Simulating the Cross-Linguistic Development of Optional Infinitive Errors in MOSAIC

Daniel Freudenthal (D.Freudenthal@liv.ac.uk)
Julian Pine (Julian.Pine@liv.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

The Optional Infinitive (OI) phenomenon in children's speech has attracted a great deal of attention due to its occurrence in a variety of languages (including English, Dutch and German), and its apparent absence in other languages (such as Spanish and Italian). Wexler (1998) explains this pattern of results in terms of a Unique Checking Constraint that interacts with cross-linguistic differences in the underlying grammar to result in Optional Infinitive errors in obligatory subject languages (which require double-checking), but not in pro-drop languages (which do not require double-checking). While Wexler's account explains the cross-linguistic data, it attributes a great deal of innate linguistic knowledge to the child, and ignores the possibility that the cross-linguistic data may be equally well explained by the interaction between a simple distributional learning mechanism and the surface characteristics of the language. This paper presents simulations of the Optional Infinitive phenomenon across 4 languages (English, Dutch, German, and Spanish) using MOSAIC, a simple distributional analyser with no built-in syntactic knowledge. MOSAIC clearly simulates the different rates of Optional Infinitive errors across the languages, suggesting (a) that it is possible to explain the basic OI phenomenon without assuming large amounts of innate linguistic knowledge, and (b) that cross-linguistic differences in the OI phenomenon may be related to differences in the surface characteristics of the languages being learned.

Introduction

Between two and three years of age, children learning English often produce utterances that appear to lack inflections, such as past tense markers or third person singular agreement markers. For example, children may produce utterances as:

(1a) *That go there**
(2a) *He walk home**

instead of

(1b) *That goes there*
(2b) *He walks home*

Traditionally, such utterances have been interpreted in terms of absence of the appropriate knowledge of inflections (Brown, 1973) or the dropping of inflections as a result of performance limitations in production (L. Bloom, 1970; P. Bloom, 1990; Pinker, 1984; Valian, 1991). More recently, however, it has been argued that they reflect the child's optional use of (root) infinitives (e.g. *go*) in contexts where a finite form (e.g. *went*, *goes*) is obligatory in the adult language (Wexler, 1994, 1998).

This interpretation reflects the fact that children produce (root) infinitives not only in English, where the infinitive is a zero-marked form, but also in languages such as Dutch (Wijnen et al. 2001) and German, where the infinitive carries its own infinitival marker (*-en*). For instance, children learning Dutch may produce utterances such as:

(3a) *Pappa eten** (*Daddy* (*to*) *eat*)
(4a) *Mamma drinken** (*Mummy* (*to*) *drink*)

Instead of

(3b) *Pappa eet* (*Daddy eats*)
(4b) *Mamma drinkt* (*Mummy drinks*)

According to Wexler (1998), the Optional Infinitive phenomenon can be explained as follows. By the time children begin to produce multi-word utterances, they have already set all the basic inflectional and clause structure parameters of their language. However, their grammars are governed by a Unique Checking Constraint that is 'genetically specified (and withering away in time)' (Wexler, 1998: 27). The Unique Checking Constraint may prevent the child from checking the D-feature of the subject DP against more than one D-feature (in this case the D-features of Tense and Agreement). As a result, Tense and Agreement can be optionally under-specified in the underlying representation of the sentence, and the child may produce non-finite verb forms (forms that are not marked for tense or agreement) in contexts in which a finite verb form is required.

The main strength of Wexler's account is that it can explain data from a range of different languages. Thus, it can explain why children produce Optional Infinitive errors at high rates in obligatory subject languages like English, Dutch and German, which require the child to check against

two D-features: Tense and Agreement. However, it can also explain why children make few Optional Infinitive errors in INFL-licensed null subject languages like Spanish and Italian, which (usually) only require the child to check against one D-feature: Tense.

On the other hand, Wexler's account also has certain weaknesses. First, while the account makes qualitative predictions about the occurrence or non-occurrence of OIs in a number of different languages, it makes no (detailed) quantitative predictions about the rate at which children will make OI errors or how these rates will change as the child's language develops. Wexler invokes the concept of maturation to explain the decline in OI errors, but the concept is relatively unspecified, and does not give rise to quantitative predictions.

Second, where the account makes qualitative predictions (e.g. about the lack of Optional Infinitive errors in pro-drop languages), it does so with reference to deep structural differences in the grammar of the languages, thereby ignoring the possibility that the interaction between an input-driven learning mechanism and the surface characteristics of the language may explain the data equally well. Freudenthal, Pine and Gobet (2002, submitted) have already shown that MOSAIC, a simple distributional analyser that learns from child-directed speech and has no built-in syntactic knowledge can provide a close quantitative fit to the basic Optional Infinitive phenomenon in Dutch and English.

This paper presents a new version of MOSAIC which addresses some weaknesses of earlier versions and explains OI errors in terms of the omission of auxiliaries or modals from constructions containing a modal/auxiliary and an infinitive. For example, the phrase *that go there* might be produced by omitting *should* from *that should go there*, and *he go home* might be produced by omitting *wants to* from *he wants to go home*. Similarly, the Dutch phrase *Pappa eten,* might be produced by omitting *wil* from *Pappa wil eten* (*Daddy wants to eat*). MOSAIC will be applied to Optional Infinitive data from Dutch and English, as well as German and Spanish. MOSAIC's ability to simulate the data across these languages, which show rather different levels of Optional Infinitive errors, serves as a strong test of its mechanisms for producing Optional Infinitive errors, and the feasibility of distributional approaches to language acquisition in general.

MOSAIC

A major change from earlier versions of MOSAIC is that the model now learns from both edges of the utterance and associates sentence-initial and sentence-final phrases, leading to the omission of sentence-internal elements. This change brings MOSAIC more in line with general psychological theorizing (MOSAIC now shows a primacy as well as a recency effect). It also allows the model to simulate a wider range of phenomena than the previous version of MOSAIC, which only learnt from the right edge of the utterance. An additional difference is that MOSAIC now distinguishes between questions and declaratives, resolving the problem that earlier versions of MOSAIC relied too heavily on questions as the source for Optional Infinitive errors (Freudenthal, Pine, & Gobet, 2005a).

MOSAIC consists of a simple network of nodes that incrementally encode words and phrases that have been presented to the model. As the model sees more input it will 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 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*.

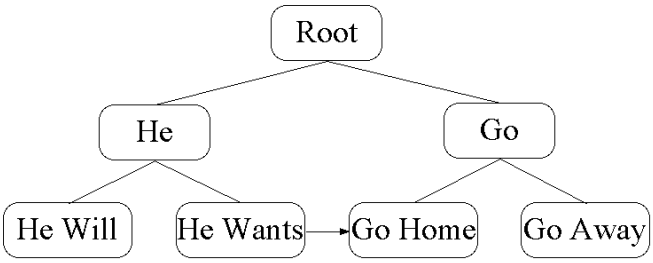

Figure 1: A partial MOSAIC model. The sentence-initial phrase *he wants*, and the sentence-final phrase *go home* have been associated, allowing the model to produce the utterance *He wants go home*.

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[1] (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[2]) identical to that in earlier versions of MOSAIC. In generation, the model traverses 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 (c.f. *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 (2005b), though the chunking mechanism described in that paper has not yet been implemented in the present version of the model.

The Simulations

All the simulations reported in this paper used the same version of MOSAIC together with corpora of realistic, child-directed speech. For English and Dutch, corpora available through the CHILDES database (MacWhinney, 2000) were used. The English child (Becky) was part of the Manchester corpus (Theakston, Lieven, Roland & Pine, 2001). The Dutch child (Peter) was part of the Groningen corpus (Bol, 1995). Additional simulations for one Dutch and one additional English child can be found in Freudenthal, Pine & Gobet, 2005a). The Spanish

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

[2] 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.

simulations were conducted using the corpus of Juan (Aguado Orea, 2004). For German, the corpus of Leo (Behrens, in press) was used. For all simulations, the same (automatic) coding scheme was used: utterances that only contained verbs matching non-finite forms were classed as non-finite. Utterances containing only finite forms were classed as simple finite. Utterances containing both finite and non-finite forms were classed as compound finites. The analyses for English deviated slightly from the other analyses. As English has an impoverished inflectional system, it is necessary to restrict the analysis to utterances containing a 3rd singular subject in order to identify Optional Infinitives. Also, since many verb forms (e.g. *walked*) are ambiguous with respect to whether they are non-finite (past-participle) or finite (past tense), utterances in which such forms were the only verb were classed as ambiguous. The children's output was analysed at different stages of increasing MLU. The Child-Directed speech for each child was then fed through the model several times. Output from the model was generated after every presentation of the input. The output files that most closely matched the child's MLU were then selected. For both the simulations and the children, the analysis was performed on utterance types. The size of the input corpora varied. For Becky, it consisted of approximately 25,000 utterances, Peter's input consisted of approximately 13,000 utterances, and the size of Juan's input was 34,000 utterances. For Leo a random sample of 30,000 utterances was chosen from the entire corpus, which consists of nearly 110,000 utterances.

English simulations

Figure 2 gives the data and simulations for English. As can be seen, the model provides a close fit to the data with respect to the rates at which Optional Infinitives are produced. However, at the lowest MLU point, the proportion of simple finites that the model produces is too high. The model generates Optional Infinitives because it is capable of omitting modals and auxiliaries from phrases such as *He wants to go*. As the model learns to produce longer utterances, such omissions become less frequent and the proportion of Root Infinitives decreases.

Fig. 2a: Data for Becky

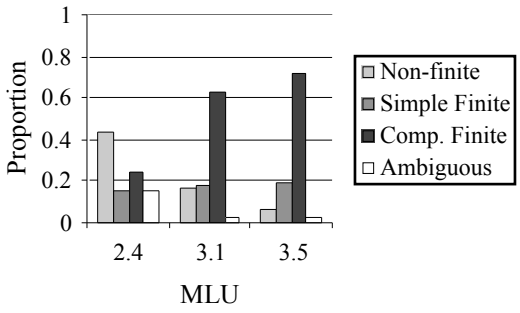

Fig. 2b: Simulations for Becky

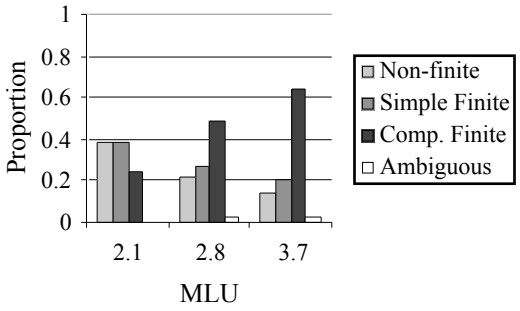

Fig. 2: Data and simulations for an English child.

Dutch simulations

Figure 2 displays the data and simulations for a Dutch child. It is apparent that the Dutch child starts out with relatively high levels of Optional Infinitives, which drop quite quickly.

Fig. 3a: Data for Peter

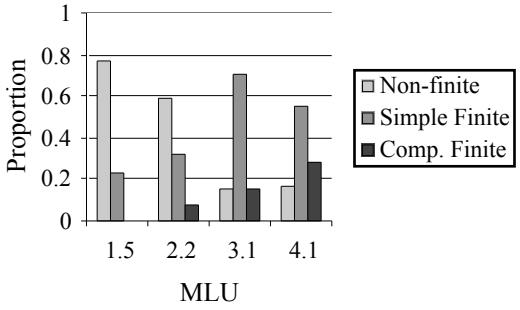

Fig. 3b: Simulations for Peter

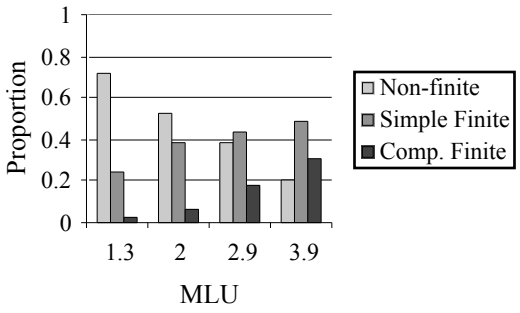

Fig. 3: Data and simulations for a Dutch child.

MOSAIC simulates the high levels of OI errors as a result of its utterance-final bias. In Dutch, non-finite verb forms take sentence-final position, while finite verbs take second position. Early in development, the model will produce mostly sentence-final phrases (and a few concatenations including sentence-initial words). The sentence-final phrases will contain many non-finite verb forms, the

sentence-initial words will mostly consist of (pro)nouns, as subjects tend to take first position in declaratives. As the model starts to produce longer utterances, finite verb forms start appearing, leading to an increase in the proportion of simple and compound finites.

German Simulations

The results for German are shown in Figure 4. German grammar is identical to Dutch as far as the relation between verb placement and finiteness is concerned. Thus, in both Dutch and German, finite verbs take second position, whereas non-finite verbs take utterance-final position. As with the Dutch data, MOSAIC simulates the patterning of the German data quite well. When comparing the results for Dutch and German, it is apparent that the rates of OI errors in the early German child data and simulations are quite a bit (16%) lower than they are for Dutch, and the decrease in levels of OI errors is not as pronounced as it is for Dutch. While this effect may reflect individual rather than cross-linguistic variation in the children's speech, it also raises the interesting possibility that although verb placement is subject to the same grammatical rules in Dutch and German, there are nevertheless subtle differences between the two languages that affect the relative frequency with which certain constructions are used. If this is the case, it suggests a greater role for input-driven learning than has so far been assumed by Wexler.

Fig. 4a: Data for Leo.

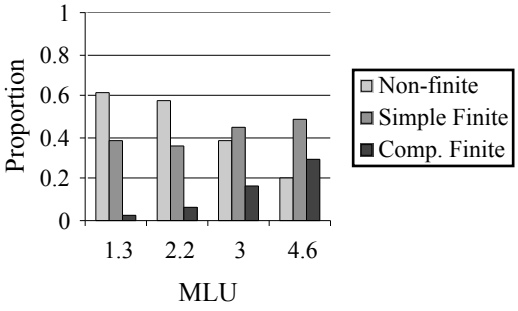

Fig. 4b: Simulations for Leo.

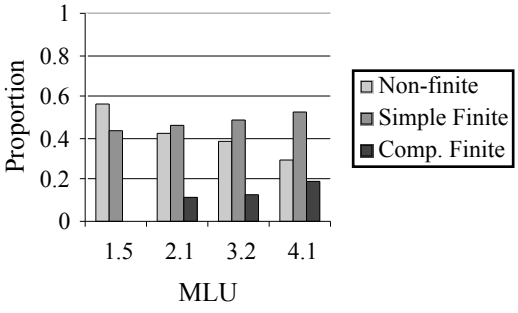

Fig. 4: Data and simulations for a German child.

Spanish Simulations

Figure 5 presents the data and simulations for Spanish. It is apparent from Figure 5 that the Spanish child produces OIs at a considerably lower rate than the children in the other languages. Again, MOSAIC simulates the basic rate of Optional Infinitives quite well. MOSAIC simulates this low rate of Optional Infinitives because of the structure of Spanish. For all languages discussed so far, Optional Infinitives are generated by the omission of modals/auxiliaries from compound finites. While these occur at roughly equal rates (.31, .22, and .25 for Dutch, German and Spanish respectively[3]), the verb forms that occur in sentence-final position (and are thus learned most easily) differ across the languages. In Spanish, a large majority of these are finite (74%). For Dutch and German, only 18% and 35% of the utterance-final verbs are finite.

The fact that Spanish is a pro-drop language also contributes to the low levels of Optional Infinitives: in situations where the subject is dropped, the utterance is likely to start with a (finite) verb. Concatenations involving a subjectless verb are therefore likely to result in a finite utterance.

Fig. 5a: Data for Juan.

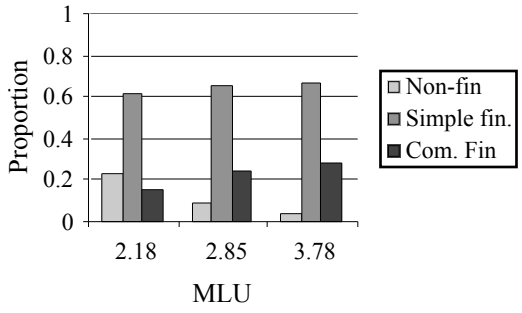

Fig. 5b Simulations for Juan

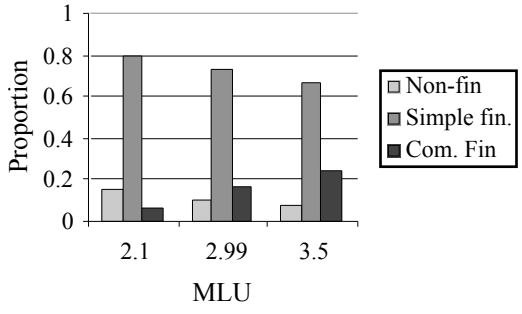

Fig. 5: Data and simulations for a Spanish child.

Conclusions

[3] The English input is ignored here as the inclusion of a subject is required in order to identify an Optional Infinitive.

MOSAIC clearly simulates the basic Optional Infinitive phenomenon in four languages that differ considerably in terms of their underlying grammar and in the rates of Optional Infinitive errors that children in these languages display. Since MOSAIC does not use any built-in linguistic knowledge, and learns from child-directed speech that has a realistic frequency distribution, this result strongly suggests that cross-linguistic differences in the Optional Infinitive phenomenon are related to the surface characteristics of the languages. Unlike Wexler's account, which invokes the relatively unspecified concept of maturation, MOSAIC also provides a plausible explanation for the gradual decrease of Optional Infinitive errors. Optional Infinitive errors are produced through the omission of modals and auxiliaries from compound finites. Early in development, MOSAIC will omit these elements at a high rate. As the model's MLU increases, omission rates decrease and Optional Infinitives are replaced by compound finites.

An interesting finding is that despite there being no difference between Dutch and German in verb placement and its relation to finiteness, the analyses of the children and simulations show a difference of 15-20% in the proportion of Optional Infinitives at the earliest stage. A similar asymmetry is apparent in the input files. Compound finites are less common in the German input (by 9%), and non-finite verbs are more common in sentence-final position in the Dutch input (by 17%). Whilst it is possible that this asymmetry reflects individual differences in the children's speech, this finding also raises the possibility that subtle differences can exist between same family languages that affect the relative frequency of certain constructions, and the subsequent rates of Optional Infinitive errors that children learning these languages display. Thus, cross-linguistic differences in the rates at which children produce Optional Infinitives appear to be graded, quantitative differences which reflect the statistical properties of the input, and can be explained without recourse to differences in the deep structural properties of the language's grammar.

Acknowledgements

This research was funded by the ESRC under grant number RES000230211. We thank Javier Aguado-Orea for making the Spanish corpus available and for aiding in the simulations and analysis of Spanish. We thank the Max Planck Institute for Evolutionary Anthropology in Leipzig for making the German corpus available.

References

Aguado Orea, J. J. (2004). The acquisition of morpho-syntax in Spanish: Implications for current theories of development. Unpublished doctoral thesis, University of Nottingham.

Behrens, H. (in press). The input-output relationship in first language acquisition. *Language and Cognitive Processes*.

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

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.

Brown, R. (1973). A first language. Boston, MA: Harvard University Press.

Freudenthal, D., Pine, J.M. & Gobet, F. (2002). 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. (2005a). Simulating Optional Infinitive Errors in Child Speech through the Omission of Sentence-Internal Elements. *This Volume*.

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*.

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*.

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

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

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

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

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