The Role of Input Size and Generativity in Simulating Language Acquisition

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 of Nottingham, University Park
Nottingham, NG7 2RD United Kingdom

Abstract

This paper presents an analysis of the role of input size and generativity (ability to produce novel utterances) in simulating developmental data on a phenomenon in first language acquisition. An existing model that has already simulated the basic phenomenon is trained on input sets of varying sizes (13,000 to 40,000 utterances). The ability of the model to produce novel utterances is also manipulated. Both input size and generativity affect the fits for later stages of development. Higher generativity improves fits for later stages, but worsens them for early stages, suggesting generativity is best increased as a function of mean length of utterance (MLU). The effect of training set is variable. Results are discussed in terms of optimal training sets for simulations, and children's developing ability to produce utterances beyond the input they have heard.

Introduction

In recent years, computational models have been shown to successfully simulate phenomena in language acquisition, thereby providing evidence for the claim that children may acquire language largely through input driven learning (Cartwright & Brent, 1997; Redington, Chater & Finch,.1998) However, as Christiansen & Chater (2001) point out, one of the major challenges facing computational approaches to syntax acquisition is to develop models that map more directly onto the task that human language learners actually face, and produce output that can be more directly compared with the output that children produce. Computational models of language acquisition should therefore ideally learn off input that closely resembles the input that children are exposed to, and produce actual utterances as output.

MOSAIC (Model of Syntax Acquisition in Children) is a model that attempts to meet this challenge by taking corpora of real, child-directed speech as input, and learning to produce progressively longer utterances that can be directly compared with individual children's speech at different points in development. In its present form, MOSAIC simulates a number of phenomena in syntactic development, including the Optional Infinitive (OI) phenomenon in English and Dutch, and Subject Omission in English (Croker, Pine & Gobet, 2000, 2001; Jones, Gobet & Pine, 2000; Freudenthal, Pine & Gobet, 2001, 2002a, 2002b). The model is a simple distributional analyser which is able to produce utterances that were not present in the input by interchanging words that occur in distributionally similar contexts. This paper focuses on how this ability to produce novel utterances and the size of the model's input affect the model's fit to data from the OI phenomenon.

The OI phenomenon revolves around the notion that children in several languages produce non-finite verbs in contexts where the adult grammar requires a finite verb form[1] (Wexler, 1994). English speaking children of around 2 to 3 years of age for instance, produce utterances such as *He go*, or *That go there*. As children grow older, they make fewer and fewer of these optional infinitive errors. While children produce such errors in several languages, the developmental dynamics of the phenomenon differ from language to language. In Dutch, children start out producing around 90% infinitive utterances, but this proportion drops to under 20% by the time the child's Mean Length of Utterance (MLU) is around three words. This large developmental variation in the provision of OI utterances makes Dutch a strong test for a model that simulates the phenomenon. Figure 1, which portrays the data[2] and simulations for one of the two children reported in Freudenthal, Pine & Gobet (2002a) shows that MOSAIC already simulates the basic developmental dynamics of the OI phenomenon in Dutch. However, the fit for the very early and late stages could be improved. For the later stages, the proportion of non-finite utterances remains too high, while the ratio of simple to compound finites is too low[3].

[1] Verb forms are classified into non-finites and finites. Finite verb forms are forms that are marked for Agreement and/or Tense (e.g. *goes*, *went*). Non-finite verb forms do not carry this marking. Non-finites include the infinitive (*go*), the past participle (*gone*) and the progressive (*going*).

[2] Figure 1a deviates slightly from the data presented in Freudenthal et al 2002a. The data from that paper were taken from Wijnen et al (2001), whose analysis differed slightly from the analysis performed on the model. The data reported in Figure 1a were analysed in the same way as the model.

[3] Non-finite utterances contain only non-finite verb forms (*That go there*). Simple finites contain only finite verb forms

MOSAIC simulates the basic phenomenon because it is a performance limited distributional analyser which is biased towards producing phrases that have occurred in a sentence final position. Since MOSAIC learns slowly, it learns to encode progressively longer utterances. In Dutch, in main clauses, non-finite verb forms take sentence final position, whereas finite verb forms take second position. Due to its slow learning, MOSAIC starts out producing very short utterances. Since these utterances have occurred in sentence final position, they are likely to contain only non-finite verb forms. As the utterances that MOSAIC produces become longer, finite verb forms, which feature early in the utterance, start coming in. Hence, non-finite utterances are slowly replaced by simple and compound finites.

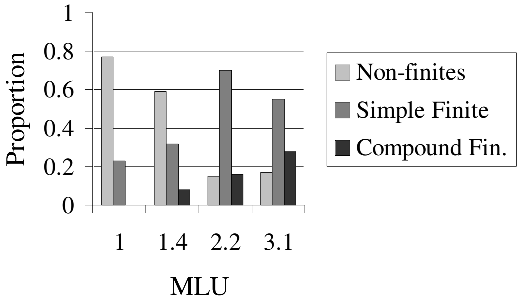

Fig. 1a: Data for Peter

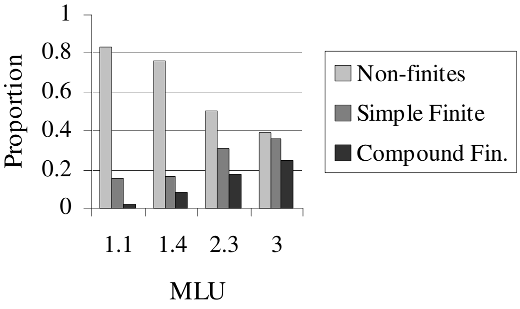

Fig. 1b: Simulations for Peter

This paper is aimed at investigating whether the fit for the later stages can be improved by a) manipulating the size of the input the model is trained on, and b) lifting constraints on the model's ability to produce novel utterances. The data portrayed in Figure 1 were obtained by training the model on an input set of roughly 13,000 utterances. This constitutes the entire available corpus of Child Directed Speech for one child. In order to analyse the output in different developmental phases, the input was fed through the model several times, and output was produced after every run of the model. Training the model on larger corpora is likely to affect the fit, as larger samples approximate the input received by the child more closely. More specifically, smaller input sets tend to overestimate the relative frequency of occurrence of infrequent items (Richards, 1987). Since (in Dutch) non-finite verb forms are less frequent than finite verb forms, a model trained on a larger input set may produce fewer non-finite utterances.

Allowing the model to produce more novel utterances is likely to increase the proportion of finite utterances, since it has already been shown that the proportion of novel utterances explains variance in the proportion of non-finite utterances when controlling for MLU (Freudenthal, Pine & Gobet, 2002a). Systematically manipulating these two variables can shed light on how a distributional learning mechanism and its ability to produce novel utterances interact with the characteristics of the input in the simulation of developmental data.

MOSAIC

The basis of the model is an n-ary discrimination net headed by a root node. Training of the model takes place by feeding utterances to the network, and sorting them. Utterances are processed word by word. When the network is empty, and the first utterance is fed to it, the root node contains no test links. When, for example, the model is presented with the utterance *He walked home*, it will create on its first pass three test links from the root. The test links hold a key (the test) and a node. The key holds the actual feature (word or phrase) being processed, while the node contains the sequence of all the keys from the root to the present node. Thus, on its first pass, the model just learns the words in the utterance. When the model is presented with the same utterance a second time, it will traverse the net, and find it has already seen the word *he*. When it encounters the word *walked* it will also recognize that it has seen this word before, and will then create a new link under the *he* node. This link will have *walked* as its key, and *he walked* in the node. In a similar way, it will create a *walked home* node under the primitive *walked* node. On a third pass, the model will add a *he walked home* node under the *he walked* chain of nodes. The model thus needs three passes to encode a three-word phrase when all of the words are new. As the model sees more input, it will encode larger and larger phrases. If word *a* is followed by different words in different contexts, this is encoded by creating multiple nodes under that for word *a*. In a similar way, the model encodes what words have preceded a word.

Apart from the standard test links between words that have followed each other in utterances previously encountered, MOSAIC employs *generative* links that connect nodes that are distributionally similar (have

(*He walks*), while compound finites contain finite as well as non-finite verb forms (*He wants to sleep, He has walked*).

occurred in similar contexts). Generative links can be created on every cycle (after an utterance has been processed). Whether a generative link is created depends on the amount of overlap that exists between nodes. The overlap is calculated as the percentage of shared nodes above and below the target nodes. This is equivalent to assessing how likely it is that the two words are preceded and followed by the same words in the input. Two words that tend to occur in similar contexts, will share a large proportion of nodes above and below them. At present, the proportion of shared nodes needs to exceed 10% in order for two nodes to be linked. Since words that are followed and preceded by the same words are likely to be of the same word class (for instance Nouns or Verbs), the generative links that develop end up linking clusters of nodes that represent different word classes. The (implicit) induction of word classes on the basis of their position in the sentence relative to other words is the only mechanism that MOSAIC uses for representing syntactic rules.

The main importance of generative links lies in the role they play when utterances are produced from the network. When the model produces utterances it will output all the utterances it can by traversing the network until it encounters a terminal node. MOSAIC will only produce an utterance when this terminal node contains an *end marker*, indicating that the final word in the utterance has occurred in a sentence final position. When the model traverses standard links only, it produces utterances or parts of utterances that were present in the input. In other words, it does *rote* production. During production, however, the model can also traverse generative links. When the model encounters a node with a generative link, it can substitute the contents (technically the *key*) of the linked node for the contents of the current node. As a result, the model is able to produce utterances that were not present in the input. Figure 2 gives an example of the production of an utterance using a generative link.

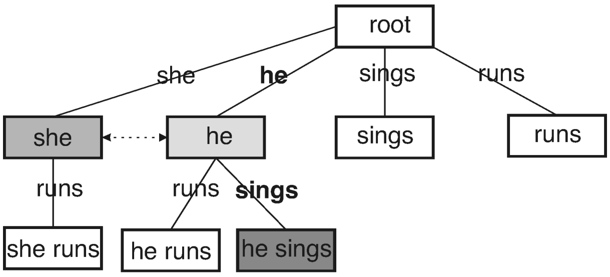

Fig. 2: Producing an utterance. Because *she* and *he* have a generative link, they can be interchanged. The model can therefore output the novel utterance *she sings* when it has only seen *he sings*. (For simplicity, preceding nodes are ignored in this figure).

So far, we have described the model as used in Freudenthal, Pine & Gobet (2001), which simulates the children's performance at one specific point in time. The model used to obtain the results portrayed in figure 1b is an extension of this model. The main difference is that the extended model learns much more slowly. By using a slow learning rate, and iteratively feeding input through the model and analysing its resulting output, it is possible 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 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 low if the network is small. As the number of nodes in the net grows, this probability increases. 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 so far), 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 occurr in utterance final position (i.e., contain an *end marker*), are treated differently from other utterances in that their length (for calculation of the NCP) is decreased by 0.5. This constitutes an *end marker bias* in learning. It has been argued that utterance final phrases are learned more easily than non-utterance final phrases (Wijnen, Kempen & Gillis, 2001; Shady & Gerken, 1999).

The Manipulations

This paper focuses on how generativity (the model's ability to produce novel utterances) and the size of the model's input affects the fit to the data for later stages of development. These manipulations are discussed next.

Generativity

Generativity in the earlier model was limited in two ways. Firstly, the model could only traverse one generative link per generated utterance. Secondly, the model could only traverse a generative link at the beginning of an utterance. For the present simulations, these limitations were lifted, so the model could take multiple generative links, at any place in the utterance. This manipulation is likely to affect the proportion of non-finites since finite verbs (in Dutch) are more frequent, and therefore more likely to have generative links. Previous analyses have also shown that, when employing the constraints on generativity, the proportion of generated utterances explains variance in the proportion of non-finites over that explained by MLU.

A specific question regarding this manipulation is whether the expected increase in finite utterances varies with MLU. Since the model is now able to traverse a generative link anywhere in the utterance, the *potential* number of utterances that can be generated off one rote-learned utterance increases exponentially with the length of the rote-learned utterance. If there is such an exponential increase in the actual proportion of generated utterances, this is likely to decrease the proportion of non-finites for the later stages only. If, on the other hand, the increase in generated utterances varies linearly with MLU, the proportion of non-finite utterances will decrease for the early stages as well. This would decrease the fit for the earlier stages. However, the effect of the manipulation is clearly dependent on the characteristics of the input, and therefore remains an empirical issue.

Input Size

The previous simulations employed all the child directed speech that was available for the children being simulated. These input sets (approximately 13,000 utterances) were then fed through the model multiple times. Children however, are obviously exposed to far greater numbers of utterances, and feeding a relatively small input set through multiple times is likely to overestimate the occurrence of low frequency items (Richards, 1987). In the present research input sets that might be more realistic in size and frequency distribution were created. Since the input sets that were used for the previous simulations consisted of all the available data for these children, input sets were created by aggregating the child directed speech for several children from the 'Groningen corpus'. Two new input sets of 27,000, and 40,000 utterances were created. These input sets constitute the aggregate sets of Peter and Matthijs (27,000), and a random sample of 40,000 utterances from all the seven children in the Groningen corpus. This sample will be referred to as the 'Half-Groningen corpus'. Note that the construction of large input sets by aggregating the data from several children precludes a comparison between the simulations and the individual children. Also, since the child directed speech for the different children may have different distributional characteristics, the manipulation is not merely a manipulation of input size, but one of variability as well. However, since larger input sets for individual Dutch children are not available in CHILDES, this is the only available method to create large input sets.

The Simulations

Simulations were run with and without the limitations in generativity. Thus, the model could either take only one link at the beginning of the utterance, or multiple links at any position in the utterance. The model with constraints was identical to the one that obtained the results in figure 1b. The simulations were run in a similar way to the earlier simulations. For Matthijs and Peter, the full input sets were fed through the model multiple times. The output sets that most closely matched the children's MLU in the last stage were then selected for analysis. For the larger input sets, a random sample of 15,000 utterances was selected for every run of the model. This was done in an attempt to make the number of runs for the different simulations comparable. Since the sample of 15,000 utterances was selected randomly, the likelihood of feeding the model the same utterances several times decreases as a function of the size of the input. The number of runs required to train the model up to the required MLU ranged from 11 (Matthijs+Peter employing multiple links) to 17 (for Peter's model employing one link).

Results

Figure 3 shows how the two manipulations affected the proportion of non-finites. The MLUs for the different simulations ranged from 2.74 to 2.89. Figure 3 clearly shows that the proportion of non-finites decreases as the constraints on generativity are relaxed, with the best results being obtained with the largest input set. The proportion of non-finites has dropped to 24%, which is quite close to the 17% that the actual child produces.

When we look at the effect of input size, however, the results are somewhat mixed. The model trained on Matthijs+Peter's input falls in between those for Peter and Matthijs, suggesting that the model for Matthijs+Peter is an average of the two. Results for the larger sample are better, especially for multiple links. In principle however, this may be due to the fact that this input set is made up of the child directed speech for several children. Matthijs and Peter's input sets may simply have a higher proportion of non-finites. Since the other children of the Groningen corpus have not been modelled separately, this explanation cannot be discounted.

The decrease in the proportion of non-finite utterances for the models with multiple links is caused by the greater generativity of these models. On average, 45% of the output of the models employing one link is novel. For the models employing multiple links, 62% of the output is novel. The proportion of generated utterances affects the proportion of non-finites, as novel utterances contain more finites. On average, 22% of the generated utterances are non-finites. For rote utterances in contrast, 64% are non-finite.

Thus, since generated utterances tend to be finite, an increase in the proportion of generated utterances necessarily means a decrease in the proportion of non-finite utterances.

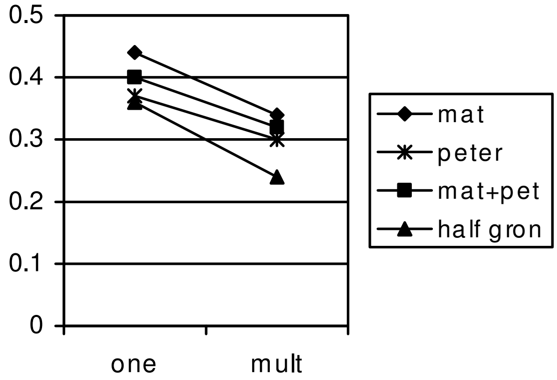

Fig.3: Proportion of non-finites in the output as a function of input set, and number of generative links allowed per utterance

Figure 4 displays the ratio of simple to compound finites for the different manipulations. Again, the results are somewhat mixed, but employing multiple links does increase the ratio of simple to compound finites for larger input sets. The ratio for the Half Groningen corpus (multiple links) is 2.45, slightly higher than it is for the child (1.96).

The reason for the higher ratio of simple to compound finites lies in the fact that the model obtains relatively low proportions of non-finites at a relatively low MLU. Since compound finites contain (a minimum of) two verbs, there can be few compound finites at an MLU of approximately 2.8.

Having improved the fit for the later stages, we can turn to the question of whether the fit for the earlier stages remains unaffected. It turns out that this is not the case. In the simulations that resulted in the best fit for the final stage (Half Groningen corpus, multiple links) the proportion of non-finites has dropped considerably, even at an MLU of 1.4. The reason for this early drop is that, when employing multiple links (and a large input set), the model becomes too generative early on. At an MLU of 1.4, the model produces 32% novel utterances, which tend to be finite. By comparison, Peter's model using one link produces 13% novel utterances at a similar MLU.

Relaxing the constraints on generativity therefore appears to decrease the proportion of non-finites across all developmental stages, rather than just in the later stages. There is however, a theoretically plausible way in which this could be remedied. Generativity at present is affected only by the distributional characteristics of the input. Using the overlap parameter that governs the creation of generative links, it is possible to gradually increase the generativity of the model. The overlap parameter could be made dependent on the size of the network or MLU. This would result in the model starting out producing rote-learned utterances only, and only starting to produce novel utterances once a reasonably large vocabulary has been established.

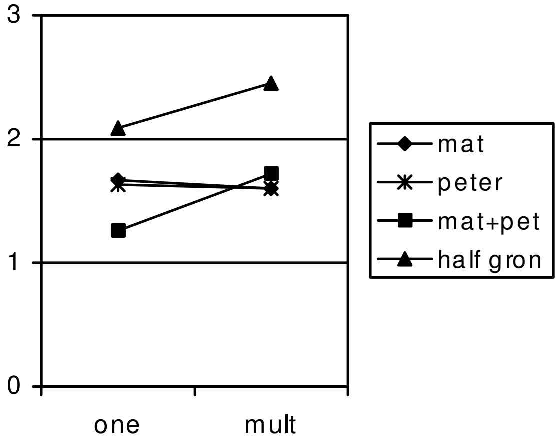

Fig.4: Ratio of simple finites to compound finites in the output as a function of input set, and number of generative links allowed per utterance.

Conclusions

This research focused on whether manipulations of input size and the ability of the model to generate novel utterances affect the fit of the simulations to the later stages of the Optional Infinitive phenomenon. Results regarding input size were somewhat mixed. While the best fits were obtained using the largest input set, increasing the size of the input sets did not unequivocally improve the fit. One possible reason for this may be that simulations using input sets for individual children were compared with simulations using aggregate input files. These aggregate files may have different distributional characteristics. In fact, the MLU for the largest input set was approximately 3.2 compared to 3.8 for the smaller sets. These findings have implications for models of syntax acquisition beyond MOSAIC, as they suggest that the use of Child Directed Speech does not in itself guarantee high input representativeness. A better approach would therefore be to obtain larger samples for individual children. This has the added advantage that it is possible to compare the model's output with the simulated child. While the corpora for Peter and Matthijs are amongst the largest Dutch corpora for individual children available in CHILDES, we have recently obtained access to considerably larger German corpora (approximately 500,000 utterances). Since German is identical to Dutch with respect to verb placement and its dependency on finiteness, we hope to address this issue in a more controlled manner in the future.

Thus, while input characteristics clearly affect the outcome of the simulations, more rigorously controlled input sets are required to unambiguously ascertain the relation between input size and fit.

Allowing the model to traverse multiple links during generation clearly improves the fit for the later stages, both in terms of proportion of non-finites and in terms of the ratio of simple finites to compound finites. For all simulations, the RMSEs with the children's data improved considerably when allowing the model to take multiple links. However, this decreased the fit for the early stages of development. While it was suggested that the interaction between the generativity mechanism and the learning mechanism might result in an exponential growth of generativity, this growth appears to be more linear in nature. The ability to generate novel utterances therefore needs to be made dependent on the actual knowledge encoded in the net. It was argued that this could be achieved by having the model start out needing a high degree of overlap between two nodes in order to create a generative link, and gradually decreasing this overlap percentage as the model learns to produce longer utterances. This suggests that the developmental patterning of children's use of (root) infinitives may be partly shaped by children's increasing ability to generalise beyond the input they have heard. Several authors have made a similar suggestion (e.g. Tomasello, 2000; Marchman & Bates, 1994).

Acknowledgements

This research was funded by the ESRC under grant number R000223954

References

Cartwright, T.A. & Brent, M.R. (1997). Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition*, *62,* 121-170

Christiansen, M.H. & Chater, N. (2001). Connectionist psycholinguistics: capturing the empirical data. *Trends in Cognitive Science* 5, 82-88.

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.* pp. 55-60. Mahwah, NJ: Lawrence Erlbaum Associates.

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. (2002a). Modelling the development of Dutch Optional Infinitives in MOSAIC. In W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Meeting of the Cognitive Science Society*. pp.328-333.

Freudenthal, D., Pine, J. & Gobet, F. (2002b). Subject omission in children's language: The case for performance limitations in learning. In W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Meeting of the Cognitive Science Society* pp. 334-339.

Jones, G., Gobet, F. & Pine, J.M. (2000). 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.

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

Marchman, V. & Bates, E. (1994). Continuity in lexical and morphological development: A test of the ciritcal mass hypotheses. *Journal of Child Language*, *21*, 339-366.

Redington, M., Chater, N. & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science 22*, 425-469

Richards, B. (1987). Type/Token Ratios: What do they really tell us. *Journal of Child Language, 14*, 201-209.

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.

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

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, *74*, 209-253.