

## The input–output relationship in first language acquisition

Heike Behrens

*University of Groningen, The Netherlands*

This study provides an account of the distributional information and the production rates in a particularly rich corpus of German child and adult language. Three structural domains are analysed: the parts-of-speech distribution for a coded corpus of circa one million words as well as the internal constituency of 300,000 noun phrases and almost 200,000 verb phrases. In all three domains, the distribution over time in the adult input is extremely homogenous. The child shows a steady approximation towards the adult distribution. It is argued that two notions of acquisition have to be distinguished: acquisition in terms of the availability of a given structure, for example in terms of first occurrence of a structure or according to various criteria of productivity, and acquisition in terms of full communicative competence, i.e., using structures in the way adults use them (cf. Slobin, 1991, 1997). The data presented here show that the child acquires not only the structural options of German but also highly conventionalised ways of encoding concepts. The amount of information about the structure and conventions of German that is available in the input has the potential of making innate stipulations unnecessary. Instead, the data support usage-based and probabilistic theories of language and language processing.

---

Correspondence should be addressed to Heike Behrens, Afdeling Duitse Taal en Cultuur, Rijksuniversiteit Groningen, Postbus 716, 9700 AS Groningen, The Netherlands. Email: h.behrens@let.rug.nl

The data presented in this paper were collected under my supervision at the Max-Planck-Institute for Evolutionary Anthropology in Leipzig, Germany. I thank Leo and his family for their good spirits in keeping up with the recording schedule, and the research assistants and students who transcribed the data and helped with the coding. Also, I wish to thank my colleagues at Leipzig and Groningen, in particular Mike Tomasello, Elena Lieven, Kirsten Abbot-Smith and Marjolijn Verspoor for fruitful discussions on the matters discussed in this study.

## INTRODUCTION

It is commonplace in many textbooks and popular science articles to read that one of the great miracles of the human language faculty is that children acquire highly complex language very fast and seemingly without effort. Two alternative solutions are offered to account for this. In the nativist tradition it is assumed that innate linguistic representations, Universal Grammar, help children to identify and acquire the linguistic rules which are relevant in their target language (e.g., Crain & Pietroski, 2002). In constructivist and emergentist approaches, no specifically linguistic innate representations are assumed. Instead, it is argued that children are very efficient pattern and intention recognisers so that they can induce linguistic structure based on the language they hear (e.g., Elman, Bates, Johnson, Karmiloff-Smith, Parisi & Plunkett, 1996; MacWhinney, 1999; Tomasello, 2003).

The nativist and emergentist approaches have opposite assumptions about the initial state of the language learner and these different assumptions lead to different learning theories. The nativist approach implies deductive learning processes. Since the core features of language are supposed to be innate, nativists have to provide an account of how and when children activate their linguistic knowledge, and how they link the innate categories or features to the language they hear. For example, evidence for the relevant *linking rules* could be provided by verb semantics (Pinker, 1984, 1989). However, so far the attempts to identify reliable, universal semantic correlates to linguistic categories have failed (Bowerman, 1990). Even though the nativist approach has been quite influential in the past decades, some central claims are as of yet not fully substantiated. First, the exact definition of the Universal Grammar is still lacking (see Pinker & Jackendoff, 2005). Second, some researchers argue that the linking problem cannot be solved because there is no a priori way by which children could know which semantic features can be linked to which formal categories (Atkinson, 1996). Third, it is still unclear how specifically linguistic representations could have evolved and where they should be represented in the brain (see Elman et al., 1996). Even if we find modularity of syntax in the adult brain, it is possible that this division of labour is the result of specialisation during development, but not evidence for their innateness (Karmiloff-Smith, 1992). For these reasons, emergentists claim that it is unwarranted to stipulate innate linguistic representations.

The emergentist approach attempts to overcome the evolutionary problem by relying on inductive learning processes. Hence it has to provide an account of how linguistic categories can be induced based on the distributional and frequency information in the input. Consequently,

this approach puts more emphasis on the structure of the input language because it is the only evidence for the linguistic structure of the target language. Emergentism thus has a strong computational motivation. Recent advances in corpus linguistics and computational modelling have led to more concise studies of the empirical correlates to linguistic structures (Baayen, 2003; Bod, Hay & Jannedy, 2003; Manning, 2003; Manning & Schütze, 1999). Computational models have been designed to prove that linguistic categories and syntactic knowledge can be derived from distributional information on the form and function of words. Furthermore, there is increasing evidence from corpus linguistics that many properties of grammar are tied to frequency distribution rather than lexeme general rules. In this view, linguistic structures and rules are seen as gradient phenomena (Boersma & Hayes, 2001). Ultimately, these data-driven approaches to language structure might help to overcome the divide between formal and functional linguistic theories (e.g., Bresnan & Aissen, 2002; Bresnan & Nikitina, 2003).

The probabilistic approaches to language have also gained in importance in language acquisition theory, where they are combined with constructivist approaches to development. Such data-driven approaches to acquisition have become known under headings such as emergentism (Elman et al., 1996; MacWhinney, 1999) or usage-based (Tomasello, 2003). Methodologically they apply methods from connectionist modelling and probabilistic parsing. It has been shown, for example, that the parts-of-speech categories for lexical items can be induced based on simple co-occurrence statistics (Redington, Chater, & Finch, 1998).

To date, large domains of acquisition as well as input data have remained unstudied with respect to frequency issues. On a practical level, substantial corpora and computer-technology to analyse them have become available only in the past decade. On a theoretical level, quantitative aspects of the input language were disregarded because much previous research focused on the question whether early child grammar is productive or not (for a summary see Tomasello, 2000). This line of research tried to identify the factors that are necessary for the child's achievement and to establish criteria for the productive acquisition of linguistic structures. The criteria for acquisition either measure onset of productivity (e.g., the use of a structure with x-number of lexical items or first morphological contrasts) or full mastery (e.g., the use of a grammatical feature in at least 90% of all obligatory contexts, see Brown, 1973). The 90% criterion is more stringent, but has the disadvantage of being applicable only when we deal with grammatically obligatory features like 3rd person agreement or plural marking in English. But whenever the phenomena under investigation are subject to (pragmatic or stylistic) variation, the 90% criterion cannot be applied. This means that acquisition

research frequently measures the success of acquisition by the onset of productivity, while less attention has been paid to the later stages of language development. This focus on the onset of productivity might have led to the “dogma” that first language acquisition is very fast because most common structures are attested by age 2–3, and more complex structures by age 6. But this account of language acquisition leaves open two critical questions: First, how fast is fast? *Slow* or *fast* are relative terms that need to be related to the scales of relevant exposures or practice trials. Second, how does the first productive use relate to adult-like use or mastery of the language? It seems that the latter question can only be answered with a fuller understanding of what the end-state of acquisition is like. While it is possible to classify individual utterances or domains as “error-free”, it is not possible to classify children’s language production on the whole as adult-like when we do not know much about *spoken* adult language.

### AIMS OF THIS STUDY

This study investigates production rates of child and input data from a detailed case study and focuses on the end-state of language development. By comparing the input–output relationship between child and adult language production, the course of language acquisition is measured against the explicit comparison to the target-state, i.e., the adult-like distribution and use of linguistic structures. Three issues are addressed, which have not figured prominently in previous research. First, what is the concrete evidence available in the input language in quantitative terms, and how does it change across development? Second, how does children’s language relate to the distributional properties of the input language? Third, can detailed case studies provide evidence regarding the question whether language learning is a slow or fast process?

The variable studied here is the frequency of selected structures. The aim is to investigate if and when children make use of the distributional information available in the input, or whether and in what domain their own language production deviates from what they hear. If language structure is a function of language use, as assumed in the emergentist perspective, children should adapt to the language they hear, i.e., they should be sensitive to the lexical items and grammatical structure and their distribution. It seems likely that a stable distribution of linguistic structures in the input might help the child to discover these structures. If the use of linguistic structures were very variable, it would provide a less reliable cue (Bates & MacWhinney, 1987). From the hypothesis that children are sensitive to distribution in the input language it follows that the distribution in the input language will act as a corrective in those domains where children’s use of language still deviates from that of their ambient

language. For example, case and tense marking are obligatory and hence frequent features of the grammar of German. Yet children start out with uninflected nouns and verbs, and acquire inflection only later. During the period when they produce uninflected forms, each utterance of the adults will provide evidence that the adult system is organised differently.

The analyses presented will provide evidence about production rates in three linguistic domains: parts-of-speech distribution, constituency of the noun phrase, and constituency of the verb complex. From a linguistic perspective, these are rather coarse domains. However, since there is hardly any information available on language production rates, this study should be viewed as a first attempt to fill this gap and provide data for the basic linguistic organisation of the input and of child language across development. Follow-up studies will then provide data with more linguistic differentiation. Before I explain the corpus that underlies this study, I will briefly summarise quantitative findings from the only two precursor studies that I am aware of.

### Precursor studies on production rates in language acquisition

Wagner (1985) used a cross-sectional design to get an estimate of how much children talk in a day. He recorded 12 monolingual German children between the ages of 1;5 and 14;10 (age in years;month.days) and recorded their speech for large portions of one day. In two cases, Wagner and his collaborators recorded a 9;7 year-old girl and a 9;6 year-old boy for two full days. Wagner (1985: 477) computed the estimated range of language production based on the assumption that children produce language for 12 hours a day and came up with a range of 11,700 words for the younger children to 37,000 words (a talkative 4-year-old). While such estimates are necessarily vague since the amount of talking may vary a lot from situation to situation and from day to day, the computation of the speech rates shows a clearer developmental trend: by age 3;5 the children in the study achieved a speech rate of a hundred words per minute. The children older than 10 years had speech rates of 149–187 words per minute and thus produce about 20,000 to 30,000 words a day.

Van den Weijer (1999) studied the ambient language of a Dutch infant when she was between 6 and 9 months old. The child's crib was equipped with a noise-activated microphone. This way, 90% of all speech spoken in her surroundings was recorded. Van den Weijer (1999) analysed all speech recorded on 18 of the 90 days and found that the child was exposed to an average of almost 3 hours of speaking time during the 8 hours she was awake each day (cf. van den Weijer, 1999: 30, 36). All input utterances were assigned to one of the following interactional situations: adult-to-

adult speech, adult-to-child speech (i.e., speech addressed to the older sibling of the infant), and adult-to-infant speech. There were a total of 79,914 utterances in these three conditions. The speech rate in utterances was 590 per hour, the speech rate for words was 1,890 (van den Weijer, 1999: 44–5). Lexical diversity in the input turned out to be a function of the interactional situation: there were twice as many word types in the adult-to-adult and adult-to-child conditions than in the adult-to-infant discourse (5,523, 4,808, and 2,081 words respectively; van den Weijer, 1999: 58).

While these studies investigated only the children's or the adults' language, the current case study will compare child language with the input directly addressed to the child. The data of the precursor studies will be used to test the representativeness of the quantitative data presented here.

## Data

### *Participants*

The participant of this study is a monolingual German boy, Leo, growing up in Leipzig, Germany. Both parents have a higher education (his father is an academic, his mother a trained bookseller) and speak dialect-free, clearly articulated standard High German. The boy's language development was recorded continuously during a 3 year-period from 1;11.13, the onset of multiword speech, up to 4;11.

### *Sampling and transcription*

The sampling intervals changed during the observation period: Two weeks before his second birthday, Leo's parents completed a vocabulary checklist modelled after the CDI for English (Fenson, Dale, Reznick, Thal, Bates, Hartung, Pethick & Reilly, 1993) because a German CDI did not exist at the time. For the following two weeks, the parents practised taking the diary notes on the newest and most complex utterances. Diary notes were spoken into a dictaphone at the time and place of the action to avoid possible misrepresentation by memory. Between 1;11.15 and 2;0, six sample recordings of varying length were made to try out the equipment and to familiarise the participants with the procedure. Between Leo's second and third birthday, five one-hour recordings per week were made and once a week the audio recording was supplemented with a video recording. In addition, the parents kept a daily diary. For another 2 years, until Leo was 4;11, five one-hour recordings were made every fourth week.

During the first year of the study, the mother was the primary caregiver of the child and was paid as a full-time research assistant for taking diary notes and making the audio recordings. Once a week, when the mother

worked in another job, the father took over child-care and the recordings. Also, once a week or more often if requested, one particular research assistant took care of the child, the diary notes, and the recordings. This division of labour guaranteed that the course of language acquisition could be followed without interruption. For the first year, when the child did not yet attend kindergarten, these three adults were his main caregivers such that the dataset offers a representative sample of what the child heard in this period. “Indirect” input as through TV did not play a role in Leo’s home environment, but the parents read to him quite a lot. From early on, Leo was passionate about books and stories. Later on, audio-recordings of children’s books like “Winnie-the-Pooh” were a frequent source of entertainment for him. As he grew older, his social circles expanded so that the recordings do no longer represent the full array of input available to Leo. During the final 2 years, Leo started to go to kindergarten for several afternoons a week and the family had a second child. Still, the five recordings every month allow reliable analyses of his progress in language learning.

The sessions were recorded with a Sony Mini-Disc recorder MZ-R35 using two wireless and portable Shure BG4.1 Unidirectional Condenser Microphones, and a Shure ETPD-NB Marcad Diversity Receiver. This meant that the parents had to turn on the stationary recording station and could take the portable microphones wherever they went with the child.

Each recording was digitised and transcribed in SONIC-Chat (cf. MacWhinney, 2000) with transcription guidelines developed for German by the author. The Sonic-Chat transcription allows easy checking of the transcription because the actual sound-segment of the speech file is directly accessible.

Table 1 gives an overview of the total sample size in terms of the number of words and utterances obtained in the 383 hours of recordings. Leo’s corpus comprises roughly half a million word forms. The corpus of the adults is almost three times as big. The same 1:3 child to adult ratio can be observed in other German acquisition corpora in the CHILDES-database,

TABLE 1  
Corpus size

	<i>Leo</i>	<i>Leo’s input</i>	
		<i>All</i>	<i>Coded sample</i>
Age range	1;11–4;11	—	—
No. recordings	383	383	151
No. utterances	158,336	258,592	98,989
No. words	495,681	1,363,955	527,930

e.g. the Miller-Corpus. Due to the increased sampling density, this corpus is about 6–10 times as large as the other corpora available. All of the child data were coded (see below), and so was a size-matched sample of the adults.

*Leo's language development in the investigation period*

The parental CDI and diary notes allow us to determine exactly the state of Leo's language development at the onset of our study. Leo was a late talker who, according to the parental report, produced his first adult-like words at about age 1;10. After that, he acquired more vocabulary quickly, and produced his first word combination at 1;11.13, when he had an active vocabulary of about 340 word forms. He also produced his first morphological contrast, a singular-plural distinction, in the same week. This confirms that the emergence of grammar requires a critical mass of active vocabulary. Bates & Goodman (1999) summarise cross-sectional data from English and Italian learners and show that the size of active vocabulary correlates with an index of grammatical complexity. Compared with the data of English and Italian children, who start to show first grammatical complexity with a vocabulary size of about 200–250 words, Leo's vocabulary size is quite large before grammatical development sets in.

As mentioned before, and as shown by his MLU (see below), the recordings started when Leo was just about to enter the two-word stage. Since MLUs are not very representative of language skills in later stages, four examples of his longest and most complex sentences will be given as an illustration of the state of his development at the end of the observation period. Example (1) shows a simple main clause with Verb-Second, in which the modal auxiliary and the infinitive are separated by eight words in the middle field. Example (2) provides a centre-embedded object relative clause, and Examples (3) and (4) exemplify that most very long sentences have a paratactic structure.

- (1) Leo 4;3.10  
 Warum sollen kleine Kinder nicht so oft auf  
 Babyflaschen rumnuckeln?  
 Why should little children not so frequently on  
 baby-bottles suck?  
 "Why mustn't little babies suck on baby-bottles?"
- (2) Leo 4;1.14  
 das Bild, das ich drucke, mach ich fuer dich, Eule .  
 the painting which I print make I for you, Owl.  
 "the painting which I print is for you, owl [stuffed animal]."

- (3) Leo 4;5.4  
 aber an besonderen Tagen, da haben die keinen  
 Sonderzug, sondern die Furchbergbahn ja ja faehrt  
 dann auch, aber sie haelt dann unterirdisch am  
 Schmiersee .  
 but on special days then have they no  
 special-train but the Furchbergtrain yes yes rides  
 then as-well but she stops then underground at-the  
 Schmiersee .  
 “but on special days they don’t have a special train, but let the  
 Fuchberg-train ride, which takes an underground stop at lake  
 Schmiersee.”
- (4) Leo 4;11.03  
 an [pause] achtzehnten zweiten gehen wir allerdings in  
 (da)s Theater und gucken uns “Peter und der Wolf” an.  
 on-the 18th February go we actually in  
 the theatre and see us “Peter and the wolf” on.  
 “actually, on the 18th of February we will go to the theatre and see  
 ‘Peter and the wolf’”

### *Morphosyntactic coding*

All child data and a size-matched sample of the input language received an extra layer of morphosyntactic coding by using the semiautomatic CLAN-programme *mor*. To this end, a lexicon file for all word forms was created which listed the citation form of the word form, as well as context independent information such as part-of-speech, inflectional class, or gender. These lexicon files were inserted into the transcripts by the mor-programme to create a %mor-tier. The mor-tier received further manual specification for context-dependent information like case and agreement, as well as part-of-speech resolution for ambiguous items. This disambiguation and coding was carried out manually by the author and two trained research assistants. A number of scripts were created to check for typing-errors and for the consistency of coding.

The coding categories not only specify the word form information for each word, but also add syntactic information which allows us to derive information about the constituency of phrases. For example, for all possible modal and auxiliary verbs, it was coded whether they indeed function as an auxiliary alongside a main verb participle or infinitive or whether they occur as the finite main verb (cf.: *I have done this* vs. *I have 50 Euros*). Likewise, all infinitives and participle forms were coded as to whether they occurred in isolation or as part of a complex tense (modal infinitive, future tense, present or past perfect, passive). Determiners and

adjectives were coded as to whether they functioned as pronouns or pronominalised adjectives, or whether they formed part of an NP. More detail about the relevant coding categories is given in the sections below.

## Analyses

### *General production measures*

The first analyses provide general measures of utterance length and production rates for words and utterances.

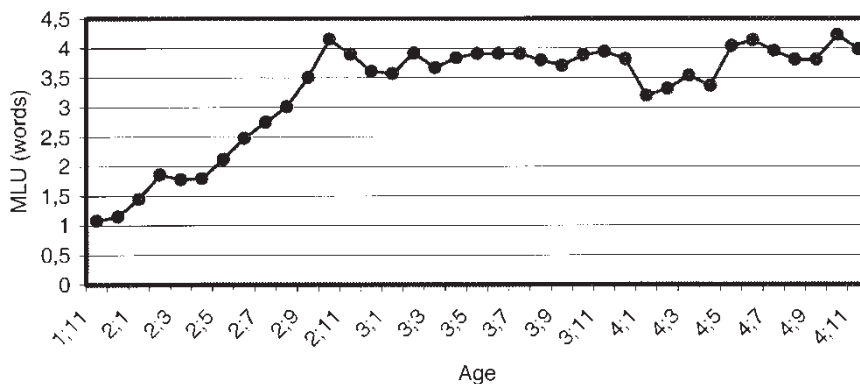
### *Mean length of utterance*

The mean length of utterance (MLU) is a quantitative measure of language development (Brown, 1973). It can be computed in words or morphemes. These days, the MLU is commonly computed in words because the counting of morphemes is problematic. Morphology is often formulaic and unproductive in the early stages of language development, and in highly inflecting language it is not always clear how many morphemes to count. Table 2 depicts the total number of utterances for the child and his three main caregivers as well as their MLUs in words for the whole investigation period; 98% of all input utterances come from the three main caregivers, his parents and the research assistant. The MLUs of the adults are 5.0 and 5.2. These values are compatible with the input data of van den Weijer (1999: 54), who measured the MLU of adult-to-adult-speech, adult-to-child speech, and adult-to-infant speech. The respective MLUs were 6.6, 4.0, and 4.0. The MLU of Leo's input thus compares with the mean of adult-to-adult and adult-to-child speech, the two types of interaction that are represented in the recordings.

Leo's MLU over the 3-year period is 2;7 with a range from 1.1 at age 1;11 to 4;2 at age 4;11 (see Figure 1). The curve shows growth up to age 2;9, after which we find stabilisation with a MLU-range between 3 and 4.2. This means that in the last 2 years of the investigation period, Leo's MLU is one to two words short of that of the adult mean. There is a plateau of the MLU between 2;2 and 2;5, when his language system reorganises. In this period, the rate of infinitives without agreement features goes down while

TABLE 2  
MLU in words (standard deviation) for child and adults

	<i>Leo</i>	<i>Mother</i>	<i>Father</i>	<i>Research assistant</i>
No. utterances	158,336	163,860	56,022	33,822
MLU words (SD)	2.7 (SD 2.6)	5.2 (SD 4.1)	5.2 (SD 3.5)	5.0 (SD 3.5)



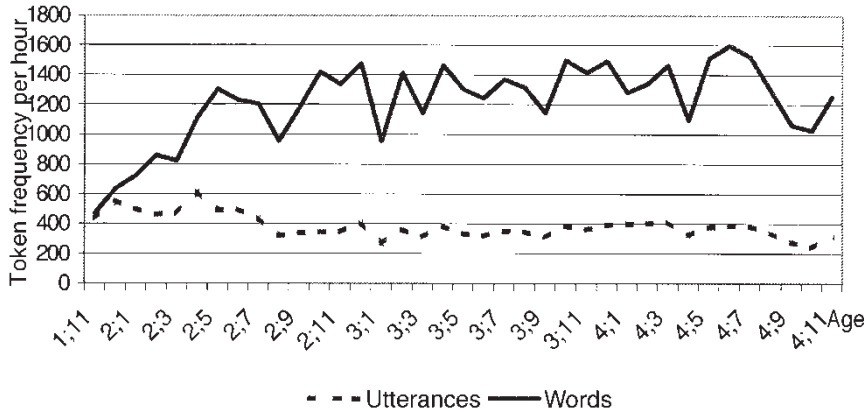
**Figure 1.** Leo's MLU (words).

more finite verbs are used, and while auxiliaries and determiners emerge (see the discussion of some of these phenomena below). In this phase, his utterances become grammatically more complex while the mean utterance length stays the same. Between 4;1 and 4;5 there is a dip in his MLU, the cause of which has not yet been identified. In the future, more detailed investigations will have to show whether structural changes occur in this period that might indicate representational redescrptions (Karmiloff-Smith, 1992).

The range of about 1.3 words found in Leo's MLU after 2;9 does not seem to be extraordinary. The range found in the adults when looking at monthly samples is also larger than one word. The MLU of the mother ranges between 4;2 and 6;2, the MLU of the father between 4;7 and 5.9, and the MLU of the research assistant between 4;1 and 5;5. While all of the lowest values come from the first month, the highest values do not necessarily come at the end.

### *Speech production rates*

This section provides data about the development of Leo's speech rate in terms of utterances and words per hour. The average speech production rate that can be inferred from Table 1 above is 413 utterances and 1,294 words per hour for Leo, and 745 utterances and 3,561 words per hour for the adults. The mean speech rates for the adults are thus considerably higher than in the van den Weijer (1999) study. The reason for this may be the different sampling technique: van den Weijer sampled full days—presumably with a mixture of more and less vivid interaction. In contrast, we sampled one hour a day in which the objective was to be communicative in order to elicit as much speech from the child as possible.



**Figure 2.** Leo's production rates of utterances and words (tokens per hour).

Figure 2 provides the data for Leo across development. In the first 6 months of development up to 2;5 we see a steady increase in the number of words produced, while the number of utterances shows a slight dip between age 2;2 and 2;4. After 2;5 the mean number of utterances decreases from 600 to a little less than 400. Corresponding to the range in his MLU values, the number of words per hour fluctuates between 1,000 and 1,600. If we multiply this by 10 hours of talk a day, Leo's speech rates compares with those found in the Wagner (1985) data for German children.

In terms of pure speech rate, this means that Leo has reached his mean production level at age 2;5, only half a year after he started to produce combinatorial speech. But of course, his language was not fully adult-like at this point. In terms of quantity, he lags behind all through the investigation period (shorter MLU, only one third of the adult speech rate). In terms of quality, it can be expected—based on earlier accounts on the acquisition of German—that different domains will be mastered at different points in time. In the remainder of this paper, I will discuss the development of different linguistic domains to test whether his performance stays behind that of the adults, or if not, when he catches up with their level of performance.

### *Production rates by linguistic domain*

In this section, distributional information about three structural domains is provided: (1) the parts-of-speech information, (2) the encoding of noun phrases (NPs), and (3) the use of different types of verbs. These domains provide an insight into general issues of constituency and syntactic organisation of child and adult speech. The aim is to provide data about

the input available to the child and about the child's approximation of the adult language. Descriptive rather than correlational statistics will be used because the categories distinguished in the figures and tables summarise a fair amount of detail. For example, the code "pronoun" subsumes personal pronouns like *he*, *she*, and *it*, as well as case-marked definite and indefinite determiner elements that are pronominalised and serve as a demonstrative pronoun (see Example 5c below). The acquisition of the individual features (definiteness, case, etc.) follows different trajectories (see Eisenbeiß, 2000). For the purposes of this analysis, they have in common that the NP-position is filled by a pronominal element rather than a simple noun or a full NP. It might be misleading to establish correlations between the general adult and child proportion of pronoun use because it could well be that at time X the child has mastered some features for a long time, while other aspects are still absent in his speech.

### *Parts-of-speech*

In the first analyses, the language production of Leo and the adults is analysed in order to provide a general picture of the morpho-syntactic properties of their speech.

### *Coding categories*

All words were coded in their syntactic context. This way, all word forms that can fulfil several syntactic functions were differentiated. In total, 11 parts-of-speech codes are distinguished for the purposes of this analysis:

- Nouns: all lexical and proper nouns
- Determiner: determiner elements which head a noun phrase
- Pronouns: pronominal elements which are the head of a noun phrase
- Adjectives (+ case): pre-nominal adjectives with case marker
- Adjectives (– case): uninflected predicative adjectives with copula
- Main Verbs: all lexical verbs and those modal verbs used as the main verb (e.g., *Ich kann das* "I can that", to mean, "I can do it")
- Copulas: all forms of the copula *sein* "to be"
- Auxiliaries: all auxiliary uses of the verbs *sein* "to be", *haben* "to have" and *werden* "to become", and all modals when used as auxiliary
- Prepositions: prepositional elements that head a Prepositional Phrase
- Adverbs: all adverbs, verbal particles, and discourse particles
- Conjunctions: all co-ordinating and subordinating elements

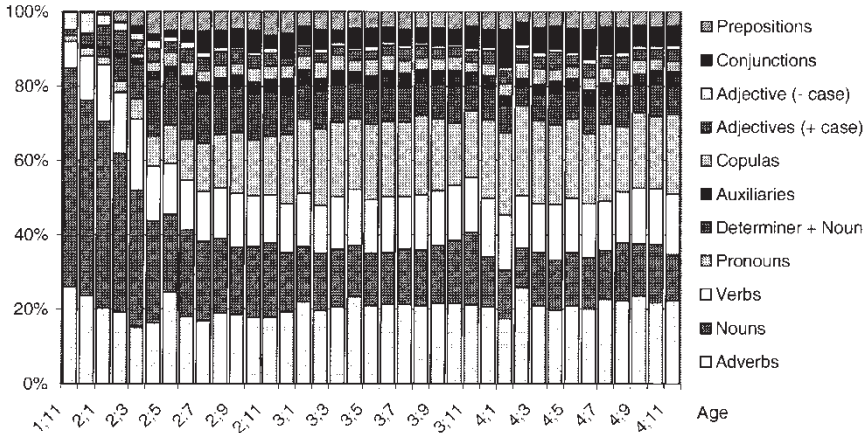


Figure 3. Leo's language production by part-of-speech.

Results

Figures 3 and 4 give the proportion of each of these categories in Leo's speech and in the adult speech per month across the three-year investigation period. The total number of words that is underlying these graphs is almost a million (403,034 for Leo and 458,564 for the adults). For two reasons, the number of coded words is lower than the total number of words in the coded corpus: First, the CHILDES-programme *mor*, which inserts the coding line, does not code repeated or retraced material (i.e., when the speaker self-corrects). Furthermore, interjections, tags, hesitation markers (*hm*, *uh*, etc.), onomatopoeia, nonce words, and non-interpretable words were excluded from these analyses.

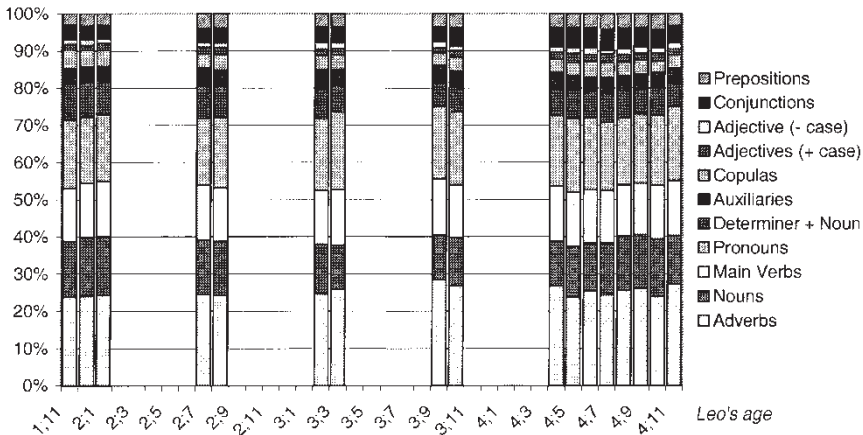


Figure 4. Adults' language production by part-of-speech.

When we first look at the adult data (Figure 4), it is striking that the relative proportion of the different parts-of-speech is extremely homogenous over time. The range of differences is less than 5% even for the large categories like nouns, adverbs, case-marked adjectives, and verbs. Leo's data show a development from a system that consists mainly of nouns and adverbs to one that approximates the adult distribution by age 2;7. After this, the distribution is also quite stable, but a little more variable than that of the adults (see for example Leo's data at 4;2).

### *Discussion*

Leo's early data confirm what we know about the development of syntactic categories in German: children start out with nouns and other uninflected elements. Case- and tense marked elements as well as elements which co-ordinate and subordinate propositions are acquired later.<sup>1</sup> What is striking in the longitudinal data presented here is the high degree of stability in the adult data and in the child's later data. While rules of grammar predict how a noun or verb phrase should be formed, they do not predict how often they should be used. The stability of structures found in the adult data shows that the input available to the child provides a reliable source of information about the basic syntactic organisation of German. The next two analyses will look at noun and verb phrases in more detail to check whether the same stability and input-output matching can be found.

### The encoding of noun phrases

Like English, German has three ways in which an NP can be encoded. There can be a full noun phrase with determiner (and adjectives, 5a), there can be a simple noun (5b), or a pronominal element (5c). Note that almost all pronouns and all determiners are inflected for gender, case, and number.

- (5) a. determiner phrase:            *das (kleine) Kind* "the (small) child"  
       b. simple noun                    *Kinder lesen Bücher* "children read books"

---

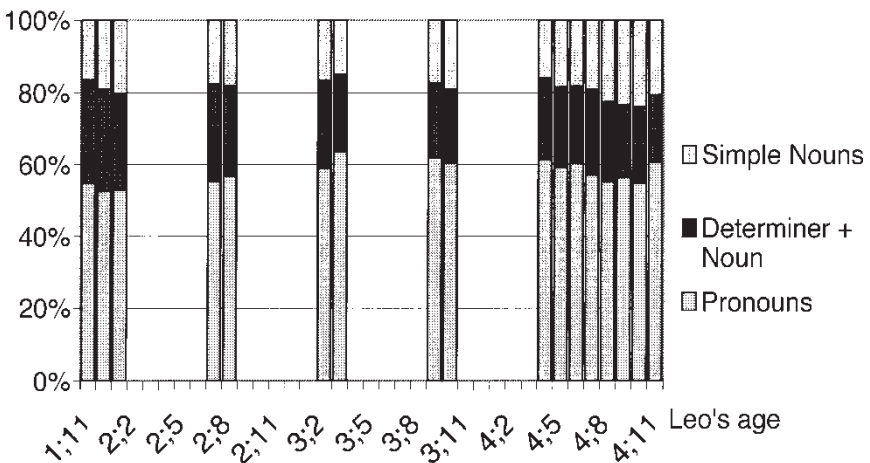
<sup>1</sup> See Mills (1985) and Clahsen (1982) for general accounts of German acquisition; Kauschke & Hofmeister (2002) for the distribution of parts-of-speech in early child German; and Behrens (2003) for more detail about the development of preposition and particles.

- c. pronominal
  - demonstrative pronominal *der kann gut malen* “the-masc can paint well”
  - *wh*-pronoun *wer kann gut malen* “who-masc can paint well?”
  - personal pronoun *er kann gut malen* “he can paint well”

Given that early child language mainly consists of content words (see above), the questions are when the child reaches the adult level of encoding the NP in these various ways, and of course what the adult level looks like. To this end, the proportion of these three types of NPs was calculated across time. Analyses are based on 148,616 NPs in the child corpus and 151,670 NPs in the coded adult corpus.

**Results**

Again, the distribution of these structures in the input remains stable over time (Figure 5). There are slightly more Determiner+Noun phrases than pronouns in the first year. The child’s data show stabilisation after age 3;0, after which we mainly find adult-like distribution (Figure 6). However, there is a steady increase of simple nouns between age 3;10 and 4;0, after which the distribution returns to the previous level. The first year of the child’s language development is characterised by a steep decrease in the proportion of simple noun phrases, while determiner and pronoun use emerges. In the second half of the first year, determiner phrases are more frequent than in the adult language, whereas pronouns are less frequent.



**Figure 5.** Adults' use of different types of noun phrases.

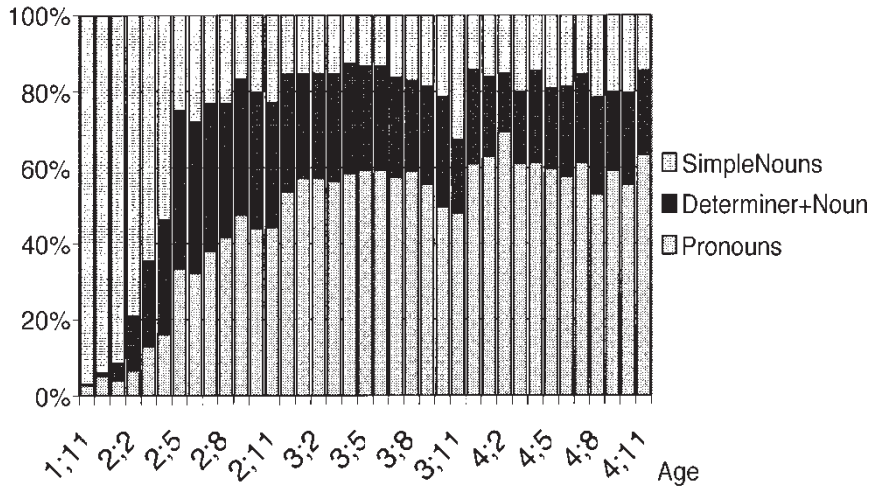


Figure 6. Leo's use of different types of noun phrases.

### Discussion

The slightly higher proportion of determiner phrases in the adult language addressed to the 2-year-old could be due to changes in the discourse contexts because early adult-child interaction is characterised by frequent pointing to and naming things. In contrast, discourse with an older child is likely to be more anaphoric such that a topic is set and talked about for a longer period. The data of the child show again that initially his language is reduced to bare referentiality, i.e., referring to things without grammatical encoding of case roles. The fact that Leo uses high proportions of complex determiner phrases and fewer pronouns in the second half of the first year indicates that the acquisition of anaphoric speech takes place later than the morphologically more complex determiner phrases.

### Verb categories

The production of all verb tokens (80,600 tokens in the child corpus and 105,550 tokens in the coded adult corpus) is analysed with respect to their syntactic function. On a lexical level, full lexical main verbs, modal verbs, and the copula are distinguished. In addition, the proportion of the auxiliaries *have*, *be*, and *become* is computed as well as the proportion of modal verbs when used as auxiliary. The proportion of auxiliaries and modal auxiliaries together represent the number of complex tenses in the data, i.e., verb phrases which consist of two components or more (e.g., *will have*, *has made*, *will have made*).

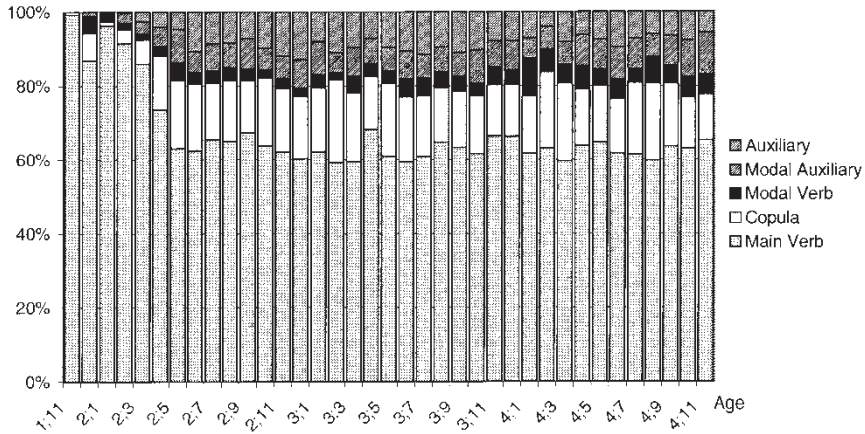


Figure 7. Leo's use of different types of verbs.

**Results**

Again, the adult data show a high degree of stability in the use of these verb types over time. About 60% of all verb tokens are main verbs and 20% (modal) are auxiliaries (see Figure 7). The child data (Figure 8) show a steady development of main verbs only at 1;11 to the approximation of the adult levels at age 2;5. In this early phase there is a noticeable peak of modal verbs and copulas at 2;0. Throughout development, Leo's use of modal auxiliaries is slightly lower than that of the adults.

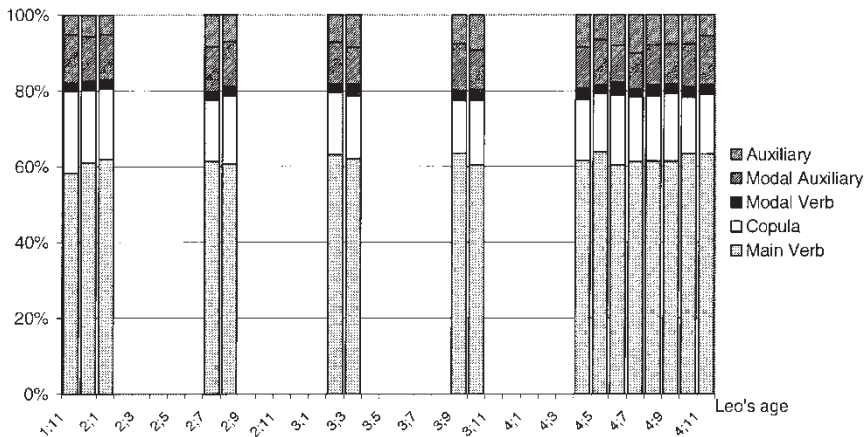


Figure 8. Adults' use of different types of verbs.

*Discussion*

Again, the first months of Leo's use of verbs are characterised by the use of simple lexical main verbs. The peak of modals and copulas at 2;0 can be attributed to a still very limited command of the verb category. The absolute number of verb tokens in this month is very low. There are only 285 verbs in more than 12,000 utterances. Moreover, many of the 36 modal and copula constructions can be linked to two speech acts: *there is/are X* and *want/would like X*. In addition, several uses of copulas and modals are (partial) imitations of the adults. Thus, the quantitative peak of modal and copula verb forms at 2;0 does not indicate a particularly fast development in this domain.

## GENERAL DISCUSSION

The analyses of three structural domains (parts-of-speech properties of lexical items, NP realisation, and use of different verb types) revealed a high degree of stability of their use in adult language over time, as well as a high degree of similarity between child and adult use in the later stages of the child's development. In sum, these data provide evidence that by age five, Leo has acquired the distributional properties in these structural domains. The onset at which these structural properties become productive was not determined in this study.

The high degree of similarity between child and adult language use could be due to sampling effects in two respects. First, the samples of the child and adult data come from the same discourse context, i.e., the child's performance is not measured against data sampled from sources like newspapers or books. Second, the child and adults talk about the same things in these data. This implies that there should be a high degree of lexical overlap (not tested here, but see Behrens, 2003). But the situational contingency guarantees enhanced comprehension because it is likely that the child's comprehension is best attuned to the specific phonotactic properties of the speech of the people he knows best, i.e., the parsing of caregiver-speech will be the easiest.

However, rules of grammar do not predict that grammatical structures should be used in the same proportion. Rather, full linguistic creativity could lead to the hypothesis that speakers are completely free in the choice of structures to encode their messages. This could lead to a high degree of variability in the use of structures over time, as well as between individuals. The data presented here suggest that this is not the case, but that language use is highly routinised and automatised (cf. Levelt, 1989). Syntactic priming effects are likely to occur in discourse interactions, and recently it

has been shown that children are susceptible to priming effects as well (Savage, Lieven, Theakston & Tomasello, 2003). The data also suggest that the children acquire much more than just the structural repertoire of German, but also the rhetorical style of their native language. This has been shown by crosslinguistic research (Slobin, 1991, 1997). It remains to be seen how much of language acquisition is determined by pure distributional and frequency information. The data presented indicate that the distributional properties of adult language indeed exercise a strong power in shaping the child's language use. Table 3 presents token frequencies per hour for parts-of-speech, based on the last 7 months of the investigation period.

The production rates of the adult show that the basic morphological properties like case or tense marking are attested several hundred times per hour. For example, there are about 500 verb tokens per hour, plus 1,000 case-marked determiners, pronouns, and adjectives, plus 450 nouns that have overt case markers in some contexts. It seems safe to conclude that both in the child's own production as in the language he hears, the psychological concept of entrenchment (cf. Tomasello, 2003), i.e., structural properties of language that are automatised through frequent repetition, has an effect.

The data presented here support earlier findings that pure distributional frequency is informative since the child obviously ends up with the same distributional properties as found in the speech addressed to him.

TABLE 3  
Production rates for parts-of-speech in tokens per hour

	<i>Leo</i> 4;4-4;11	<i>Adults</i> 4;4-4;11
Hours analysed	35	35
Nouns	200	448
Pronouns	335	622
Determiners	112	231
Adjectives (+case)	46	65
Adjectives (- case)	15	49
Main verbs	209	469
Copulas	58	120
Auxiliaries	49	136
Conjunctions	91	171
Prepositions	60	126
Adverbs	308	832

However, since he starts out differently from the adults, frequency cannot be the only factor to account for language development. From an emergentist perspective, one looks for perceptual, functional, and social factors to explain children's early language because children have access to more information than just type/token-information (social cues, semantic, and pragmatic cues). Also, their working memory is smaller than that of adults. These factors might act as filters on the uptake of information from the input (Ellis, 2002; Freudenthal, Pine, & Gobet, 2002; Wijnen, Kempen, & Gillis, 2001).

Future research will have to investigate how perceptual salience, semantic and pragmatic informativeness and memory development interact with frequency and distribution.

Finally, the analyses presented here provide indirect evidence regarding the speed of acquisition. Contrary to textbook knowledge, the speed of acquisition is slow, when measured against the rich input: Even if we estimate the amount of speech that Leo has heard conservatively, he has heard several hundred thousand tokens of nouns and verbs before he starts to talk. For example, the production rate for verbal elements in the adult data is about 500 per hour (Table 3). If we assume that children have acquired basic parsing skills at age one (Jusczyk, 1997) and take into account that Leo starts to use verbs just before his second birthday, he has heard more than 550,000 verb tokens even if he heard only 5 hours of speech a day. In future research, the amount of "rehearsal" it takes the child to reach the adult level of performance will have to be determined. Lieven, Behrens, Speares, and Tomasello (2003) looked at the dynamics of change and analysed an equally dense corpus of a child learning English regarding what was new in her speech on a given day. They demonstrated that early child language is highly conservative: about 63% of the utterances had been said before in exactly that form, another 27% differed from earlier productions with only small changes (e.g., substitution, addition, or deletion of a word or chunk). Only 10% of the utterances showed more substantial creativity.

In the future, advances in corpus design will allow us to address psycholinguistic issues with more methodological rigour. In terms of sampling, large annotated corpora will provide the distributional information necessary to compute the corpus size one needs in order to study the acquisition of rare structures (Tomasello & Stahl, 2004). Moreover, large corpora will allow us to investigate the causes of intra-individual variation and developmental spurts and plateaus (van Geert & van Dijk, 2002). If we know that the distribution of structures in the adult language is very stable, we can test reliably whether plateaux in development indicate phases of re-organisation.

## REFERENCES

- Atkinson, M. (1996). Now, hang on a minute: Some reflections on emerging orthodoxies. In H. Clahsen (Ed.), *Generative perspectives on language acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Baayen, H. (2003). Probabilistic approaches to morphology. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 229–287). Cambridge, MA: MIT Press.
- Bates, E. A., & Goodman, J. (1999). On the emergence of grammar from the lexicon. In B. MacWhinney (Ed.), *The emergence of language* (pp. 29–79). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Bates, E. A., & MacWhinney, B. (1987). Competition, variation and language learning. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 157–193). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Behrens, H. (2003). Verbal prefixation in German child and adult language. *Acta Linguistica Hungarica*, 50, 37–55.
- Bod, R., Hay, J., & Jannedy, S. (2003). Introduction. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 1–10). Cambridge, MA: MIT Press.
- Boersma, P., & Hayes, B. (2001). Empirical tests of the gradual learning algorithm. *Linguistic Inquiry*, 32, 45–86.
- Bowerman, M. (1990). Mapping thematic roles onto syntactic functions: Are children helped by linking rules? *Linguistics*, 28, 1253–1290.
- Bresnan, J., & Nikitina, T. (2003). *On the gradience of the dative alternation*. Submitted manuscript, Stanford University, CA, USA.
- Bresnan, J., & Aissen, J. (2002). Optimality and functionality: Objections and refutations. *Natural Language and Linguistic Theory*, 20, 81–95.
- Brown, R. (1973). *A first language: the early stages*. Cambridge, MA: Harvard University Press.
- Clahsen, H. (1982). *Spracherwerb in der Kindheit: Eine Untersuchung zur Entwicklung der Syntax bei Kleinkindern*. Tübingen: Narr.
- Crain, S., & Pietroski, P. (2002). Why language acquisition is a snap. *The Linguistic Review*, 19, 163–183.
- Eisenbeiß, S. (2000). The acquisition of the determiner phrase in German child language. In M.-A. Friedemann & L. Rizzi (eds.), *The acquisition of syntax: Studies in comparative developmental linguistics* (pp. 26–62). London / New York: Longman.
- Ellis, N. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition*, 24, 143–188.
- Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.
- Fenson, L., Dale, P., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., Pethick, S., & Reilly, J. S. (1993). *The MacArthur communicative development inventories: User's guide and technical manual*. San Diego, CA: Singular Publishing Group.
- Freudenthal, D., Pine, J., & Gobet, F. (2002). Modelling the development of Dutch optional infinitives in MOSAIC. *Proceedings of the 24th Meeting of the Cognitive Science Society*, 328–333.
- Jusczyk, P. (1997). *The discovery of spoken language*. Cambridge, MA: MIT Press.
- Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science*. Cambridge, MA: MIT Press.
- Kauschke, C., & Hofmeister, C. (2002). Early lexical development in German: A study on vocabulary growth and vocabulary composition during the second and third year of life. *Journal of Child Language*, 29, 735–757.

- Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.
- Lieven, E. V. M., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: a usage-based approach. *Journal of Child Language*, 30, 333–370.
- MacWhinney, B. (Ed.) (1999). *The emergence of language*. Mahwah, NJ: Lawrence Erlbaum Associates Inc.
- MacWhinney, B. (2000). *The CHILDES-Project: Tools for analyzing talk* (2 volumes). (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Inc.
- Manning, C. D. (2003). Probabilistic syntax. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 289–341). Cambridge, MA: MIT Press.
- Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.
- Mills, A. E. (1985). The acquisition of German. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition. Vol. 1: The data*. (pp. 141–254). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
- Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.
- Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, MA: MIT Press.
- Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's so special about it? *Cognition*, 95, 201–236.
- Redington, M., Chater, N., & Finch, S. (1998). Distributional information: a powerful cue for acquiring syntactic categories. *Cognitive Science*, 22, 425–469.
- Savage, C., Lieven, E. V. M., Theakston, A., & Tomasello, M. (2003). Testing the abstractness of children's linguistic representation: Lexical and structural priming of syntactic constructions in young children. *Developmental Science*, 6, 557–567.
- Slobin, D. I. (1991). Learning to think for speaking: Native language, cognition and rhetorical style. *Pragmatics*, 1, 7–25.
- Slobin, D. I. (1997). The origins of grammaticizable notions: Beyond the individual mind. In D. I. Slobin (Ed.), *The crosslinguistic study of language acquisition. Vol. 5: Expanding the contexts* (pp. 265–323). Mahwah, NJ: Lawrence Erlbaum Associates Inc.
- Tomasello, M. (2000). Do you children have adult syntactic competence? *Cognition*, 74, 209–253.
- Tomasello, M. (2003). *Constructing a language: A usage-based account of language acquisition*. Cambridge, MA: Harvard University Press.
- Tomasello, M., & Stahl, D. (2004). Sampling children's spontaneous speech: How much is enough? *Journal of Child Language*, 31, 101–121.
- van den Weijer, J. (1999). *Language input for word discovery*. Doctoral dissertation, Nijmegen: Katholieke Universiteit (MPI Series in Psycholinguistics 9, Max-Planck-Institute for Psycholinguistics, Nijmegen).
- van Geert, P., & van Dijk, M. (2002). Focus on variability: New tools to study intra-individual variability in developmental data. *Infant Behavior and Development*, 25, 340–374.
- Wagner, K. R. (1985). How much do children say in a day? *Journal of Child Language*, 12, 475–487.
- Wijnen, F., Kempen, M., & Gillis, S. (2001). Root infinitives in early Dutch child language: An effect of input? *Journal of Child Language*, 28, 629–660.

Copyright of Language & Cognitive Processes is the property of Psychology Press (UK). The copyright in an individual article may be maintained by the author in certain cases. Content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

Copyright of *Language & Cognitive Processes* is the property of Psychology Press (UK) and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.