Children's Grammars Grow More Abstract with Age – Evidence from an Automatic Procedure for Identifying the Productive Units of Language

Gideon Borensztajn (gideon@science.uva.nl)
Willem Zuidema (jzuidema@science.uva.nl)
Rens Bod (rens@science.uva.nl)
Institute for Logic, Language and Computation, University of Amsterdam
Plantage Muidergracht 24, 1018 TV, Amsterdam, The Netherlands

Abstract

We develop an approach to automatically identify the most probable multi-word constructions used in children's utterances, given syntactically annotated utterances from the Brown corpus of CHILDES. The found constructions cover many interesting linguistic phenomena from the language acquisition literature, and show a progression from very concrete towards abstract constructions. We show quantitatively that for all children of the Brown corpus grammatical abstraction, defined as the relative number of variable slots in the productive units of their grammar, increases globally with age.

Keywords: First language acquisition; Usage Based Grammar; Constructions; Data-oriented Parsing

Introduction

Many contemporary theories of language acquisition assume that the basic units of language acquisition are *constructions*: associations between a semantic frame and a syntactic pattern, for which the meaning or form is not strictly predictable from its component parts. Learning, in this framework, consists of the gradual acquisition of a structured inventory of constructions, a *constructicon*, where the constructions are of various sizes and varying degrees of complexity and abstractness (Goldberg, 2006; Tomasello, 2003).

Empirical studies in this tradition (e.g., Peters, 1983; Tomasello, 2003) show that, first, the primary units of speech of children in their first stage of language acquisition are not words but complete utterances, or *holophrases*. Second, in the earliest stages the child's language is item-based in nature (Tomasello, 2000). Verb constructions are typically learned case-by-case (so-called *verb islands*), without reference to a general verb-class. The scope of the syntactic rules is limited to specific constructions, and system-wide syntactic rules or categories are mostly lacking. Third, in subsequent stages the child breaks down the item-based constructions, introducing variables, such as in *Where's the X?*, *I wanna X*, etc.

The acquisition of constructions with variable slots forms the beginning of abstraction and category formation, and it marks the beginning of grammar. Such *Usage Based* theories of language acquisition assume a dynamically changing grammar that follows a route from simple and concrete to complex and abstract constructions. This view is in sharp contrast with the view on language acquisition taken in many versions of generative grammar. Here, grammar rules and categories are assumed to be universally and innately specified by a Universal Grammar. The reason that children do not produce adult-like grammatical sentences is their limited memory and attentional abilities (Chomsky, 1980); most of the knowledge of language is in place from birth, and only their parametrization needs to be set by the environmental triggers (Clahsen, 1996). Hence, in this tradition children are assumed to have, at least in competence, the same syntactic categories and rules as adults; this is referred to as the *continuity assumption* (e.g. Crain & Thornton, 2005).

These opposing views on the units of language acquisition are, of course, best investigated empirically, on the basis of actual language usage. Several in-depth case-by-case analyses on the productive units in children's corpora have been reported. For instance, Hodges, Krugler, and Law (2004) analyzed the item-based nature of the acquisition of the complex construction 'I V (NP) to VP-INF', as in *I want (you) to play*. Lieven, Behrens, Speares, and Tomasello (2003) traced back the sources of creativity of target utterances in the child's speech. The target utterances were reconstructed (manually) from a set of utterances used in the previous 6 weeks. They found that 74% of all novel target utterances produced in one day by a 2 year old child could be reduced to previously produced utterances by using a single combinatorial operation. Their finding supports the hypothesis that the smallest units used in language production are often memorized multi-word constructions, rather than single words.

To resolve the controversy, however, we believe it is essential to move beyond the typical handful of linguistic examples that support one view over the other. In the current study, we develop computational tools for automatically identifying the most likely primitive units that were used by the child to produce the utterances in a given corpus. We apply these tools to a well-known English-language corpus (the Brown corpus in CHILDES) with longitudinal data from three children. We then present a qualitative and quantitative analysis of the productive units that these children employ in progressive stages of acquisition. Note that we do not model the actual process of language acquisition or attempt to directly choose between usage-based and generative theories of language acquisition. Rather, we aim at providing a new way to evaluate predictions from theories of language performance in either tradition about the productive units in child language.

Choosing the right representation

In this section, we develop a formal definition of the productive units of language and a probability model that defines the likelihood of various hypotheses on the units used. The for-

mal model of choice needs to have the flexibility to allow for elementary syntactic units of variable size, form and level of abstraction. Moreover, the model should not assume a priori that the syntactic units the child uses coincide with the units used in adult language. The grammar framework should therefore be data-oriented: potential syntactic units should be derived from the corpus itself. For instance, the construction "I V to VP-INF" should be a possible building block, even though it contains multiple words, separated by variable slots (i.e., it is *discontiguous*), as well as "I going V NP-OBJ", even though it is agrammatical (it lacks the conjugated "am").

A formalism with the required flexibility is that of Tree Substitution Grammar (TSG), which forms a generalization over the well-known context-free grammars (CFG) and a subclass of the Tree Adjoining Grammars (Joshi, 2004). TSGs can model complex multi-word syntactic primitives as well as single unit primitives (see Figure 1 for an illustration). The generative components of a TSG are tree fragments of arbitrary size and depth, which can be (partly) lexicalized or abstract. In the latter case the fragments contain variable slots for syntactic categories (*nonterminals*), making them suitable for representing abstract constructions or abstract rules.

TSGs are used extensively in the framework of Data Oriented Parsing (DOP) (Bod, 2003; Scha, Bod, & Sima'an, 1999), which provides the techniques to parse new sentences using fragments from sentences observed in a corpus. In DOP, the elementary tree fragments of the TSG can in principle be any subtree occurring in an annotated corpus (the *treebank*). Two elementary tree fragments can be combined by means of the substitution operator $\circ$ if the left-most nonterminal leaf node of the first fragment is identical to the root node of the other fragment. A derivation of a sentence in DOP is a sequence of elementary tree fragments $t_1 \circ t_2 \circ \ldots \circ t_n$ such that the root of the first fragment is S and the leaves of the resulting tree are terminals (see Figure 1).

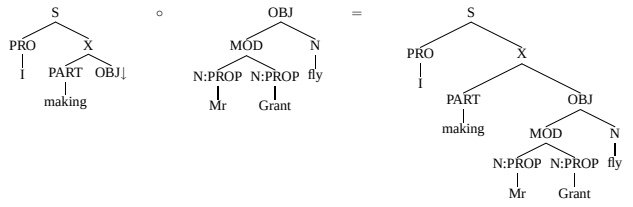

Figure 1: Derivation for 'I making Mr Grant fly' (Adam, 3;3.04). The substitution site is marked with $\downarrow$.

In the DOP-framework, several probability models have been worked out. In the simplest set-up, it is assumed that the probability of any substitution is independent of the context; the probability of a derivation is therefore the product of probabilities of the fragments used:

$$P(t_1 \circ t_2 \circ \ldots \circ t_n) = \prod_i P(t_i|r(t_i))$$

where $P(t_i|r(t_i)$ is the probability of a single substitution of a leaf node with a fragment t_i, of which the root $r(t_i)$ has the same label as the leaf node. Enriched with probabilities, TSGs become *probabilistic tree substitution grammars*.

Several alternative methods (*estimators*) for finding the probabilities $P(t_i|r(t_i))$ of the fragments (the parameters of the DOP grammar) have been proposed. The earliest estimator is known as DOP1 (Bod, 1998), and assigns probabilities based on relative frequencies. For the current work we adopted a recent estimator, "push-n-pull" (Zuidema, 2006, 2007), which yields linguistically more plausible results.

Formal details of push-n-pull fall outside the scope of this paper, but the basic idea is as follows. The algorithm uses the discrepancy between the observed frequency of the subtrees in the treebank and their expected frequency (as predicted by the current parameters of the grammar) to either *push* probability mass from a subtree to the elementary trees involved in its derivation, or *pull* probability mass from the elementary trees to the subtree. The algorithm includes a parameter that regulates the strength of a bias toward smaller subtrees. The difference between observed and expected frequency is highest for subtrees in the corpus that are most overrepresented relative to what should be expected based on the frequencies of their components; many of these subtrees correspond to linguistically interesting constructions. By iterating the process, probability mass is shifted between subtrees until expected frequencies approach observed frequencies.

Given a TSG as described above and a probability distribution $P(t_i|r(t_i))$ as found by push-n-pull, we can use standard statistical parsing techniques to find the most probable derivation of any sentence in a corpus. This yields a decomposition of the sentence into those elementary tree fragments that together constitute a hypothesis on how the sentence was generated. This way, we can use DOP as a statistical approach for discovering the constructions in child language.

All the analyses reported here were conducted with the push-n-pull algorithm, with the bias parameter set to 0.3. We have also performed tests with different settings of the bias and with the DOP1 estimator, and found the same, and sometimes even more pronounced trends than are reported here.

Method

The studies were conducted on the Brown corpus (Brown, 1973) from the CHILDES database (MacWhinney, 2000). This corpus contains transcribed longitudinal recordings of three children, Adam, Eve and Sarah. We split each of these subcorpora into three parts of roughly equal size, representing three consecutive time periods (see Table 1). We removed the parental speech and any annotation or comments. We also removed from the child's speech incomplete and interrupted sentences ('+...', '+/.' and '+,'), and sentences containing pauses ('#'). These account for approximately 20% of the sentences. Furthermore, we discarded the final punctuation.

In splitting the data, we did not attempt to match children on either age, Mean Length of Utterance (MLU) or traditional "stages" of language development (Brown, 1973); we

Table 1: Statistics of input (P1=Period 1; MLU= range of numbers of morphemes per utterance, averaged per file; a.s.l.= number of words per utterance, averaged per period; vocab.= number of distinct words; t/t = type/token frequency-ratio of words).

	files	age range	#sent.	MLU	a.s.l.	vocab.	t/t
Adam							
P1	1-16	2:3-2:11	11184	1.83-2.90	2.23	1407	.056
P2	17-32	2:11-3:6	11578	2.44-4.06	3.29	2010	.053
P3	33-48	3:6-4:5	9071	3.63-4.97	4.0	2006	.055
Eve							
P1	1-7	1:6-1:9	3485	1.53-2.28	1.88	669	.102
P2	8-14	1:9-2:0	3395	2.51-3.22	2.80	785	.083
P3	15-20	2:1-2:3	3535	2.60-3.41	3.13	958	.087
Sarah							
P1	1-45	2:3-3:2	11693	1.48-2.70	1.87	1389	.063
P2	46-90	3:2-4:1	8384	2.23-3.70	2.71	1706	.075
P3	91-135	4:1-5:0	8525	2.98-4.86	3.2	1944	.071

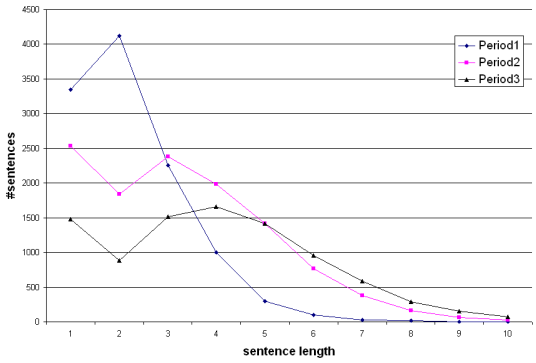

Figure 2: Sentence length distribution for Adam.

can thus only compare grammatical development within each child, and not between them. Table 1 summarizes the input used for our studies, after all preprocessing steps. Note that, unsurprisingly, average sentence lengths increase markedly in each child; in Figure 2 we plot the number of sentences of each length for each of the three parts of the Adam corpus.

Push-n-pull was trained on syntactically annotated sentences from each of the subcorpora. Recently the Brown corpus has been augmented with syntactic dependency annotations by Sagae, Davis, Lavie, MacWhinney, and Wintner (2007). The authors labeled the dependencies using 37 distinct grammatical relations (details of the procedure and a complete list of the labels can be found in (Sagae et al., 2007)). Their parser uses the parts of speech from the MOR-tagger, described in (MacWhinney, 2000). In Table 2 we list the most frequent grammatical relations and PoS tags.

We converted the dependency annotation and labels of Sagae et al. (2007) to a constituency annotation for further processing[1]. The conversion heuristic we used is similar to that of (Xia & Palmer, 2001). From Table 3 it can be seen, that at times the conversion introduced a dummy node (labeled **X**), to fill up a gap in the (binary) parse tree, where the dependency annotation did not provide this.[2]

[1]Approximately 10% of the sentences failed to convert, mostly due to a dependency having more than a single root, or the postag (MOR) and syntax (XSYN) sequence being of unequal length.

[2]Details at http://staff.science.uva.nl/∼gideon/cogsci/

Table 2: Frequent PoS tags and Grammatical Relations.

Parts of Speech	Category
N, N:PROP	Noun, Proper Noun
V, V:AUX	Verb, Auxiliary verb, including modals
DET, DET:NUM	Determiner (*the, a*), Number
ADJ, ADV	Adjective, Adverb
PRO, PRO:DEM, PRO:WH	Pronoun, Demonstrative Pronoun (*this, that*), Interrogative Pronoun (*who, what*)
CONJ	Conjunction
INF	Infinitive marker (*to*)
PREP	Preposition
Gramm. Relation	**Category**
ROOT	Special relation for the top node
SUBJ, OBJ	Subject, Object
PRED	Predicative (I am not *sure*)
COMP, XCOMP	Clausal complements, finite (I think *I saw Paul*) and non-finite (you have *to put it in your truck*)
JCT	Adjunct (optional modifier of verb)
COORD	Coordination, dependents of the conjunction (*go* and *get it*)
AUX, NEG	Auxiliary and negation
LOC	Locative arguments of verbs (*in your truck*)

Results

Qualitative analysis

In the current setup, the syntactic categories (nonterminals) are pregiven; our method only determines the size of the productive units involved in the generation of each sentence. We are interested in those cases where larger fragments seem necessary than implicit in the existing corpus annotations; for our analysis, we therefore focus on elementary trees of depth larger than 1 and will refer to these as the *constructions*.

Our method found linguistically very informative constructions in all children. In Table 3 we give Adam's 15 most frequently used constructions of each period, as well as the 15 most frequent discontiguous ones. In the figure, part of speech tags are indicated in capitals, and grammatical relations appear in bold capitals. Explanations of the labels are in Table 2. A few things may be noted from Table 3:

- Whereas in Period 1 most constructions are very concrete, starting from Period 2 constructions become abstract (as can be seen from the increased number of substitution sites). We further support this observation by quantitative results in the next section.

Table 3: Adam's most frequent multi-word constructions (shown are only the leaf nodes). To facilitate reading, we have restored some of the lexical items from the MOR tagger with their original form. For instance, we replaced *be-3s* by *is*, *go-prog* by *going*, *go-past* by *went*, and zero-forms, such as *put-zero* by *put*.

#	Period 1	#	Period 2	#	Period 3
88	right there	82	what is this	127	I do not know
48	where go	80	PRO:WH is PRO:DEM	69	what is this
45	why not	74	do you want PRO **COMP**	51	PRO:WH is that
42	where is	53	I do not know	46	I going **XCOMP**
36	play toy	52	do you want **X**	44	it is **PRED**
33	where N go	41	I going **XCOMP**	33	I want INF **X**
30	what happen	36	open it	27	it is **X**
28	read that	32	PRO:WH is it	27	I am going INF **X**
24	nineteen twelve	30	it is **PRED**	27	what is it
21	N go	28	you want INF **X**	23	I think I **X**
20	busy bulldozer	26	I going V **OBJ**	22	I can not
19	in there	25	what you want	22	that is **PRED**
19	PRO:DEM a N	24	let me **COMP**	21	here is **SUBJ**
19	that N	23	how do you know	20	is a N
18	that is right	22	I **AUX NEG X**	20	V it

Most frequent constructions

#	Period 1	#	Period 2	#	Period 3
33	where N go	9	you V it	10	you V it
11	I V it	8	I do **NEG** want INF **X**	5	you **X** and PRO **X**
6	what that N doing	7	I V it **JCT**	5	will you V it
5	take N off	6	you V it	4	can PRO put **X**
4	who N that	6	where PRO went	4	a ADJ one
4	do **NEG** V it	6	let me V it	4	do **NEG** know PRO:WH PRO V
4	have N on	5	I can **NEG** V it	4	do **NEG** know PRO:WH PRO:DEM V
4	you V it	5	going INF make DET N	4	and PRO:WH is that
3	what N doing	5	let us play DET game	4	can PRO put **OBJ LOC**
3	put N on	5	what kind N that	4	what is PRO:DEM for
3	put **OBJ** on	4	going put **OBJ** in it	4	I can not V it
3	do not V me	4	a N cake	3	how **AUX** you V PRO:DEM
3	take **OBJ** off	4	I V him	3	maybe PRO is **X**
3	where N N go	4	in DET kitchen	3	you V this ADV
3	I V some	4	you V me **COMP**	3	I going **X** off

Most frequent discontiguous constructions

- The lists cover many linguistically interesting constructions, such as the progressive, use of auxiliaries, clausal constructions with *want* and *think*, and particle verbs (*take **OBJ** off*, *going put **OBJ** in it*).
- Constructions including non-finite clausal complements (XCOMP) start to appear in Period 2, but become more frequent in Period 3 (sometimes annotated *INF **X***) .
- There is a tendency to progressively use verb constructions in combination with variable pronomina, as is particularly notable from the increased use of pronominal tags among discontiguous constructions.
- The use of do-support in questions and negations starts in the P2 and becomes more abstract in the top discontiguous constructions of P3 (*how **AUX** you V PRO:DEM*).

Another way of looking at the output of our method is by going through individual sentences from each of the corpora and checking how sentences get decomposed into their hypothesized building blocks. Figure 3 gives some typical examples of derivations, but note that there are also examples of linguistically less plausible decompositions. The decompositions of the entire Brown-corpus are available as supplementary material to this article (see footnote 2).

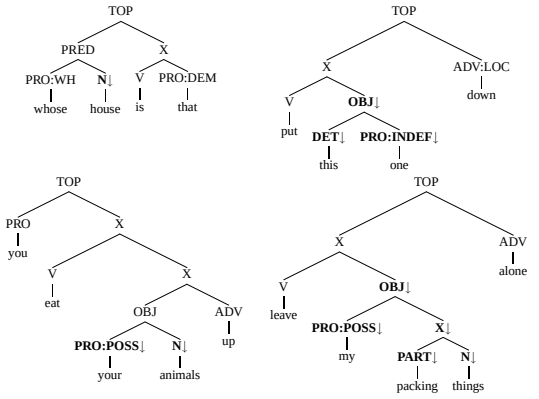

Figure 3: Examples of derivation trees as found by our method. Substitution sites are in bold and marked with ↓.

Quantitative analysis

Once we have determined the most probable derivations of a child's utterances as recorded in a corpus, it becomes possible to quantify the properties of the child's grammar at various stages in terms of properties of the used elementary trees, such as node count and depth. This, in turn, allows us to start answering questions like:

- What is the size of the primitive building blocks used?
- Does the size of the building blocks decrease with age, as would be expected if constructions are broken down into their parts? To operationalize size, we simply counted the number of nodes in the elementary trees.
- Do the productive units of the grammar become more abstract with age? Abstraction correlates with the number of variable slots in linguistic constructions, which can be operationalized as the ratio between the number of substitution sites (non-terminals) and the number of lexical items (terminals) in the elementary trees of each derivation.
- Are all elementary trees contiguous? The occurrence of discontiguous constructions, where a substitution site is preceded and followed by a lexical item would help explain long distance dependencies (such as agreement of number and tense) between the lexical items.

A first observation from Table 4 is that constructions become ubiquitous with age (see the column #constructions/sentence). For all children there is a sharp increase in the number of constructions between P1-P2, and for all except Eve also between P2-P3. The overall averages of most of the relevant quantities show an increase with age for all children (for instance for the number of nodes, nonterminals, terminals, depth, discontiguity: see Table 4). However, before drawing conclusions about changes of the nature of constructions with time, it is important to rule out the possibility that the effects are only due to sentence length distributions, which are shifting toward longer sentences for the later peri-

Table 4: Overall averages of the most important measures on constructions.

Quantity	Adam			Eve			Sarah		
	Period 1	Period 2	Period 3	Period 1	Period 2	Period 3	Period 1	Period 2	Period 3
average #nodes in cxs	7.36	8.80	9.20	7.14	8.32	8.76	7.64	8.65	8.86
#terminals	2.22	2.46	2.47	2.21	2.50	2.45	2.30	2.35	2.34
#non-terminals	5.14	6.35	6.73	4.93	5.83	6.32	5.34	6.3	6.52
#leaf non-terminals	0.41	0.93	1.16	0.32	0.59	0.93	0.45	0.98	1.15
ratio leaf-non-terminals/leaf-nodes	0.13	0.25	0.30	0.11	0.18	0.26	0.14	0.27	0.31
average depth	4.44	4.74	4.79	4.35	4.61	4.74	4.52	4.75	4.78
#constructions/sentence	0.39	0.59	0.67	0.25	0.46	0.41	0.24	0.39	0.50
#discont. cxs/sentence	0.050	0.085	0.093	0.020	0.051	0.106	0.041	0.064	0.071
construction coverage	0.33	0.40	0.37	0.26	0.35	0.28	0.29	0.32	0.35
#construction types	1409	2658	2543	324	665	685	967	1294	1732

ods. This is a major methodological challenge, because most of the quantities of interest, such as size and depth, depend on the length of the sentence in which the construction appears.

Therefore, in all the following studies, we neutralized the MLU factor by comparing sentences across periods according to their length. We computed the average quantity (e.g., depth, #nodes) for (constructions belonging to) different sentence lengths separately. Averages were computed over at least 30 constructions or discarded otherwise. We then computed, still for each sentence length separately, growth rates of those quantities. These were averaged, to obtain an average growth rate for the quantity between any two periods (note that average growth rate is different from the growth rate of the average, as computable from Table 4).

After sentence length has been factored out, there is hardly any effect left of age on construction size (the total number of nodes in a construction). As can be seen in Table 5, most growth rates are just above one, so the size of the construction within sentences of a certain length remains close to constant (the variance is written within the parentheses). The same is true for construction depth, which (unsurprisingly) correlates well with construction size.

Table 5: Growth rates of construction size and depth. Shown are averages of the growth rates per sentence length.

	P1→P2	P2→P3	P1→P3
Construction size			
Adam	1.019 (.042)	1.002 (.016)	1.024 (.054)
Eve	1.020 (.045)	1.002 (.042)	1.013 (.052)
Sarah	0.998 (.033)	1.006 (.029)	0.992 (.023)
Construction depth			
Adam	1.016 (.019)	1.001 (.009)	1.022 (.021)
Eve	1.047 (.029)	1.033 (.033)	1.059 (.037)
Sarah	0.988 (.015)	1.008 (.009)	0.996 (.009)

Whereas the number of nodes of the constructions remains constant with age, the number of nonterminals increases with age for all sentence lengths independently; the number of nonterminals in the leaves of constructions increases even more quickly. At the same time, the number of terminals in constructions decreases. In Table 6 we show the time development of the average 'abstraction' of the constructions, which is defined as the ratio between leaf non-terminals and leaf nodes in a construction. It can be seen, that abstraction of the constructions increases with age for all children. The big variance is due to sentence length 2 ("holophrases"), for which the constructions remain very concrete in all stages; if we leave these out, abstraction still increases significantly with age. Note that there is no simple explanation for increasing abstraction in terms of the type/token frequency of the vocabulary, since these quantities are neither positively nor negatively correlated (see Table 1).

Table 6: Growth rates of abstraction of constructions.

	P1→P2	P2→P3	P1→P3
Adam	1.15 (.20)	1.06 (.17)	1.32 (.39)
Eve	1.41 (.37)	1.30 (.14)	1.68 (.39)
Sarah	1.33 (.35)	1.01 (.22)	1.25 (.19)

In Figure 4 we plot the average abstraction per sentence length for Sarah; results for Adam and Eve are similar.

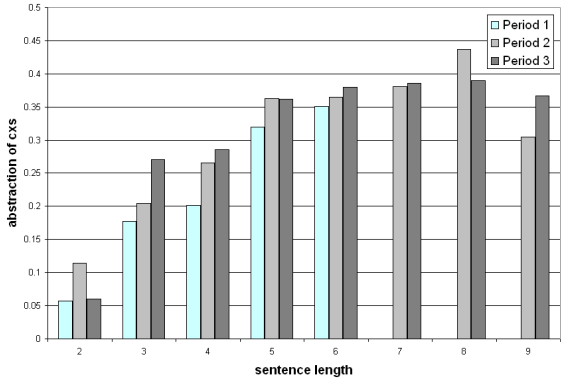

Figure 4: Abstraction (ratio leaf non-terminals/leaf nodes in constructions) against sentence length for Sarah.

This result is striking, because when we look at the parse trees in their entirety, the ratio between non-terminals and ter-

minals is equal to 2 and does not vary with age (because the input parse trees are binarily branching). This explains the fact that for the nodes in depth 1 subtrees we found a strong opposite effect of decreasing abstraction:

Table 7: Growth rates of abstraction of depth one subtrees.

Adam	0.86 (.071)	1.00 (.053)	0.85 (.082)
Eve	0.94 (.031)	0.98 (0.056)	0.92 (.049)
Sarah	0.90 (.051)	0.95 (.032)	0.86 (.071)

Conclusions and discussion

This study presented a novel and automatic procedure for the discovery of multi-word constructions. Our approach is a promising alternative to the evaluation of some of the core assumptions in theories of language acquisition. We believe there is much useful information available in the distributional patterns in corpora of child language that remains heretofore underexplored; this information can be accessed using sophisticated statistical methods, such as used here, that are flexible enough to accommodate for multi-word constructions. Note that the fact that our method works without information about the semantics and pragmatics should not be interpreted as implying a minor role for semantics in language acquisition. In fact, we believe semantics is central in acquisition as well as use, but we aimed at developing techniques that work with the information present in current corpora.

A fundamental problem for research on the continuity hypothesis, and language acquisition in general, is that no consensus exists about reliable methods to identify the productive units of language. Here, we explored an approach to identifying the basic building blocks of language based on distributional patterns alone, but alternative sources of information are also available, such as those explored in approaches based on processing data (e.g., reading times, errors) or relations to linguistic input (tracing back sentences to child-directed speech, (Lieven et al., 2003)). Our method, applied to the Brown corpus, confirms the *progressive abstraction* hypothesis: abstraction, defined as the relative number of non-terminal leaves in multi-word constructions, increases with age. We show that it does so independently of sentence length. Complex constructions lose their lexical parts to specialized lexical rewrite rules, and in the process the constructicon becomes more abstract.

This finding is in line with the theory of item-based learning and clearly point to an incremental learning path, but we believe it goes further than the state-of-the-art. By making available the found constructions of the entire Brown corpus, our version of the progressive abstraction hypothesis now becomes *falsifiable*: using other approaches to identify the elementary units of language, researchers can evaluate the quality of the hypothesized constructions. Although we cannot exclude the possibility that the performance-competence distinction might rescue the continuity hypothesis, the onus is now on its defenders to demonstrate that either our hypothesized constructions are incorrect, or that a generative performance theory makes identical predictions.

Acknowledgments We thank Alon Lavie and colleagues for supplying us with the syntactic annotation of the Brown corpus in an early stage, and three anonymous reviewers for valuable comments. This research was funded by the Netherlands Organization for Scientific Research (NWO), through a Vici-grant "Integrating Cognition" (277.70.006) to RB and a Veni-grant "Discovering Grammar" (639.021.612) to WZ.

References

Bod, R. (1998). *Beyond grammar: An experience-based theory of language*. Stanford, CA: CSLI Publications.

Bod, R. (2003). An efficient implementation of a new DOP model. In *Proceedings EACL'03*.

Brown, R. W. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences*, *3*, 1-61.

Clahsen, H. (1996). *Generative perspectives on language acquisition*. Amsterdam, The Netherlands: Benjamins.

Crain, S., & Thornton, R. (2005). Acquisition of syntax and semantics. In M. Traxler (Ed.), *Handbook of psycholinguistics.* Oxford: Elsevier.

Goldberg, A. E. (2006). *Constructions at work. the nature of generalization in language.* Oxford University Press.

Hodges, A., Krugler, V., & Law, D. (2004). A corpus study on the item-based nature of early grammar acquisition. *Colorado Research in Linguistics*, *17*.

Joshi, A. K. (2004). Starting with complex primitives pays off: complicate locally, simplify globally. *Cognitive Science*, *28*(5), 637-668.

Lieven, E., Behrens, H., Speares, J., & Tomasello, M. (2003). Early syntactic creativity: a usage-based approach. *Journal of Child Language*, *30*, 333-370.

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

Peters, A. (1983). *The units of language acquisition*. Cambridge, UK: Cambridge University Press.

Sagae, K., Davis, E., Lavie, A., MacWhinney, B., & Wintner, S. (2007). High-accuracy annotation and parsing of CHILDES transcripts. In *Proc. ACL-2007 workshop on cognitive aspects of computational language acquisition* (p. 25-32).

Scha, R., Bod, R., & Sima'an, K. (1999). A memory-based model of syntactic analysis: data-oriented parsing. *J. of exp. and theoretical artificial intelligence*, *11*, 409-440.

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Science*, *4*(4), 156-163.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Xia, F., & Palmer, M. (2001). Converting dependency structures to phrase structures. In *Proceedings of HLT 2001*.

Zuidema, W. (2006). What are the productive units of natural language grammar? A DOP approach to the automatic identification of constructions. In *Proc. CONLL-X, pp.29-36*

Zuidema, W. (2007). Parsimonious data-oriented parsing. In *Proc. EMNLP-CONLL 2007, pp. 551-560*