Simulating Optional Infinitive Errors in Child Speech through the Omission of Sentence-Internal Elements

Daniel Freudenthal (D.Freudenthal@Liverpool.Ac.Uk)
Julian Pine (Julian.Pine@Liverpool.Ac.Uk)
School of Psychology, University of Liverpool
L69 7ZA Liverpool, UK

Fernand Gobet (Fernand.Gobet@Brunel.Ac.Uk)
School of Social Sciences and Law, Brunel University
Uxbridge, Middlesex, UB8 3PH, UK

Abstract

A new version of the MOSAIC model of syntax acquisition is presented. The modifications to the model aim to address two weaknesses in its earlier simulations of the Optional Infinitive phenomenon: an over-reliance on questions in the input as the source for Optional Infinitive errors, and the use of an utterance-final bias in learning (recency effect), without a corresponding utterance-initial bias (primacy effect). Where the old version only produced utterance-final phrases, the new version of MOSAIC learns from both the left and right edge of the utterance, and associates utterance-initial and utterance-final phrases. The new model produces both utterance-final phrases and concatenations of utterance-final and utterance-initial phrases. MOSAIC now also differentiates between phrases learned from declarative and interrogative input. It will be shown that the new version is capable of simulating the Optional Infinitive phenomenon in English and Dutch without relying on interrogative input. Unlike the previous version of MOSAIC, the new version is also capable of simulating cross-linguistic variation in the occurrence of Optional Infinitive errors in Wh-questions.

The Characteristics of Early Child Speech

Early child speech is often telegraphic, and (in many languages) lacks inflections that are required in the adult grammar. For example, English-speaking children produce utterances such as *Play car* and *He go* and Dutch-speaking children produce utterances such as *Pappa eten* (Daddy eat) and *Trein spelen* (Train play). As children grow older, the length of their utterances increases, their speech becomes less telegraphic, and they provide the appropriate inflections more frequently. However, there is a period in which children use verbs in both their correct (inflected) and incorrect (uninflected) forms in contexts in which inflected forms are required. The apparent lack of inflection in child speech has been the subject of considerable linguistic and Nativist theorizing. Wexler (1994) proposes the Optional Infinitive hypothesis, which states that young children know the full grammar of their language but optionally use nonfinite forms where the adult grammar requires a finite form[1]. Wexler's hypothesis explains the data from a variety of languages. However, there are two main weaknesses associated with the account. First, it fails to provide any quantitative predictions regarding the rate at which children will use nonfinite forms in finite contexts, and, second, it ignores the possibility that children's early language use may reflect the operation of an input-driven learning mechanism as opposed to rich innate linguistic knowledge.

Simulating Child Language in MOSAIC

MOSAIC is an attempt to investigate the extent to which children's early language use can be explained by an input-driven learning mechanism. MOSAIC learns from Child-Directed speech and produces output that can be directly compared to children's speech. MOSAIC has already been used to simulate the basic Optional Infinitive phenomenon in English and Dutch (Freudenthal, Pine & Gobet, 2002a, submitted), as well as phenomena related to Subject Omission (Freudenthal, Pine & Gobet, 2002b) and the Modal Reference effect (Freudenthal, Pine & Gobet, 2004). MOSAIC is a simple discrimination network that incrementally learns and stores utterances that are presented to it. An important restriction on MOSAIC's learning mechanism is that it builds up its representation of an utterance by starting at the end of the utterance and slowly working its way to the beginning. MOSAIC is therefore capable of producing an utterance such as *He go* by producing the final phrase of *Does he go?* Similarly, it can produce the Dutch utterance *Trein spelen* by producing the ending of the phrase *Ik wil met de trein spelen* (*I want with the train play*). As MOSAIC sees more and more input, it learns to produce progressively longer utterances. As utterances become longer, they are more likely to contain finite verb forms. (Both in Dutch and English finite verb forms tend to occur near the beginning of the utterance. A model that produces utterance-final phrases will therefore

[1] Data from languages like Dutch, which has an infinitival morpheme, suggest that, rather than dropping inflections, children are using non-finite verb forms in finite contexts.

produce more utterances containing finite verb forms as the length of these utterances increases.)

While MOSAIC successfully simulates the quantitative patterning of the Optional Infinitive phenomenon in English and Dutch, its reliance on learning from the end of the utterance also gives rise to certain weaknesses. First, the model is overly reliant on questions in the input as the source of Optional Infinitive errors with (third singular) subjects. While some Optional Infinitive errors with third singular subjects (e.g. *Daddy eat* or *Pappa eten*) can be learned as sequences from (relatively infrequent) declarative double verb constructions (e.g. *I see Daddy eat/Ik zie Pappa eten*), others (e.g. *He eat* and *Hij eten)* never occur as sequences in declarative utterances. MOSAIC simulates such errors by learning them from questions such as *Does he eat?* or *Gaat hij eten?* (*Goes he eat?*). However, given the obvious differences in the intonation contours of declaratives and questions, learning declaratives from questions might be regarded as somewhat implausible, especially if MOSAIC is seen as implementing a constructivist model of language development in which children's early knowledge consists of a repertoire of unanalyzed wholes and lexically specific constructions learned directly from the input (e.g. Pine, Lieven & Rowland, 1998; Tomasello, 2000, 2003). Developing a way of learning Optional Infinitives from declarative contexts would therefore not only increase the plausibility of the model, but also bring it more in line with general constructivist theorizing.

One way in which MOSAIC could learn Optional Infinitives from declaratives is through the omission of sentence-internal elements. An utterance such as *He go*, for example, could be produced by omitting the modal *can* from *He can go*, or omitting *wants to*, from *He wants to go*. In Dutch, *Hij eten* (*He eat*) could be learned from *Hij wil eten* (*He wants (to) eat*). The omission of sentence-internal elements may also enable MOSAIC to simulate children's Optional Infinitive errors in Wh-questions. English-speaking children often produce utterances such as *What he do?* or *Where he going?* At present, MOSAIC is unable to produce such utterances as it is not capable of omitting the sentence-internal *is* or *does*. Developing a way of simulating such errors would therefore also be a step forward. Moreover, the occurrence of Optional Infinitive errors in Wh-questions is an interesting domain for simulation in itself, as English and Dutch speaking children appear to produce such errors at rather different rates.

At a more general level, the strict utterance-final bias in MOSAIC is not very plausible in terms of general learning theory. There is a wealth of evidence that human subjects display a primacy as well as a recency effect. The addition of an utterance-initial bias (a requirement for implementing sentence-internal omission) to MOSAIC may therefore resolve the weaknesses associated with the reliance on questions and the omission of sentence-initial phrases, as well as bring the model more in line with general psychological theorizing. This paper describes a new version of MOSAIC that aims to accomplish this by learning from both edges of the utterance and associating sentence-initial and sentence-final fragments.

The remainder of this paper is organized as follows: First, the new version of MOSAIC and its mechanism for associating utterance-initial and utterance-final phrases is described. Next, two new simulations on a Dutch and an English child are compared to simulations with the earlier version of MOSAIC in terms of the fit to the Optional Infinitive phenomenon. It will be shown that the new version still simulates the basic Optional Infinitive phenomenon. Importantly, however, the new analyses are performed on output learned from declarative phrases. Next, a more detailed analysis is performed on MOSAIC's ability to simulate Optional Infinitive errors in Wh-questions.

MOSAIC

MOSAIC consists of a simple network of nodes that encode words and phrases that have been presented to the model. As the model sees more input it will incrementally encode more and longer phrases and will consequently be able to generate more and longer output. Figure 1 shows a sample MOSAIC network. Learning in MOSAIC is anchored at the sentence-initial and sentence-final positions: MOSAIC will only encode a new word or phrase when all the material that either follows or precedes it in the utterance has already been encoded in the network. When presented with the utterance *He wants to go to the shops* for instance, the model may in the first instance encode the words *He* and *shops*. At a later stage it may encode the phrases *He wants* and *the shops*, until the point where it has encoded the entire phrase *He wants to go to the shops*. When the model processes an utterance, and a sentence-final and sentence-initial phrase for that utterance have already been encoded in the network, MOSAIC associates the two nodes encoding these phrases, to indicate the two phrases have co-occurred in one (longer) utterance. In Figure 1, the model has associated the phrases *He wants* and *Go home*.

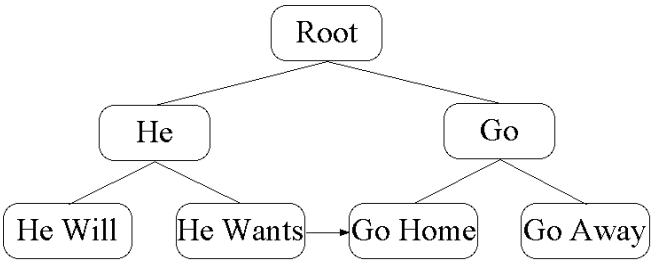


Figure 1: A partial MOSAIC model. The sentence-initial phrase *he wants*, and the sentence-final phrase *go home* have been associated, allowing the model to produce the utterance *He wants go home*.

Learning in MOSAIC takes place by adding nodes that encode new words and phrases to the model. Learning is relatively slow. The formula governing the probability of creating of a node in MOSAIC is as follows:

$$NCP = \left(\frac{1}{1 + e^{0.5((m*c)-u)}} \right)^{d}$$

where: ncp = node creation probability
m = a constant, set to 20 for these simulations.
c = corpus size (number of utterances)
u = (total number of) utterances seen
d = distance to the edge of the utterance

The formula results in a basic sigmoid function, with the probability of creating a node increasing as a function of the number of times the input has been presented. The input corpus (which consists of realistic child-directed speech) is fed through the model iteratively, and output can be generated after every presentation of the input corpus. Making the node creation probability dependent on the number of times the corpus has been seen allows for comparison across corpora of differing sizes. The distance to the edge (or length of the utterance being encoded) features in the exponent in the formula, and lowers the likelihood of encoding long utterances. As a result, MOSAIC will initially only learn sentence-initial and sentence–final words. Only when the base probability in the formula starts to increase (as a result of seeing more input), will longer phrases start being encoded. Due to node-creation being probabilistic, a word or phrase must normally be seen several times before it will be encoded. Frequent words or phrases therefore have a higher probability of being encoded than infrequent words or phrases.

MOSAIC maintains an utterance-final bias in that learning from the right edge of the utterance is faster than learning from the left edge. This is accomplished by adding 2 to the length of a left edge phrase[2] (the parameter d) that is considered for encoding (The parameter d designates distance from the left edge of the utterance for left edge learning, and distance to the right edge of the utterance for right edge learning). This learning mechanism results in a model that is biased towards learning sentence-initial words and a few (high-frequency) sentence initial phrases coupled with comparatively long utterance-final phrases. As a result, the sentence-internal elements that MOSAIC omits will tend to be located near the left edge of the utterance.

Generating output from MOSAIC

MOSAIC has two mechanisms for producing (rote) output. The first mechanism is (almost[3]) identical to that in earlier versions of MOSAIC. In generation, the model traverses the branches of the network, and generates the contents of branches that encode sentence-final phrases. (Sentence-initial fragments are not generated as these may end in the middle of the sentence, and often do not resemble child speech).

The second mechanism which is new to this version of MOSAIC is the concatenation of sentence-initial and sentence-final phrases. When MOSAIC builds up the network, it associates the sentence-initial and sentence-final fragments from each utterance (cf *He wants go home* in Figure 1). Since the concatenation of phrases could result in many implausible utterances, not all possible concatenations are produced. A source utterance like *Give the man a hand*, for example, could potentially give rise to the concatenated phrase *Give the a hand*. This utterance is awkward (and not typical of child speech) because it breaks up the unit *the man*. MOSAIC prevents such concatenations by only concatenating phrases that are anchored: a sentence-initial phrase can only be used for concatenation if the last word in that phrase has occurred in a sentence-final position. Likewise, a sentence-final phrase can only be concatenated if the first word in that phrase has occurred in sentence-initial position. Since the word *the* will not occur in sentence-final position, the phrase *Give the a hand* will not be generated. The rationale behind this restriction is that, to the extent that children concatenate phrases/omit sentence-internal elements, they will rarely break up syntactic units. Restricting concatenation to phrases where the internal edges are anchored effectively achieves this, as an anchored word is unlikely to be a partial unit.

The rote output of MOSAIC thus consists of a mixture of sentence-final phrases and concatenations of sentence-initial and sentence-final phrases. Both types of utterances are apparent in child language. An example of a phenomenon that might be explained through omission of sentence-initial elements is the omission of subjects from the sentence-initial position (Bloom, 1990). Due to MOSAIC's learning mechanism and faster right-edge learning, MOSAIC's output will initially contain a large proportion of sentence-final fragments. As the Mean Length of Utterance (MLU) of the model increases, concatenations will become more frequent. The concatenations themselves will be slowly replaced by complete utterances.

The two mechanisms described so far produce output that directly reflects the utterances present in the input (with the potential omission of sentence-initial or sentence-internal material). These two mechanisms are complemented by a third mechanism which is responsible for the generation of novel utterances through the substitution of distributionally similar words. When two words tend to be followed and preceded by the same words in the input, they are considered equivalent, and can be substituted for each other. Thus, the model is capable (in principle) of producing the utterance *She run* by omitting *will* from *He will go*, and substituting *She* for *He*, and *run* for *go*. A more in-depth discussion of MOSAIC's mechanism for substituting distributionally similar items is given in Freudenthal, Pine and Gobet (2005a), though the chunking mechanism described in that paper has not yet been implemented in the present version of the model.

[2] The utterance-final bias applies to phrases, but not words. Sentence-initial and sentence-final words are equally likely to be encoded.

[3] In line with the restriction discussed under concatenation, only utterance-final phrases that start with a word that has occurred in utterance-initial position are produced.

The Simulations

The main aim of the simulations was to replicate the simulations of the Optional Infinitive phenomenon as reported in Freudenthal, Pine, and Gobet (submitted). In these simulations, a good fit to the data was achieved, but these simulations relied too strongly on interrogative input. For the present simulations, questions and declaratives were marked separately in the input (using the punctuation present in the raw input files). Every word in the interrogative input utterances was marked for being part of a question (creating a separate entry in the model for the occurrence of a word in a declarative and an interrogative context). This made it possible to filter out utterances learned from interrogative input and only generate output that was learned from declarative input.

Figure 2a: Data for Matthijs.

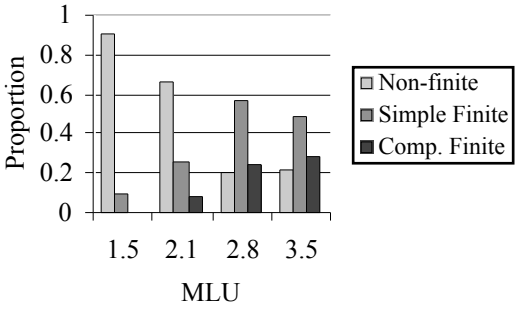


Figure 2b: Data for Anne

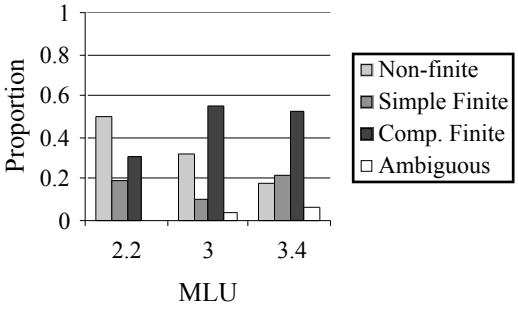


Figure 2: Development of finiteness marking for a Dutch and English child.

The simulations were run using the child-directed speech for one Dutch, and one English child, both taken from the CHILDES database (MacWhinney, 2000). The Dutch Child (Matthijs) was part of the Groningen Corpus (Bol, 1995), the English child (Anne) was taken from the Manchester corpus (Theakston, Lieven, Rowland & Pine 2001). The size of the input was approximately 14,000 utterances for Matthijs, and 35,000 utterances for Anne. Additional simulations for one Dutch and English child, as well as a German and Spanish child can be found in Freudenthal et al. (2005b). As detailed in Freudenthal, Pine & Gobet (submitted), Dutch children show considerable developmental variation in their use of Optional Infinitives. Early in development, nearly all their utterances with verbs contain non-finite verb forms. By the time they approach an MLU of 4, this has decreased to roughly 20% (see Fig. 2a). For English, the data are less clear. Since English uses an impoverished inflectional system, it is necessary to restrict the analysis to utterances with a third singular subject. Doing so suggests a rate of Optional Infinitive errors around 50% at MLU 2, which rapidly declines as the MLU increases (see Fig. 2b).

For all analyses the following classification of utterances was used. Utterances that only contained non-finite verb forms were classed as non-finite. Utterances that only contained finite verb forms were classed as simple finites. Utterances containing both finite and non-finite verb forms were classed as compound finites. Utterances with the copula as the main verb were excluded from the analysis. The same classification scheme was used for English and Dutch, with the exception that the analysis on English was restricted to utterances containing a third singular subject, and that English verb forms which could either be finite or non-finite (e.g. *bought*), were classed as ambiguous.

Figure 3a: Old simulations for Matthijs.

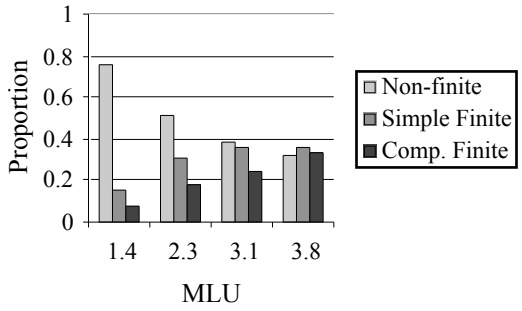
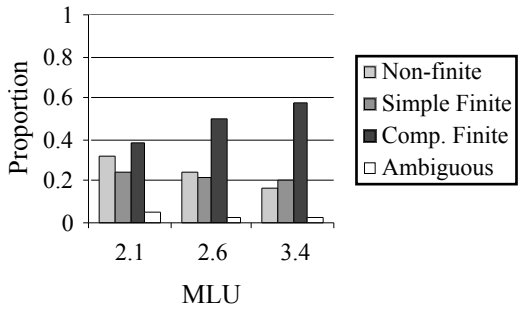


Figure 3b: Old simulations for Anne



Results for the old simulations are shown in Figures 3a and 3b. The main reason why the model simulates the developmental pattern apparent in the children is because the model generates utterance-final phrases of increasing length. As was mentioned, inflected verb forms tend to

occur near the beginning of the utterance, while uninflected verb forms tend to occur nearer the end of the utterance (especially for Dutch where non-finite verb forms are placed in utterance-final position). A model that produces utterance final phrases of increasing length will therefore show a decreasing proportion of utterances containing only non-finite verbs with increasing MLU.

Figures 4a and b show the results for the new simulations. In these simulations, MOSAIC could only produce utterances from nodes learned from declarative contexts. Output was made up of concatenations as well as utterance-final phrases. The main thing to note about Figure 4 is that the concatenation mechanism which results in the omission of sentence-internal elements is capable of producing Optional Infinitive errors at rates that are sufficiently high to match the child, even when Optional Infinitive errors are restricted to third singular contexts (for English). As such, sentence-internal omission appears to be a successful mechanism for the production of Optional Infinitive errors.

Fig 4a: New simulations for Matthijs

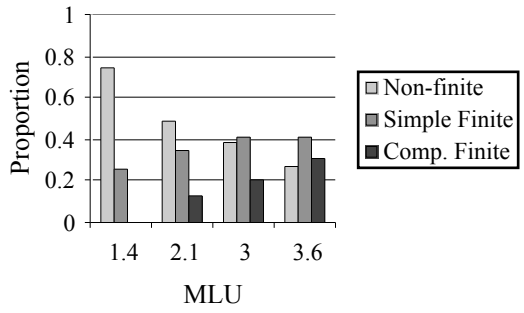
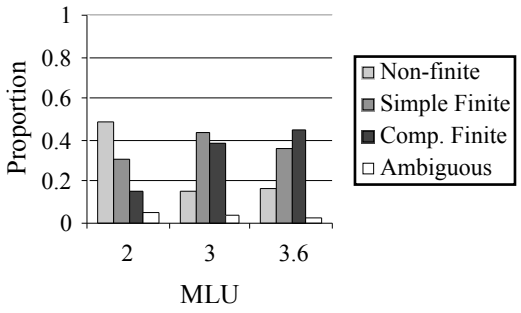


Figure 4b: New simulations for Anne.



For Dutch, the results are comparable to those for the earlier simulations, though a slightly better fit is achieved for the final stage. For English, the proportion of non-finites early on is increased relative to the earlier simulations, resulting in a slightly better fit. One thing that stands out in the English simulations is that, while the model produces Optional Infinitives at rates that match the child reasonably well, the model produces more simple finites than compound finites at the first two data points. This is a weakness that we hope to address through further refinement of the concatenation mechanism.

Optional Infinitives in Wh-questions

Having established that MOSAIC is capable of simulating the basic Optional Infinitive phenomenon in declaratives, we can now turn to Optional Infinitives in Wh-questions. English-speaking children sometimes omit inflection in Wh-questions, resulting in utterances such as *Why he go?* In Dutch, and other V2 languages, Optional Infinitives appear to be quite rare in Wh-questions (Wexler, 1998). In order to establish the rates of Optional Infinitive errors in the two children, the corpora of Anne and Matthijs were searched for the occurrence of Wh-words in interrogative contexts. In order to unambiguously identify root infinitives in these utterances, only utterances with a subject and main verb were included. Anne's corpus contained 111 such questions, of which 41 (37%) were non-finite. The corpus of Matthijs contained relatively few Wh-questions (11), but none of these were non-finite, confirming Wexler's (1998) observation. In order to establish whether MOSAIC simulates this pattern of results, a sample of Wh-questions was generated from the model (at MLU 3.0). Analysing these utterances in the same way as the children's utterances yielded 20% non-finite Wh-questions for Anne's model compared to 8% non-finite Wh-questions for Matthijs's model (see Table 1). While not as pronounced as the difference between the children, this difference was statistically significant $X^2 = 8.88$, $p < .01$

Table 1: Finite and Non-finite Wh-questions for Anne and Matthijs's simulations

	Non-Finite	Finite
Anne	30	123
Matthijs	15	164

MOSAIC simulates this distinction between English and Dutch because, despite omitting sentence-internal elements, MOSAIC's output still adheres to the basic word order for the language it is learning. In English, Wh-questions include a non-finite main verb preceded by a finite auxiliary (e.g. *Where does he go?)*. English Wh-questions will therefore always contain a non-finite verb form and omission of the auxiliary will result in an Optional Infinitive error. Dutch on the other hand, allows for finite Wh-questions, such as *Wat eet hij?* (*What eats he?*). While modal plus non-finite constructions are possible in Dutch Wh-questions (*Wat wil hij eten?/What wants he (to) eat?)*, Wh-questions in Dutch are less likely to contain a non-finite verb. Thus, Optional Infinitives are less likely to occur in the Dutch simulations since frames that can give rise to Optional Infinitives through omission of sentence-internal elements make up a smaller proportion of the Wh-questions in the input.

Conclusions

This paper set out to address some weaknesses in earlier versions of MOSAIC: an over-reliance on questions as the source for Optional Infinitives, and the lack of an utterance-initial bias in learning. A new mechanism was proposed which allows MOSAIC to concatenate the beginnings and ends of sentences, resulting in the omission of sentence-internal elements. The simulations presented in this paper show that the model is still capable of simulating the Optional Infinitive phenomenon without relying on questions as the source for Optional Infinitive errors. In the present version, declaratives are the source of Optional Infinitive errors. Declaratives in the input therefore appear to include a sufficiently high number of frames that can give rise to Optional Infinitive errors to offset the loss of Optional Infinitives learned from questions.

The omission of sentence-internal elements, coupled with a distinction between questions and declaratives, has also made it possible to simulate Optional Infinitive errors in Wh-questions. Thus, MOSAIC now not only produces Optional Infinitive errors in Wh-questions, but also simulates the difference in the rate of Optional Infinitive errors in Wh-questions in English and Dutch. This suggests that differences in the way that questions are formed in English and Dutch may be the cause of the differential rates of Optional Infinitives in Wh-questions in the two languages. Contrary to Wexler's (1998) claims, the present simulations show that differential rates of Optional Infinitive errors may arise from a simple distributional analysis of the input, and therefore do not provide evidence for rich innate linguistic knowledge on the part of the child.

One possible weakness of the present model is that some of the fine detail of the simulations (the ratio of simple to compound finites in the English declarative simulations) does not match the child as well as it might. Further experimentation with the implementation of the concatenation mechanism may improve this more detailed fit. However, the finding that the omission of sentence-internal elements from declaratives can still result in high rates of Optional Infinitives is encouraging as it brings MOSAIC's mechanism for the production of Optional Infinitive errors more in line with general constructivist theorizing. Likewise, the primacy effect that is implemented with the left-edge learning resolves an inconsistency with a large body of general learning research, thus making MOSAIC more credible as a general learning mechanism.

Acknowledgements

This research was funded by the ESRC under grant number RES000230211

References

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

Bol, G.W. (1995). Implicational scaling in child language acquisition: the order of production of Dutch verb constructions. In M. Verrips & F. Wijnen, (Eds.), *Papers from the Dutch-German Colloquium on Language Acquisition*, Amsterdam Series in Child Language Development, 3, Amsterdam: Institute for General Linguistics.

Freudenthal, D., Pine, J.M. & Gobet, F. (submitted). Modelling the development of Children's use of Optional Infinitive in Dutch and English using MOSAIC. Submitted to *Cognitive Science*.

Freudenthal, D., Pine, J.M. & Gobet, F. (2005a). 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. (2005b). Simulating the cross-linguistic development of Optional Infinitive Errors in MOSAIC. *This Volume*.

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: Erlbaum.

Freudenthal, D., Pine, J.M. & 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* (pp. *322-327*). Mahwah, NJ: LEA.

Freudenthal, D., Pine, J.M. & 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*, (pp. *328-333*). Mahwah, NJ: LEA.

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

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.

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

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

Wexler, K. (1994). Optional infinitives, head movement and the economy of derivation in child grammar. In N. Hornstein & D. Lightfoot (Eds.), *Verb Movement*. Cambridge: Cambridge University Press.

Wexler, K. (1998). Very early parameter setting and the unique checking constraint: a new explanation of the optional infinitive stage. *Lingua*, *106*, *23-79.*