Modelling the Development of Dutch Optional Infinitives in MOSAIC

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

Abstract

This paper describes a computational model which simulates the change in the use of optional infinitives that is evident in children learning Dutch as their first language. The model, developed within the framework of MOSAIC, takes naturalistic, child directed speech as its input, and analyses the distributional regularities present in the input. It slowly learns to generate longer utterances as it sees more input. We show that the developmental characteristics of Dutch children's speech (with respect to optional infinitives) are a natural consequence of MOSAIC's learning mechanisms and the gradual increase in the length of the utterances it produces. In contrast with Nativist approaches to syntax acquisition, the present model does not assume large amounts of innate knowledge in the child, and provides a quantitative process account of the development of optional infinitives.

The Optional Infinitive Stage

One phenomenon which has received considerable attention in the area of syntax acquisition is the so-called *Optional Infinitive (OI) stage* (Wexler, 1994, 1998). Children in the OI stage of development use a high proportion of (root) infinitives, that is, verbs which are not marked for tense or agreement. In English, root forms such as *go*, or *eat* are infinitive forms, whereas *ate* or *goes* are marked for tense and agreement + tense respectively. Verbs which are marked for agreement or tense are known as *finite* verbs. (Technically, infinitives are a subclass of the class of *non-finite* verb forms, which also includes past participles and progressive particles).

Another feature of the OI stage is that children often omit subjects from their sentences. That is, children will produce utterances such as *throw ball* from which the subject (*I*) is absent. While the proportion of infinitives is (considerably) higher than for adult speech, children in the OI stage do show competence regarding other syntactic attributes of the language. Typically, children will not make errors in the basic verb-object order. English-speaking children, for instance, will say *throw ball*, but not *ball throw*. One puzzling feature of the OI stage is that children produce both inflected and uninflected forms in contexts requiring the inflected form, but do not produce finite forms in nonfinite contexts. The fact that children use both inflected and uninflected forms shows that it is not the case that they simply don't know the inflected forms.

The optional infinitive stage has been shown to occur in many different languages, which can differ considerably in their underlying syntactic properties, and children do show competence regarding these syntactic properties. Different languages also differ with respect to how pronounced the OI stage is. Since most verb forms in English are not distinguishable from non-finite forms, it is relatively difficult to distinguish optional infinitives from grammatically correct utterances. In other languages (e.g. Dutch), the number of unambiguously finite forms is larger, and as a result the optional infinitive stage is more pronounced.

Wexler (1998) has proposed a Nativist account of why children in the optional infinitive stage produce a large number of non-finite forms. In accordance with Chomsky's theory of Universal Grammar (Chomsky 1981), he theorizes that children in the optional infinitive stage actually know the full grammar of the language. The only thing they do not know is that Agreement and Tense are obligatory. This approach accounts for the fact that children produce both correct finite forms and incorrect (optional) infinitives. It also explains why children rarely produce other types of errors. Finally, its great strength is that it unifies across languages where children clearly use optional infinitives despite differences in their underlying grammar. However, there are also a number of problems with Wexler's account.

Firstly, Wexler's theory does not give a process account of developmental change in the use of optional infinitives. He assumes this to be due to *maturation*.

Secondly, the theory makes very limited quantitative predictions. It only predicts that the optional infinitive stage occurs, and that children will stop making optional infinitive errors at some point. It makes no specific predictions regarding the time course of this development, or related changes in other attributes.

Thirdly, the theory assumes a large amount of innate knowledge in the child (the theory assumes that the child does not know that inflection is obligatory, but otherwise knows the full grammar of the language).

An obvious alternative to Wexler's theory is that children learn the grammar of a language through exposure to that language. Wexler discounts this kind of learning-based approach on the grounds that the grammar is too difficult to learn, that the optional infinitive stage lasts too long (years), and that, although children produce both correct and incorrect forms, when they use finite forms, they use them correctly (Wexler, 1994).

In this paper, we aim to show that the dynamics of the optional infinitive phenomenon can be simulated using a simple learning mechanism which performs a distributional analysis of naturalistic input. Earlier versions of the model have already been shown to simulate the basic optional infinitive phenomenon in both English (Croker, Pine & Gobet, 2001) and Dutch (Freudenthal, Pine & Gobet, 2001). Whereas the earlier versions modelled one specific stage in development, the present model aims to simulate the *developmental change* that is apparent in the use of optional infinitives.

There are a number of reasons for choosing Dutch as the target language. Firstly, as was mentioned, in adult speakers' Dutch, unambiguous finite forms are far more frequent than they are in English. In English, in the present tense, only the third person singular can be distinguished from the infinitive form. In Dutch, the first, second and third person singular are unambiguously finite. If, for instance, an English speaking child produced *I throw ball*, it would be unclear whether the verb *throw* was an infinitive form. The Dutch equivalent *ik gooi bal* would be classified as a finite form, because *gooi* is different from the infinitive *gooien*. Thus, the number of unambiguously finite forms is larger in Dutch than in English. (This suggests that developmental change in the use of optional infinitives is likely to be more pronounced in Dutch than it is in English, which makes the simulation of Dutch child language more informative as a modelling exercise.) A second reason for using Dutch is that detailed data regarding this development are available. Wijnen, Kempen & Gillis (2001) have analysed the corpora of two Dutch speaking children and have shown that the proportion of root infinitives decreases from around 90% to roughly 10% between the ages 1;6 and 3;0. By comparison, root infinitives are used in less than 10% of adults' utterances. Wijnen et al. concluded that the frequency of occurrence of optional infinitives in the child's speech was related to frequency, and utterance position, as well as lexical transparency.

A third reason for choosing Dutch as the target language is that Dutch grammar is relatively complex when considering finiteness of verb forms. Dutch is what is known as an SOV/V2 language. This means that the verb in Dutch can take one of two positions, depending on its finiteness. A non-finite verb takes the sentence final position, whereas finite verbs take the second position. Therefore, in the sentence

Ik gooi een bal (1)
(I throw a ball)

the verb *gooi* (*throw)* is finite and takes second position. In the construction

Ik wil een bal gooien (2)
(I want a ball throw/ I want to throw a ball)

the verb *gooien* is a non-finite form, and takes sentence final position. (The auxiliary *wil* is finite and takes second position.) In English, which is an SVO language, verb position is not dependent on the finiteness of the verb. If a model is to learn from the distribution of naturalistic speech input, then the production of a large number of infinitives while respecting the overall grammar would appear to represent a greater challenge in Dutch than in English.

MOSAIC

MOSAIC (Model of Syntax Acquisition In Children) is an instance of the CHREST architecture, which in turn is a member of the EPAM (Feigenbaum & Simon, 1984) family of models. CHREST models have successfully been used to simulate novice-expert differences in chess (Gobet & Simon, 2000), as well as several phenomena in language acquisition (Jones, Gobet & Pine, 2000a, 2000b; Croker, Pine & Gobet, 2001, 2002; Freudenthal, Pine & Gobet, 2001, 2002). We will now give a brief description of MOSAIC. A more detailed description of the model can be found elsewhere in this volume (Freudenthal, Pine & Gobet 2002). The model we have used in these simulations is identical to the one that Freudenthal et al. (2002) used for the simulation of a different phenomenon (Subject Omission) in another language (English).

The basis of the model is a discrimination net, which is used to store the input that is fed to the model. The network is an n-ary tree which is headed by a root node. Utterances that the model sees are encoded by sequences of nodes in the network.

The model encodes the fact that word *a* has been followed by word *b* in the input by creating a node for word *b* under the node for word *a*. The fact that word *a* has preceded word *b* is similarly encoded. Fig. 1 may illustrate the basic MOSAIC network. Apart from the standard links between words that have followed each other in utterances previously encountered, MOSAIC also employs *generative links*. Generative links connect nodes that are distributionally similar. When two nodes (phrases) have a high likelihood of being preceded and followed by the same words in the input, a generative

link is created between them. Since distributionally similar phrases are likely to belong to the same word class, generative links that develop end up linking clusters of nodes that represent different word classes. The induction of word classes on the basis of co-occurrence statistics is the only mechanism that MOSAIC employs for representing syntactic rules. The main importance of generative links lies in the generation of utterances from the model. In generation, words that share a generative link can be substituted, thus allowing the model to generate novel utterances. Again, the reader is referred to Freudenthal, Pine & Gobet (2002) for details regarding generation. One point worth mentioning here is that the model will only output utterances that contain an end marker (i.e. where the utterance final phrase has occurred in a sentence final position in the input). Several authors have suggested that sentence final position is particularly salient, and that children are more likely to produce utterances that have occurred in sentence final position (Shady & Gerken, 1999; Naigles & Hoff-Ginsberg, 1998).

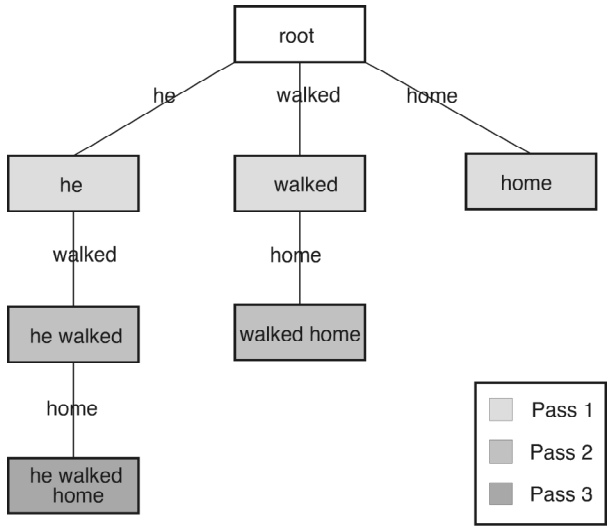

Fig. 1: MOSAIC learning an input

The model we used for these simulations is an extension of that used in Freudenthal, Pine & Gobet, (2001), which simulates the children's performance in Dutch at one specific point in time. This version of the model has also been shown to produce both root infinitives and correct inflected forms in English (Croker, Pine & Gobet, 2001). The main difference between this and the previous version of the model is that the present model learns much more slowly. By using a slow learning rate, and iteratively feeding input to the model and analysing its resulting output, we were able to model consecutive stages of development. In the previous version, a word was encoded on the first occasion it was seen, which resulted in a model with an MLU (Mean Length of (output) Utterance), that was comparable to that of a child that has passed the OI stage. In the present version, the probability of creating a node is dependent on the size of the network (a measure of the linguistic knowledge or vocabulary size of the child), and the length of the phrase that is being encoded. More specifically, the probability of creating a node is given by the following formula:

$$NCP = \left(\frac{*nodes\_in\_net*}{50{,}000} \right)^{length\_phrase}$$

It will be apparent from the formula above that the probability of creating a node is very low if the network is small (i.e., the number of nodes in the net is low). As the number of nodes in the net grows, this probability will increase. A second point to note is the occurrence of the length of the phrase (number of words) in the exponent. This has the effect of lowering the probability of creating nodes that encode longer phrases. The value 50,000 has been chosen somewhat arbitrarily. Its main role is to ensure that the difference in node creation probability for short and long utterances decreases as a function of the size of the net. As the number of nodes in the net approaches 50,000 (a typical number for a *saturated* model given the Dutch input used here), the base number in the formula approaches one, and thus the weight of the exponent diminishes. One additional remark must be made about this formula: phrases that occurred in utterance final position (i.e., contained an *end marker*), were treated differently from other utterances in that their length (for calculation of the NCP) was decreased by 0.5. This constitutes an *end marker bias* in learning, rather than at production. It has been argued that utterance final phrases are learned more easily than non-utterance final phrases (Wijnen, Kempen & Gillis, 2001).

The Simulations

The data that were simulated were taken from Wijnen, Kempen & Gillis (2001). Wijnen et al. analysed two Dutch corpora of child and adult speech (the corpora of Matthijs and Peter and their mothers). The corpora consisted of transcribed tape recordings of speech between mother and child. For Matthijs, the recordings were made between the ages 1;9 and 2;11. For Peter they were made between 1;7 and 2;3. The children's MLU (Mean Length of Utterance) ranged from 1 to roughly 3. Wijnen et al. analysed the corpora with respect to the presence of the optional infinitive phenomena in both the mother's and the children's speech. On the basis of the children's data, four developmental stages were identified, and the proportion of finite, non-finite and discontinuous finites

(see below) was assessed. Since the corpora that Wijnen et al. analysed are available in the CHILDES data base (MacWhinney & Snow, 1990), we had access to the same corpora, and used these (maternal corpora) as input for the model.

In order to compare the output of the model to the children's speech, we ran the input through the model several times. After each run of the model, we generated output, and compared the MLU of the model with the child's MLU in the developmental stages that Wijnen et al. identified. We then selected for further analysis those output files that most closely matched the children's MLU for the four developmental stages. The actual analysis performed was similar to that of Wijnen et al. Firstly, we selected those utterances that contained one or more verb forms. We then classified these utterances as finite, non-finite or discontinuous finite. In doing so, we used the following criteria:

- An utterance is considered *non-finite* if it contains only non-finite verb forms.
- An utterance is considered *finite* if it contains only finite verb forms.
- An utterance is considered a *discontinuous finite* if it contains both a non-finite, and a finite form (e.g. a finite auxiliary).

There were some small differences from Wijnen et al.'s analysis. The most notable difference is that Wijnen et al. removed all forms resembling imperatives, starting with the early two word stage. When coding actual speech, this is relatively easy to do, since context allows one to disambiguate. Since the model's output does not provide this context, the classification remains somewhat ambiguous. We therefore decided not to remove forms resembling imperatives.

Results

Figure 1 shows the data and the simulations for Matthijs and Peter. The model shows a considerable drop (around 50%) in the proportion of non-finites for both input sets. For the children, the corresponding drop is 80-85%. Given the fact that we are using naturalistic input to model the development of children's speech, and the fact that we used an identical model for both children (i.e. no parameters were adjusted) we consider

Fig. 2a: Data for Matthijs

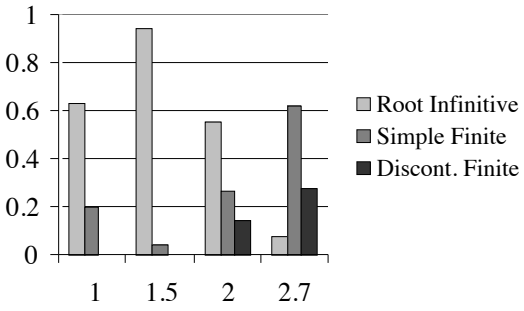

Fig. 2b: Model for Matthijs

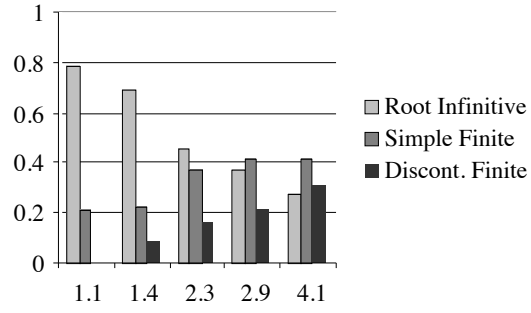

Fig. 2c: Data for Peter

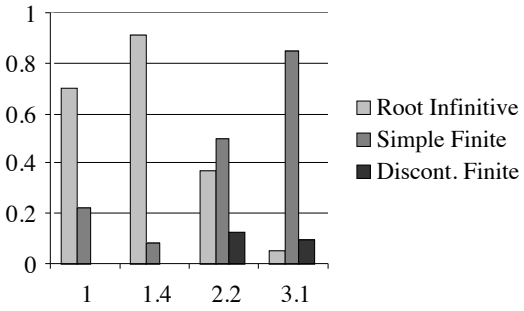

Fig. 2d: Model for Peter

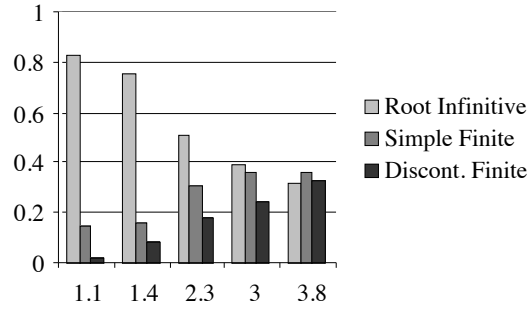

Figure 2: Distribution of root infinitives and (discontinuous) finites as a function of MLU for Matthijs, Peter, and their respective model.

this figure promising. (Note however, that we report five rather than four data points for the models. The last data point reflects an MLU larger than that for the children in the final stage, and is included to show that the proportion of non-finites continues to decrease.)

What mechanism is responsible for this drop in the model's output? The thing to note is that non-finite forms take sentence-final position in Dutch, and that the model is biased towards generating (and encoding) phrases that occurred in sentence-final position. The formula for calculating the node creation probability ensures that early on, the model will encode relatively short utterances that occurred in sentence-final position. If these utterances contain a verb, it will (in Dutch) most likely be a non-finite form. These non-finite forms may have been part of an auxiliary + verb construction (e.g. *He wants to build a house*). Since the model can generate partial utterances, it can learn the root infinitive *build a house* from this (discontinuous) finite form. Therefore, a high proportion of non-finite forms is expected in the early stages of the model's development. As the model sees more and more utterances, the number of nodes in the net will increase, and the probability of creating a node will also increase. As a result, longer and longer utterances will be encoded in the network. As the encoded utterances increase in length, they will be more likely to include words that occur early in the utterance. Since finite forms take second position in Dutch, the number of finite forms will increase as the model starts generating longer utterances. Note that this also means that root infinitives will slowly be replaced by discontinuous finites. Where the model may have output the root infinitive *build a house* early on, it will be able to output the discontinuous finite *he wants to build a house* as the size of the net increases.

Table 1: Proportion of correct Object-Verb orderings for the model as a function of finiteness (averaged over developmental phase).

	Finites	Non-Finites
Matthijs	.94	.91
Peter	.96	.93

Given that the model simulates the basic optional infinitive phenomenon, we now need to assess whether it conforms to the other criteria of the optional infinitive stage. Tables 1 and 2 show the proportion of correct verb placement and the position of the object relative to the verb. It is evident, that, in the majority of cases, the model uses the correct placement, indicating that it is sensitive to basic Dutch grammar.

The fact that the model gets the basic word order right in the majority of the cases is perhaps not very surprising. After all, the input that the model learns from has the correct word order. This is not a trivial result however, as the fact the children correctly produce the correct word order has been taken as evidence by Wexler (1994, 1998) that the child knows the actual grammar.

Table 2: Proportion of correct verb placement for the model as a function of finiteness (averaged over developmental phase).

	Finites	Non-Finites
Matthijs	.85	.95
Peter	.88	.97

Though these results are very promising, especially considering the fact that we are using naturalistic input to simulate actual children's speech, some issues require attention. For both children, the proportion of non-finites is underestimated for stage 2, and overestimated for the later stages. Possible causes for the underestimation in the early stages may lie in the fact that Wijnen et al. removed forms resembling imperatives as of stage two (which may also explain the relatively low proportion of non-finites in stage one in the data). We did not do this. This underestimation may be exacerbated by the fact that the model produces relatively few utterances early on, thus making it relatively sensitive to small changes. A second, possibly more likely cause may be that there are additional factors that cause the high proportion of non-finites in the children. Wijnen et al. claim, on the basis of a regression analysis, that frequency of occurrence alone is not enough to explain the high incidence of non-finite forms. They suggest that non-finite forms are learned more easily and attribute this to lexical transparency. Since MOSAIC does not employ any semantics, we cannot model this effect. Regarding the later stages, one possible cause for the overestimation is the fact that MOSAIC has a limited ability to unlearn. That is, at any stage, when the model generates output, it will generate all the utterances it can. Thus, once the model has learnt to generate *he wants to build a house*, it will also (still) generate *build a house*.

Mechanism for change

The model shows a drop in the proportion of non-finites of roughly 50%. We can now ask ourselves what has caused this change. Two possible explanations come to mind. Firstly, as the model learns, the MLU of the generated utterances increases. As explained earlier, if the generated utterances adhere to Dutch grammar, an increase in the proportion of finites is expected. A second possible cause lies in the proportion of generated (rather than rote learned) utterances. As the model's MLU increases, so does the proportion of generated utterances. This may result in a

disproportionate growth in the number of finite utterances. (Since finite forms are more frequent, a relatively large proportion of the generated utterances contain finite verbs.) While a regression analysis showed that the increase in MLU alone explained 90% of the variance in the proportion of finite utterances, and the proportion of generated utterances explained an additional 6%, the correlation between generativity and MLU was relatively large, which might decrease the sensitivity of this analysis. We therefore assessed the proportion of non-finites in rote utterances only. This increased the proportion of non-finites in the last stage by 10% for Peter's model, and by 20% for Matthijs' model. Apparently, the role of generativity is greater than the regression analysis suggests.

Conclusions

The model described in this paper clearly captures the development that is evident in Dutch children's use of infinitive verb forms. In doing so, the model provides both a process model, and a quantitative account of this transition. Furthermore, it shows that a considerable portion of the drop in non-finite forms can be explained by a learning mechanism that emphasizes utterance final phrases, and an increase in MLU, although the process is likely to be augmented by other considerations (as witnessed by the relatively poor fit for the very early and late stages). While it does not solve the learnability problem, and as such is probably too simplistic a model of syntax acquisition, the present simulations clearly show that the Optional Infinitive phenomenon does not, in itself constitute evidence for the innateness of syntactic knowledge. As such, it supports the suggestion that children's sensitivity to the distributional characteristics of their linguistic environment may aid them in learning their native language. In order to further test this suggestion, it will be necessary to assess to what extent the present findings generalise to other languages. This may also suggest possible extensions to the model.

Acknowledgements

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

References

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

Croker, S., Pine, J.M., & Gobet, F. (2000). Modelling optional infinitive phenomena: A computational account. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the Third International Conference on Cognitive Modelling*. Veenendaal: Universal Press.

Croker, S., Pine, J.M. & Gobet, F. (2001). Modelling children's case-marking errors with MOSAIC. In E.M. Altmann, A. Cleeremans, C.D. Schunn & W.D. Gray (Eds.), *Proceedings of the Fourth International Conference on Cognitive Modeling*. Mahwah, NJ: LEA.

Feigenbaum, E.A. & Simon, H.A. (1984). EPAM-like models of recognition and learning. *Cognitive Science*, 8, 305-336.

Freudenthal, D., Pine, J. & Gobet, F. (2001). Modeling the optional infinitive stage in MOSAIC: A generalisation to Dutch. In E.M. Altmann, A. Cleeremans, C.D. Schunn & W.D. Gray (Eds.), *Proceedings of the Fourth International Conference on Cognitive Modeling*. pp. 79-84. Mahwah, NJ: LEA.

Freudenthal, D. Pine, J. & Gobet, F. (2002). Subject omission in children's language: The case for performance limitations in learning. *This Volume*.

Gobet, F. & Simon, H.A. (2000). Five seconds or sixty: Presentation time in expert memory. *Cognitive Science*, *24*, 651-682.

Jones, G., Gobet, F. & Pine, J.M. (2000a). A process model of children's early verb use. In L.R. Gleitman & A.K. Joshu (Eds.), *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*. pp. 723-728. Mahwah, N.J.: LEA.

Jones, G., Gobet, F. & Pine, J.M. (2000b). Learning novel sound patterns. In N. Taatgen & J. Aasman (Eds.), *Proceedings of the Third International Conference on Cognitive Modelling* (pp.169-176). Veenendaal: Universal Press.

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

Naigles, L. & Hoff-Ginsberg, E. (1998). Why are some verbs learned before other verbs. Effects of input frequency and structure on children's early verb use. *Journal of Child Language*, 25, *95-120*.

Shady, M. & Gerken, L.(1999). Grammatical and caregiver cue in early sentence comprehension. *Journal of Child Language*, 26, *163-176*.

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.