Subject Omission in Children's Language: The Case for Performance Limitations in Learning

Daniel Freudenthal (DF@Psychology.Nottingham.Ac.Uk)
Julian Pine (JP@Psychology.Nottingham.Ac.Uk)
Fernand Gobet (FRG@Psychology.Nottingham.Ac.Uk)
School of Psychology, University of Nottingham
University Park, Nottingham, NG7 2RD UK

Abstract

Several theories have been put forward to explain the phenomenon that children who are learning to speak their native language tend to omit the subject of the sentence. According to the pro-drop hypothesis, children represent the wrong grammar. According to the performance limitations view, children represent the full grammar, but omit subjects due to performance limitations in production. This paper proposes a third explanation and presents a model which simulates the data relevant to subject omission. The model consists of a simple learning mechanism that carries out a distributional analysis of naturalistic input. It does not have any overt representation of grammatical categories, and its performance limitations reside mainly in its learning mechanism. The model clearly simulates the data at hand, without the need to assume large amounts of innate knowledge in the child, and can be considered more parsimonious on these grounds alone. Importantly, it employs a unified and objective measure of processing load, namely the length of the utterance, which interacts with frequency in the input. The standard performance limitations view assumes that processing load is dependent on a phrase's syntactic role, but does not specify a unifying underlying principle.

Subject Omission

Children who are acquiring English often produce sentences with missing subjects, like those shown below.

Hug Mummy
Play Bed
Writing Book
See Running

While these examples clearly do not adhere to adult English grammar, many contemporary theories of child language assume that children produce their sentences on the basis of an abstract grammar. Theories differ with respect to how much the hypothesized grammar differs from the adult grammar. According to the *pro-drop hypothesis* (Hyams, 1986; Hyams & Wexler, 1993), children represent a grammar that is different from the adult grammar in that it allows *null subjects*. In this respect, children's grammar resembles that of adult Italian and Spanish speakers. Other authors have argued that children actually possess the correct adult grammar, but drop subjects because they have difficulty expressing the (correct) underlying form due to some kind of processing bottleneck (L. Bloom, 1970; L. Bloom, Miller & Hood, 1975; Pinker, 1984; P. Bloom 1990; Valian, 1991). Thus, a child producing an utterance is thought to represent a grammatically correct underlying structure, but, due to performance limitations, some elements have a lower probability of being expressed than others.

A number of phenomena have been cited as evidence for the performance limitations view. P. Bloom (1990) showed that, in utterances with a subject, the length of the Verb Phrase (VP) is shorter than it is in utterances without a subject. The load associated with the provision of a subject is thought to decrease the likelihood of expressing a longer verb phrase. Along similar lines, the length of the VP is greater when the subject is a pronoun, than when it is a noun. This is thought to result from the fact that pronouns are phonetically shorter, and the fact that non-pronominal subjects may be longer than pronominal ones. L. Bloom (1970) has also found that subject omission is more likely in negated sentences or in sentences with relatively new (unfamiliar) verbs. Presumably, the load associated with negation and novel verbs is such that it induces subject omission.

While the performance limitations view makes sense from an information-processing point of view, it is not very precise in its predictions (Theakston, Lieven, Pine & Rowland, 2001). Performance limitations accounts also tend to be rather ad hoc in nature. Given the imprecise nature of performance limitations, it becomes all too easy to posit a greater processing load whenever the provision of a certain element leads to a greater likelihood of the omission of another, especially when there is an interaction with frequency. Furthermore, it is not clear whether an explanation of the patterns in the data requires a limitation in production coupled with full knowledge of a language's grammar (as the performance limitations view typically has it). In fact, as Theakston et al. point out, performance limited *learning of lexical items* (independent of syntactic complexity) may well give rise to the same pattern of

results without the need to assume a full representation of the grammar, and a different processing load for various types of grammatical roles. The present paper aims to test these claims by seeing to what extent a performance limited distributional analysis of naturalistic input can account for the pattern of omission and provision of grammatical categories that is found in children's speech. To this end, we aim to simulate the effects that P. Bloom (1990) attributes to performance limited production. We will now introduce the model we have used for these simulations.

MOSAIC

MOSAIC (Model of Syntax Acquisition In Children) is an instance of the CHREST architecture, which in turn is a member of the EPAM family of models. CHREST models have successfully been used to model phenomena such as novice-expert differences in chess and computer programming. In language acquisition, MOSAIC has been applied to the modelling of the use of optional infinitives in English and Dutch, the learning of sound patterns and the Verb Island phenomenon. Due to space limitations, we refer the reader to another paper in this volume for the relevant references (Freudenthal, Pine & Gobet, 2002).

The basis of the model is a discrimination net, which can be seen as an index to Long-Term Memory. The network is an n-ary tree, headed by a root node. Training of the model takes place by feeding utterances to the network, and sorting them (see Figure 1). Utterances are processed word by word. When the network is empty, and the first utterance is fed to it, the root node contains no test links. When the model is presented with the utterance *He walked home*, it will create on its first pass three test links from the root. The test links hold a key (the test) and a node. The key holds the actual feature (word or phrase) being processed, while the node contains the sequence of all the keys from the root to the present node. Thus, on its first pass, the model just learns the words in the utterance. When the model is presented with the same sentence a second time, it will traverse the net, and find it has already seen the word *he*. When it encounters the word *walked* it will also recognize it has seen this word before, and will then create a new link under the *he* node. This link will have *walked* as its key, and *he walked* in the node. In a similar way, it will create a *walked home* node under the primitive *walked* node. On a third pass, the model will add a *he walked home* node under the *he walked* chain of nodes. The model thus needs three passes to encode a three-word phrase with all new words. (For expository purposes, here we assume that a node is created with a probability of 1. As is explained under *learning rate*, this probability is actually lower and dependent on a number of factors). Figure 1 shows the development of the net through the three presentations of the sentence.

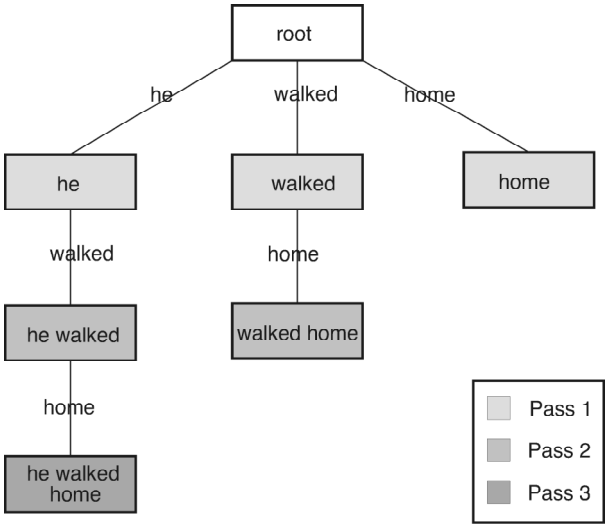

Figure 1: MOSAIC learning an input.

As the model sees more input, it will thus encode longer and longer phrases. Apart from the standard test links between words that have followed each other in utterances previously encountered, MOSAIC employs *generative* links that connect nodes that have a similar context. Generative links can be created on every cycle. Whether a generative link is created depends on the amount of overlap that exists between nodes. The overlap is calculated by assessing to what extent two nodes have the same nodes directly above and below them (two nodes need to share 10% of both the nodes below and above them in order to be linked). This is equivalent to assessing how likely it is that the two words are preceded and followed by the same words in an utterance. Since words that are followed and preceded by the same words are likely to be of the same word class (for instance Nouns or Verbs), the generative links that develop end up linking clusters of nodes that represent different word classes. The induction of word classes on the basis of their position in the sentence relative to other words is the only mechanism that MOSAIC uses for representing syntactic classes.

The main importance of generative links lies in the role they play when utterances are generated from the network. When the model generates utterances, it will output all the utterances it can by traversing the network until it encounters a terminal node. When the model traverses standard links only, it produces utterances or parts of utterances that were present in the input. In other words, it does *rote* generation. During generation, however, the model can also traverse generative links. When the model traverses a generative link, it can

supplement the utterance up to that point with a phrase that follows the node that the current node is linked to. As a result, the model is able to generate utterances that were not present in the input. Figure 2 gives an example of the generation of an utterance using a generative link.

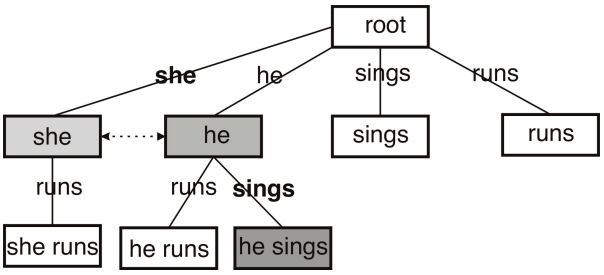

Figure 2: Generating an utterance. Because *she* and *he* have a generative link, the model can output the novel utterance *she sings*. (For simplicity, preceding nodes are ignored in this figure.)

Learning Rate

MOSAIC does not simply learn all the utterances it encounters. The probability of the creation of a node is dependent on the size of the net and the length of the utterance it encodes. This has the effect of making the learning process frequency sensitive. If an utterance is seen more often, it has a higher probability of being created. Finally, phrases that occur in an utterance final position in the input (have an *end marker*) have a higher probability of being encoded. The precise formula governing learning rate is given elsewhere in this volume (Freudenthal, Pine & Gobet, 2002).

Performance Limitations in MOSAIC

The only performance limitations in MOSAIC are the following:

- Frequency: high frequency items have a higher likelihood of being encoded, and thus feature in longer utterances
- Short phrases have a higher likelihood of being encoded than long phrases
- Utterance final phrases have a higher likelihood of being encoded.
- An utterance will only be produced (generated) if its final phrase has occurred in sentence final position in the input.

It may be appropriate to point out that these performance limitations are plausible from general theorizing in the cognitive psychology and learning literature. Huttenlocher et al. (1991) provide evidence for the effect of frequency on vocabulary learning. Evidence for the importance of sentence final position has been provided by Naigles & Hoff-Ginsberg (1998) and has been attributed to prosodic highlighting of the sentence final position (Shady & Gerken, 1999). In contrast to the standard performance limitations view, processing load in MOSAIC does not vary as a function of grammatical role. Also note that the version of MOSAIC used for these simulations is identical to that which Freudenthal, Pine & Gobet (2002) used for the simulation of the optional infinitive phenomenon in Dutch. No free parameters were fitted to obtain these results.

Subject Omission in MOSAIC

MOSAIC creates utterances without subjects because the model can output partial utterances, provided that the utterance final element has occurred in a sentence final position in the input. As a consequence, constituents that take a position early in the sentence, have a higher probability of being omitted than those that take a position further *downstream*. Since the subject takes first position in English, it has the highest likelihood of being omitted. However, this prediction is not tied to the English language. MOSAIC would generate utterances with omitted subjects in all languages that have the subject as the first element in their underlying word order.

Method

In order to simulate the data presented by P. Bloom (1990), two MOSAIC models were trained using corpora of maternal speech available in the CHILDES database (MacWhinney & Snow, 1990). We used the files of Anne and Becky. The mean length of utterance (MLU) in the output generated from the models was 2.87 for Anne's model, and 3.41 for Becky's model. In line with Bloom's analysis, we limited our analysis to utterances which could not be interpreted as imperatives. This is necessary as subjectless sentences in English are grammatical as imperatives (e.g. *Put it down*). Bloom selected a list of *nonimperative verbs* and *past tense verbs* for his analysis. Since these verbs cannot be used in an imperative form, sentences which contain a verb from these lists, and do not contain a subject, are true examples of subject omission. Tables 1 and 2 give the lists of verbs that were used for these analyses.

Table 1: Nonimperative verbs used for analysis

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

In line with Bloom's analysis, we removed from our samples all questions, all utterances that contained the words *not* or *don't*, all utterances where the verb was not used in a productive way, and all utterances where the target verb was part of an embedded clause.

Table 2: Past tense verbs used for analysis

Ate	Fixed	Saved
Bit	Folded	Saw
Bought	Forgot	Sent
Broke	Found	Sharpened
Brought	Gave	Spilled
Came	Goed	Stepped
Caught	Ironed	Stopped
Closed	Left	Thought
Cooled	Lost	Throwed
Covered	Made	Took
Cried	Melted	Tored
Drinked	Pee-peed	Tripped
Dropped	Pulled	Turned
Dropt	Rode	Washed
Falled	Said	Went
Fell	Sat	Wrote

Table 3 gives the data for three children that Bloom reports and the two simulations (Anne's and Becky's model). It can be seen that for the children, the Verb Phrase length in utterances with a subject is shorter than in utterances without a subject. It can also be seen that MOSAIC readily simulates this result, and the size of the effect is quite comparable to that in the children. The difference in verb phrase length is statistically significant for both Anne's model ($t(330) = 4.82$, $p < .001$), and Becky's model ($t(314) = 4.64$, $p < .001$).

Table 3: Mean length of Verb Phrases in sentences with and without subjects

Child	With Subject	Without Subject
Adam	2.33	2.60
Eve	2.02	2.72
Sarah	1.80	2.46
Anne's Model	2.14	2.76
Becky's Model	2.58	3.31

MOSAIC obtains this result because the probability of learning an item in MOSAIC is dependent only on its frequency and length, and not on its grammatical role. There is thus no reason (apart from differences in frequency), why sentences with subjects should, on average, be longer (or shorter) than those without. The fact that verb phrases in utterances with subjects should be longer than verb phrases in utterances without a subject is a straightforward consequence of this fact.

A second analysis performed by Bloom was to look at the length of the verb phrase as a function of the type of subject (no subject, pronoun or non-pronoun). The reasoning was that, since the processing load of a subject is higher than that of a missing subject, and the processing load of a non-pronoun subject is higher than that of a pronoun (since the pronoun is both phonetically shorter as well as shorter in word length), this should again result in length effects on the Verb Phrase. The results of this analysis are shown in table 4.

Table 4: Mean length of Verb Phrase as a function of subject size

	No Subject	Pronoun	Non-Pronoun
Adam	2.60	2.55	2.25
Eve	2.75	2.30	2.00
Sarah	2.45	1.90	1.50
Anne's Model	2.76	2.45	1.60
Becky's Model	3.31	2.93	1.67

Again, it is clear that MOSAIC has no difficulty in simulating these results (though the size of the effect in MOSAIC appears to be slightly larger than in the children that Bloom analysed). The difference in verb phrase length between utterances with a pronoun subject and those with a non-pronominal subject is statistically significant for both Anne's model ($t(64) = 3.45$, $p < .001$) and Becky's model ($t(104) = 4.40$, $p < .001$). There are two possible reasons why MOSAIC might simulate this result. Firstly, non-pronoun subjects are on average slightly longer than pronoun subjects. Pronouns are by definition one word long, while non-pronoun NP's can contain determiners and adjectives. In fact, Bloom indicates that the average non-pronoun subject for the children he analysed was 1.16 words long. Secondly, pronouns have a higher frequency of occurrence than non-pronominal subjects. In MOSAIC, this increases the likelihood that they will be learnt, and the likelihood that they will feature in longer utterances. We decided to test these two explanations in MOSAIC by performing the analysis on non-pronominal subjects of length one and greater separately. As it turned out, only a small proportion of the non-pronominal subjects had a length greater than one. For non-pronominal subjects of length one, the size of the VP was 1.58 for Anne's model, and 1.88 for Becky's model[1]. Both values are smaller than the VP length for pronoun

[1] One would expect the length of the verb phrase to increase when limiting this analysis to subjects of length 1. This is the case for Becky's model, but not for Anne's model. This is due to the fact that, for Anne's model, there were relatively few long non-pronominal subjects, but one of those that did occur had a particularly long verb phrase.

subjects. Given the low incidence of long non-pronominal subject in both these and Bloom's data, this clearly indicates that the lower complexity effect that Bloom attributes to the fact that pronouns are phonetically shorter, can be explained by frequency in the input. Note that MOSAIC does not employ a phonetic component. Phonetic differences can therefore not have contributed to MOSAIC's simulation of the effect.

The importance of frequency in the input as an explanation for the difference between pronouns and non-pronouns is also highlighted by a point made by Hyams & Wexler (1993). Though pronouns may be phonetically shorter, the process of assigning the referent to a (potentially ambiguous) pronoun may actually result in its processing load being higher, rather than lower. This would predict a shorter Verb Phrase length for pronominal than for non-pronominal subjects.

Subject versus Object Omission

It has often been shown that subjects are omitted more often than objects. In order to test how often objects are omitted, Bloom selected utterances which contain verbs that require an object, and calculated the proportion of object omission from these obligatory contexts. Table 5 shows this list of verbs.

Table 5: Verbs that take obligatory objects.

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

Table 6 compares the proportion of omitted subjects and objects from obligatory contexts (verbs from tables 1 and 2 for subjects, verbs from table 5 for objects). It can be seen that the proportion of subject omission is considerably higher than the proportion of object omission. The subject-object asymmetry was significant for both Anne's model (X^2 (1, N = 560) = 98.83, p < .001), and Becky's model (X^2(1, N = 548) = 125.97, p < .001). Bloom suggests several possible causes for this asymmetry. Firstly, it may be due to pragmatic factors. Since subjects typically convey given information, while objects convey new information, it may be more pragmatically appropriate to omit subjects when processing capacity is limited. A second possible cause might be that there is a 'save the heaviest for last' bias. This would result in subjects having a higher processing load than objects, and as a result, in them being omitted more often.

Table 6: Omission from obligatory contexts

	Subjects	Objects
Adam	57%	8%
Eve	61%	7%
Sarah	43%	15%
Anne's Model	64%	21%
Becky's Model	60%	14%

The explanation for the effect in MOSAIC is simple. As a result of MOSAIC's performance limitations, a constituent is less likely to be omitted when it occurs further toward the end of the sentence. Since subjects take first position, and objects usually come after the verb, the probability of omitting an object is smaller than the probability of omitting a subject. Bloom goes on to suggest that the hypothesized processing asymmetry should cause other differences between subjects and objects. For example, since pronouns exert less of a processing load, more pronouns will occur in subject position than in object position. Table 7 shows the relevant data, both for Bloom's analysis, and MOSAIC's simulations. Again, the assymmetry is significant for Anne's model (X^2 (1, N = 243) = 8.08, p < .01), and Becky's model (X^2(1, N = 292) = 27.53, p < .001).

Table 7: Proportion of overt pronominal Noun Phrases

	Subjects	Objects
Adam	41%	25%
Eve	36%	14%
Sarah	91%	33%
Anne's Model	47%	27%
Becky's Model	72%	40%

There is no specific reason why MOSAIC would predict this effect, but the pragmatic factors that Bloom mentions may well explain this result. Subjects tend to convey given information, and objects tend to convey new information. It certainly makes sense to introduce new information using a non-pronoun NP. The use of a pronoun requires the listener to resolve the referent of the pronoun. The use of a non-pronoun NP is usually less ambiguous, which aids the resolution process. In fact, several authors have argued that this is the preferred argument structure for English (Clancy, 2001). As such, it is not just a feature of child language, but is actually the preferred structure in adult language. The fact that MOSAIC simulates this result is simply a

reflection of the fact that it mimics the distribution of the input.

Conclusions

MOSAIC clearly simulates all the results that Bloom reports. MOSAIC is not an ad hoc model of subject omission, as it has already been shown to account for several phenomena in children's speech, and is firmly grounded in the CHREST/EPAM framework. Though MOSAIC has performance limitations, these reside mainly in the learning mechanism. Unlike the standard performance limitations view, MOSAIC does not assume full competence. In fact, MOSAIC has no built in knowledge regarding grammatical categories or roles. The effects arise in MOSAIC through a combination of performance limited distributional learning, and frequency sensitivity. Effects that are present in the input (such as a higher proportion of pronominal subjects than objects), are mimicked in MOSAIC's output because of the fact that it is a distributional analyser.

On a theoretical level, MOSAIC has two main strengths over the standard view of performance limitations. Firstly the definition of processing load in the standard view is somewhat ad hoc. If the provision of certain elements leads to a higher rate of subject omission, this is seen as evidence for a relatively high processing load of these elements. The actual reason for this high processing load then varies from effect to effect. Within MOSAIC, processing load is a function of the interaction of two objectively measurable variables: frequency in the input, and length of the phrase being encoded. When an item is more frequent in the input, it has a higher likelihood of being encoded, and therefore features in longer utterances that have a higher likelihood of being grammatical (i.e. including the subject). If two utterances have equal overall frequency, and one of the two includes a longer element (verb phrase), then some other element will necessarily be omitted. Since the subject is the first element in the sentence, this has a higher likelihood of being omitted.

Secondly, the standard performance limitations view assumes a large amount of innate knowledge in the child. For the simulation of these results, MOSAIC assumes no innate syntactic knowledge.

Acknowledgments

This research was funded by the Leverhulme Trust under grant number F/114/BK.

References

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*.

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.

Freudenthal, D. Pine, J. & Gobet, F. (2002). Modelling the development of Dutch optional infinitives in MOSAIC. *This Volume*.

Huttenlocher, J. Haight, W., Bryk, A. Seltzer, M. & Lyons, T. (1991). Early vocabulary growth: relation to language input and gender. *Developmental Psychology,* 27, *236-248*.

Hyams, N. (1986). *Language Acquisition and the Theory of Parameters*, Dordrecht: Reidel.

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

MacWhinney, B. & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language*, 17, 457-472.

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*.

Pinker, S. (1984). *Language Learnability and Language Development*, Cambridge, MA: Harvard University Press.

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*.

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