Unifying Cross-linguistic and Within-language Patterns of Finiteness Marking 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 and Law, Brunel University

Abstract

MOSAIC, a model that has already simulated cross-linguistic differences in the occurrence of the Optional Infinitive phenomenon, is applied to the simulation of the pattern of finiteness marking within Dutch. This within-language pattern, which includes verb placement, low rates of Optional Infinitives in Wh-questions and the correlation between finiteness marking and subject provision, has been taken as evidence for the view that children have correctly set the clause structure and inflectional parameters for their language. MOSAIC, which employs no built-in linguistic knowledge, clearly simulates the pattern of results as a function of its utterance-final bias, the same mechanism that is responsible for its successful simulation of the cross-linguistic data. These results suggest that both the cross-linguistic and within–language pattern of finiteness marking can be understood in terms of the interaction between a simple resource-limited learning mechanism and the distributional statistics of the input to which it is exposed. Thus, these phenomena do not provide any evidence for abstract or innate knowledge on the part of the child.

The Optional Infinitive phenomenon

One of the key features of children's early multi-word speech is that, in many languages, children often produce utterances that contain non-finite verb forms in contexts where a finite verb form is obligatory in the adult language. Thus, English-speaking children may produce utterances such as *he go* instead of *he goes*. While the English example may suggest that the child has simply omitted the inflectional morpheme (*-es*), data from languages such as German and Dutch where the infinitive carries its own morphological marker (*-en*) make it clear that children actually produce an infinitive instead of a finite verb form. Thus, German-speaking children may produce utterances such as *Vater spielen* (Daddy play-inf) instead of *Vater spielt (*Daddy plays-fin).

While such *Optional Infinitive errors* are quite frequent in obligatory subject languages such as English, Dutch and German, they are quite rare in pro-drop languages like Spanish and Italian. An influential theory by Wexler (1994, 1998) explains this cross-linguistic pattern of results (and the relative sparseness of other types of errors) by assuming that children have correctly set all the clause structure and inflectional parameters for their language from a very early age, but optionally under-specify Agreement and/or Tense. Due to cross-linguistic differences between the grammars of pro-drop and obligatory subject languages, this leads to the provision of non-finite verb forms in contexts where a finite verb form is required in obligatory subject languages such as English and Dutch, but not in pro-drop languages such as Spanish.

The great strength of Wexler's account, which is consistent with the Universal Grammar approach to language acquisition (Chomsky, 1981; Pinker, 1984) is that it provides a unified explanation of the cross-linguistic occurrence of Optional Infinitive errors. Recently, however, Freudenthal, Pine and Gobet (2006, submitted) have shown that the cross-linguistic differences in the occurrence of Optional Infinitive errors may reflect sensitivity to differences in the distributional statistics of the input children are subject to. Thus, Freudenthal et al. showed that MOSAIC, a performance limited distributional analyser which receives child-directed speech as input and learns to produce progressively longer utterance-final phrases provides a good fit to the developmental characteristics of the Optional Infinitive phenomenon in English, Dutch, German and Spanish. The key to MOSAIC's simulation of the differential rates of Optional Infinitive errors across the four languages is its bias towards learning material that occurs near the end of an utterance. Since finite and non-finite verbs pattern differently in the four languages, utterance-final phrases have radically different proportions of non-finite verbs. MOSAIC was therefore able to provide a good quantitative fit to the cross-linguistic data as a result of the interaction between its utterance-final bias and the distributional properties of the languages to which it was exposed.

While MOSAIC's ability to deal with the cross-linguistic pattern of finiteness marking based on one, relatively simple processing constraint is encouraging, a clear challenge for MOSAIC remains the simulation of the pattern of regularities in finiteness marking that has been found *within* languages that display the OI phenomenon. Thus, Poeppel and Wexler (1993) have identified a number of constraints relating to verb placement, differential rates of OI-errors in declaratives and Wh-questions, and the correlation between the provision of subjects and the finiteness of an utterance. These regularities, which are particularly noticeable in

languages such as German and Dutch, where verb position is dependent on finiteness, place considerable additional constraints on models of children's early multi-word speech. This is particularly true, as these regularities appear to be consistent with Wexler's claim that children have correctly set all the clause structure and inflectional parameters for their language from a very early age, whilst being problematic for rival Nativist theories of the Optional Infinitive phenomenon (e.g. Rizzi, 1994).

Given the apparent support that it provides for Wexler's account, the within-language pattern of finiteness marking can be considered a strong test for models like MOSAIC that embody the notion that children's early multi-word speech reflects an interaction between a performance-limited learning process and the statistical regularities of the input children receive.

Within-language regularities in the pattern of finiteness marking in German and Dutch

German and Dutch are particularly interesting languages with respect to the development of finiteness marking as verb position in these languages is dependent on finiteness: finite verbs take second position, whereas non-finite verbs take sentence-final position. Poeppel and Wexler (1993) show that (in German) the child overwhelmingly places finite and non-finite verbs in their correct position from a very early age. From this finding, Poeppel and Wexler conclude that the child knows the rules for verb placement.

Poeppel and Wexler also show that, while Optional Infinitive errors are quite frequent in German and Dutch children's early declarative multi-word speech, they are virtually absent from Wh-questions. Wexler (1998) argues that it is a property of V2 languages (like German and Dutch) that Wh-questions should contain a finite verb. By implication, the absence of OI-errors in Dutch Wh-questions provides evidence for the notion that Dutch-speaking children have correctly set the V2 parameter (see also Poeppel & Wexler, 1993).

Finally, there is a correlation in German and Dutch between the occurrence of finite verb forms in an utterance and the likelihood that the utterance contains a subject. It is well known (see e.g., Bloom, 1990) that children often produce utterances with missing subjects (e.g. *Need a toy*). Several authors have shown that utterances that contain finite verbs are more likely to contain a subject than utterances that contain only non-finite verb forms. Poeppel and Wexler (1993) explain this correlation by claiming that non-finite verbs *license* subject omission; it is grammatical for non-finite verbs to occur without subjects. Thus, in the utterance *He wants to go*, the non-finite verb *go* has no subject. Finite verbs, on the other hand, do require a subject. Wexler thus claims that children's lower rates of subject provision on non-finite verbs reflects children's knowledge that non-finite verbs can occur without overt subjects, whereas finite verbs cannot.

In summary, the Optional Infinitive phenomenon has attracted a great deal of attention due to its occurrence across a range of languages. MOSAIC has already been shown to provide a close quantitative fit to the basic Optional Infinitive phenomenon in English, Dutch, German and Spanish. Nativist accounts such as Wexler's (1998), explain the Optional Infinitive phenomenon by assuming that children misrepresent a small portion of the grammar, but have largely set the clause structure and inflectional parameters for their language correctly. The within-language regularities are consistent with such a view, and therefore provide considerable constraints for any process model of children's early multi-word speech. This paper aims to establish whether MOSAIC captures these regularities, thereby providing support for the notion that children's early multi-word speech can be understood in terms of the interaction between psychologically plausible constraints on learning and the statistical structure of the input.

MOSAIC

MOSAIC is a simple distributional learner that learns off child-directed speech and produces utterances that can be directly compared to children's speech. The basis of MOSAIC is a discrimination net consisting of nodes and arcs that store the utterances shown to it. Learning in MOSAIC takes place by adding nodes that encode new words and phrases to the network. The version of MOSAIC used in these simulations departs slightly from that described in Freudenthal et al. (2006, submitted). These versions of MOSAIC were too reliant on questions in the input as a source for Optional Infinitive errors: a phrase like *he go* was generated by producing the final words of *can he go*. Freudenthal et al. (2005a) therefore developed a version of the model that is capable of producing Optional Infinitive errors by omitting sentence-internal elements. This version can produce a phrase like *he go* by omitting the sentence internal modal *can* from the declarative *He can go* and has been shown to provide a good quantitative fit to the developmental data from English, Dutch, German and Spanish (see Freudenthal et al., 2005b). Learning in the new version of MOSAIC is anchored at the sentence-initial and sentence-final position: MOSAIC will only encode a word or phrase when all the material that either precedes or follows it in the utterance has already been encoded in the network. Learning in MOSAIC is slow, and governed by the following formula:

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

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

The formula results in a basic sigmoid curve, with the probability of creating a node increasing as a function of the number of times the input has been seen. Initially, MOSAIC will only learn short, sentence-initial and sentence-final phrases. Longer phrases will only be learnt when the base number in the formula starts to increase (as a result of seeing more input).

MOSAIC employs two mechanisms for generating (rote) output. The first mechanism traverses all branches of the network, and generates the contents of branches that encode sentence-final phrases. This first mechanism thus results in phrases with missing sentence-initial elements. The second mechanism involves the concatenation of sentence-initial and sentence-final phrases. When MOSAIC builds up the network, it associates sentence-initial and sentence-final fragments of utterances. The concatenation of utterance-initial and utterance-final phrases results in phrases with missing sentence-internal elements. Sentence-initial elements that are associated with sentence-final elements are normally limited to one word phrases. Longer sentence-initial phrases can be concatenated only if the phrase is frequent enough to have been 'chunked up' by the model's chunking mechanism[1]. MOSAIC therefore maintains an utterance-final bias: concatenations consist of relatively long utterance-final phrases with relatively short utterance-initial phrases.

MOSAIC's output thus consists of utterances with missing sentence-internal or sentence-initial elements. Both utterance-types are apparent in child speech. Utterances with missing sentence-internal elements may give rise to Optional Infinitive errors (e.g. *He* (*wants to*) *go home*). Utterances with missing sentence-initial elements may give rise to missing subjects (e.g. (*He) wants to go to the shops*). MOSAIC's early output will consist largely of sentence-final fragments. As the mean length of utterance (MLU) of the output increases, concatenations will become more frequent. With increased learning, incomplete utterances are slowly replaced by complete utterances.

The two mechanisms described so far are complemented by a mechanism that allows for the production of novel utterances through the substitution of distributionally similar items. A description of this mechanism can be found in Freudenthal et al. (2005c).

The Simulations

The version of MOSAIC used for the simulations described in this paper has changed slightly from the version described in Freudenthal et al. (2005a, 2005b) in that the chunking mechanism described in Freudenthal et al. (2005c) has now been implemented as well. For this reason, we will first examine whether MOSAIC still provides a good fit to the basic Dutch data. For these simulations, the input corpora (consisting of the maternal speech directed at Peter and Matthijs; Wijnen, 1993) were fed through the model several times, and output (of increasing average length) was generated after every run of the model. The output files that most closely matched the MLU (Mean Length of Utterance) of the child at 4 different points in development were selected. The output was then divided into utterances containing only non-finite verb forms (non-finites), utterances containing only finite verb forms (simple finites), and utterances containing both finite and non-finite verb forms (compound finites). The child data are shown in fig. 1. The results of MOSAIC's simulations are shown in fig. 2.

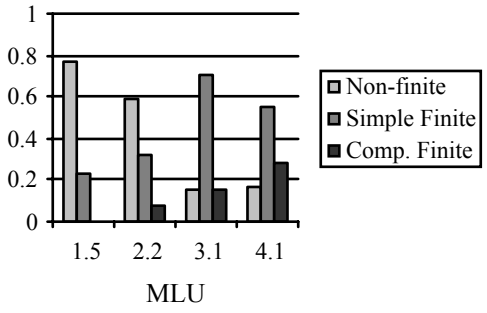

Fig. 1a: Data for Peter

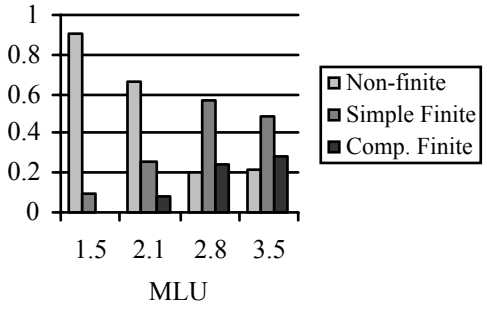

Fig. 1b: Data for Matthijs

As can be seen, MOSAIC clearly simulates the basic OI phenomenon for both children. While the model tends to overestimate the occurrence of Optional Infinitive errors during later stages of development, it successfully simulates the initial near exclusive use of Optional Infinitive errors followed by a substantial increase in the number of (compound) finites.

MOSAIC simulates the large decrease in Dutch Optional Infinitive errors as a result of the interaction between its utterance-final bias and the rules for verb placement in Dutch grammar. Non-finite verbs take sentence-final position in Dutch, whereas finite verbs take second position. Since MOSAIC's early output consists mostly of utterance-final phrases, it will contain large numbers of non-finite verb forms. As the MLU of the model increases (with successive exposures to the input), finite verb forms, which occur earlier in the utterance, start appearing and non-finite

[1] MOSAIC's chunking mechanism results in frequent multi-word phrases being treated as one unit. It is discussed in more detail in Freudenthal et al. (2005c).

utterances are slowly replaced with the compound finites from which they are learnt.

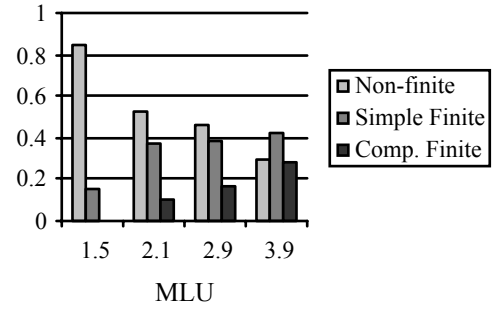

Fig. 2a: Model for Peter

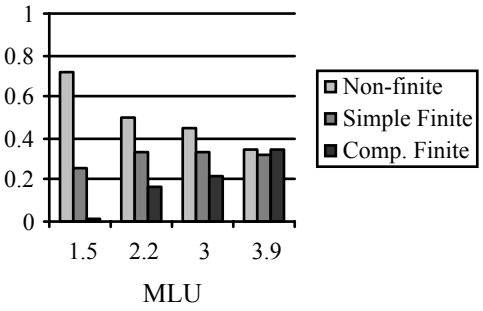

Fig. 2b: Model for Matthijs

Having established that MOSAIC successfully simulates the basic Optional Infinitive phenomenon, we can now turn to MOSAIC's success at simulating the finer detail associated with it. This will be done by investigating whether a) MOSAIC correctly places finite verbs in second position and non-finites in sentence-final position; b) Optional Infinitive errors are less frequent in Wh-questions than in declaratives; and c) rates of subject omission in utterances containing only non-finite verb forms are higher than in utterances containing finite verb forms. It should be stressed that these additional analyses were performed on the same output files that were analysed for the basic OI phenomenon. Thus, the output from one, identical model is analyzed with respect to four phenomena.

Verb Placement

MOSAIC's success at placing verbs in their correct sentential position was investigated by coding two samples (from MLU point 3 in fig. 2) of finite and non-finite utterances from MOSAIC's output with respect to verb placement. This analysis is similar to that performed on a German[2] child by Poeppel and Wexler (1993). Poeppel and Wexler showed that the German child placed 90% of the verbs correctly, and conclude from this analysis that 'the finiteness distinction is made correctly at the very earliest stages of development' (p. 7). Two-word utterances, where second and final position overlap, were excluded from this analysis[3]. As can be seen from table 1 and 2, the model overwhelmingly places finite and non-finite verbs in the correct position. MOSAIC simulates this effect because it preserves the word order it sees in the input. This finding makes it clear that correct verb placement is not necessarily evidence for knowledge of the rules of verb placement. MOSAIC makes no distinction between finite and non-finite verbs (in fact, MOSAIC has no built-in linguistic knowledge at all). MOSAIC places verbs correctly because it is sensitive to the statistical properties of the input.

Table 1: Verb placement for Matthijs's model

	V2/not final	Final/not V2
Finites	65	3
Non-finites	4	56

Table 2: Verb placement for Peter's model

	V2/not final	Final/not V2
Finites	60	4
Non-finites	4	44

Wh-questions

Poeppel and Wexler (1993) show that Optional Infinitive errors are rare in topicalized structures with a non-subject in first position in a German child. Such structures include Wh-questions. Other authors have also shown OI-errors to be rare in Wh-questions in German and Dutch. This is in contrast to declarative structures, where OI-errors are quite frequent. Poeppel and Wexler claim that German (and Dutch) grammar dictates that such topicalized structures contain a finite verb form. The lack of OI-errors in Wh-questions and related structures is therefore seen as evidence that the child has acquired the relevant portion of the grammar. Table 3 gives the rates of OI-errors in declaratives and Wh-question for the simulations at MLU point 3 in Fig. 2. For the analysis of Wh-questions, a sample was drawn from the output file[4]. While the model produces relatively high rates of Wh-questions with only non-finite verb forms compared to the rates reported in the literature, it clearly produces few of them compared with the rate at which it produces OI-errors in declaratives. This result is not an obvious consequence of MOSAIC's utterance-final bias, as modal/auxiliary + non-finite constructions (which are the source of OI-errors in MOSAIC) occur in Wh-questions at

[2] Dutch and German have identical rules regarding the placement of finite and non-finite verbs. Analyses of Dutch and German can therefore be directly compared.

[3] Utterances with a missing subject and a finite verb in first position were coded as having the finite verb in V2.

[4] In keeping with the coding scheme used for declaratives, as verbs matching the infinitive were counted as infinitives. As a result, some plural, present tense finites were counted as non-finites. This may result in these analyses showing relatively high rates of non-finites compared to those reported in the literature. The present analysis however, is comparable to that performed on declaratives.

rates that are comparable to declaratives. Instead, it reflects an interaction between utterance length and frequency which results in Wh-questions being learned more quickly than declaratives. The average MLU for Wh-questions in the input (Matthijs: 5.39; Pet: 4.88) is lower than it is for declarative structures (Matthijs: 6.09; Pet: 5.84) Wh-questions also appear to be more formulaic than declaratives. In declaratives, the finite verb form may be preceded by a range of subjects, which may lead to finite verb forms being encoded relatively late. In Wh-questions, only a handful of Wh-words precede the finite verb form. This may lead to relatively fast learning and early encoding of the finite verb form in Wh-questions.

Table 3: Proportion of non-finite Wh-questions and declaratives in Matthijs and Peter's models

	Dec.	Wh-questions
Matthijs model	.46	.27
Peter model	.45	.24

Finiteness and Subject Omission

Several authors have reported a correlation between the finiteness of children's utterances and the likelihood of that utterance containing a subject. Children are less likely to omit the subject from an utterance that contains a finite verb form than from utterances that only contain non-finite verb forms. Poeppel and Wexler (1993) claim this pattern of results provides support for the notion that children are aware of the distinction between finite and non-finite verb forms. Children have high rates of subject omission on non-finite utterances because the infinitive *licenses* subject omission: it is grammatical for the infinitive to occur without a subject (for example, in the English utterance *he wants to go*, the infinitive *go* has no subject). Since children are aware of this, they are relatively likely to omit subjects from non-finite utterances. Finite utterances on the other hand do require the inclusion of a subject. Children's lower rates of subject omission on finite utterances are thought to reflect this knowledge. Fig. 3 shows the levels of subject provision for the two Dutch children for simple finite and non-finite utterances (at the same MLU points as portrayed in Fig. 1).

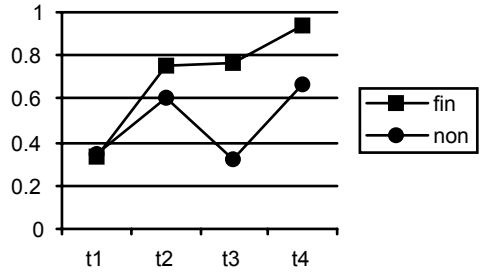

Fig. 3a: Data for Peter

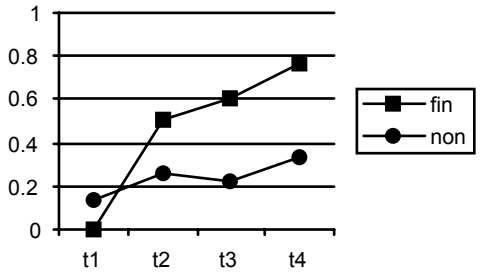

Fig. 3b: Data for Matthijs

Note that, while the levels of subject omission on finite utterances are lower than for non-finite utterances, they are by no means zero, even at relatively late stages of development. Poeppel and Wexler explain this result, which is not predicted on their account of full competence, by assuming these errors are not instances of subject drop, but rather instances of *topic drop*, a pragmatic effect that bears little relevance to the child's grammatical development.

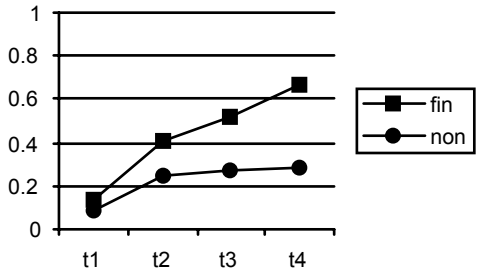

Fig. 4a: Model for Peter

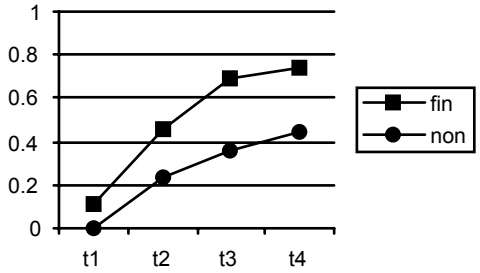

Fig. 4b: Model for Matthijs

As can be seen in Fig. 4, MOSAIC clearly simulates the correlation between finiteness and subject omission. MOSAIC simulates this effect as a result of the interaction between its utterance-final bias, and the position of finite and non-finite verb forms relative to the subject in the utterance. Subjects in Dutch usually take sentence-initial position, with finite verbs taking second position and non-finite verbs taking sentence-final position. A result of this is that the positional distance between finite verb forms and the subject is smaller than that between non-finite verb

forms and the subject. Thus, by the time a learning mechanism with an utterance-final bias has encoded a finite verb form, it is quite likely to have encoded the subject as well. When the same learning mechanism has encoded a non-finite verb form, it may still be some time before the subject is encoded. Note that, unlike Poeppel and Wexler's account, MOSAIC does not require separate mechanisms to explain omission of subjects from finite and non-finite utterances. One mechanism (the omission of sentence-initial elements) is responsible for subject omission from both types of utterances. It is the interaction between this mechanism and the distributional statistics of the input that results in the differential rates of omission.

Conclusions

This paper set out to simulate the regularities that exist within the patterning of finiteness marking in Dutch. MOSAIC, a model that has already been applied to the simulation of cross-linguistic differences in the development of finiteness marking, was applied to the simulation of 1) correct placement of finite and nonfinite verbs, 2) low levels of OI-errors in Wh-questions, and 3) the correlation between finiteness and subject omission.

MOSAIC clearly simulates all the phenomena as a result of the interaction between the psychologically plausible constraints on its learning mechanism and the distributional statistics of the input. MOSAIC places finite and non-finite verbs correctly as it preserves the word order it sees in the input. MOSAIC produces fewer Optional Infinitive errors in Wh-questions than it does in declaratives. This appears to be caused by Wh-questions, on average, being shorter and more formulaic than declaratives. This may lead to the finite verb in Wh-questions being learned more quickly than in declaratives. Finally, the correlation between finiteness and subject omission is caused by the differential positional distance between the subject and finite and non-finite verbs. Thus, the mechanism that is responsible for MOSAIC's successful simulation of the cross-linguistic patterning of the OI phenomenon (MOSAIC's utterance-final bias) can also explain the within-language pattern of regularities associated with the OI phenomenon.

These findings are significant, as these phenomena have been taken as support for the notion that children have correctly set the clause structure and inflectional parameters for their language. By implication, they are also viewed as evidence for the wider role of innate linguistic knowledge as exemplified by the Universal Grammar approach to language acquisition.

The successful simulation of both the cross-linguistic and the within-language phenomena within MOSAIC, which employs no built-in linguistic knowledge, strongly suggests that these phenomena can be explained through children's sensitivity to the distributional statistics of the language to which they are exposed. It therefore suggests that the fact that children largely respect the patterning of their language is not evidence for abstract (or innate) knowledge. In fact, it suggests that the apparent overlap between the patterning of the adult and child language may not be the most appropriate focus for theories of language acquisition, as it may simply reflect sensitivity to the distributional statistics of this input.

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.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht: Foris.

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). Simulating the cross-linguistic development of Optional Infinitive errors in MOSAIC 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. (2005c). 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 Infinitive in Dutch and English using MOSAIC. *Cognitive Science*, *30*, *277-310*.

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

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

Poeppel, D. & Wexler, K. (1993). The full competence hypothesis of clause structure in early German. *Language*, *69*.

Rizzi, L. (1994). Early null-subjects and root null-subjects. In: T. Hoekstra & B. Schwartz (eds.) *Language acquisition studies in generative grammar*. Amsterdam: John Benjamins.

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*.

Wijnen, F. (1993a), Verb placement and morphology in child Dutch: do lexical errors flag grammatical development? *Antwerp Papers in Linguistics, 74,* 79–92.