

Modelling the optional infinitive stage in MOSAIC: A generalisation to Dutch

Daniel Freudenthal (DF@Psychology.Nottingham.Ac.Uk)

Julian M. Pine (JP@Psychology.Nottingham.Ac.Uk)

Fernand Gobet (FRG@Psychology.Nottingham.Ac.Uk)

Department of Psychology, University of Nottingham

University Park Nottingham, NG7 2RD, UK.

Abstract

This paper presents a model of a stage in children's language development known as the optional infinitive stage. The model was originally developed for English, where it was shown to provide a good account of several phenomena. The model, which uses a discrimination network, analyzes the distribution of words in the input, and derives word classes from them by linking words that are used in a similar context. While the earlier version of the model is sensitive only to characteristics of phrases that follow target words, the present version also takes preceding input into consideration. Also, the present version uses a probabilistic rather than a deterministic learning mechanism. Generalisation of the model to Dutch is considered a strong test of the model, since Dutch displays the optional infinitive phenomenon, while its syntax differs substantially from that of English. The model was presented with child-directed input from two Dutch mothers, and its output was compared to that of the respective children. Despite the fact that the model was developed for a different language, it captures the optional infinitive phenomenon in Dutch as it does in English, while showing sensitivity to Dutch syntax. These results suggest that a simple distributional analyzer can capture the regularities of different languages despite the apparent differences in their syntax.

Introduction

Theories of language acquisition can be roughly divided into *nativist* and *constructivist* theories. A central tenet of nativist theories is that children come into the world equipped with universal knowledge about grammars, and they then have to learn parameter settings for the specific language they are exposed to (Chomsky, 1981). One reason for assuming this innate knowledge is the fact that the input to the child is *underspecified*. That is, the number of legal utterances in a grammar is limitless, yet the child learns to produce legal utterances with exposure to only a limited set of utterances. Since children are able to generate new legal utterances, the reasoning is, they must have represented the rules that govern the legality of an utterance. It is furthermore assumed that these rules are too complex for a child to learn; therefore, they must be innate.

Constructivist theories, on the other hand, do not assume a large amount of knowledge being present at

birth, but assume that most of the syntactic knowledge is acquired as a result of exposure to a specific language. A challenge to constructivist theories is to provide general-purpose learning mechanisms which can acquire the grammars of different languages despite their apparent differences.

This paper aims to show that MOSAIC, a constructivist model of syntax acquisition which was developed to model and explain certain phenomena in English, can do a good job of modelling similar phenomena in Dutch, despite the syntactic differences between these two languages. The model takes as its input child-directed speech from mothers, and builds a representation of the syntax of the language by analysing the distribution of instances of words in the language. After the model has processed the input, it can generate utterances which were not present in the original input. The output of the model is then compared to children's speech. This paper addresses the adequacy of the model in simulating the *optional infinitive* stage in Dutch.

The Optional Infinitive Stage

One phenomenon which has received a considerable amount of attention in the area of syntax acquisition is the so-called optional infinitive stage (Wexler, 1994, 1998). Children in the optional infinitive stage 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 *jump* are infinitive forms, whereas *goes* or *jumped* are marked for agreement and 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* verbs forms, which also includes past participles and gerunds). The optional infinitive stage is furthermore characterized by the fact that the subject of the sentence is often dropped. That is, children will say things such as *throw ball*, deleting the subject (*I*). While the proportion of infinitives is (considerably) higher than for adult speech, children in the optional infinitive stage show competence regarding other syntactic attributes of the language. Typically, children will get the basic verb-object order right. English-speaking children, for instance, will say *throw ball*, but not *ball throw*. One puzzling feature of the optional infinitive

stage is that children produce both the inflected and infinitive forms, in a context requiring the inflected form without substituting finite forms in infinitive contexts.

Wexler (1998) proposes a nativist account of why children in the optional infinitive stage produce a large number of non-finite forms. 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 inflections for agreement and tense are obligatory. This approach accounts for the fact that children produce both correct finite forms and the incorrect (optional) infinitive. It furthermore explains why children rarely produce other types of errors. An obvious alternative to Wexler's account is a learning theory. On this account, children learn the grammar of a language through exposure to this language. Wexler discounts learning-based approaches on the grounds that the optional infinitive stage lasts too long (years), the fact that children produce both the correct and the incorrect form, and the claim that when children do use finite forms, they use them correctly (Wexler, 1994).

The optional infinitive stage is an interesting phenomenon to model, since it exists in many languages which may differ considerably in terms of other syntactic attributes. A strong test of a model of the optional infinitive stage, is to see whether the model correctly predicts the occurrence of the optional infinitive phenomenon in a language where the phenomenon occurs, but which differs in other syntactic attributes. Dutch is such a language where the optional infinitive stage occurs, but which differs considerably from English in its Object-Verb order. 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. Dutch furthermore differs from English in the fact that finite forms are far more numerous 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 meant to say *I throw ball*, but dropped *I*, the resulting *Throw ball* would be counted as an infinitive in analysis. 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 it is in English. If a model is to learn from the distribution of naturalistic speech input, then the production of a large number of infinitives would appear easier in English than in Dutch.

Given these differences between the languages, generalisation of an optional infinitive model from English to Dutch provides a strong test of the generality of the mechanisms incorporated in the model. The remainder of this paper is devoted to a description of the model, and the results of the simulation of the optional infinitive stage in Dutch.

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. CHREST models have successfully been used to model phenomena such as novice-expert differences in chess (Gobet & Simon, 2000) and computer programming as well as phenomena in diagrammatic reasoning (Lane, Cheng & Gobet 1999) and language acquisition (Jones, Gobet & Pine, 2000a, 2000b). The basis of the model is a discrimination net which can be seen as an index to Long-Term Memory. The network is a n-ary tree, headed by a root node. Training of the model takes place by feeding utterances to the network, and sorting these (see Figure 1). 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 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 sentence 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 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 with all new words. Figure 1 shows the development of the net through the three presentations of the sentence.

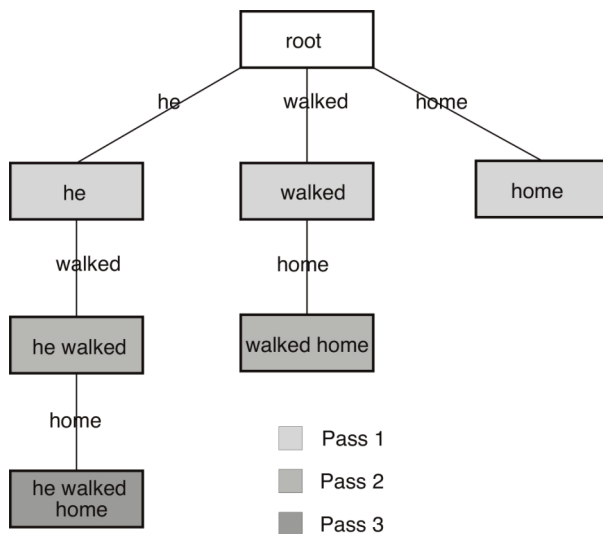


Figure 1: MOSAIC learning an input.

As the model sees more input, it will thus encode larger and larger phrases. 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 similar. 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 by assessing to what extent two nodes have the same nodes directly above and below them (two nodes need to share 10% of both the nodes below and above them in order to be linked). This is equivalent to assessing how likely it is that the two words are preceded and followed by the same words in an utterance. 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 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. Note that MOSAIC does not have access to any morphological information concerning words or phrases. All the morphological information it acquires is based on a simple distributional analysis of the input.

The main importance of generative links lies in the role they play when utterances are generated from the network. When the model generates utterances it will output all the utterances it can by traversing the network

until it encounters a terminal node. Once it encounters a terminal node, it will output the contents of the nodes it encountered, thus producing utterances. 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* generation. During generation, however, the model can also traverse generative links. When the model traverses a generative link, it can supplement the utterance up to that point with a phrase that follows the node that the current one is linked to. As a result, the model is able to generate utterances that were not present in the input. Typically, the output of a MOSAIC model will consist of more than 50% generated (non-rote) utterances. The model thus is highly generative. Figure 2 gives an example of the generation of an utterance using a generative link.

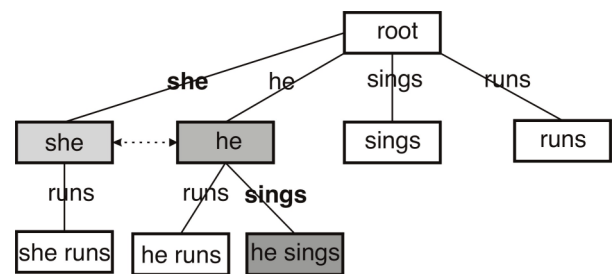


Figure 2: Generating an utterance. Because *she* and *he* have a generative link, the model can output the novel utterance *she sings*. (For simplicity, preceding nodes are ignored in this figure).

As was mentioned earlier, generativity is certainly a characteristic that children display. In fact, proponents of nativist theories of language acquisition have argued that since the number of grammatical utterances in a language is infinite, the child can never hear them all. Seeing that children are able to create utterances they have never heard is seen as evidence for the existence of a grammar-like representation in the child.

The Simulations

An earlier version of the model described above has been shown to provide a good account of optional infinitive phenomena in English (Croker, Pine & Gobet, 2000). The present model differs from the Croker et al. version in two ways. Firstly, when deciding whether two nodes should have a generative link, the previous version only assessed whether two words were likely to be *followed* by the same words. The present version is sensitive to both the words preceding and following the two words. Secondly, the present model calculates the overlap as a percentage of the nodes preceding and following the nodes that are considered. The previous model only considered the absolute number of nodes. These changes were not required to simulate the Dutch

data, but constitute a refinement of the earlier model on theoretical grounds. Simulations have shown that the newer version of MOSAIC also provides a good account of the optional infinitive stage in English. Apart from the two newer preconditions for creating generative links, the version of MOSAIC used for these simulations is identical to the one used by Croker et al.

The data that were simulated were taken from Wijnen, Kempen & Gillis (in press). 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 between mother and child. For Matthijs, the recordings were made between the ages 1;9 and 2;11. For Peter this was 1;7 and 2;3. Wijnen et al. analysed the children's and mothers' utterances with respect to the presence of the optional infinitive phenomena in both the mother's and the children's speech. 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 analysed, and used these as input for the model.

It seems appropriate to point out at this juncture that the corpora used to train the model are just samples of the mother's speech, which are taken to be representative of the mother's speech towards the child. Obviously, the child is subject to other sources of speech as well, but the mother's speech is considered a fairly representative sample. Also, the sample from the child covers a period during which the child develops as well. In fact, between the ages one and three the child moves through four phases (one word, early two word, optional infinitive and end phase). By the time the child reaches the end phase, its speech is fairly similar to the mother's speech in terms of basic syntax. The present model is a model of the child's performance in the optional infinitive stage only. Thus, in the analyses performed, the model was trained on the entire corpus of maternal speech, and the model's output was compared to the speech of the child during the optional infinitive stage. Potential ways of extending the model to other stages are explored in the discussion.

The samples of the mother's speech are 14,000 utterances for Matthijs, and 12,500 utterances for Peter. Two separate analyses were run for the two corpora. For both analyses, the model was trained using all the mother's utterances. After the model was trained, all the utterances that the model could produce (both rote and generated) were collected. This resulted in a sample of 35,000 utterances for Matthijs, and 26,000 utterances for Peter. This relatively large difference in output given the small quantitative difference in input is caused by differences in lexical diversity and mean utterance length in the two corpora. The proportion of rote utterances was .30 for Matthijs and .37 for Peter.

Thus, the majority of the utterances that the model created was not present in the mother's speech.

As was mentioned earlier on, the optional infinitive stage is characterized by 3 phenomena:

1. The child produces a large number of non-finite verb forms.
2. The basic pattern of verb placement is correct.
3. The child drops the subject of the sentence relatively often.

Of the generated utterances, those which contained one or more verbs were collected, and divided into utterances with a finite and a non-finite verb form. Cases where the utterance contained a finite auxiliary verb plus non-finite form (e.g. *He wants to go*) were counted as non-finite forms. This same procedure was used by Wijnen et al.

Table 1 shows the proportions of non-finites that were present in the corpora of the children in the optional infinitive stage, the mothers, and the models of the two children. It is clear from table 1 that the proportion of non-finites for the children is higher than it is in the adult speech.

Table 1: Proportion of non-finites for mothers, children and simulations.

	Matthijs	Peter
Mother	.40	.35
Child	.73	.62
Model	.62	.47

Table 1 also shows that the scores for the model are higher than those for the mothers. From the mother's input, the model has generated output that looks more like the child's output. The model does underestimate the proportion of non-finites, though. Another way of assessing whether the model's output resembles the child's output is to look at the proportion of *root infinitives*. Formally, non-finite forms include all verb forms that are not marked for agreement or tense. This includes past participles, gerunds and auxiliary-plus-infinitive constructions. A special form of the infinitive is the root infinitive, where the infinitive (root) form of the verb is the only verb in the sentence. An example of a root infinitive in English is:

He build house (3)

Root infinitives are relatively rare in adult speech, and only acceptable in special cases. In children's speech (during the optional infinitive stage), they are fairly common, though. In the speech of Mathijs' mother, root infinitives only occurred in 5% of the utterances containing a verb. For Peter's mother this figure was 10%. The simulation shows that the model

of Matthijs produced 40% root infinitives, while Peter's model produced 22% root infinitives. Thus the models have learned constructions which are quite infrequent in adult speech, and resemble the children's data more closely (unfortunately, exact proportions for the children are not available).

An obvious question now is how the model has learnt to produce these utterances that were not so prominent in the mother's speech. A possible source of these utterances lies in the auxiliary + infinitive construction (which is used in around 30% of the mother's utterances). Suppose the model has seen an utterance like:

Wil je met de blokken spelen? (4)
(Want you with the blocks play?)

Because the model can output partial utterances, it may well produce the last two words of the sentence, i.e. *blokken spelen*, which is a root infinitive. Needless to say, if the node for *blokken* has a generative link to another word, say, *trein* (*train*), the model could also produce *trein spelen*, a generated (new) root infinitive.

As was mentioned earlier, a second feature of the optional infinitive stage is that, while children produce a relatively large number of non-finites they do place them in the right position in the sentence. In order to check whether the model has done so, samples of the utterances containing finite and infinitive verbs were coded with respect to verb placement. Table 2 gives the relevant data for Mathijs' and Peter's model.

Table 2: Percentages of correct verb placement for Matthijs and Peter's model as a function of the verb's finiteness.

	Finite	Infinitives
Matthijs	.88	.89
Peter	.95	.87

Table 2 clearly shows that the model has learnt the basic rules of verb placement, and, coupled with the relatively large number of infinitives, the model thus conforms to the definition of the optional infinitive stage.

A third analysis performed by Wijnen et al. was to examine to what extent the children's placement of the object relative to the verb conformed to the mother's placement. Klein (1974) observed that for Dutch children in the optional infinitive stage the Object-Verb order was dominant over the Verb-Object ordering. In order to compare the model's output to that of Matthijs and Peter's, two samples of 1,500 utterances were examined for utterances containing a possible object and a verb. In these utterances, the order of object and

verb was assessed using the semantics of the verb and the potential object. This resulted in some 300 utterances per sample where we were fairly confident what constituted the object. (Note that not all utterances contain a phrase that could be considered an object, and some utterances are ambiguous with respect to object placement.) Table 3 gives the proportions Object-Verb orderings for the mothers, children and the model's samples.

Table 3: Proportion of Object-Verb orderings for mothers, children and simulations.

	Matthijs	Peter
Mother	.65	.60
Child	.90	.68
Model	.65	.57

Table 3 confirms Klein's observation that Dutch children in the optional infinitive stage use the OV order more than their mothers, though the effect is more pronounced for Mathijs than it is for Peter. The model does not conform to this prediction however, as it resembles the mother's data more than the children's data. In fact, it might be argued that the model looks too much like an adult. The general underestimation of the data analysed earlier also seems to point in this direction. One possible cause for this relative maturity of the data might be found in the fact that the model may learn too quickly. The fact that the model is learning too quickly is certainly true when comparing the amount of input for the model and the actual children. The input for the two models consist of 12,500 and 14,000 utterances respectively, which is to simulate the exposure to a language that a child has had in slightly over two years. The high speed of learning is also apparent in another measure; the mean length of an utterance (MLU). The MLUs for Matthijs and Peter are 2.0 and 2.3 respectively. For both models, the MLUs are around 3.1, roughly 50% too high. These relatively high MLUs may be a cause of the low incidence of object-verb orderings in the models' output since long sentences may contain subordinate clauses which have a verb-object ordering (e.g. *I know it, you want an ice cream*).

In order to assess whether short utterances conform more closely to the data, a sample of utterances containing three words or less was selected, and object placement was again coded. Though this is not equivalent to decreasing the learning rate, it does give some insight into properties of shorter sentences which would be more frequent in the output of a model with a lower learning rate. The resultant MLUs for the new sample were 2.44 and 2.45, still slightly higher than the data-MLU, but considerably lower than for the full output. Re-analysis of the sample showed the Object-

Verb order proportion to be .82 for Matthijs' model and .64 for Peter's model. These figures are actually quite close to the values of .90 and .68 in table 3. For Matthijs's model anyway, the figure is closer to Matthijs's data than to his mother's data. This suggests that the fit for Object placement would be better for a model with a lower learning rate. (Overall, changes to the proportions reported in earlier tables tended to be negligible, and/or in the direction of the children's data rather than the mother's data).

Conclusions

Results show that MOSAIC, which was developed as a model of English speaking children, gives a good account of the performance of Dutch speaking children. As such, it supports the contention that general purpose learning mechanisms can account for cross-linguistic variation. It also shows that phenomena in different languages can result from a simple distributional analysis of input from that language. Comparing this account to Wexler's (1998) approach, it clearly shows that a distributional analysis can be sensitive to the broader syntactic properties of a language and at the same time produce the correct inflected form as well as the incorrect infinitive form. Importantly, it does so without postulating innate knowledge about the grammar in the child.

A final note might be added regarding the speed of learning. At present, MOSAIC is seen as a model of a child in the optional infinitive stage. From the results presented here, it is apparent that the data it produces appear to be too adult. Limiting the output to shorter sentences results in a closer fit to the children's data. It was argued that decreasing the learning rate might improve the performance of the model. We have attempted to decrease the learning rate by increasing the number of times a word has to be seen before being encoded, but this did not have the desired effect. One other way in which the learning rate might be decreased is by increasing the number of times *sequences* must be seen before being encoded in the network. At present, a two-word sequence only has to be seen once before it is encoded (provided the two words have been seen in another context). Future work will address the issue of learning rates and the effect this has on the length and characteristics of generated utterances. Investigations into ways of decreasing learning rates (and manipulating the amount of input) may also allow us to examine more closely the developmental patterns that are evident in the model's output. That is, analyzing the model's performance after it has seen varying amounts of input may allow us to model developmental stages that precede and follow the optional infinitive stage.

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.
- Crocker, 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.
- Feigenbaum, E.A. & Simon, H.A. (1984). EPAM-like models of recognition and learning. *Cognitive Science*, 8, 305-336
- 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. Joshi (Eds.), *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*. pp. 723-728. Mahwa, 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.
- Klein, R.M. (1974). Word order: Dutch children and their mothers. *Publications of the Institute of General Linguistics* 9. University of Amsterdam.
- Lane, P.C.R., Cheng, P.C-H., & Gobet, F. (1999). Learning perceptual schemas to avoid the utility problem. In M. Bramer, A. Macintosh and F. Coenen (Eds.) *Research and Development in Intelligent Systems XVI*, (pp. 72-82) Cambridge, UK: Springer-Verlag.
- MacWhinney, B. & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language*, 17, 457-472.
- 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. (in press). Root infinitives in Dutch early child language. To appear in *Journal of Child Language*.