Meter based omission of function words in MOSAIC

Daniel Freudenthal (D.Freudenthal@liv.ac.uk),
Julian Pine (Julian.Pine@liv.ac.uk)
School of Psychology, University of Liverpool

Fernand Gobet (Fernand.Gobet@Brunel.ac.uk)
School of Social Sciences, Brunel University

Abstract

MOSAIC (Model of Syntax Acquisition in Children) is augmented with a new mechanism that allows for the omission of unstressed function words based on the prosodic structure of the utterance in which they occur. The mechanism allows MOSAIC to omit elements from multiple locations in a target utterance, which it was previously unable to do. It is shown that, although the new mechanism results in Optional Infinitive errors when run on children's input, it is insufficient to simulate the high rate OI errors in children's speech unless combined with MOSAIC's edge-first learning mechanism. It is also shown that the addition of the new mechanism does not adversely affect MOSAIC's fit to the Optional Infinitive phenomenon. The mechanism does, however, make MOSAIC's output more child-like, both in terms of the range of utterances it can simulate, and the level and type of determiner omission that the model displays.

Keywords: MOSAIC, Syntax Acquisition, Optional Infinitives, Determiner Omission.

Introduction

Child speech differs from adult speech in a number of ways. Apart from the average child utterance being noticeably shorter than adult utterances, child speech often lacks inflection where this is appropriate in the adult language. It is also rather telegraphic (i.e., is marked by the relative absence of function words). These characteristics are illustrated in utterances (1) and (2), which are plausible child utterances.

(1) He go home.
(2) I want cookie.

The lack of inflection in child speech has been the subject of considerable Nativist theorizing in recent years. Early theories suggested that utterances like (1) reflect the omission of an inflectional morpheme (-s) from a finite form. More recent theories (Wexler, 1998), however, claim, on the basis of cross-linguistic data, that such utterances actually reflect the use of a non-finite form (the infinitive) in place of a finite form (in this case the 3rd singular present tense). Following Wexler's work utterances like (1) have become known as *Optional Infinitive* (OI) errors. Wexler proposes that children have, from a very early age, correctly set all the inflectional and phrase structure parameters for their language, but are subject to a 'Unique Checking Constraint' which results in them optionally producing infinitives in contexts where a finite form is required. As a result of maturation, children will provide the correct, inflected form increasingly often as they get older, leading to a decrease in OI errors.

Alternative accounts claim that OI errors can be understood in terms of input-driven learning mechanisms without the need to assume innate knowledge. In particular, it is claimed that OI errors can be explained as compound (auxiliary/modal + infinitive) constructions with a missing modal or auxiliary (Ingram & Thompson, 1996). Thus, an utterance such as *he go home* might result from omitting the modal *will* from *he will go home*. According to these accounts, OI errors disappear as children's utterances become longer and missing modals and auxiliaries are realized more and more often.

MOSAIC is a computational model that implements the view of OI errors as truncated compound constructions. Freudenthal et al. (2006, 2007) have shown that the rates at which children produce OI errors can be closely simulated through an input-driven learning mechanism that produces partial utterances. Freudenthal et al. were able to show that MOSAIC provides a close quantitative fit to the OI data from four different languages: English, Dutch, German and Spanish. They were also able to trace the differential rates with which children produce OI errors in these languages back to characteristics of the input from these languages: the frequency of compound constructions and the position of the infinitive in compound constructions.

The particular mechanism used by Freudenthal et al., however, suffers from some weaknesses as the model only produces utterance-final phrases. That is, the model learns the input it receives by building up its representation from the right edge of the utterance. The OI errors with third singular subjects that this model produces are largely learned from questions (e.g. (*Can*) *he go?*). Since children produce OI errors as declaratives, it seems somewhat implausible they should learn such constructions from interrogative contexts. A further problem with the simulations reported by Freudenthal et al. is that child language is far more telegraphic than MOSAIC's output. That is, children will often produce utterances with many omitted constituents (e.g., *Play train*). Since such constructions do not occur (as utterance-final phrases) in the input, they cannot be produced by MOSAIC.

Freudenthal et al. (2005a) report preliminary simulations with a version of MOSAIC that alleviates this problem. This version learns from the left as well as right edge of an utterance and associates sentence-initial and sentence-final phrases. Given an utterance like *He wants to go to bed* the

model is capable of associating the phrase *go to bed* with the sentence-initial word *he* resulting in the OI error *he go to bed*. This version of MOSAIC, however, is still unable to produce certain structures that children frequently produce. In particular, children often appear to omit material from multiple locations in an utterance. Thus, an utterance like (3) appears to involve the omission of both a modal or auxiliary and an article from an utterance like *he can go to the shops*.

(3) He go to shops

Since MOSAIC is capable of omitting only one sentence-internal phrase from an utterance it cannot produce an utterance like (3). Modifying MOSAIC so that it is able to produce such utterances will therefore greatly increase its credibility as a model of children's early multi-word speech.

Omission Errors in Child Speech

It has long been recognised that omission errors are an important characteristic of child speech (Brown, 1973). Moreover, it is clear that children can make multiple errors of omission within the same sentence. Such errors have often been interpreted as resulting from performance limitations in production (Bloom, 1990; Valian, 1991). According to this view, the child is thought to have full competence (a correct underlying representation), but some elements of this representation fail to surface due to a processing bottleneck in production. In the words of Bloom, 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…' (Bloom, 1990, p. 492).

An elegant performance limitations account of the pattern of omission errors in child speech is provided by Gerken (1991, 1996). Gerken's account focuses on the prosodic structure of the target utterance, in particular the occurrence of an element with respect to metrical feet. The metrical foot is a basic prosodic unit, which is described by the nature and number of syllables it contains. Gerken's account focuses on the position of unstressed syllables relative to trochaic feet (which have a strong-weak stress pattern). The majority of English (di-syllabic) words are trochaic in nature: primary stress is placed on the first syllable (e.g. PAper, TAble). Gerken argues that children have a preference for trochaic meter to the extent that unstressed syllables that are not part of a trochaic foot are more likely to be omitted. Thus, children are likely to omit the first (unstressed) syllable from *banana*, resulting in *nana*. The omission of unstressed (or weak) syllables that are not part of a trochaic foot also goes some way towards explaining the omission of elements from sentential contexts. Thus, Gerken (1996) shows that children are more likely to omit the object article *the* from sentence (4b) (where it is unfooted), than from sentence (4a) where it is part of a trochaic foot (An asterisk denotes an unfooted element, S and W stand for Strong and Weak. Dashes connect items that combine to form a foot).

(4) a. he KICKS the PIG
* S-----w S(-w)

(4) b. he CATCHes the PIG
* S----w * S(-w)

Gerken (1991) also explains the finding that children are more likely to omit pronominal subjects than objects in terms of the stress pattern. Unstressed sentence-initial subjects are likely to be omitted as they are unfooted. Sentence-internal objects are less likely to be omitted as they can be part of a trochaic foot. Further support for Gerken's account comes from recent work by Demuth et al. (in press), who show that children are less likely to omit determiners from footed than from unfooted contexts.

Gerken's account is appealing, as it provides a unified explanation of function word omission in child speech that is largely independent of grammatical class. A mechanism based on this account could therefore be readily combined with MOSAIC's input-driven learning mechanism (which does not assume knowledge of grammatical categories) to simulate the pattern of sentence-internal omission in children's speech. However, since within Gerken's account, modals, like other function words, can be unfooted and therefore omitted from modal + verb structures, it is also possible that a prosody-based omission mechanism may itself be sufficient to explain the OI phenomenon.

The aim of this paper is therefore to investigate the utility of Gerken's metrical template account as a means of increasing the levels of omission that MOSAIC displays, while at the same time considering the possibility that a prosody-based omission mechanism might be sufficient to simulate the level of OI errors in children's output. To this end, prosodic structure was assigned to MOSAIC's output, and unstressed items were probabilistically deleted from the output based on their location relative to trochaic feet. As suggested by Gerken, this mechanism was implemented as a limitation in production[1]. Thus, MOSAIC's learning mechanism (association of sentence-initial and sentence-final phrases) remained unaltered and omission of unstressed elements only occurred in production. Output generated in this way was then compared with output generated by applying the prosody-based omission mechanism to the input samples on which the model was trained. This allowed us to establish whether MOSAIC's learning mechanism was necessary to simulate the rate of OI errors in children's speech.

The Simulations

The simulations were conducted using the version of MOSAIC described in Freudenthal et al. (2005a) augmented with the chunking mechanism described in Freudenthal et

[1] Clearly this does not address the question of how this bias is acquired. However, given that, at present, MOSAIC's learning mechanism operates at the level of the word rather than the syllable, this question is currently beyond the scope of the model.

al. (2005b). MOSAIC learns from realistic input (child-directed speech) and combines a strong utterance-final bias (recency effect) with a smaller primacy effect. MOSAIC's basic learning mechanism slowly builds up a representation of the utterances it is shown by starting at the right edge of the utterance and slowly working its way to the beginning of the utterance. This mechanism is complemented by a (slower) learning mechanism that builds up its representation of the utterance by starting at the left edge of the utterance, and slowly working its way to the end of the utterance. MOSAIC associates these utterance-final and utterance-initial phrases and is therefore capable of producing utterances with missing sentence-internal phrases. This is illustrated in Fig. 1. Since the utterance-final phrases MOSAIC learns tend to be longer than the utterance-initial phrases (as utterance-final learning is faster) the omitted phrases tend to be located near the left edge of the utterance.

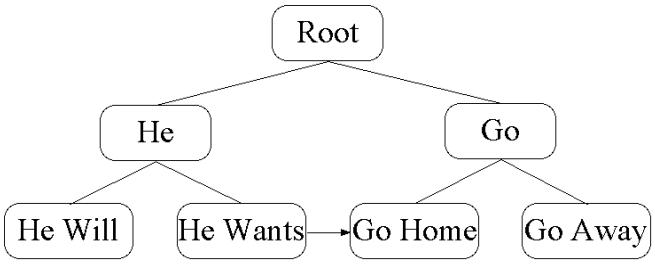

Figure 1: A partial MOSAIC network. The sentence-initial phrase *he wants*, and the sentence-final phrase *go home* have been associated, allowing the model to produce the utterance *He wants go home*.

Output is generated from MOSAIC by traversing all the branches in the model and outputting the (utterance-final) phrases they encode. Where these phrases have been associated with utterance-initial phrases a concatenation of these phrases is also produced. MOSAIC can produce output with an increasing Mean Length of Utterance (MLU), thereby simulating developmental change. Learning in MOSAIC is relatively slow, and the input is shown to the model several times. With every exposure to the input MOSAIC represents more and longer phrases that were present in the input. Output is generated from the model after each exposure to the input, which results in output files of increasing MLU. For the present simulations, models were run using the maternal speech directed at two English children (Anne and Becky). The child-directed speech for these children consists of ~33,000 and ~25,000 utterances. Where relevant the output from the model was compared to the actual speech produced by Anne and Becky.

Determining the Stress Pattern

The input that MOSAIC learns from is transcribed in an orthographic format that does not include any prosodic information. Likewise, the output from MOSAIC consists of simple text files that lack prosodic information. In order to probabilistically delete unstressed elements the stress pattern for an output utterance thus needs to be assigned. This was done on a word-by-word basis using the stress pattern detailed in the dictionary entry for the individual words. The Unilex dictionary (Fitt & Isard, 1999) was used for this purpose. The Unilex dictionary contains some 100,000 lemmas and details their phonetic form, syllabification and stress pattern. For all utterances in MOSAIC's output, the stress pattern was determined by concatenating the stress patterns for the individual words[2]. Mono-syllabic function words (articles, determiners, pronouns etc.) as well as modals and auxiliary verbs (including the copula) were assigned weak (no) stress. All content words were considered stressed. After the stress pattern had been determined it was decided which unstressed elements were not part of a trochaic foot (i.e. were unfooted). This was done in the following manner:

1. All elements preceding the first stressed syllable in an utterance were deemed unfooted.
2. Every stressed syllable was considered the start of a new foot.
3. An unstressed element that was preceded by a stressed syllable was considered part of a trochaic foot.
4. An unstressed element that was preceded by an unstressed syllable was deemed unfooted.

This procedure results in utterances (4a) and (4b) being assigned the indicated stress pattern. In both (4a) and (4b) the subject *he* is unfooted as it precedes the first stressed syllable. The object article *the* in utterance 4a is part of a trochaic foot as it is preceded by the stressed syllable *kicks*. The article in 4b is unfooted as it is preceded by the unstressed syllable *–es*. A further example is given in (5).

(5) a. he can GO to the SHOPS
* * S—w * S(-w)

(5) b. PETE can GO to the SHOPS
S-------w S—w * S(-w)

Once the stress pattern for an utterance was determined unstressed (mono-syllabic) words were probabilistically deleted from the utterance. The asymmetry in the omission of footed and unfooted items was modelled by setting the probability of deleting an unstressed item to different values for footed and unfooted items.

Results

[2] In instances where a word had no entry in the dictionary, no stress pattern was applied to the utterance, and no omission from this utterance was possible. Such utterances were maintained in the analyses presented as their omission affected the MLU distribution, which precluded MLU matching across simulations.

As was mentioned earlier, the omission of unstressed elements can lead to modal omission, and thus result in OI errors. This raises the possibility that the omission mechanism itself may be sufficient to explain the OI phenomenon. This possibility was investigated by running the omission mechanism on the input files (maternal speech) for Anne and Becky, and comparing the rates of OI errors as well as simple and compound finites in the resulting output with the child data at around MLU 2.1. The results of this analysis were compared to those obtained from MOSAIC models with and without omission. This allowed us to compare the performance of the omission mechanism with the learning mechanism of MOSAIC. Comparing the performance of MOSAIC with and without omission allowed us to establish if the omission mechanism had any effects (positive or negative) on MOSAIC's output.

Running the omission mechanism on the maternal speech resulted in utterances that were considerably longer than the child speech they were compared against. For this reason, the rates of OI errors were also determined for the subset of utterances that were not longer than three words. This resulted in output files with an MLU of approximately 2.1. The results of the analyses on short and long utterances are presented in Fig. 2. These results were obtained by setting the omission probability to 0.5 for unfooted items, and 0.1 for footed items. The omission mechanism was also run with probabilities of 0.8 and 0.2 respectively. This gave very similar results.

As can be seen in Fig. 2, the omission mechanism did result in the production of OI errors, in particular when the analysis was restricted to short utterances (0.19 for Anne's input and 0.12 for Becky's input). These proportions are higher than those that occur in the maternal speech directed at these children (~ 5%). However, they are considerably lower than the rates of OI errors that the English children display early in development.

These results suggest that omission of weak elements can account for some of children's OI errors, particularly when combined with an additional mechanism that restricts the length of the utterances children produce. The mechanism implemented for these analyses (only selecting short utterances) however, is not sufficient to produce OI errors at rates comparable to the rates that actual children produce.

Fig. 2a. Data and input analysis for Anne.

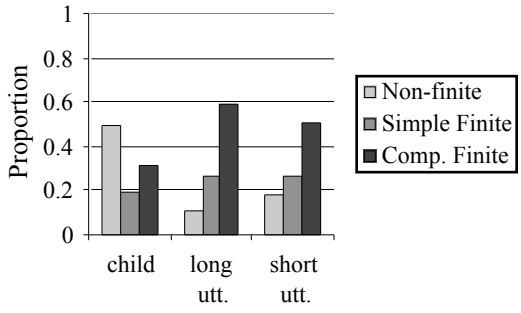

Fig 2b. Data and input analysis for Becky.

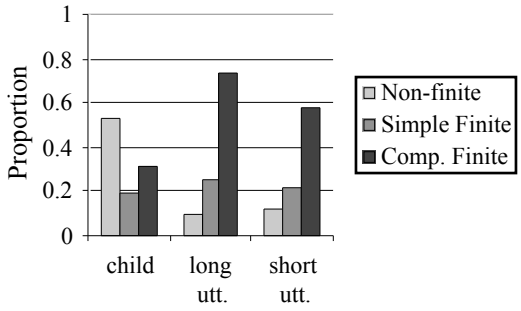

Fig. 2: Rates of OI errors, simple and compound finites for children and maternal speech with omission.

The next set of analyses was aimed at establishing if MOSAIC's mechanism for restricting the length of utterances (omission of sentence-internal material through the concatenation of utterance-initial and utterance-final phrases) is more successful in producing OI errors at rates comparable to English children. For these simulations standard MOSAIC models were run and output at an MLU of 2.1 was generated.

Fig. 3a. Data and simulations for Anne.

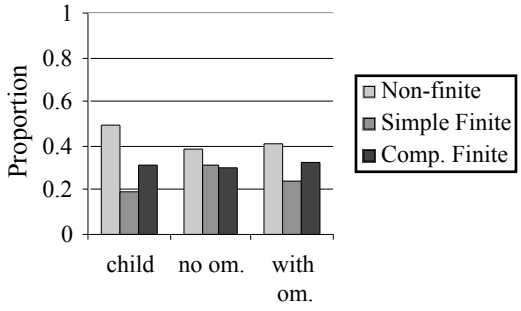

Fig 3b. Data and Simulations for Becky.

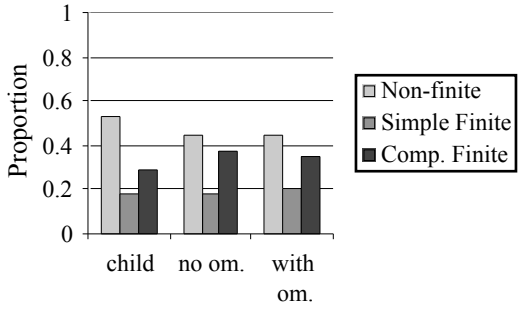

Fig. 3. Rates of OI errors, simple and compound finites for children and simulations with and without omission.

Next, the omission mechanism was run on MOSAIC's output. Where necessary, output from slightly more mature models was selected in order to match the MLUs in the simulations without omission (omission of words from utterances reduces the MLU for the output). The omission mechanism was run with an omission probability of 0.5 for unfooted syllables and 0.1 for footed syllables. Fig. 3 gives the results for these analyses. The rates of OI errors in MOSAIC's output clearly provide a closer match to the children's data than the omission of weak elements from complete utterances (both short and long ones). The prosody-based omission mechanism had very little effect on the model's fit to the data.

Omissions errors in MOSAIC's output

Having established that the addition of the prosody-based omission mechanism does not adversely affect MOSAIC's fit to the OI data, we can now assess whether the model produces any utterances that it previously could not. The examples in Table 1 show that this is the case.

Table 1: Examples of utterances with multiple sentence-internal omissions in MOSAIC's output.

He go to shop
He sit on chair
She give you kiss
She go hospital
That going sleep

All the examples in Table 1 constitute OI errors where the modal has been omitted through the association of an utterance-final and utterance-initial phrase. Additionally, the prosody-based omission mechanism has resulted in unstressed words like *the*, *a*, and *to* being omitted. Thus, the phrase *He go to shop* may have been learned from the input utterance *He wants to go to the shop*. During learning, MOSAIC has associated the utterance-final phrase *go to the shop* with the utterance-initial phrase *he*. The omission mechanism has resulted in the unstressed and unfooted determiner *the* being omitted. In four out of the five examples an unfooted item has been omitted. In the phrase *She go hospital* the particle *to* which forms a trochaic foot with *go* is missing.

One further measure of how well the model's output approximates children's speech relates to the levels of determiner omission. Demuth et al. (in press) provide an analysis of 5 English children which shows that 4 of these children omit determiners from unfooted contexts at higher rates than from footed contexts. In order to determine how well MOSAIC approximates this pattern we assessed the levels of determiner omission from footed and unfooted contexts in the simulations as well as in the actual speech of Anne and Becky at different MLU points. This was done in the following manner. First, a list of nouns that are predominantly used with a determiner was compiled by searching the maternal speech for nouns that are used with a determiner in at least 75% of the cases. Next, the child speech and model output were searched for utterances containing one of these nouns. Allowing for the occurrence of common adjectives, it was then decided if a determiner (*a*, *an* or *the*) was provided, and whether the context was footed or unfooted. Utterances that contained other determiners (e.g. *my*) were disregarded. Note that the assignment of the metrical pattern was done in an identical (automated) manner for the child speech and model output. That is, all function words were considered to be unstressed. For all other items, the stress pattern given by the Unilex dictionary was used. Tables 2a and b compare the child data with MOSAIC's output before the omission mechanism was run. Apart from Anne's earliest stage, the children omit determiners from unfooted contexts at higher rates than from footed contexts. This is not the case for the simulations. Provision levels in footed contexts exceed those in unfooted contexts in just 2 of the 6 simulations (by a maximum of 8 percentage points), while provision levels in unfooted contexts are higher (by 14 percentage points) in one of the simulations.

Table 2a: Determiner provision in footed and unfooted contexts for Anne and Anne's model without omission.

	Anne		Anne's model	
MLU	Footed	Unfooted	Footed	Unfooted
2.2	.13	.14	.50	.42
3.0	.70	.47	.65	.66
3.5	.80	.62	.76	.76

Table 2b: Determiner provision in footed and unfooted contexts for Becky and Becky's model without omission

	Becky		Becky's model	
MLU	Footed	Unfooted	Footed	Unfooted
2.2	.53	.20	.31	.45
3.0	.79	.60	.66	.66
3.7	.86	.60	.78	.72

Table 2c: Determiner provision in footed and unfooted contexts for Anne and Becky's model with omission.

	Anne's model		Becky's model	
MLU	Footed	Unfooted	Footed	Unfooted
2.2	.43	.29	.29	.28
3.0	.59	.40	.65	.39
3.5	.65	.40	.69	.41

Table 2c presents the results of this analysis on MOSAIC's output after the omission mechanism was run. These results look much improved over the simulations without omission. Apart from the early simulation for Becky, determiner omission clearly occurs more frequently in unfooted contexts. The developmental pattern (increase in provision rates) in the models is not as pronounced as it is in the children. The simulations, however, were run with fixed

omission probabilities (0.5 for unfooted items and 0.1 for unfooted items) for all developmental stages. A simple solution to this problem would be to vary these probabilities with developmental stage.

Conclusions

This paper set out to establish the value of a prosody-based omission mechanism aimed at making the output of MOSAIC more child-like. One particular aim was to allow MOSAIC to produce utterances with multiple sentence-internal omissions. The prosody-based omission mechanism clearly increases the range of utterances that MOSAIC can produce and thus makes the model's output more child-like and increases its credibility as a model of childen's early multi-word speech. It is also apparent that the model without the omission mechanism does not simulate the differential rates of determiner omission from footed and unfooted contexts. The addition of the omission mechanism rectifies this divergence between the model output and child speech, and thereby increases the child-likeness of the model's output on this measure as well.

Obviously, it is not very surprising that the mechanism produces these results, as this is what it has been designed to do. However, future, (cross-linguistic) work may provide a more stringent test of the mechanism. In particular, the mechanism predicts that the pattern of omission of function words will be different for languages that predominantly display iambic feet (e.g. French). Some evidence for this claim is provided by Tremblay & Demuth (in press).

It could be argued that the present mechanism is somewhat crude in that all function words are considered unstressed. The mechanism could however, easily be made more sophisticated by specifying stress patterns for different types of (frequent) constructions. Some possible refinements have already become apparent as a result of the simulations reported here. Inspection of the pattern of determiner omission in the two children suggests that omission levels after pronouns with a contracted copula (e.g. *that's*) are lower than in the model's output. Such an effect could easily be incorporated in the present mechanism on the plausible assumption that a pronoun with a contracted copula receives higher stress (and therefore forms a trochaic foot with a determiner that follows it) than a bare pronoun. Another possible refinement would be to specify different stress patterns for interrogative and declarative utterances.

The analyses reported here also have theoretical implications. The simulations which determined the levels of OI errors when the omission mechanism was run on the input showed that prosody-based omission alone is not sufficient to explain the OI phenomenon even when restricting the analysis to short utterances. Thus, an (unspecified) learning mechanism which produces short complete utterances (one possible instantiation of full competence) coupled with prosody-based omission does not provide an adequate fit to the child data. In order to obtain such a fit, omission needs to be co-determined by other factors. The simulations reported here suggest that a learning mechanism that is subject to a primacy and recency effect is such a factor.

Acknowledgements

This research was funded by the Economic and Social Research Council under grant number RES000230211.

References

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

Brown, R. (1973). A first language: The early stages. London: George Allen & Unwin Ltd.

Demuth, K., McCullough, E. & Adamo, M. (in press). The prosodic (re)organization of determiners. To appear in *Proceedings of the 31st Boston University Conference on Language Development*.

Fitt, S & Isard, S. (1999). Synthesis of regional English using a keyword lexicon. *Proceedings: Eurospeech 99*, Vol. 2, pp. 823-6.

Freudenthal, D., Pine, J.M., Aguado-Orea, J. & Gobet, F. (2007). Modelling the developmental patterning of finiteness marking in English, Dutch, German and Spanish using MOSAIC. *Cognitive Science*, *31*, *311-341*

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

Freudenthal, D., Pine, J.M. & Gobet, F. (2005a). Simulating optional infinitive errors through the omission of sentence-internal elements. In B.G. Bara, L. Barsalou & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society.* Mahwah NJ: LEA.

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

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

Gerken, L. A. (1996). Prosodic structure in young children's language production. *Language*, *72*, *683-712*.

Ingram, D & Thompson, W. (1996). Early syntactic acquisition in German: evidence for the modal hypothesis. *Language*, *72*, *97-120*.

Tremblay, A. & Demuth, K. (in press). Prosodic licensing of determiners in children's early French. In A. Belikova, L. Meroni and M. Umeda (Eds.). *Proceedings of the Conference on Generative Approaches to Language Acquisition*. Somerville, MA: Cascadilla Press.

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

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