Understanding the Developmental Dynamics of Subject Omission: The Role of Processing Limitations in Learning*

Daniel Freudenthal

Julian M. Pine

University of Liverpool

Fernand Gobet

Brunel University

*This research was funded by the Economic & Social Research Council under grant number R000223954. A preliminary version of this research (which involved simulating the pattern of subject omission in Adam, Eve and Sarah's data) is reported in Freudenthal, Pine & Gobet (2002b). Address for correspondence: Daniel Freudenthal, School of Psychology, University of Liverpool, L69 7ZA, United Kingdom. Email: D.Freudenthal@Liverpool.ac.uk

Abstract

P. Bloom's (1990) data on subject omission are often taken as strong support for the view that child language can be explained in terms of full competence coupled with processing limitations in production. This paper examines whether processing limitations in learning may provide a more parsimonious explanation of the data without the need to assume full competence. We extended P. Bloom's study by using a larger sample (12 children) and measuring subject-omission phenomena in three developmental phases. The results revealed a Verb Phrase-length effect consistent with that reported by P. Bloom. However, contrary to the predictions of the processing limitations account, the proportion of overt subjects that were pronominal increased with developmental phase. The data were simulated with MOSAIC, a computational model that learns to produce progressively longer utterances as a function of training. MOSAIC was able to capture all of the effects reported by P. Bloom through a resource-limited distributional analysis of child-directed speech. Since MOSAIC does not have any built-in linguistic knowledge, these results show that the phenomena identified by P. Bloom do not constitute evidence for underlying competence on the part of the child. They also underline the need to develop more empirically grounded models of the way that processing limitations in learning might influence the language acquisition process.

Introduction

A central feature of children's early multi-word speech is that it includes utterances with missing constituents. Several researchers have argued that this phenomenon is best explained in terms of full competence coupled with processing limitations in production (Pinker, 1984; P. Bloom, 1990; Valian, 1991, Valian & Eisenberg, 1996; Valian, Hoeffner & Aubry, 1996). According to this view, children represent the correct syntactic structure of the sentences they are producing. However, their ability to express this structure in their speech is limited by some kind of processing bottleneck in production. The result is that, when the demands of producing a particular sentence exceed a certain level, some elements from the underlying structure are not expressed and errors of omission occur. As P. Bloom points out, this kind of analysis is '…one way to reconcile a nativist theory of language acquisition with the fact that most of young children's sentences are less than three words long…' (P. Bloom, 1990: 492).

The strongest support for a processing limitations explanation of the pattern of errors in children's speech comes from work on subject omission in English. It is a well-established fact that young language-learning children frequently omit subjects in contexts in which a subject would be obligatory in the adult language (e.g. *Want tea, Went home*). In an early analysis of this phenomenon, L. Bloom (1970) found that subject omission errors were more common in negated than non-negated sentences, and in sentences containing relatively new (unfamiliar) verbs (see also L. Bloom, Miller & Hood, 1975). She concluded that subject omission was a response to the increased processing load associated with the production of sentences with these particular properties. In a later study, P. Bloom (1990) tested this kind of explanation more directly by comparing the length of the Verb Phrase (VP) in utterances in which a subject was provided and utterances in which a subject was omitted. P. Bloom hypothesized that the load associated with the production of longer VPs would decrease the likelihood of subject provision, and hence that the VPs of sentences with subjects would be shorter than the VPs of sentences without subjects. His results confirmed this prediction. They also showed that sentences with pronominal subjects tended to have longer VPs than sentences with lexical subjects, and that children omitted subjects at higher rates than they omitted objects. These results were interpreted as consistent with a processing limitations account according to which pronominal subjects carry a lower processing load than lexical subjects (because they are phonetically shorter and tend to contain fewer lexical items), and subjects carry a higher processing load than objects (because they occur nearer to the beginning of the sentence where the processing demands associated with sentence production are particularly high).

At first sight, these findings appear to provide strong support for the view that subject omission errors are a consequence of processing limitations in production. In fact, however, this conclusion is problematic for a number of reasons. First, although it is clearly possible to explain these phenomena in terms of processing limitations in production, it is important to realise that P. Bloom's account of these phenomena is actually rather ad hoc. Thus, although the finding of a VP-length effect does seem to suggest that there is a relation between subject omission and the overall processing load of the sentence that the child is trying to produce, no attempt is made to specify how processing load interacts with the sentence production mechanism. Moreover, each of the additional phenomena identified is dealt with by making an additional assumption about differences in the processing load encountered at different points in the sentence, or the processing load exerted by different types of constituent. Thus, the asymmetry between subject and object omission is explained by assuming that the processing load is heavier at the beginning of the sentence, and the fact that VP length varies as a function of subject type is explained by assuming that lexical NPs exert a heavier processing load than pronouns. The implication is that the plausibility of P. Bloom's account relies very heavily on the plausibility of these additional assumptions.

A second problem is that there is increasing evidence that the second of these assumptions is actually incorrect. Thus, as Hyams & Wexler (1993) point out, if pronouns exert a lower processing load than lexical NPs, then a processing limitations account would seem to predict that children's preference for pronominal over lexical subjects should decrease as their processing resources increase. That is to say, the proportion of lexical as opposed to pronominal subjects should increase as a function of development. In a developmental analysis of the data from Adam and Eve, Hyams & Wexler show that, in fact, the opposite is true, with both children showing a decrease in the proportion of lexical subjects over the period in question. This result suggests that, if anything, children find pronominal subjects more difficult to produce than lexical subjects. Moreover, there is further support for this conclusion from the results of elicited imitation studies. Thus, both Gerken (1991) and Valian et al. (1996) show that young children are significantly more rather than less likely to omit pronominal than lexical subjects from their utterances in elicited imitation tasks[1]. The implication is that pronominal subjects exert a higher processing load than lexical subjects — and hence that the pattern of VP length effects in P. Bloom's data is actually more difficult to explain in terms of processing limitations in production than it might at first appear.

Third, P. Bloom's account appears to be built on the assumption that the VP-length effect can only be explained by processing limitations in production. However, it seems likely that this kind of length effect could also be explained in terms of processing limitations in learning. For example, if one assumed that children were building syntactic knowledge gradually, and that the mechanism responsible for building this knowledge was subject to processing limitations that restricted how much could be learned from each of the utterances to which it was exposed, then one would expect the average length of the structures that the child was capable of representing to increase gradually as a function of learning. This would not only result in children's sentences being shorter on average than those of their parents, but would also result in children operating with partial structures at intermediate stages in development. Assuming that the length of these structures was primarily determined by processing limitations in learning, one would expect the length of structures including subjects to be the same on average as the length of structures that did not include subjects, and hence that subjectless VPs (i.e. utterances with no subject) would be longer on average than the VPs of utterances with subjects (i.e. partial utterances from which the subject had been removed by the researcher). Note that this kind of explanation is not necessarily incompatible with the view that children represent the correct syntactic structure of the sentences they are producing from the beginning. However, if shown to be viable, it would suggest that the VP-length effect does not in itself provide evidence for such a position, and hence that this effect is also consistent with models of grammatical development that do not assume that children are operating with adult-like grammatical knowledge during the early stages (e.g. Bowerman, 1973; Braine, 1976; MacWhinney, 1982; Pine, Lieven & Rowland, 1998; Tomasello, 2000a, 2000b).

This paper has two main aims, each of which relates to one or more of the problems identified above. The first aim is to replicate P. Bloom's results on a larger sample of children, and to investigate the extent to which the developmental patterning of the phenomena is consistent with his processing limitations account. The second aim is to provide a well-specified account of the way in which processing limitations might interact with the language-learning process and investigate the extent to which this account is able to explain both the phenomena identified by P. Bloom and any developmental changes in these phenomena. The overall aim is to use the analyses outlined above to investigate what kinds of processing limitations and what kinds of grammatical knowledge are required to explain the developmental data.

One way to test whether processing limitations in learning are sufficient to explain the phenomena attributed to processing limitations in production is to implement a process-limited learning mechanism as a computational model. In doing so, one has to specify exactly what processing limitations are assumed. In order to increase the ecological validity of the simulations, one would ideally also use a model that learns from input that is similar to that to which the child is exposed, that is, Child Directed Speech.

In this paper, we use MOSAIC, a model that has already been used to simulate several other aspects of grammatical development. MOSAIC has three processing limitations — sensitivity to frequency, sensitivity to utterance length and sensitivity to sentence position — all of which are also assumed in the standard processing limitations account, although they are construed as limitations in production rather than learning. MOSAIC does not contain ancillary assumptions regarding processing limitations, such as a lower processing load for pronouns than for lexical NPs. Our reasoning is that, by assessing whether the simulation's output mimics the child data, it is possible to test whether such ancillary assumptions are required. A further important attribute of MOSAIC is that it uses no built-in linguistic knowledge. The only grammatical knowledge that develops in MOSAIC arises from a distributional analysis of the input it receives. Because MOSAIC uses no built-in linguistic knowledge, the extent to which it mimics the child data may serve as a test of the assumption that the effects described by processing limitations theorists imply underlying competence on the part of the child.

MOSAIC has already been used to simulate child language data in English and Dutch, German and Spanish (Freudenthal, Pine & Gobet, 2002a, 2002b, 2004, 2005, 2006, in press). It learns from Child Directed Speech, and produces output that consists of actual utterances that can be directly compared to children's speech. With training, MOSAIC learns to produce progressively longer utterances, which allows for a comparison of the developmental trends displayed by the model and by language-learning children. MOSAIC will be used to simulate the VP-length effect and related phenomena as described by P. Bloom (1990).

The organisation of this paper is as follows. In the empirical section, we will assess whether P. Bloom's results are replicated in a larger sample, and how the phenomena change over time. After this, MOSAIC will be described, and we will assess the extent to which the model captures the phenomena apparent in the children.

Empirical Study

Method

Analyses were performed on all 12 children in the Manchester corpus (Theakston, Lieven, Pine & Rowland, 2001), which is available in the CHILDES database (MacWhinney, 2000). The Manchester corpus consists of transcripts of mother-child interaction recorded while the child was playing with toys in the normal home environment. Each child was recorded for an hour, twice every three weeks over a one-year period. This resulted in 34 or 35 tapes being available per child. At the beginning of the study the children's ages ranged from 1;8.22 to 2;0.25. The average MLU for the children increased from 1.58 to 3.49 over the period of the study. For the present analyses, each child's transcripts were aggregated into three batches to give three developmental phases (tapes 1-10, 11-20, and 21-34/35). The analysis was performed in a similar way to that performed by P. Bloom. In order to screen out grammatical subjectless utterances, analysis was confined to utterances including verbs that do not occur in imperative frames. Thus, only utterances containing one of the non-imperative verbs identified by P. Bloom were included in the analysis (See Table 1). Utterances containing the non-imperative verbs *See* and *Like* were also excluded from the analysis. Utterances containing *See* were excluded because *See* was found to occur in imperative frames in the input of several of the children. Utterances containing *Like* were excluded because the dual status of *Like* as both a verb and a preposition meant that utterances including this word were sometimes difficult to interpret. From those utterances that contained one of the target verbs, questions, utterances containing the words *no* or *don't*, and utterances where the target verb occurred in an embedded clause were excluded. VP length was calculated as the number of words in the utterance, starting at the target verb. Thus, the utterance *He wants to eat* had a VP of length 3. In line with P. Bloom's analysis, the analysis was carried out on utterance types rather than utterance tokens. All analyses were performed by extracting relevant output from the output files through lexical search. The resultant utterances were then hand-checked in order to determine their status. NPs were classed as subjects/objects if they would have been regarded as subjects/objects if the utterance was treated as an adult utterance.

INSERT TABLE 1 ABOUT HERE

Results

Presence or absence of a subject

Table 2 provides data on the children's average MLU in words for the three developmental phases and on the children's average VP length for utterances with and without a subject. The VP in subjectless utterances is, on average, around 0.2 words longer than the VP in utterances with a subject, which is in line with P. Bloom's results. There is also a clear increase in VP length with developmental stage, though the size of the VP-length effect seems to remain relatively constant over time.

INSERT TABLE 2 ABOUT HERE

To test the statistical significance of these results, the data were submitted to a 3 x 2 ANOVA, with developmental stage (3) and presence of a subject (2) as within-subjects measures. The ANOVA revealed a significant main effect of presence of a subject ($F(1,11) = 6.76$, $p = .025$; partial eta squared = .38), indicating that the VP is longer for subjectless utterances, and a significant main effect of developmental phase ($F(2,10) = 84.15$, $p < .001$; partial eta squared = .94), indicating that VP length increases with developmental phase (or MLU). However, the interaction was not significant ($F(2,10) = .82$, $p = .47$; partial eta squared = .14). There is thus no suggestion that the size of the VP-length effect changes over the developmental period studied here. Interestingly, there was considerable variability in the data. For all 36 comparisons (three per child), 10 comparisons showed the VP to be longer for utterances with a subject. Nevertheless, no children consistently showed longer VPs for utterances with a subject, and on average, subjectless utterances had longer VPs, which is consistent with the results reported by P. Bloom.

Pronominal vs. lexical subjects

P. Bloom also compared VP length for utterances with a pronominal and a lexical subject. The rationale was to distinguish a processing limitations account from an alternative explanation, namely that children omit subjects only when their meaning can be inferred from the context. Since longer VPs might supply more of this context, subjects might be omitted more frequently with longer VPs. This pragmatic account would, however, not predict any differences in VP length between utterances that contain different types of subjects. In contrast, the processing limitation account predicts that, since pronouns are phonetically shorter than lexical NPs, they carry a lower processing load, and will therefore occur with longer VPs. P. Bloom restricted his analysis to the pronouns *I* and *You*, since other pronouns are referentially ambiguous, and might therefore lead to longer VPs. He found that, for all three of the children, utterances with pronominal subjects had significantly longer VPs than utterances with lexical subjects, and interpreted these results as evidence for a processing limitation account.

Table 3 provides data on the average VP length for the children in the Manchester corpus for utterances with a pronominal and a lexical subject. These data suggest that pronominal subjects do tend to occur with longer VPs than lexical subjects.

INSERT TABLE 3 ABOUT HERE

The data were entered into a 3 x 2 ANOVA with developmental phase and type of subject (pronominal subject or lexical subject) as within-subjects variables. The ANOVA revealed a significant main effect of type of subject ($F(1,10) = 25.35$, $p = .001$; partial eta squared = .72)[2], and a significant main effect of developmental phase ($F(2,9) = 23.16$, $p < .001$; partial eta squared = .84). Again, the interaction was not significant ($F(2,9) = .31$, $p = .74$; partial eta squared = .06), indicating that the difference in VP-length between the two utterance types is stable over developmental phase. These results are consistent with those reported by P. Bloom, and suggest that the VP-length effect cannot be explained in pragmatic terms.

Subject vs. Object Omission

In a third set of analyses P. Bloom tested the prediction that children tend to omit subjects more often than they omit objects. This was done by comparing children's rates of subject omission with their rates of object omission for the set of verbs listed in Table 4 (all of which take obligatory objects).

INSERT TABLE 4 ABOUT HERE

P. Bloom found that all three of the children omitted subjects significantly more often than they omitted objects and interpreted this finding as consistent with the view that subjects carry a heavier processing load than objects. Table 5 provides data on the levels of subject and object provision for the children in the Manchester corpus. The object provision data were obtained using the same method as P. Bloom by measuring the level of object provision in utterances that contained one of the verbs with obligatory objects listed in Table 4. Note that, while these verbs are considered obligatory object verbs by Bloom, some of them can, on occasion occur without objects (e.g. *the cup broke*). This complication was dealt with by excluding from the analysis any objectless utterances that were grammatical in the opinion of the coder.

INSERT TABLE 5 ABOUT HERE

These data were entered into a 3 x 2 ANOVA with developmental phase and type of provision as within-subjects variables. The ANOVA revealed a significant main effect of type of provision (subject vs. object, ($F(1,10) = 255.33$, $p < .001$; partial eta squared = .96) and a significant main effect of developmental phase ($F(2,9) = 12.82$, $p = .002$; partial eta squared = .74). It also revealed a significant interaction between these two factors ($F(2,9) = 6.91$, $p = .015$; partial eta squared = .61), indicating that the level of subject provision rises at a higher rate than the level of object provision (although there may be a ceiling effect for the level of object provision). These results are also consistent with those reported by P. Bloom. However, as he himself acknowledges, there are a number of possible interpretations of this particular asymmetry.

Proportion of overt pronominal subjects and objects

In a final set of analyses, P. Bloom attempted to provide further support for a processing limitations account of the subject-object asymmetry by comparing the proportion of pronouns in subject and object position. The reasoning was that if differences in subject and object provision reflect a processing asymmetry between subjects and objects, one would expect to find other differences between these different kinds of arguments. More specifically, given that pronouns are assumed to carry a lower processing load than lexical NPs, one would expect children to use a higher proportion of pronouns in subject than in object position in order to offset the greater processing demands exerted by subjects. All three of the children in the study showed such an effect. However, as Hyams & Wexler (1993) point out, if children produce a higher proportion of pronominal than lexical subjects because pronouns exert a lower processing load than lexical NPs, then a processing limitations account would seem to predict that children's preference for pronominal as opposed to lexical subjects should decrease as their processing resources increase — and there is some evidence that the opposite is true. In this section we therefore examine differences in the proportion of pronouns in subject and object position and differences in how these proportions change over time.

Table 6 provides data on the proportion of pronouns in subject and object position for the children in the Manchester corpus across three developmental phases. It is clear from Table 6 that there is a higher proportion of pronouns in subject than in object position during all three developmental phases. However, it is also clear that the size of this effect increases rather than decreases over time.

INSERT TABLE 6 ABOUT HERE

These data were entered into a 3 x 2 ANOVA with developmental phase and type of provision as within-subjects variables. The ANOVA revealed a significant main effect of type of provision (object vs. subject) ($F(1,9) = 138.29$, $p = .001$; partial eta squared = .94), but no significant main effect of developmental phase ($F(2,8) = 1.93$, $p = .21$; partial eta squared = .33). It also revealed a significant interaction between type of provision and developmental phrase ($F(2,8) = 22.63$, $p < .001$; partial eta squared = .85), indicating a significant increase in the difference between the proportion of pronouns in subject and object position over time. These results replicate the finding of a significant difference in the proportion of pronouns in subject and object position. However, they also replicate Hyams & Wexler's finding that the size of this effect increases as a function of development. They therefore raise doubts about P. Bloom's interpretation of differences in the proportion of pronouns in subject and object position and suggest the need for an alternative explanation of this effect.

Discussion

The results of the analyses reported above largely replicate those reported by P. Bloom. Thus, they show that the VP length of subjectless utterances was significantly longer on average than the VP length of utterances with subjects, and that the VP length of utterances with pronominal subjects was significantly longer than the VP length of utterances with lexical subjects. They also show that subjects were significantly less likely to be provided in obligatory contexts than objects, and that subjects were significantly more likely than objects to be pronouns.

These results suggest that the effects identified in Adam, Eve and Sarah's data do generalise to the population of English-speaking children as a whole. However, they also extend P. Bloom's analysis by showing how these effects change over time. More specifically, they show that while the size of the VP-length effect remains relatively constant, the difference in the proportion of pronouns in subject and object position actually increases with development. This last finding raises problems for P. Bloom's interpretation of his results since it is inconsistent with the idea that differences in the proportion of pronouns in subject and object position reflect differences in the processing load associated with these different argument types. It thus illustrates the potential value of using patterns of developmental change to assess the plausibility of processing limitation accounts of performance at particular points in development. It also raises the question of whether it is possible to explain this finding as a function of processing limitations in learning, and if so, what mechanisms are responsible.

Simulating the children's speech in MOSAIC

The data from the 12 children from the Manchester corpus were compared with simulations of MOSAIC trained on the maternal speech of three children from the Manchester corpus. The model produces output of increasing MLU, thus allowing comparison of the developmental trends in the children and the simulations. The model will be described first, followed by the actual simulations.

MOSAIC

MOSAIC (Model of Syntax Acquisition In Children) has already been used successfully to simulate several phenomena in language acquisition including the Verb Island phenomenon (Jones, Gobet & Pine, 2000), patterns of pronoun case marking error (Croker, Pine & Gobet, 2000, 2001), and the Optional Infinitive phenomenon (Freudenthal, Pine & Gobet, 2002a, 2004, 2006, in press). The version used for the present simulations is identical to the one used in Freudenthal et al (in press), which successfully simulates the developmental patterning of Optional Infinitive error across four different languages (Dutch, English, German and Spanish) without fitting any parameters, or the need to make any changes to the model. The basis of the model is an n-ary discrimination network, consisting of nodes connected by (test)-links. The network is headed by a root node that has no contents. The other nodes in the network encode words or phrases. Test links encode the difference between the contents of two nodes. MOSAIC employs two learning mechanisms. The first mechanism, based upon discrimination, adds new nodes and test links to the network in a probabilistic fashion. The second mechanism, based upon similarity, adds *generative* links between nodes encoding phrases encountered in similar contexts. In its present form, the model learns from text-based input (i.e. corpora of orthographically transcribed child directed speech). The model therefore assumes that the phonological stream has been segmented into words. Given that the majority of publicly available child-directed speech corpora are orthographically rather than phonetically transcribed, MOSAIC's ability to accept such corpora as input obviously has certain advantages. However, it should be noted that the use of orthographically transcribed data also means that the model is insensitive to information that is not included in this format, such as information about intonation and relative stress. As a result, MOSAIC is unable to simulate aspects of the data that depend on such factors. For example, MOSAIC is insensitive to the difference between stressed and unstressed morphemes and will learn sequences including unstressed function words (e.g. "kick the ball") as readily as sequences of stressed content words (e.g. "Anne likes chocolate").

Adding new nodes to the network

The model encodes utterances by processing them in a left to right fashion. When the network is given input, it creates new nodes under the root node. We call the nodes immediately under the root *primitive nodes*. When additional input is received, new nodes at increasingly deeper levels are created. The model therefore encodes phrases of increasing length.

An important restriction on the creation of nodes in MOSAIC is that, while it parses the utterance in a left to right fashion, learning is anchored at the right edge of the utterance. That is, MOSAIC will only create a node encoding a word or phrase when everything that follows this word in the utterance has already been encoded in the model. Thus, MOSAIC has a strong utterance-final bias in learning. Several authors have argued, on the basis of empirical data, that sentence-final position carries more weight than other positions. That is, children learn phrases that occurred in sentence-final position more easily than phrases that occur in other positions (Wijnen, Kempen & Gillis, 2001; Shady & Gerken, 1999; Naigles & Hoff-Ginsberg, 1998).

The processing of an utterance in MOSAIC can be likened to a moving window or buffer, with the size of the window being determined by how much of the utterance has already been encoded by the model. Whenever the model encounters a word or word transition it has not yet encoded, it clears the contents of the window, and deposits the new word in it. Only when the rest of the utterance has already been encoded in the model will the new word remain in the buffer, thus making it eligible for encoding. Thus, MOSAIC processes the utterance in a left-to-right fashion, but builds up its representation of the utterance by starting at the back and slowly working its way to the front. In terms of a child attending to the speech stream, the occurrence of an unknown word will effectively clear the contents of the speech stream encountered so far, leaving the new word and the rest of the utterance for analysis. Thus, children's language learning is viewed as a process that is strongly biased towards the most recent elements in the speech stream. This process is now illustrated with an example. To keep the example simple, we assume that a node is created with a probability of 1. In fact, the probability of creating nodes, which depends on a number of factors, is initially much lower than this (see section *Node creation probability*).

An example

Assume that an empty network receives the utterance *did he go*. Since the model processes the utterance from left to right, it will first encounter the word *did*, and deposit this in the buffer. Upon reaching the second word (*he*), which has not yet been encoded in the network, the buffer is cleared, and the word *he* deposited in it. Upon encountering the unencoded word *go*, this will replace the word *he* in the buffer. Thus, when the model reaches the end of the utterance, the buffer will contain the word *go*. Since this has not yet been encoded in the model and the end of the utterance has been reached, a primitive node for the word *go* will be created. Upon a second presentation of the phrase *did he go*, the buffer will contain the phrase *he go* when reaching the end of the utterance. The model will now attempt to create a branch for the phrase *he go*. However, since no primitive node for the word *he* exists yet, it will have to create this first. Only upon a third presentation will a branch for the phrase *he go* be created. Likewise, a fourth presentation will result in the creation of a primitive node for *did*, and a fifth presentation will see the phrase *did he go* being encoded in the model. Fig. 1 shows the model after 5 presentations of the phrase *did he go*.

INSERT FIGURE 1 ABOUT HERE

Suppose the model now sees the phrase *he walks*. It first recognizes the word *he*. When it comes to the utterance-final word *walks*, there is no primitive *walks* node and therefore the model creates one. When encountering the phrase *he walks* again, it creates the test link *walks* (and node *he walks*) underneath the *he* node. At this point, the *he* node has two test links, encoding the fact that *he* has been followed by *go* and *walks*. Figure 2 shows the network at this point.

INSERT FIGURE 2 ABOUT HERE

Probability of creating a node

So far, we have assumed that nodes are created whenever there is an opportunity to do so. In fact, however, the creation of nodes is directed by a *node-creation probability* (NCP) that can vary between 0 and 1. When this probability is equal to 1, a node is always created when the opportunity arises, as in our example. However, when the NCP is less than 1, a node may or may not be created; the outcome is decided by the model generating a random number between 0 and 1 and determining whether or not this number is less than or equal to the NCP. Thus, the lower the NCP, the less likely it is that a node will be created.

Making node creation probabilistic in this way has two important consequences. First, using lower NCP values reduces the rate at which the model learns from its input and hence prevents it from learning to produce long utterances too quickly. Second, using lower NCP values makes the model more frequency sensitive. Thus, because nodes are no longer created whenever the opportunity arises, nodes for high frequency items are more likely to be created than nodes for low frequency items because opportunities to create nodes for high frequency items arise more frequently (and the probability of generating a random number that is less than or equal to the required node creation probability in several attempts is obviously higher than the probability of generating such a number in one attempt).

In order to simulate the range of MLUs displayed by young children, and generate enough output to perform meaningful analyses, a decision was made to set the NCP to gradually increasing values as a function of the size of the network (i.e. to make it easier to create nodes as the size of the network increases). A further constraint on learning is the relative ease with which phrases of various lengths can be learned. In the present simulations, nodes that encode long phrases have a lower likelihood of being created.

The specific formula for calculating the node creation probability is the following:

$$NCP = \left(\frac{1}{1 + e^{m-u/c}} \right)^{\sqrt{d}}$$

where: NCP = Node Creation Probability.

m = a constant, set to 20 for these simulations.

c = (input) corpus size (number of utterances).

u = total number of utterances seen.

d = distance to the end of the utterance.

The formula, which results in a basic sigmoid curve (when plotting node creation probability against amount of input seen), contains the number of utterances seen and the size of the corpus. Early in training, when few utterances have been seen, the term m-u/c will be close to 20, and the resulting node creation probability low. With increased training, the value of the term m-u/c will decrease, and the resulting node creation probability will increase, as will the learning rate.

The first term of the equation is raised to the power of *d* (the distance to the end of the utterance, or length of the phrase being encoded for this node). This exponent makes it more difficult to create nodes that encode longer phrases. Moreover, with increased training, the relative difficulty of encoding longer phrases decreases. When the first term approaches 1, the weight of the exponent diminishes, since the ratio between 0.9 raised to the power 2 or 3 is smaller than the ratio between 0.1 raised to the power 2 or 3.

Creation of generative links

The second type of learning used by MOSAIC is the creation and removal of *generative* links. Generative links are created between phrases[3] that share a certain percentage overlap between both the preceding and following context. Since new nodes are constantly created in the network, the percentage overlap between two phrases is likely to vary over time. As a result, the percentage overlap between two nodes may drop below the threshold and the generative link be removed. Thus, unlike nodes, generative links can be unlearned.

The rationale behind generative links is the following. When two words belong to the same *word class*, they are likely to take the same position in the sentence, and hence to be preceded and followed by similar kinds of words. For example, in English, nouns are likely to be preceded by articles and adjectives and followed by verbs. MOSAIC will pick up on this similarity by linking words that are preceded and followed by overlapping sets of words. Note that MOSAIC does not know anything about the class of nouns as a linguistic construct; it only knows that there are words that tend to take the same position relative to other words in the sentence. The development of a class of nouns is thus an emergent property of MOSAIC's distributional analysis of the input.

The percentage overlap between nodes necessary to create a generative link is an important parameter in MOSAIC. A typical value for this parameter is 20%. Setting this parameter to a lower value results in more generative links being created. Setting it to a higher value results in fewer generative links being created.

Producing utterances

MOSAIC produces output by traversing the network from the root node and outputting the contents of the test links. If the model only traverses test links, the utterances it produces must have been present in the input either as entire utterances or as utterance-final fragments (and can be seen as *rote-learned* utterances). However, MOSAIC is also able to traverse generative links during production. When the model traverses a generative link, it is able to supplement the utterance produced up to that point with the contents of the test links following the generative link. It is thus able to produce novel or *generated* utterances. Figure 4 illustrates the mechanism by which generated utterances are produced.

INSERT FIGURE 4 ABOUT HERE

Note that since there is a distinction in the model between test links and generative links, it is possible to separate the model's output into utterances that were produced by traversing only test links and utterances that were produced by traversing test and generative links. This makes it possible to distinguish between rote-learned and generated utterances in the model's output.

Processing Limitations in MOSAIC

The theoretical model that is implemented in MOSAIC is one in which children start learning the input from the right edge of the utterance. Learning is also sensitive to utterance length, in that short sequences are more likely to be learned than longer sequences, and to frequency, in that high frequency constructions are more likely to be learned than low frequency constructions. Thus, according to the model, as children develop, they learn to produce longer and longer utterances, but these utterances grow from 'right to left' in a frequency-sensitive manner. In the current version of the model, these processing limitations in learning (i.e. sensitivity to sentence position, sensitivity to utterance length and sensitivity to frequency) are implemented in two ways: 1) by manipulating the probability that certain kinds of sequences will be learned (see section on Node Creation Probability) and 2) by constraining the kind of sequences that the model is able to learn (through learning being anchored at the right edge of the utterance). Note that the utterance-final bias in learning affects the output in a very similar way to P. Bloom's assumption that there is a higher memory load associated with producing elements earlier in the sentence. Elements that occur further from the right edge of the sentence (such as subjects) are more likely to be omitted than elements that occur close to right edge of the sentence (such as objects).

The simulations

Methods

The model was trained on three input sets, consisting of maternal speech directed to Dominic, Gail and Warren, three children from the Manchester corpus (Theakston et al., 2001). These three children were chosen as a target for simulation, as the analysis of the child data showed that these children consistently showed the VP-length effect across the three developmental phases in the experimental study.

Each of these input sets consisted of approximately 30,000 utterances. Input sets were created by extracting all the maternal utterances from a given child's transcripts, concatenating these utterances into a single input file, and removing incomplete utterances (i.e. false starts and interrupted utterances) and unintelligible or partially intelligible utterances (i.e. cases where the transcriber had been unable to identify one or more of the words in the utterance). The input corpora were also (automatically) filtered to remove the following kinds of material: filler words such as "oh" and "um"; the repeated and corrected material in retraced utterances such as "That's a … that's a dog" and "I want … I need a coffee"; and words and sequences tagged onto the end of utterances, such as the vocative in "What would you like, Anne?" and the tag in "You like chocolate, don't you?" (although it should be noted that vocatives and tags were not removed if they occurred as isolated utterances).

As MOSAIC learns slowly, input files were presented to the model several times. Output files were then created by outputting all of the utterances that the model was able to produce after each presentation of the input file. This included all the rote-learned utterances that the model was able to produce (i.e. all the utterance-final phrases encoded in the model at each point in development), together with all the generated utterances that the model was able to produce (i.e. all the utterances that could be produced by substituting a word into an utterance-final phrase across a generative link).

The MLU of the model's output increases steadily with consecutive runs of the model. In order to compare the model's output with that of the children, the output files that matched the children most closely in terms of MLU were selected. The 'utterances' in these files were analysed in exactly the same way as the utterances in the child data.

Results

Descriptive statistics for Dominic, Gail and Warren's models are shown in Table 7. As can be seen, the MLU, amount of output and the proportion of novel utterances all increase with additional runs of the model.

INSERT TABLE 7 ABOUT HERE

Presence or absence of subject

Table 8 shows the VP length for the three models, for utterances with and without a subject. The relevant child data are shown in Table 9. There is a clear difference in the VP length of subjectless utterances and utterances with subjects in all of the models. In 7 out of 9 comparisons the VP length for utterances without a subject is higher than for utterances that include a subject. This effect appears to be more pronounced in the models than in the children. However, the critical finding is that the model simulates the VP-length effect.

INSERT TABLE 8 ABOUT HERE

INSERT TABLE 9 ABOUT HERE

The reason why MOSAIC captures this effect is that its output is based on sentence-final representations that are becoming progressively longer as a function of learning in a way that is independent of the grammatical role played by particular words or phrases. This means that, all other things being equal, the average length of utterances including subjects will be the same as the average length of utterances that do not include a subject, and hence that subjectless VPs (i.e. utterances with no subject) will tend to be longer on average than the VPs of utterances with subjects (i.e. partial utterances from which the subject has been removed).

Pronominal vs. lexical subjects

Table 10 shows the VP length for the three models, for utterances with pronominal (restricted to *I* and *You*) and lexical subjects. The relevant child data are shown in Table 11. There is a clear difference in the VP length of utterances with pronominal and lexical subjects. In 7 out of 8[4] comparisons the VP length for utterances with a pronominal subject is higher than for utterances with a lexical subject. The model therefore simulates the pronominal/lexical effect.

INSERT TABLE 10 ABOUT HERE

INSERT TABLE 11 ABOUT HERE

The reason why MOSAIC captures this effect is that pronouns are more frequent than lexical NPs, and lexical subjects are, on average, longer than pronouns. This leads to structures with pronominal subjects being learned more easily. As a result, pronominal subjects tend to occur in longer utterances.

Subject vs. Object Omission

Table 12 shows the levels of subject and object omission for the three models. The child data are shown in table 13. While the model somewhat underestimates the levels of subject and object provision, the subject-object asymmetry is clearly reflected in the models' output as the levels of object provision are higher than the levels of subject provision for all comparisons. Both the levels of subject and object provision increase with developmental phase. The models omit subjects more often than objects because subjects occur earlier in the utterance. Since the model produces incomplete, sentence-final phrases, its output is more likely to include objects than subjects in constructions where they are required.

INSERT TABLE 12 ABOUT HERE

INSERT TABLE 13 ABOUT HERE

Proportion of Pronominal Subjects and Objects

Table 14 shows the proportion of overt subjects and objects that are pronominal in the three models. The relevant child data are shown in Table 15. Again, all comparisons show a difference in the expected direction. However, it should be noted that in this case the effect is not due to any processing asymmetry between pronouns and lexical NPs since there is no such asymmetry implemented in the model. What the difference appears to reflect is an asymmetry in the input on which the model has been trained. The obvious explanation for this asymmetry in the input is that subjects tend to convey given information, whereas objects tend to convey new information, and it is pragmatically more appropriate to introduce new information using a lexical NP than a pronoun. Indeed, several authors have argued that using pronouns in subject position and lexical NPs in object position is the preferred argument structure for English (e.g. Clancy, 2001).

INSERT TABLE 14 ABOUT HERE

INSERT TABLE 15 ABOUT HERE

The obvious implication of this finding is that one does not need to assume a processing asymmetry between pronouns and lexical NPs in order to explain this particular effect. However, it will also be remembered that in the children's data, the difference in the proportion of subjects and objects that were pronouns became more pronounced with developmental phase. This developmental pattern is apparent in Dominic and Gail (but not in Warren), and this is mimicked in their respective models. Thus the difference increases from .26 to .47 in Dominic's model, from .07 to .42 in

Gail's model, and decreases from .82 to .68 in Warren's model. The absence of this effect in Warren and his model (given a strong effect across the children in the Manchester corpus), may be caused by a potential ceiling effect in the child which in turn may reflect relatively large proportions of pronominal subjects and lexical objects in the input for that child.

The reason why Gail and Dominic's model show an increasing ratio of pronominal subjects over objects is that as the models develop they tend to become more generative around pronouns and lexical NPs that occur in similar positions in the input. As subjects tend to be pronouns and objects tend to be lexical NPs, increasing generativity tends to result in an increase in the proportion of subjects that are pronominal and an increase in the proportion of objects that are lexical NPs.

MOSAIC thus not only simulates the effect predicted by P. Bloom, but also the way in which this effect changes over time, which, as was noted earlier, is actually inconsistent with his interpretation of the data.

Discussion

The research reported in the present paper had two principal aims. The first aim was to replicate and extend the findings reported by P. Bloom (1990) on patterns of subject provision in young English-speaking children's speech. The second aim was to investigate the extent to which these results could be explained in terms of processing limitations in learning, as opposed to processing limitations in production.

With respect to the first of these aims, it is clear that the results from the children's data largely replicated the results reported by P. Bloom. Thus, the VP length of subjectless utterances was significantly longer on average than the VP length of utterances with subjects; the VP length of utterances with pronominal subjects was significantly longer than the VP length of utterances with lexical subjects; subjects were significantly less likely to be provided in obligatory contexts than objects; and subjects were significantly more likely than objects to be pronouns. These results suggest that the effects identified in Adam, Eve and Sarah's data do generalise to the population of English-speaking children as a whole. However, there is one aspect of the data that raises doubts about Bloom's interpretation of these effects. This is the fact that, although the children did tend to use pronouns in subject position significantly more often than in object position, this asymmetry actually increased rather than decreased with development. This pattern is inconsistent with the assumption that differences in children's use of pronouns and lexical NPs in subject and object position reflect a tendency to offset the greater processing demands associated with subject position by producing pronominal as opposed to lexical subjects. If this were the case, one would expect differences in the proportion of pronouns in subject and object position to decrease rather than increase with development as children's processing resources increase. However, the pattern is consistent with the results of a developmental analysis of Adam and Eve's data reported by Hyams and Wexler (1993). P. Bloom (1993) dismisses Hyams and Wexler's results on the grounds that they are based on only one (arbitrarily chosen) transcript per data point. However, the results of the present study are based on speech from 12 children, with at least 10 one-hour transcripts contributing to each developmental phase for each child. It is therefore unlikely that the developmental changes observed are due to sampling error. The implication is that P. Bloom's account of differences in the proportion of pronouns in subject and object position is on the wrong track, and that a different explanation is required, at least for this particular phenomenon.

The second aim of the paper was to investigate the extent to which both the phenomena identified by P. Bloom and the developmental patterning of these phenomena could be explained in terms of processing limitations in learning. This was done by using a computational model of language learning (MOSAIC) to simulate data on subject and object provision across three developmental phases. The results show that MOSAIC was able to simulate all of the effects identified by P. Bloom in terms of a process-limited distributional analysis of real child-directed speech.

They also show that, for the two children that show this effect, MOSAIC was able to simulate the developmental changes in the proportion of pronouns in subject and object position that are problematic for P. Bloom's account.

These results have a number of important implications both for processing limitation accounts in general and for P. Bloom's processing limitation account in particular. First, they show that by focusing on the interaction between processing limitations and the distributional characteristics of the language to which children are exposed, it is possible to provide a simpler, more unified explanation of the phenomena identified by P. Bloom. Thus, sensitivity to input frequency, utterance length and utterance position, all of which are also assumed within the standard processing limitation account, appear to be sufficient to explain all of the phenomena identified by P. Bloom. Moreover, this is despite the fact that these phenomena include effects that, according to P. Bloom's account, require the additional assumption that lexical NPs carry a heavier processing load than pronouns. Interestingly, these pronominal/lexical effects both appear to be due to differences in the distributional properties of pronouns and lexical NPs in the input language. Thus, it is possible to explain differences in the VP length of utterances with pronominal and lexical subjects in terms of differences in the frequency of pronouns and lexical NPs. On the other hand, it is possible to explain differences in the proportion of pronouns in subject and object position in terms of similar differences in the input language that appear to reflect the preferred argument structure of English. These results suggest that P. Bloom's failure to consider the distributional characteristics of English as it is actually spoken led him to develop an unnecessarily complex account of the phenomena identified in his study. They thus underline the importance of considering the potential role of the distributional statistics of the language being learned before invoking hypothetical processing asymmetries in order to explain the data.

Second, the results show that it is possible to simulate effects that have traditionally been interpreted as evidence for full competence coupled with processing limitations in production as a function of a resource-limited learning process that is constructing much less abstract representations. Thus, length effects such as those reported by P. Bloom have traditionally been interpreted as evidence that children represent the full underlying structure of the target sentence, but omit elements of this structure from the utterances that they produce because the overall processing load of the target sentence exceeds their limited processing capacity. However, the results of the present study show that it is possible to simulate such effects as a function of increases in the knowledge required to produce longer utterances, regardless of the level at which this knowledge is represented. Thus MOSAIC simulates the VP-length effect because, all other things being equal, the utterances that it produces with and without subjects tend to be equally long on average.

This means that the VPs of subjectless utterances (i.e. utterances that consist only of a VP) tend to be longer than the VPs of utterances with subjects (i.e. utterances that consist of a subject and a VP). The implication is that the VP-length effect is a straightforward consequence of the fact that the average length of children's utterances tends to increase with development, and hence tells us little about the nature of the child's underlying knowledge. Note that this conclusion is not necessarily incompatible with the nativist view that children represent the correct syntactic structure of the sentences they are producing from the beginning. However, it does suggest that the kind of length effects reported by P. Bloom should not be taken as evidence for such a position. That is to say, these effects are also consistent with the constructivist view that children's early grammatical knowledge is rather less abstract than that of adults (e.g. Bowerman, 1973; Braine, 1976; MacWhinney, 1982; Pine, Lieven & Rowland, 1998; Tomasello, 2000a, 200b).

In fact, the model described in the present paper does a remarkably good job of simulating the phenomena identified by P. Bloom, despite using less abstract representations than would be assumed even by most constructivist accounts. However, it is important to acknowledge that, although the model captures all of the effects reported by P. Bloom, the fit between the output of the model and the output of young English-speaking children is still far from perfect. Thus, while MOSAIC produces longer verb phrases for subjectless utterances than for utterances with subjects (as well as for utterances with pronominal compared to lexical subjects) it rather overestimates the size of this effect. This discrepancy suggests that MOSAIC provides an overly simplistic account of children's early language processing. However, it is worth emphasising that the aim of the present study was not to build a realistic model of early language acquisition, but to replicate the phenomena identified by P. Bloom and investigate the kinds of processing limitations and the kinds of grammatical knowledge that are required to explain these phenomena. What the results of the study show is that it is possible to simulate all of the phenomena identified by P. Bloom in terms of a resource-limited distributional analysis of real child-directed speech. These results suggest that the effects reported by P. Bloom do not constitute evidence for underlying competence on the part of the child, and are at least as compatible with an explanation in terms of processing limitations in learning. However, they also underline the need to develop more empirically grounded models of the way that processing limitations in learning might influence the language acquisition process. One obvious way of doing this is to use artificial language-learning tasks to investigate limitations in children's early distributional learning abilities (e.g. Santelmann & Jusczyk, 1998; Gomez & Gerken, 1999). Another is to use computational modelling techniques to explore how such limitations might interact with the distributional properties of real child-directed speech to determine the nature and scope of children's early grammatical knowledge.

References

Bates, E. & Carnavale, G.F. (1993). New directions in research on child development. *Developmental Review*, *13*, 436-470.

Bloom, L. (1970). *Language development: Form and function in emerging grammars*. Cambridge, MA: MIT press.

Bloom, L., Miller, P. & Hood, L. (1975). Variation and reduction as aspects of competence in language development. In A. Pick (Ed.), *The 1974 Minnesota Symposium on Child Psychology*. Minneapolis: University of Minnesota Press.

Bloom, P. (1990). Subjectless sentences in child language. *Linguistic Inquiry*, *21*, 491-504.

Bloom, P. (1993). Grammatical continuity in language development: the case of subjectless sentences. *Linguistic Inquiry*, *24*, 721-734.

Bowerman, M. (1973). Structural relationships in children's utterances: syntactic or semantic? In T. E. Moore (Ed.), *Cognitive development and the acquisition of language*. New York: Academic Press.

Braine, M. D. S. (1976). Children’s first word combinations. *Monographs of the Society for Research in Child Development*, *41*, 164.

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

Clancy, P. (2001). The lexicon in interaction: Developmental origins of preferred argument structure in Korean. In J.W. DuBois, L.E. Kumpf & W.J. Ashby (Eds.), *Preferred argument structure: Grammar as architecture for function*. Amsterdam:

John Benjamins.

Croker, S., Pine, J.M., & Gobet, F. (2000). Modelling optional infinitive phenomena: A computational account. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the Third International Conference on Cognitive Modelling*. Veenendaal: Universal Press.

Croker, S., Pine, J.M. & Gobet, F. (2001). Modelling children's case-marking errors with MOSAIC. In E.M. Altmann, A. Cleeremans, C.D. Schunn & W.D. Gray (Eds.), *Proceedings of the Fourth International Conference on Cognitive Modeling*. Mahwah, NJ: Erlbaum.

Freudenthal, D., Pine, J. & Gobet, F. (2002a). Modelling the development of Dutch Optional Infinitives in MOSAIC. In: W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Freudenthal, D., Pine, J. & Gobet, F. (2002b). Subject omission in children's language: the case for performance limitations in learning. In W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Freudenthal, D., Pine, J.M. & Gobet, F. (2004). Simulating the temporal reference of Dutch and English Root Infinitives. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 410-415). Mahwah, NJ: Lawrence Erlbaum Associates.

Freudenthal, D., Pine, J.M. & Gobet, F. (2005). On the resolution of ambiguities in the extraction of syntactic categories through chunking. *Cognitive Systems Research*, 6, 17-25.

Freudenthal, D., Pine, J.M. & Gobet, F. (2006). Modelling the development of children's use of Optional Infinitives in Dutch and English using MOSAIC. *Cognitive Science*, *30*, 277-310.

Freudenthal, D., Pine, J.M., & Gobet, F. (in press). Modelling the developmental patterning of finiteness marking in English, Dutch, German and Spanish using MOSAIC. To appear in *Cognitive Science.*

Gerken, L. (1991). The metrical basis for children's subjectless sentences. *Journal of Memory and Language*, *30*, 431-451.

Gomez, R. L. & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109-135.

Hyams, N. & Wexler, K. (1993). On the grammatical basis of null subjects in child language. *Linguistic Inquiry*, *24*, 421-59.

Jones, G., Gobet, F. & Pine, J.M. (2000). A process model of children's early verb use. In L.R. Gleitman & A.K. Joshu (Eds.), *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

MacWhinney, B. (1982). Basic syntactic processes. In S. Kuczaj (Ed.), *Language acquisition: Volume 1. Syntax and semantics*. Hillsdale, NJ: Erlbaum.

MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk (3rd Edition).* Mahwah, NJ: Erlbaum.

Naigles, L. & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs. Effects of input frequency and structure on children's early verb use. *Journal of Child Language*, *25*, 95-120.

Pine, J. M., Lieven. E. V. M. & Rowland, C. F. (1998). Comparing different models of the development of the English verb category. *Linguistics*, *36*, 807-830.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Santelmann, L. & Jusczyk, P. (1998). Sensitivity to discontinuous dependencies in language learners: Evidence for limitations in processing space. *Cognition*, *69*, 105-134.

Shady, M. & Gerken, L. (1999). Grammatical and caregiver cue in early sentence comprehension. *Journal of Child Language*, *26*, 163-176.

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

Tomasello, M. (2000a). Do young children have adult syntactic competence? *Cognition*, *4*, 209-253.

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

Valian, V. (1991). Syntactic subjects in the early speech of American and Italian children. *Cognition*, *40*, 21-81.

Valian, V. & Eisenberg, Z. (1996). The development of syntactic subjects in Portuguese-speaking children. *Journal of Child Language*, *23*, 103-128.

Valian, V., Hoeffner, J. & Aubry, S. (1996). Young children's imitation of sentence subjects: Evidence of processing limitations. *Developmental Psychology*, *32*, 153-164.

Wijnen, F. Kempen, M. & Gillis, S. (2001). Root infinitives in Dutch early child language. *Journal of Child Language*, *28*, 629-660.

Figure 1: MOSAIC after it has seen the utterance *did he go* five times.

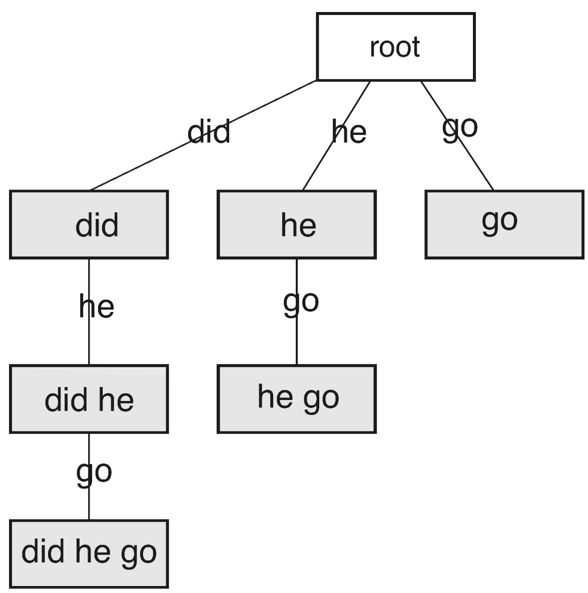

Figure 2: MOSAIC after it has seen the utterance *did he go* five times, and the utterance *he walks* twice.

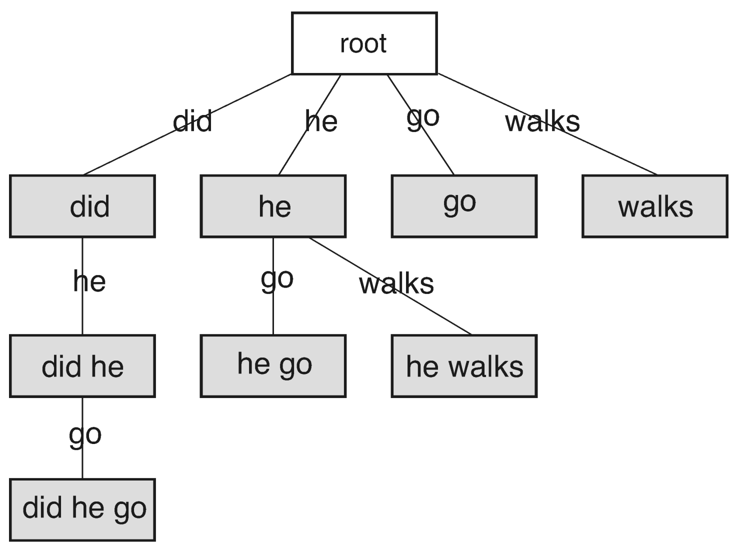

Figure 3: Generating an utterance. Because *she* and *he* have a generative link, the model can output the novel utterance *she sings*. (For simplicity, preceding nodes are ignored in this figure.)

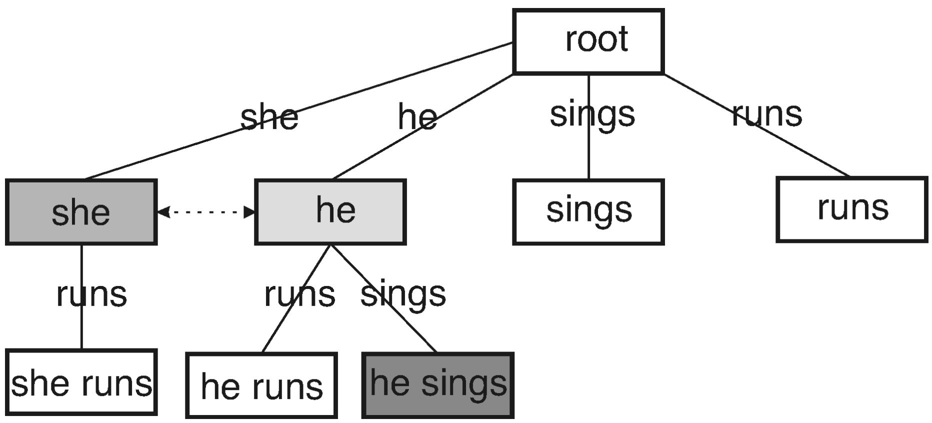

Table 1: Non-Imperative Verbs used in the analysis.

Care	Know	Loves
Cry	Laugh	Miss
Fall	Laughs	Need
Falls	Live	Sneeze
Forget	Lives	Want
Grow	Love	Wants

Table 2: Mean Length of Utterance and VP length as a function of Developmental Phase and Presence or Absence of a Subject (standard deviations in parenthesis).

Dev. Phase	MLU	+Sub	-Sub
1	2.22 (.26)	2.36 (.44)	2.55 (.58)
2	2.94 (.29)	3.06 (.45)	3.30 (.50)
3	3.60 (.32)	3.77 (.41)	3.89 (.45)

Table 3: Mean Length of Verb Phrase for utterances containing a Pronominal or Lexical subject as a function of Developmental Phase (standard deviations in parenthesis).

Dev. Phase	Pronominal	Lexical
1	2.46 (.45)	2.08 (.65)
2	3.27 (.43)	2.67 (.68)
3	3.82 (.37)	3.27 (.79)

Table 4: Obligatory Object Verbs used by P. Bloom (1990).

Bought	Drinked	Ironed	Miss	Saved	Throwed
Broke	Fix	Like	Need	Saw	Took
Brought	Folded	Love	Pulled	See	Want
Caught	Found	Loves	Rode	Sharpened	Wants
Covered	Gave	Made	Said	Thought	Washed

Table 5: Proportion of Overt Subjects and Objects as a function of Developmental Phase (standard deviations in parenthesis).

Dev. Phase	Subject	Object
1	.39 (.15)	.85 (.06)
2	.55 (.14)	.94 (.02)
3	.64 (.13)	.94 (.02)

Table 6: Proportion of Overt Pronominal Subjects and Objects as a function of Developmental Phase (standard deviations in parenthesis).

Dev. Phase	Subject	Object
1	.70 (.16)	.26 (.11)
2	.79 (.15)	.23 (.10)
3	.87 (.15)	.24 (.09)

Table 7: Descriptive statistics for Dominic, Gail and Warren's models.

Phase	Dominic's Model			Gail's Model			Warren's Model		
	MLU	Total Utt.	Prop. novel	MLU	Total Utt.	Prop. novel	MLU	Total Utt.	Prop. novel
1	2.02	2119	.10	2.35	7965	.15	2.45	7851	.20
2	2.48	6042	.22	3.05	20555	.33	3.40	23249	.37
3	3.58	26651	.49	3.73	28604	.46	4.24	30878	.47

Table 8: Models' Verb Phrase lengths for utterances with and without a Subject as a function of Developmental Phase.

Phase	Dominic's Model		Gail's Model		Warren's Model	
	+sub	-sub	+sub	-sub	+sub	-sub
1	1.63	2.17	1.71	2.39	1.73	2.52
2	2.88	2.69	2.39	3.64	3.30	3.29
3	3.10	3.52	3.00	4.20	3.24	4.48

Table 9: Children' Verb Phrase lengths for utterances with and without a Subject as a function of Developmental Phase.

Phase	Dominic		Gail		Warren	
	+sub	-sub	+sub	-sub	+sub	-sub
1	1.95	2.20	3.12	3.99	2.38	2.78
2	2.50	2.89	3.73	4.36	3.63	4.07
3	3.25	3.60	4.32	4.45	4.04	4.36

Table 10: Models' Verb Phrase lengths for utterances with a Pronominal and Lexical Subject as a function of Developmental Phase.

Phase	Dominic's Model		Gail's Model		Warren's Model	
	Pron.	Lex.	Pron.	Lex.	Pron.	Lex.
1	1.67	n.a.	2.25	1.00	1.88	2.00
2	3.37	2.00	2.70	1.44	3.57	2.00
3	3.49	1.90	3.15	2.29	3.26	2.35

Table 11: Children's Verb Phrase lengths for utterances with a Pronominal and Lexical Subject as a function of Developmental Phase.

Phase	Dominic		Gail		Warren	
	Pron.	Lex.	Pron.	Lex.	Pron.	Lex.
1	2.14	1.60	2.77	1.67	2.52	1.67
2	2.64	2.10	2.91	2.25	3.93	2.92
3	3.42	2.65	3.43	2.70	4.10	3.62

Table 12: Models' rates of Subject and Object provision as a function of Developmental Phase.

Phase	Dominic's Model		Gail's Model		Warren's Model	
	Sub	Obj	Sub	Obj	Sub	Obj
1	.26	.75	.44	.68	.25	.72
2	.45	.82	.34	.83	.39	.83
3	.56	.94	.51	.84	.45	.95

Table 13: Children's rates of Subject and Object provision

as a function of Developmental Phase.

Phase	Dominic		Gail		Warren	
	Sub	Obj	Sub	Obj	Sub	Obj
1	.36	.83	.25	.93	.71	.89
2	.50	.97	.36	.93	.48	.97
3	.49	.95	.56	.92	.50	.98

Table 14: Model's Proportions of Overt Subjects and Objects that are Pronominal as a function of Developmental Phase.

Phase	Dominic's model		Gail's model		Warren's model	
	Sub	Obj	Sub	Obj	Sub	Obj
1	1.00	.74	.71	.64	.91	.09
2	.98	.59	.75	.49	.98	.26
3	.90	.43	.91	.49	.92	.24

Table 15: Children's Proportions of Overt Subjects and Objects that are Pronominal as a function of Developmental Phase.

Phase	Dominic		Gail		Warren	
	Sub	Obj	Sub	Obj	Sub	Obj
1	.74	.38	.81	.45	.84	.15
2	.78	.39	.87	.19	.70	.15
3	.78	.30	.91	.14	.87	.22

Footnotes

---

[1] Gerken interprets differences in young children's ability to imitate pronominal and lexical subjects in terms of a metrical template model of the omission of pronouns and function words (Gerken, 1991). However, although they did find differences in young children's ability to imitation pronominal and lexical subjects, Valian et al. (1996) failed to find any support in their data for 4 further predictions of Gerken's metrical template hypothesis. The implication is that although metrical structure may play a role in determining the pattern of subject provision in young children's data, the metrical hypothesis is unlikely to provide a comprehensive account of the subject omission data.

[2] For this (and some of the subsequent) analyses, one of the children did not produce any relevant utterances in developmental phase one. Since the data for this child do not contribute to the ANOVA, the degrees of freedom add up to 11.

[3] Strictly speaking, generative links are created between *nodes* encoding phrases that have the property mentioned above. When the context is clear, we will use the simpler construction

[4] No lexical subjects were produces by Dominic's model in the first developmental phase. A comparison of the VP length effect is therefore not possible here.