *sa*

Table 2. The distribution by verb class of the ten most frequently used verb in past tense in the corpora respectively

	CDS		ADS		WRI	
	N	%	N	%	N	%
WL	0	0	244	2.3	0	0
WS	2 058	31.7	4 455	41.9	9 529	27.9
HW	380	5.8	3 278	30.8	5 958	17.4
S	4 058	62.5	2 666	25	18 674	54.7
Total	6 496	100	10 643	100	34 161	100

"said" and *gjorde* "did", may be responsible for such a set of verbs. In order to clarify this question the ten most frequently used past tense forms in each of the three corpora were analyzed.

The ten most frequently used verbs

Some differences between the corpora were revealed when looking at the distribution of the ten most frequently used verbs (hereafter "top ten" verbs) in past tense in the corpora by verb class. In the CDS-corpus the "top ten" verbs comprise 70% of all past tense tokens in the corpus. In the ADS-corpus and in the written corpus the "top ten" verbs comprise 51%. This fact indicates that a small set of verbs is more commonly used in past tense in child-directed speech. Table 2 shows the distribution by verb class of the ten most frequently used verbs in past tense in each of the corpora.

The table shows that past tense tokens from the strong verb class constitute 62.5% of all past tense tokens among the top ten in the CDS-corpus. This is to be compared with

25 % in the ADS corpus and 54.7 % in the written corpus. The WS-verb class also makes up a high proportion in the CDS-corpus: 31.7 % compared to 41.9 % and 27.9 % in the other corpora respectively.

As we can see in Table 3 below, partly responsible for the high proportion of past tense tokens from the strong verb class is the past tense form of the copula verb *var* "was", which alone represents 42% of all past tense tokens among the "top ten" verbs in the CDS-

Table 3. The „top 10" verbs in past tense

CDS	N	%	ADS	N	%	WRI	N	%
var	2 726	42	*skulle*	3 278	30.8	*var*	11 270	33
sa	865	13.3	*hade*	2 260	21.2	*hade*	6 509	19
gjorde	703	10.9	*var*	1 703	16	*skulle*	3 795	11.1
hade	490	7.5	*sa*	1 610	15.1	*sa*	3 020	8.9
fick	478	7.3	*gick*	418	4	*kunde*	2 163	6.3
skulle	380	5.9	*tyckte*	314	3	*kom*	1 741	5.1
kom	261	4	*blev*	299	2.8	*fick*	1 484	4.3
gick	218	3.3	*borde*	271	2.5	*blev*	1 436	4.2
sĺg	191	3	*tog*	246	2.3	*gick*	1 388	4.1
blev	184	2.8	*prata-de*	244	2.3	*sĺg*	1 355	4
	6 496	100		10 643	100		34 161	100

Table 4. The distribution by verb class of the ten most frequently used verb in past tense in the corpora respectively (*var* 'was', *sa* 'said' and *gjorde* 'did' excluded)

	CDS		ADS		WRI	
	N	%	N	%	N	%
WL	0	0	244	3.3	0	0
WS	490	22.3	2 845	38.8	6 509	32.75
HW	380	17.3	3 278	44.7	5 958	30
S	1 332	60.4	963	13.2	7 404	37.25
Total	2 202	100	7 330	100	19 871	100

corpus. In the written corpus var represents 33% and in the ADS-corpus 16% of all past tense tokens among the "top ten" verbs. Partly responsible for the high proportion of past tense tokens from the WS-subclass are the two verbs *sa* "said" and *gjorde* "did". Together they make up 24.2 % of all past tense tokens among the "top ten" verbs in the CDS-corpus. This is to be compared to the other two corpora where *gjorde* "did" is not represented among the "top ten". *Sa* "said" contributes with 15.1 % and 8.9 % in the ADS-corpus and in the written corpus respectively.

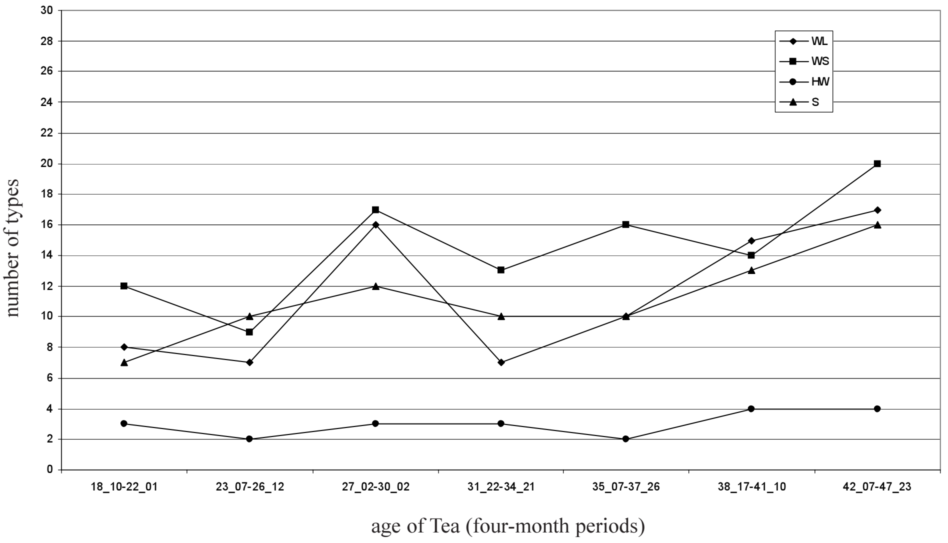
So the corpora are similar to some extent but when comparing the ten most frequently used verbs in the respective corpora excluding (*var* "was", *sa* "said" and *gjorde* "did" (Table 4) we see that the difference between the corpora mainly depends on the fact that in the CDS-corpus the strong verbs are much more frequent (60.4 %) than in the ADS-corpus (13.2 %) and the written corpus (37.3 %). The proportion of past tense tokens from the WS-class is smaller in the CDS-corpus (22.3 %) than in the other corpora (38.8% and 32.7 % respectively).

Are the same verbs from the different verb classes used in the corpora?

In the CDS-corpus we find the verb *gjorde* "did" among the ten most frequently used verbs in past tense. This verb did not occur among the ten most frequent verbs in past tense in the written corpus (and neither in the other spoken corpus). The only other difference between the CDS-corpus and the written corpus is the "half-weak" verb kunde "could" in the written corpus. The strong verbs among the "top ten" were the same in the CDS-corpus and in the written corpus. In the ADS-corpus, on the other hand, there are some verbs from all verb classes (except for the "half-weak class") among the "top ten verbs" which are not represented in the other corpora ; from the WL-class *prata-de* "talk-ed", from the WS-class *tyck-te* "thought" and *bor-de* "should" and from the strong class *tog* "took". The ADS-corpus is different from the other corpora as regards the strong verbs; in this corpus only four strong verbs are represented among the "top ten" verbs; in the other corpora there are six strong verbs.

Concluding this part of the study we see that the corpora are similar to a large extent and that the CDS-corpus shows greater similarity to the written corpus than to the other spoken corpus. These results indicate, at the same time, that a small set of verbs are used more often in the corpus of child-directed speech than in the other corpora, namely three verbs, the past tense form of the copula verb *var* "was" which belongs to the strong verb class, and two verbs from the smaller weak subclass: *gjorde* "did" and *sa* "said". It is known from other studies (Veres, 1999) that those verbs are commonly used in questions and that questions frequently occur in speech directed to children.

Figure 3. Number of past tense types by verb class in the input to Tea



Does this imply that it is of no importance what kind of corpus we use as input data? I will discuss some early results from an ongoing study of the input and the production in a longitudinal corpus which may shed some light on this issue.

The use of past tense in two longitudinal case studies

The data for this part of the study are derived from a Swedish longitudinal case study of four children recorded from 18 to 47 months (Richthoff, 2000). Two children are studied for the purpose of this study. Both are girls, Tea and Bella, aged 18 months to 41 months. They are from two different families and they interact mainly with one of their parents, i.e the input to the two girls comes from different adults. The activities going on include playing, conversation, looking at picture books and story-telling.

For each child 24 datapoints have been used, and the recording sessions are about the same length, around 25 minutes. For readability of the graphs they are shown in periods of four months, each period containing the results of four recordings. They are age-matched as far as possible.

Figures 3 and 4 show past tense types by verb class in Tea's input and produced by Tea respectively, over time in periods of four months.

In the input to Tea verbs in past tense from all four verb classes had been introduced already during the first four-month period. The relatively high number of past tense verbs in this period derives from the first of the four recordings for this period. (The reason could be that, since this is the first recording, the adult is anxious to use an elaborate language in the interaction with the child). In the following periods the input seems to follow the child's development. Most verbs in past tense in the input are from the WS- class throughout most of the periods, which is also the case in Tea's own production when she starts to use the past tense forms. Tea's first WS-verbs (9 different verbs) in the past tense occur for the first time in the third period, i.e 27 to 30 months. In this period verbs from the strong verb class and from the "half-weak" class also occur in past tense.

Figure 4. Number of past tense types by verb class produced by Tea

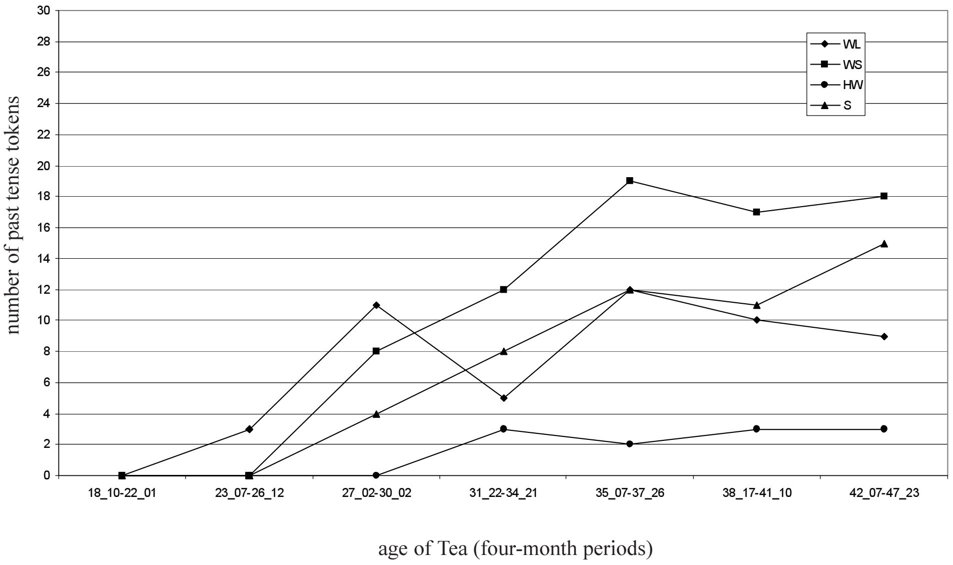


Figure 5. Number of past tense tokens in the input to Tea

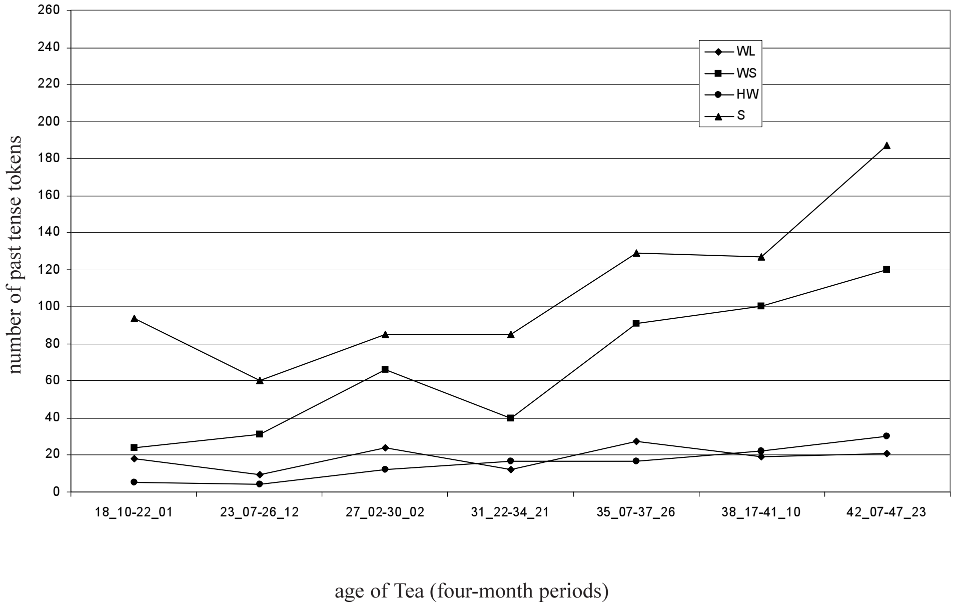


Figure 7. Number of past tense types by verb class in the input to Bella

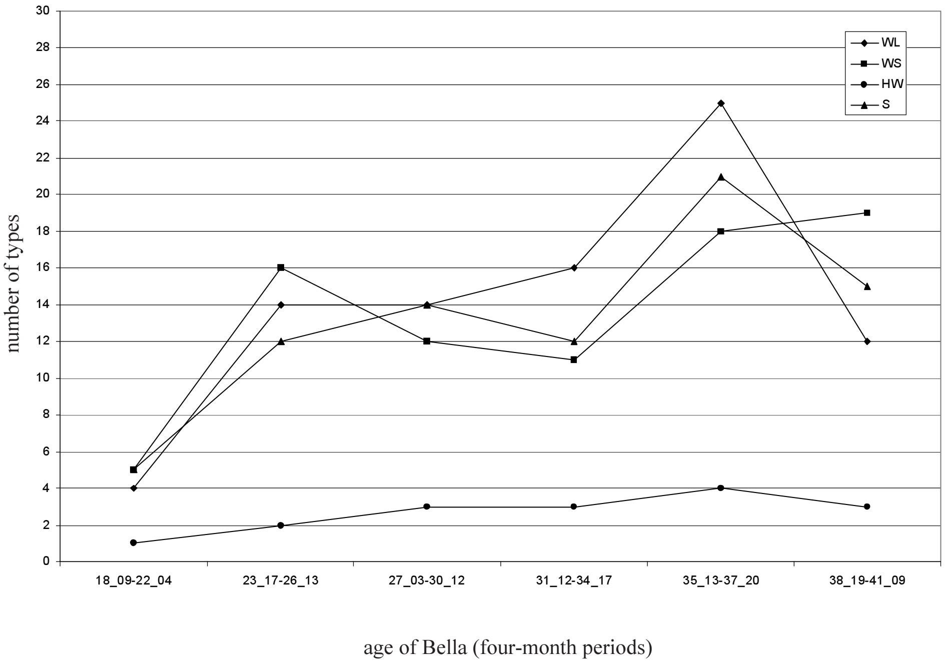


Figure 6. Number of past tense tokens by verb class produced by Tea

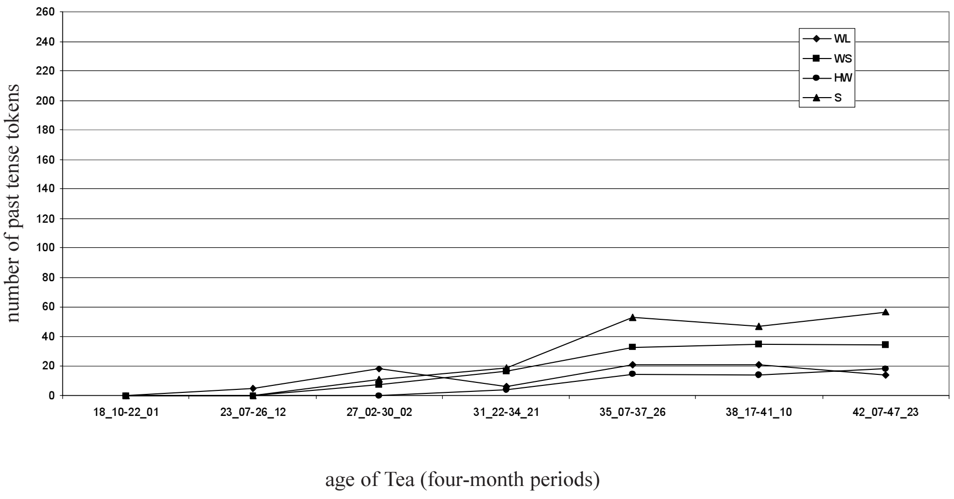


Figure 8. Number of past tense types by verb class produced by Bella

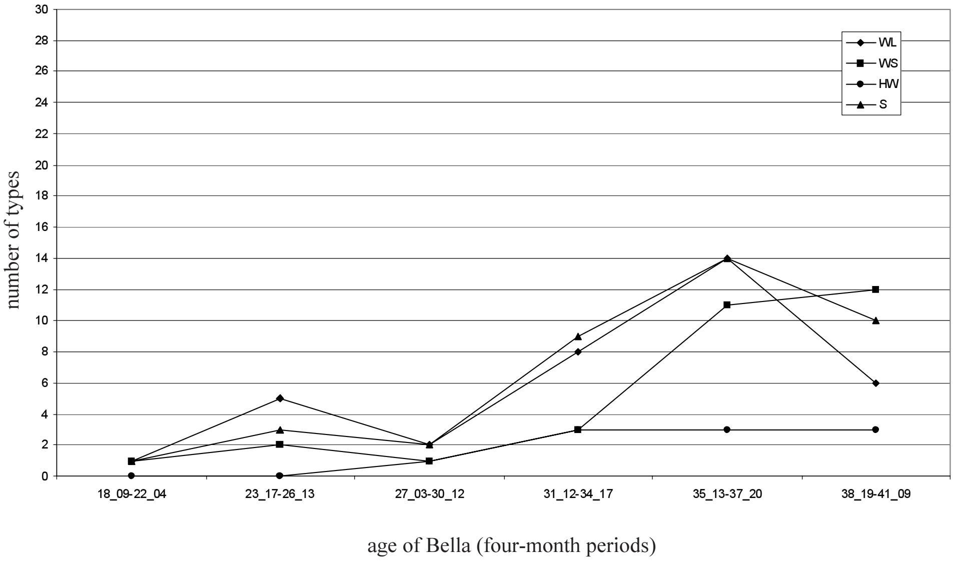


Figure 9. Number of past tense tokens by verb class in the input to Bella

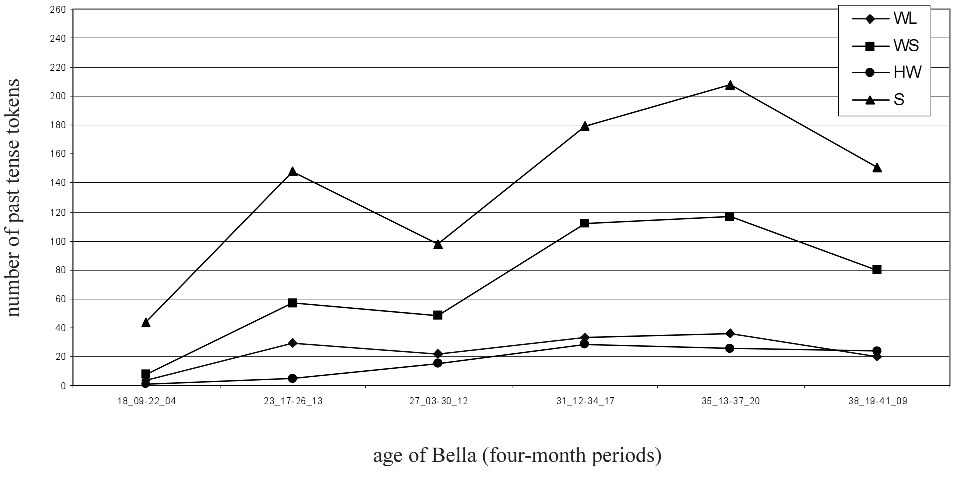


Figure 10. Number of past tense tokens by verb class produced by Bella

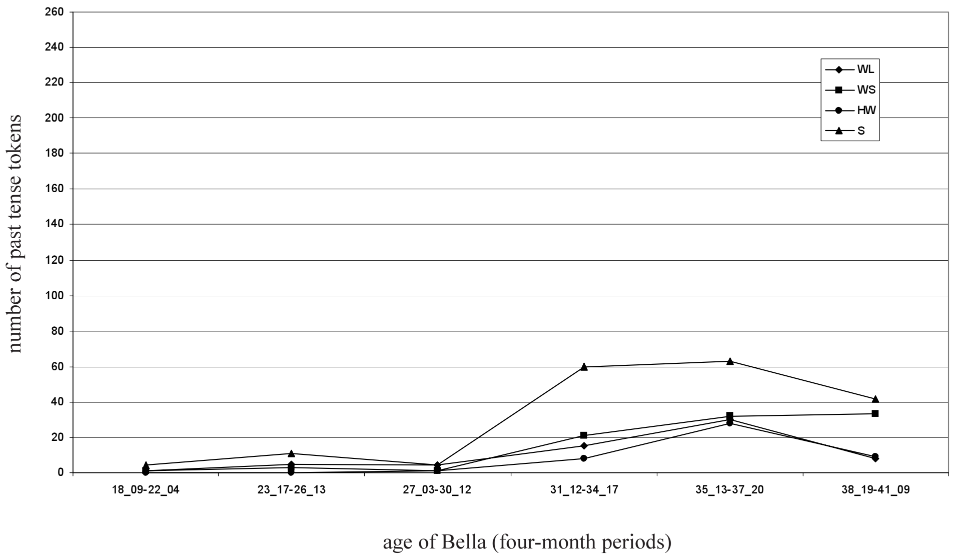


Figure 11. Number of past tense types by verb class (input Bella)

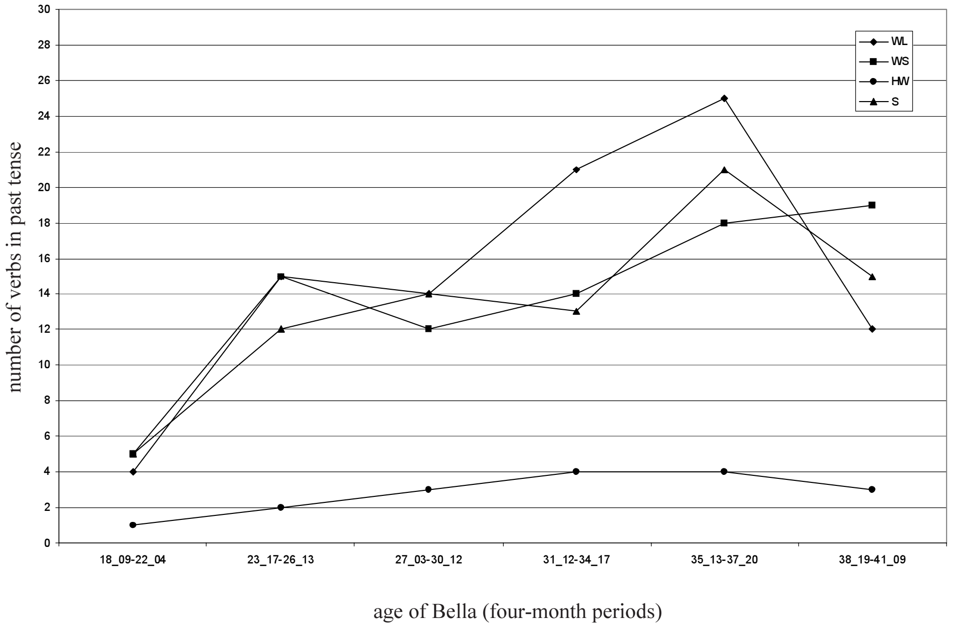


Figure 12. Number of past tense types by verb class (input Tea)

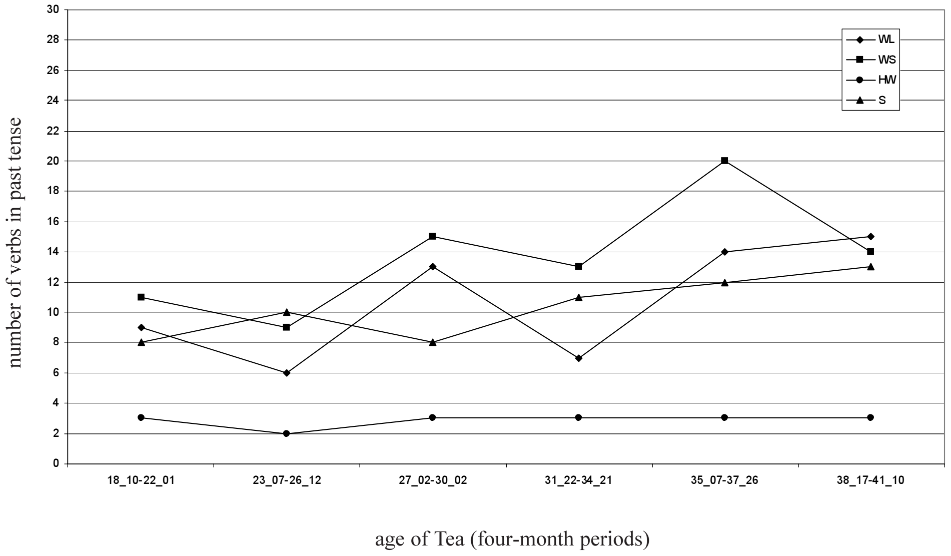


Figure 13. Number of past tense tokens by verb class (input Bella)

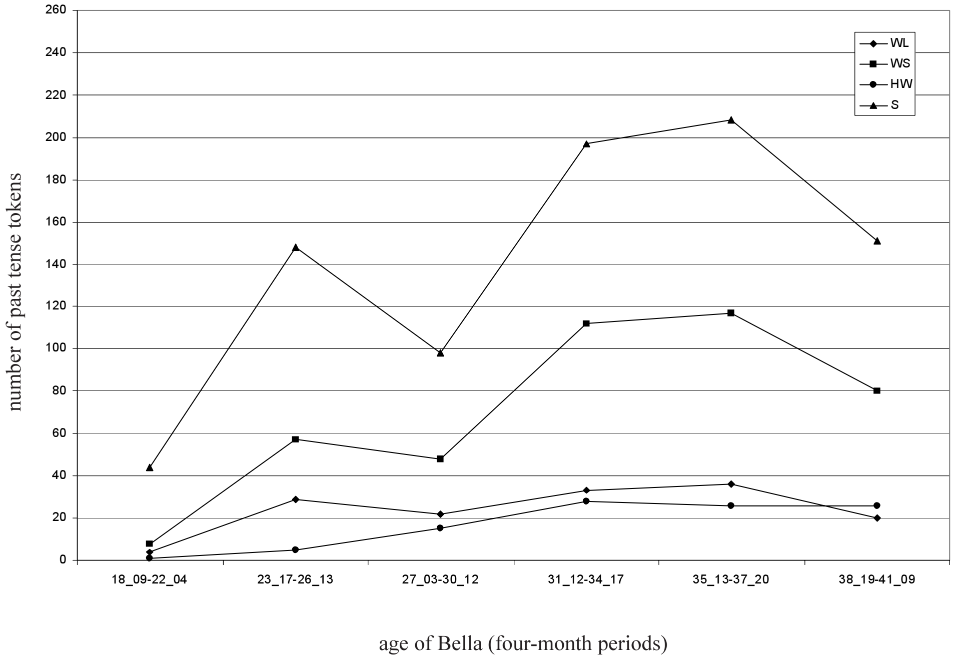
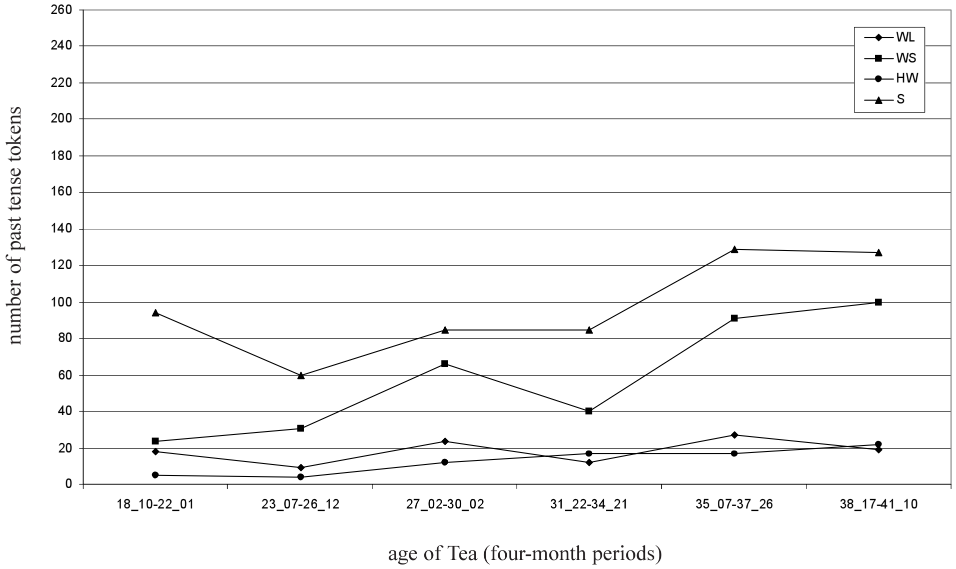


Figure 14. Number of past tense tokens by verb class (input Tea)



Figures 5 and 6 show the past tense tokens in Tea's input and produced by Tea respectively, over time in the same periods of four months. As for the past tense types in the input, there is also a large number of past tense tokens in the input in the first period, mainly of verbs from the strong class. The other periods show a correspondence between the input and the child's production of the past tense tokens from all verb classes. This could be a general pattern in the use of forms, but, as we shall see later when comparing the two children's production, it seems to be a function of parental adjustments to the child's level of development.

Figures 7 and 8 show the number of past tense types by verb classes in Bella's input and produced by Bella respectively, over time, in periods of four months.

In the input to Bella it is even more obvious that the adult is following the child's development, already from the first four-month period. In this period there are a fairly equal number of past tense types from each verb class (WL, WS and S). This is also the case in Bella's own production. Throughout all periods there is a correspondence between the input to Bella and her own production of different past tense forms.

Figures 9 and 10 show the number of past tense tokens in Bella's input and past tense tokens produced by Bella over time in same periods of four months. In all periods there is a correspondence between the input and the child's production of past tense tokens from all verb classes. There is a large number of past tense tokens from the strong verb class in the input in the three last periods and this is the case also in Bella's own production.

When comparing the input to the children regarding the past tense types by verb class, we find that the input is different in the sense that in the case of Bella most types are from the WL- class. In the case of Tea most types are from the WS-class (Fig 11 and 12).

When comparing the input concerning the number of past tense tokens by verb class

(Fig 13 and 14) it is similar in the sense that in the case of both children the greatest number of past tense tokens are from the strong verb class, across all periods. There are differences between the adults, however. In the input to Bella there are more past tense tokens of strong verbs than in the input to Tea, across all periods, and also somewhat more past tense tokens from the WS-class during all periods with the exception of the last period. This is reflected in Bella's own production.

The children's production of both past tense types and past tense tokens differ. More past tense types from the WS- class occur in Tea's production during the later periods and more from the WL- class in Bella's production which is also reflected in the input to the children.

This finding suggests that the graphs not only show a general pattern of use of forms but that this might be a function of parental adjustment to the child's language development.

The input to both children is dynamic in the sense that it follows the child's way of communicating with the adult to a great extent. (There is a difference between the two children as to when the first past tense forms occur. This will be reported in another paper concerning the children's production of Swedish past tense).

A comparison of the ten most frequent verbs in past tense in the longitudinal input with the ten most frequently used verbs in the corpora

In a comparison of the ten most frequently used verbs in past tense, it was found that in the longitudinal input to the children the strong verbs most often used are almost the same as in the corpora. But what is interesting in this comparison is that the input to the children changes over time. With the exception of the strong verbs, the other most frequent verbs vary from datapoint to datapoint. More verbs from both the WS-class and the WL- class are represented as the child gets older.The verbs are not the same in the input to the two children The verbs reflect the kind of conversation and the activity going on. Past tense forms from the WS-class such as *ĺkte* "went by car/bus/boat", *lekte* "played", *hände* "happened" and from the WL-class verbs such as *ramla-de* "fell", *hoppa-de* "jumped", *kasta-de* "threw" occurred among the most frequent past tense forms. None of those verbs are among the most frequent in past tense in the corpora.

Summing up the results from this part of the study, we see that the input of past tense verb forms to children varies over time both concerning the type/token frequencies and verbs used in past tense from the different subclasses. The input in the longitudinal studies is hardly comparable to the approximated input in the language corpora, with the exception of the strong subclass.

Conclusions

The results from this study show that the three language corpora are similar to a large extent when taking into account only the frequency distributions of past tense forms. The corpus of child-directed speech shows in this respect somewhat more correspondence to the written corpus than to the other spoken corpus, which may show that speech directed to children can be seen as more simple and clean than speech addressed to adults. (Snow, 1995).

In the corpus of CDS, however, a small set of verbs is used more frequently in past tense than in the other corpora. Looking at the top ten verbs in the respective corpora, excluding this small set of verbs, the difference between the corpora mainly depends on the fact that the strong verbs are more frequent in the CDS-corpus than in the other corpora. This may partly be explained by the fact that they belong to the so-called nuclear verbs (Viberg, 1993). Those are verbs with very central meaning, and they are often to be found at the top of frequency lists in different languages. Nevertheless, the CDS-corpus still shows more resemblance to the written corpus than to the other spoken corpus.

Does this findings – that the differences between different corpora are relatively small, imply that it is of no importance what kind of language corpora we use to approximate input data? The results from the two longitudinal case studies have cast some light on this issue. As was shown above, in speech addressed to children the use of past tense forms changed from datapoint to datapoint and there was also differences in the use of past tense between the adults who participated in the interaction with the two girls in the study. This indicates that there are significant individual differences as to which kind of input a child gets and it is hard to generalize over different cases. The results also indicate that the input is dynamic in the sense that it is changes over time.

Language corpora based on communication between adults, spoken or written, are static in the sense that they do not change over time. The dynamic interplay between the child's own production and the input to the child will be neutralized in a model that uses static corpora as input.

References

Allwood, J. (1999). *Talspraksfrekvenser. Gothenburg Paper of Theoretical Linguistics*. Göteborg: Göteborg University, Department of Linguistics.

Andersson, E. (1993). *Grammatik fran grunden*. Uppsala: Hallgren & Fallgren Studieförlag.

Bleses, D. 1998. The role of input, productivity and transparency in Danish children's acquisition of past tense morphology. (Dissertation). Odense Working Papers in Language and Communcation.

Jörgensen, N. & Svensson, J. (1987). *Nu svensk grammatik*. Lund: Gleerups.

MacWhinney, B. (1991). *The CHILDES projeckt. Tools for analyzing talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

Nivre, J. (1999). *MSO 6 Modified Standard Orthography*. Göteborg: Göteborg University, Department of Linguistics.

Plunkett, K & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perception. *Cognition*, 38, 43-102.

Ragnarsdottir, R., Simonsen, H.G. & Plunkett K. (1999). The acqusition of past tense morphology in Icelandic and Norwegian children. An experimental study. *Journal of Child Languag*e, 26, 3

Richthoff, U. (2000). *En svensk barnsprĺkskorpus. Uppbyggnad och analyser*. Göteborg: Göteborg University, Department of Linguistics.

Snow, C. (1995). Issues in the study of input: Finetuning, universality, individual and developmental differences, and necessary causes. In P. Fletcher, & B. MacWhinney (Eds.), *The handbook of child langu*age. Oxford: Blackwell

Snow, C. (1977). Mothers speech research: from input to interaction. In C. Snow, & C. Ferguson (Eds.), *Talking to children. Language input and acquisition*. Cambridge: Cambridge University Press.

Strömqvist, S. & Richthoff, U. (forthcoming). Linguistic feedback, input and analyses in early language development. In K. Meng, & S. Strömqvist, (Eds.), *Discourse markers in language acquisition*. To appear in *Journal of Pragmatics*

Teleman, Hellberg & Andersson (1999). *Svenska Akedemiens Grammatik*. Stockholm: Svenska Akademien.

Thorell, O. (1987). *Svensk grammatik*. Stockholm: Scandinavian University Books.

Veres, U. (1999). A contrastive study of linguistic corpora used to approximate children's input. Paper delivered at the Eighth International Congress for the Study of Child Language, San Sebastian

Veres, U. (2000). De svenska verbens 1:a konjugation – en jämförelse mellan SAG och nagra andra svenska grammatikor. In E. Engdahl & K. Norén (Eds.), *Att använda SAG – 29 uppsatser om Svenska Akademiens Grammatik. MISS 33.* Göteborg: Göteborg University, Department of Swedish Language.

Viberg, A. (1993). Crosslinguistic perspectives on lexical organization and lexical progression. In K. Hyltenstam, & A. Viberg (Eds.), *Progression and regression in language*. Cambridge: Cambridge University Press.