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

Modelling the developmental patterning of finiteness marking in English, Dutch, German and Spanish using MOSAIC

Daniel Freudenthal, Julian M. Pine, Javier Aguado-Orea

School of Psychology

University of Liverpool

Fernand Gobet

School of Social Sciences and Law

Brunel University

ABSTRACT

In this paper we apply MOSAIC (Model of Syntax Acquisition in Children) to the simulation of the developmental patterning of children's Optional Infinitive (OI) errors in four languages: English, Dutch, German and Spanish. MOSAIC, which has already simulated this phenomenon in Dutch and English, now implements a learning mechanism that better reflects the theoretical assumptions underlying it, as well as a chunking mechanism which results in frequent phrases being treated as one unit. Using one, identical model that learns from child-directed speech, we obtain a close quantitative fit to the data from all four languages, despite there being considerable cross-linguistic and developmental variation in the OI phenomenon. MOSAIC successfully simulates the difference between Spanish (a pro-drop language where OI errors are virtually absent), and Obligatory Subject languages that do display the OI phenomenon. It also highlights differences in the OI phenomenon across German and Dutch, two closely related languages whose grammar is virtually identical with respect to the relation between finiteness and verb placement. Taken together, these results suggest that (a) cross-linguistic differences in the rates at which children produce Optional Infinitives are graded, quantitative differences that closely reflect the statistical properties of the input they are exposed to and (b) theories of syntax acquisition need to consider more closely the role of input characteristics as determinants of quantitative differences in the cross-linguistic patterning of phenomena in language acquisition.

1. Introduction

In many languages, children go through a stage in which they produce utterances that appear to lack tense and agreement markers that are obligatory in adult speech. For example, English-speaking children often produce utterances such as *That go there* instead of *That goes there* or *He go to school* instead of *He went to school*.

Traditionally such errors have been interpreted in terms of lack of knowledge of the appropriate inflections (Brown, 1973), or performance limitations in production (Bloom, 1990; Valian, 1991). However, Wexler (1994) argues that, in such cases, rather than dropping inflections, children are producing non-finite verb forms (in this case infinitives) in contexts in which a finite verb form is obligatory. While this may not be apparent from the English data (where the infinitive is indistinguishable from all present tense forms except the third person singular), data from languages such as German and Dutch (where the infinitive carries its own morphological marker: *-en)* suggest that children do indeed produce infinitive verb forms in finite contexts. For example, Dutch-speaking children often produce utterances such as *Hij spelen* (He play-INF) instead of *Hij* speelt (He play-FIN) and *Ik eten* (I eat-INF) instead of *Ik eet* (I eat-FIN).

Wexler (1994) explains this pattern of errors by assuming that, although children have correctly set all the inflectional and phrase structure parameters of their language, there is a stage of development in which the abstract features of Tense and Agreement may be optionally underspecified in the underlying representation of the sentence. This results in children using both non-finite and finite verb forms in contexts in which a finite verb form is required, and hence in the occurrence of "Optional Infinitive" (OI) errors like those mentioned above. It also explains why, when children do use finite verb forms,

they tend to use them correctly. For example, children rarely produce finite verb forms that fail to agree in person or number with the subject of the sentence (e.g. *I goes* instead of *I go*, or *Ik loopt (I walks)* instead of *Ik loop*) (Harris & Wexler, 1996).

The great strength of Wexler's analysis is that it provides a unified account of the occurrence of OI errors in a wide range of different languages, including English, Dutch, German, French, Swedish, Danish and Russian (Wexler, 1994). However, Freudenthal, Pine and Gobet (2006) have recently shown that it is possible to simulate the OI phenomenon in two of these languages (English and Dutch) in terms of the interaction between a computational model with an utterance-final processing bias (MOSAIC) and the distributional properties of real child-directed speech. The aim of the present paper is to present a new version of MOSAIC that eliminates some of the weaknesses of the previous version of the model, and to assess the extent to which this new version of the model can simulate the developmental patterning of finiteness marking in two additional languages: a third OI language (German), which is structurally similar to Dutch (though also subtly different in certain important respects), and an INFL-licensed null-subject language (Spanish), which is structurally different from both English and Dutch, and in which OI errors are not predicted to occur by Wexler's theory.

1.1. Simulating the OI phenomenon in English and Dutch

MOSAIC is a simple distributional learning mechanism with a strong utterance-final processing bias, which takes as input corpora of orthographically transcribed child-directed speech and learns to produce as output utterances that become progressively longer as learning proceeds. As a result of these characteristics, MOSAIC can be used to

model the behaviour of children learning different languages across a range of Mean Length of Utterance (MLU) values.

Freudenthal et al. (2006) have shown that MOSAIC is able to simulate the developmental patterning of the OI phenomenon in two languages: English and Dutch. More specifically, they have shown that the same version of MOSAIC provides a good fit to the developmental data on the rate at which English and Dutch children produce OI errors as their average utterance length increases.

MOSAIC simulates OI errors as a result of its utterance-final processing bias. This bias results in the production of partial utterances that were present as utterance-final phrases in the input on which the model was trained. The utterances in the input that give rise to OI errors are *compound finites*: utterances that contain a (finite) modal or auxiliary plus an infinitive form (e.g. *Can he go*). Omission of the modal *can* from the English utterance *Can he go* results in the Optional Infinitive *He go*. Similarly, omission of the modal *wil* from the Dutch utterance *Wil hij spelen* (Wants he play-INF/ Does he want to play) results in the Optional Infinitive *Hij spelen* (He play-INF).

MOSAIC simulates the developmental patterning of OI errors because it learns to produce progressively longer utterance-final phrases as a function of the amount of input to which it is exposed. Children start out producing OIs at relatively high rates, and produce fewer OIs as the length of the utterances they produce increases. MOSAIC simulates this phenomenon because of the way that compound finites pattern in English and Dutch. In compound finites, the finite modal or auxiliary precedes the infinitive. Since MOSAIC produces increasingly long utterance-final phrases, the early (short) phrases it produces are likely to contain only non-finite verb forms. As the phrases

MOSAIC produces become longer, finite modals and auxiliaries start to appear, and OIs are slowly replaced by compound finites.

The idea that OI errors are learned from compound finites in the input is not a new one. For example, Jordens (1990) argues that OIs in Dutch are incomplete compound verb forms, which are gradually absorbed into compound predicates as the child begins to produce more and more modals and auxiliaries; Ingram and Thompson (1996) suggest that the modal reading of Optional Infinitives in German can be explained by the association between infinitive forms and (compound) modal constructions in the input; and Wijnen, Kempen and Gillis (2001) argue that OIs in early child Dutch reflect the interaction between an utterance-final bias in learning and the position occupied by infinitives in compound finites in Dutch. What Freudenthal et al. show, however, is that, although compound finites are relatively rare in Dutch (making up only approximately 30% of all the utterances including verbs in the input), MOSAIC's utterance-final bias is sufficient to result in virtually exclusive use of OIs during the early stages. They also show that the same version of the model that captures this phenomenon provides a good fit to the developmental data on English.

These results suggest that it is possible to closely simulate the OI phenomenon in two different languages as a function of the same kind of resource-limited distributional analysis of the input. However, there are two reasons why they might be regarded as potentially problematic. The first reason is that, for certain implementational reasons, the version of MOSAIC used by Freudenthal et al. (2006) only approximates the theory that is assumed to underlie the model. This raises the question of whether it is possible to develop a version of the model that implements the underlying theory more directly while

still providing a good fit to the developmental data. The second reason is that the model has so far only been used to simulate data from two closely related languages. This raises the question of whether it is possible to simulate data from a wider range of languages using the same mechanisms that provide a close fit to the developmental data on English and Dutch.

1.2. Three weaknesses of the earlier version of MOSAIC

There are three ways in which the version of MOSAIC used by Freudenthal et al. departs from the theory assumed to underlie the model. The first of these is the fact that the model's utterance-final bias is not implemented as a constraint in learning, but as a restriction on what the model is allowed to output in production. This means that in practice there are many words and phrases that are represented within the network (and hence influence the way in which the model encodes further information), but are not produced when generating output. Given the central role of the utterance-final bias in simulating the developmental data, it would obviously be preferable if this was not the case, and MOSAIC's utterance-final bias was implemented more directly as a constraint on learning. It is also important to show that changing the implementation in this way does not adversely affect the fit between model and child data.

A second weakness of the version of MOSAIC used by Freudenthal et al. is the fact that, in order to restrict the amount of output produced by the model to manageable proportions, the model's mechanism for generating novel utterances was unrealistically constrained. MOSAIC generates novel utterances by computing a measure of distributional similarity across the words it encodes. Words that tend to occur in similar

contexts (i.e. are preceded and followed by overlapping sets of words) are linked together in the network and can be substituted when the model produces output. This leads to the production of utterances that were not present in the model's input. In the version of MOSAIC used by Freudenthal et al. this mechanism was constrained to allow the substitution of only one word per utterance. This constraint is clearly unrealistic, particularly at later stages of development. It is therefore important to show that removing this constraint does not result either in a substantial increase in the level of error in MOSAIC's output or in a substantial decrease in the model's ability to simulate the developmental data.

A third weakness of the version of MOSAIC used by Freudenthal et al. is the fact that the model's ability to *unlearn* OI errors is extremely limited. Thus, although the model learns to produce increasingly large numbers of compound finites and hence lower and lower proportions of OI errors, it never actually learns to stop making OI errors. Indeed the short incomplete utterances that the model learns early in development continue to be produced even when the model is capable of producing the complete sentence frames from which the short utterances were originally acquired. This state of affairs is clearly at odds with the idea that OI errors ultimately become compound finites. It would therefore be preferable if some mechanism could be added to the model that allowed it to *unlearn* OI errors by replacing them with the compound finites from which they were originally acquired. It is also worth noting that since the version of MOSAIC used by Freudenthal et al. tended to overestimate the proportion of OI errors at high MLUs, the addition of a mechanism for unlearning OI errors might actually improve the fit between the model's and the children's output at later stages of development.

In summary, although the version of MOSAIC used by Freudenthal et al. provides a good fit to the developmental patterning of the OI phenomenon in English and Dutch, this version of the model has a number of weaknesses, which, if corrected, might adversely affect the model's ability to simulate the developmental data. One of the aims of this paper is therefore to present a new version of MOSAIC that implements the underlying theory more directly, and to show that this new version of the model is still able to provide a close fit to the developmental data on English and Dutch.

1.3. Simulating the OI phenomenon in a wider range of languages

In addition to the implementational weaknesses outlined above, there are also some clear challenges facing MOSAIC with respect to the range of languages that it is able to simulate. Freudenthal et al. simulated data from only two relatively similar languages, and it is unclear to what extent MOSAIC's success in simulating the OI phenomenon will extend to a wider range of different languages. This is of interest as the OI phenomenon occurs across a range of languages that are likely to differ in terms of their distributional characteristics. As a result of these differences, the fine detail of the OI phenomenon may show considerable variation across these languages. For example, German is an interesting language to simulate as it has the same rules for verb placement as Dutch. This includes the feature that finite verbs take second position and non-finites take final position. However, there are some subtle differences between German and Dutch that favour the occurrence of non-finites in sentence-final position in Dutch. For example, in Dutch progressive aspect is expressed using a construction which includes a sentence-final infinitive, whereas German does not have a progressive construction. Similarly,

Dutch uses compound constructions with the verb *gaan* (*go)* + *infinitive* to express future events or intentions. This construction is semantically more restricted and less frequent in German. These differences between German and Dutch may well impact on the rates at which both children and MOSAIC produce OI errors in the two languages. It is therefore of interest to assess whether there are any differences between German and Dutch children with respect to the OI phenomenon, and, if there are, whether the model is able to simulate the German data as well as it simulates the Dutch data.

An even greater challenge facing MOSAIC is the simulation of so-called INFL-licensed null subject languages like Spanish and Italian. In these languages, children rarely produce bare infinitives in finite contexts. However, they do produce utterances with other non-finite forms such as progressive and past participles as the only verb form (Wexler, 1998). MOSAIC's ability to simulate this pattern of results together with data from languages where OI errors do occur at high rates is of particular interest as Wexler (1998) has reformulated his 1994 theory to take the data from Spanish and Italian into account. According to the 1998 formulation of Wexler's theory, Tense or Agreement may be missing from the underlying representation of the sentence, because children's grammars are governed by a 'Unique Checking Constraint'. The Unique Checking Constraint impacts on the child's ability to check 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 in contexts in which a finite verb form is required.

The 1998 formulation of Wexler's theory can explain why children produce OI errors at high rates in obligatory subject languages like English, Dutch and German. This is because such languages require the child to check against two D-features: Tense and Agreement. It can also explain why children make few OI errors in INFL-licensed null subject languages like Spanish and Italian. Since these languages (usually) only require the child to check against one D-feature (Tense), the unique checking constraint does not result in OI errors in these languages. Finally, it can explain the finding that, like children learning obligatory subject languages, children learning null subject languages do produce utterances with other non-finite verb forms such as perfect and progressive participles in finite contexts. This is because constructions containing an auxiliary and a perfect or progressive participle (e.g. *He has gone, He is going*) require checking of the D-feature for both Tense and Agreement on the Auxiliary. Since children learning Spanish or Italian are subject to the same Unique Checking Constraint as children learning Obligatory Subject languages, the auxiliary fails to surface in such cases leading to the production of bare perfect or progressive participles in both null subject languages and obligatory subject languages.

The 1998 formulation of Wexler's theory explains differences in the developmental patterning of finiteness marking in Obligatory Subject and Null Subject languages with reference to deep structural differences between languages. MOSAIC's ability to simulate these cross-linguistic differences using one simple learning mechanism would suggest a simpler explanation in terms of graded quantitative differences in the surface structure of the languages. However, the attempt to simulate early child Spanish also provides a particularly strong test of the notion that children's early multi-word speech is

shaped by the interaction between the statistical properties of the input to which they are exposed and a learning mechanism that is biased towards learning elements that occur near the right edge of the utterance. Compound finites (the main source of OIs) occur in Spanish at rates that are roughly comparable to those in OI languages such as German and Dutch. However, unlike children learning OI languages, Spanish children produce very few OI errors. The interaction between the utterance-final bias and the position in which finite and non-finite verbs occur in the input will therefore be crucial if MOSAIC is to simulate the Spanish data successfully.

In view of the challenges outlined above, the second aim of this paper is to assess the extent to which MOSAIC's success in simulating the OI phenomenon in English and Dutch extends to two other languages: German and Spanish. German was chosen as a target for simulation because it is structurally similar to Dutch but subtly different in terms of its use of modal constructions. Spanish was chosen because it does not display the OI phenomenon despite showing comparable rates of compound finites in the input. The simulation of English, Dutch, German and Spanish using one identical learning mechanism serves as a strong test of the notion that children's early multi-word speech may be understood in terms of the interaction between an utterance-final bias in learning and the distributional properties of the input language. A basic description of the four languages as well as a number of example utterances that illustrate the differences and commonalities among them are provided in Table 1. These commonalities and differences can be summarised as follows: German and Dutch have the same word order (SOV/V2), which differs from the word order for Spanish and English (SVO); German,

Dutch and Spanish have roughly equal proportions of compound finites[1], which are considerably lower than the proportion of compound finites in English; Spanish differs from English, Dutch and German in that it allows simple finite constructions where a pronominal object precedes an utterance-final main verb (e.g. (*Yo*) *lo quiero* ((I) it want)).

---------- Insert Table 1 about here ----------

2. MOSAIC

2.1. The MOSAIC network

MOSAIC consists of a simple network of nodes and arcs that incrementally stores utterances to which it is exposed. The network is headed by a root node that has no contents. The other nodes in the network store words or phrases shown to the model. Nodes immediately beneath the root node are called primitive nodes, and encode single words[2]. Nodes below the primitive level encode phrases, or sequences of words encoded at the primitive level. The arcs or 'test-links' that connect nodes are used to store the difference between the nodes they connect. A MOSAIC network is slowly built up from exposure to the input it receives. As MOSAIC sees more input, it will create more primitive nodes encoding the different words in the input. In addition, it will create nodes at deeper and deeper levels in the network thus encoding longer and longer utterances. MOSAIC learns from orthographically transcribed child-directed speech with whole words being the unit of analysis. That is, MOSAIC assumes that the speech stream has

[1] Proportions of compound finites were calculated from the maternal speech in the corpora used for the simulations. Details of these corpora are provided later.

[2] Primitive nodes encoding multi-word phrases can be created by the chunking mechanism that will be described later. For clarity of exposition, this possibility is ignored here.

been segmented into words. Learning in MOSAIC is anchored at the right edge of the utterance; a word or phrase is only encoded when everything following that word or phrase in the utterance has already been stored in the network. MOSAIC thus has a strong utterance-final bias in learning. In previous versions of MOSAIC learning took place from the left edge of the utterance, and the utterance-final bias was implemented by restricting the generation of utterances to utterance-final phrases (phrases with an end-marker). Learning from the right edge of the utterance brings the model more in line with the theoretical assumption that the learning of language is a process that is heavily biased towards the most recent elements in the speech stream.

The processing of an utterance in MOSAIC can be likened to a moving window or buffer, with the size of the window being determined by how much of the utterance has already been encoded by the model. Whenever the model encounters a word or word transition it has not seen before, the contents of the window are cleared, and the new word is deposited in the empty buffer. Only when the rest of the utterance has already been encoded in the model will the new word remain in the buffer, thus making it eligible for encoding. Thus, MOSAIC processes the utterance in a left-to-right fashion, but builds up its representation of the utterance by starting at the back and slowly working its way to the front. In terms of a child attending to the speech stream, the occurrence of an unknown word will effectively clear the contents of the speech stream encountered so far, leaving the new word and the rest of the utterance for analysis. Thus, children's language learning is viewed as a process that is strongly biased towards the most recent elements in the speech stream. Such an utterance-final bias (or recency effect) is psychologically plausible, and several authors have argued that children learn or comprehend material

that occurs towards the end of the utterance more easily than material that occurs further to the left (Shady & Gerken, 1999; Wijnen, Kempen & Gillis, 2001).

As an example of how MOSAIC builds up its representation of an utterance, consider an empty network being shown the utterance *he goes away*. At a first presentation of this utterance the model has not yet encoded any of the words in the utterance. When the model reaches the end of the utterance, the buffer will contain the word *away*. Since it has reached the end of the utterance (as signified by an end marker), the model will now encode the word *away*. At a second presentation, the model will reach the end of the utterance with the words *goes away* in the buffer. The model will now attempt to encode the phrase *goes away*. However, as the word *goes* has not yet been encoded at the primitive level, the model will create this node first. Only after a third presentation, will the model create a branch encoding the phrase *goes away*. A fourth presentation will result in the creation of the primitive node *he*. A fifth presentation will result in the model encoding the fact that the phrase 'goes away' has been preceded by 'he' (by copying the *goes away* branch underneath the *he* node). Fig. 1 shows what the MOSAIC network looks like after five presentations of the utterance *He goes away*.

---------- Insert Fig. 1 about here ----------

2.2 Node Creation Probability

So far, we have assumed that a node is created whenever the possibility arises. In reality, creation of nodes in MOSAIC is probabilistic and learning proceeds slowly. The probability of a node being created is given by the following formula:

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

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 end of the utterance.

The formula results in a basic sigmoid curve when plotted as a function of the number of utterances the model has seen. Early in training, when the model has seen few utterances, node-creation probability will be low. The node-creation probability goes up as the number of utterances the model has seen increases. Learning thus speeds up as the amount of knowledge encoded in the model grows. The size of the corpus used for the simulation is also included in the formula. The reason for this is that the size of the corpora used for different languages differs considerably. The use of the term *m-u/c* ensures that after an equal number of presentations of the entire input corpus, the node-creation probability is identical for corpora of different sizes. The formula also includes the distance to the end of the utterance in the exponent. This results in the probability of encoding material that occurs near the beginning of the utterance being lower than for material near the end of the utterance (as the base number in the formula is bounded between 0 and 1). Since learning in MOSAIC is slow, the input corpus is fed through the model several times, and output is generated from the model after every presentation of

the corpus. In this way, output of increasing length can be generated from the model, allowing for the simulation of developmental variation. Early in training, the model will tend to encode only utterance-final words. As the model sees more and more input, it will encode more and longer (utterance-final) phrases.

A further consequence of node-creation being probabilistic is that it makes learning in MOSAIC frequency-sensitive. Since a word or phrase has a certain probability of being encoded on every encounter, it normally has to be seen several times before being encoded. For words that occur frequently in the input, this will generally be the case at earlier stages in development than it will be for words that are encountered infrequently.

2.3 Generating output from MOSAIC

MOSAIC employs two mechanisms for generating output. The first mechanism simply traverses all the branches of the network, and generates all the (utterance-final) phrases they encode. Output generated using this mechanism is *rote* output (i.e. it is made up entirely of phrases or partial utterances that were present in the input corpus). This mechanism is complemented by a second mechanism that generates novel output through the substitution of distributionally similar words. MOSAIC stores the context (preceding and following words), in which a word has been encountered in the input for all words encoded in the network. Words that tend to be followed and preceded by the same words become connected through a *generative link*, and can be substituted for each other in production. The rationale for this is that words that occur in distributionally similar contexts tend to be of the same word class (Redington, Chater & Finch 1998; Mintz 2003). Substitution of words that share a generative link results in MOSAIC having the

ability to produce novel utterances that were not present in the model's input. For the present simulations two words need to share 20% of the words immediately preceding and following them[3] in order for them to be substituted in production.

In previous versions of MOSAIC substitution of words was subject to a maximum of one substitution per utterance. For the present version of MOSAIC this restriction has been lifted. MOSAIC can now substitute multiple words in an utterance.

2.4 Unlearning

One weakness of the earlier version of MOSAIC was that it had a limited ability to *unlearn*. Even at late stages of development, the model continued to produce short, incomplete utterances that were learned early in training, despite the model having encoded the longer, complete utterances from which they had been learned. The new version of MOSAIC implements a mechanism for unlearning. When a longer version of an incomplete utterance is learned, the shorter utterance is marked for being part of a longer utterance that has been encoded in the model. When output is generated from the model, utterances that are marked in this way are omitted.

2.5 Chunking

An important change to MOSAIC is the addition of a novel chunking mechanism. According to the chunking theory (Chase & Simon, 1973; Gobet & Simon, 1998; Gobet

[3] In previous versions of MOSAIC two words needed to share 10% of the preceding and following context in order to become connected by a generative link. The changes made to MOSAIC have greatly increased the model's ability to produce novel utterances, thereby drastically increasing the amount of output the model produces. A value of 20 for the required overlap percentage keeps output size at a manageable level, and decreases the likelihood of poor quality substitutions.

et al., 2001), frequently encountered stimuli are grouped into larger structures that can be retrieved as one unit. MOSAIC's chunking mechanism results in frequent phrases being treated as one unit by the generativity mechanism. This prevents certain substitutions, as the individual words making up a 'chunked up' phrase are no longer considered for substitution in the chunked context.

The nodes encoding words or phrases contain a slot that stores the frequency with which the word or phrase has been encountered in the input. For nodes at the primitive level, this slot encodes how often the word encoded in that node has been encountered in the input. For non-primitive nodes (e.g. a node encoding the word *walks* underneath the node encoding the word *he*), the slots store the number of times the phrase encoded in that node has been encountered. When the frequency for a phrase exceeds a pre-determined threshold[4] a new, single node encoding the phrase is created at the primitive level. This new node replaces the sequence of two nodes that originally encoded the phrase. Since the phrase in question may be encoded as a sequence of two nodes at deeper levels in the network (i.e. in other contexts), all sequences of nodes encoding this phrase are replaced by single nodes encoding the phrase. Chunking is an important mechanism in constraining the substitutions that are made through the generativity mechanism. As detailed in Freudenthal, Pine and Gobet (2005), a potential problem with the extraction of syntactic categories through co-occurrence statistics is that substitutions that are correct in one context may be inappropriate in other contexts. The verbs *do* and *put* for example, may share considerable context due to their occurrence as main verbs,

[4] For the present simulations the chunking threshold was set to 1/4 of the square root of the number of nodes in the net. Due to the differing sizes of the input corpora, a threshold expressed relative to the number of nodes in the net/amount of knowledge encoded was considered more appropriate than an absolute frequency count.

and substituting them in a context where they are used as main verbs may not result in utterances that are syntactically anomalous. The verb *do,* however, is also used as a (dummy) modal in question formation. Substituting *put* for *do* in this context will result in anomalous utterances such as *Put you want an ice cream*. The chunking mechanism is designed to prevent such inappropriate substitutions. Since the phrase *Do you* is very frequent, it will quickly get chunked in the model. One result of this is that if the words *do* and *put* share a generative link, they will no longer be substituted in the *Do you* context, since the phrase *do you* rather than its constituent words is now the target for substitution. Thus, a phrase that has been chunked up may be substituted for other distributionally similar phrases (e.g. *don't you*), but its constituent words cannot be substituted in the context of the chunk.

The chunking mechanism is illustrated in Fig. 2, which shows a sample MOSAIC network before and after a phrase has been chunked up. A more detailed description of the chunking mechanism and the effect it has on error rates in MOSAIC's output is given in Freudenthal et al. (2005).

---------- Insert Fig. 2 about here ----------

3. The Simulations

3.1 Preparation of the input

All input to the model consisted of child-directed speech collected during mother/father-child interactions recorded over several sessions. The corpora used were: two English children from the Manchester corpus (Theakston, Lieven, Pine & Rowland, 2001), two Dutch children from the Groningen corpus (Bol, 1995), the German corpus of Leo (Behrens, 2006) and the Spanish corpus of Juan (Aguado-Orea, 2004). Recording details differ for the different corpora (further details of the corpora/recording regimes are given in section 3.5). However, all corpora consisted of at least fortnightly recordings of one hour over a period of one year. Files making up the individual sessions (containing both child-directed and child speech) were transcribed using CLAN format. Simulations were run in the same manner for all languages. First, input files for the model were created by extracting all maternal speech from the files making up the corpus for a given child. A limited amount of (automated) filtering was carried out on the maternal speech. Utterances that were incomplete (e.g. false starts or interrupted utterances), or contained words that were unintelligible to the transcriber, were excluded. The following material was removed from the remaining utterances: filler words such as 'uh' and 'oh'; repeated and corrected material in retracings (as evident from the transcriber's coding); vocatives and tags occurring at the end of utterances. All corpora contained a wide range of utterances including fully formed utterances, single-word utterances as well as sentence fragments (where these occurred as complete utterances in the original transcripts).

3.2 Preparation of the child data

The child data were extracted from the same files and prepared in the same manner as the child-directed speech. The child data were subsequently split into batches of increasing

age/MLU. For Dutch and German, four different MLU points between 1.5 and 4.0 were distinguished. For English, no meaningful data were available at an MLU of less than 2.0 as a result of the decision (expanded on below) to limit analysis to utterances with third person singular subjects. Thus, only three data points were used for English. As the Spanish child did not produce many utterances at low MLUs, three MLU points were also used in the Spanish simulations.

3.3 Running the simulations

Simulations were run by feeding the relevant child-directed speech for each child through the model several times. Output was generated after every presentation of the input, and consisted of all the utterances the model was capable of producing. This included rote (sentence-final) phrases that were present in the input, as well as novel utterances produced through the model's generativity mechanism. The output files that most closely matched the MLU points distinguished in the child data were subsequently selected for analysis.

3.4 Coding and data analysis

Data analysis was carried out in an identical (automated) manner for the output of the model and the child data. Data analysis was restricted to utterances containing (at least) one verb (excluding the copula), and was restricted to utterance types, rather than utterance tokens. That is, duplicate utterances were removed before analysis. The remaining utterances were assigned to one of three categories: simple finite, compound finite and non-finite.

Simple finite utterances were defined as utterances that only included unambiguously finite verb forms (for example utterances containing first person singular, second person singular or third person singular verb forms in Dutch or German, and utterances containing third person singular verb forms and irregular past tense verb forms in English.

Compound finite utterances were defined as utterances containing both an unambiguously finite verb form and a verb form that was not unambiguously finite (for example, utterances containing a singular present tense verb form and a form matching the infinitive in Dutch or German, and utterances containing a modal and an infinitive or an auxiliary and a perfect or progressive participle in English).

Non-finite utterances were defined as utterances that did not include an unambiguously finite verb form (for example, utterances containing infinitive or plural present tense verb forms in Dutch and German and utterances containing zero-marked verb forms in English).

For Spanish, an additional distinction was made within the utterances that were classed as non-finite. According to Wexler (1998), utterances with bare infinitives (i.e. utterances containing an infinitive as the sole verb in the utterance) are not produced by children learning Spanish, while utterances with bare participles or progressives do occur. In line

with this distinction, utterances containing bare infinitives and utterances containing bare participles or progressives were counted separately for the Spanish analysis.

Note that a disadvantage of the above coding scheme is that, for many verb forms, it is not possible to unambiguously decide if they are finite or non-finite. Thus, in Dutch and German, present tense plurals cannot be distinguished from the infinitive. In English, of all the present tense verbs, only the 3rd singular can be distinguished from the infinitive. In the present coding scheme, all such verb forms are treated as if they are non-finite. This feature of the coding scheme will result in the analyses underestimating the children's and the model's ability to produce correct finite utterances. It does not, however, affect the validity of the comparisons between the children's and the model's data since in both cases the analysis performed is identical. In actual fact, the occurrence of finite verb forms in Dutch and German is unlikely to be seriously underestimated as the relevant ambiguity is restricted to relatively low frequency plural verb forms. With respect to English, however, it could be argued that, since the ambiguity involves high frequency singular present tense forms, the levels of finiteness would be underestimated to such an extent that the model's ability to simulate the phenomenon would become almost trivial. In order to deal with this problem it was decided to restrict the analysis of English to utterances containing a third singular (pronominal) subject (e.g. *He go(es)*), as the provision of a zero-marked form in such a context is clearly incorrect. Analysis was restricted to pronominal third singular subjects (*He, she, it, this, that*) as this allows an automated lexical search and therefore an automated analysis.

However, even when restricting the analysis to third singular contexts, a certain level of ambiguity remains due to English regular past tense forms being indistinguishable

from past participles. Thus an utterance such as *he dropped* can either reflect the use of a correct past tense or a past participle with a missing auxiliary. Utterances with a verb form matching a regular past tense/past participle and no other finite verb forms were therefore classed as ambiguous and counted separately.

3.5 Datasets used in the simulations

For the Dutch and English simulations, the same input corpora were used as in Freudenthal et al. (2006). For Dutch, these were the corpora of Peter and Matthijs from the Groningen corpus (Bol, 1995). The Groningen corpus consists of a series of one-hour recordings of parent-child interaction made at regular fortnightly intervals over a period of approximately two years. Recording started for Matthijs at the age of 1;10 and for Peter at the age of 1;5. For English, they were the corpora of Anne and Becky, from the Manchester corpus (Theakston, Lieven, Pine & Rowland, 2001). The Manchester corpus consists of a series of one-hour recordings of parent-child interaction made approximately twice every three weeks over a period of approximately one year. Recording started for Anne at the age of 1;10 and for Becky at the age of 2;0. The corpora of Peter, Matthijs, Anne and Becky are available through the CHILDES database (MacWhinney, 2000). The Dutch input corpora consisted of approximately 14,000 (Matthijs) and 13,000 (Peter) utterances. The input corpora of Anne and Becky consisted of approximately 33,000 and 27,000 utterances.

Only one corpus was used for the German and Spanish simulations since very few dense data sets are available for these languages. For German, the corpus of Leo (Behrens, 2006) was used. This corpus consists of 5 one-hour recordings per week in the

home environment between the ages of 2;0 and 3;0 and 4 one-hour recordings per month between the ages of 3;0 and 4;11. Since this corpus is much larger than the English and Dutch corpora described above, for Leo's simulations, a random sample of 30,000 utterances was taken from the entire corpus of approximately 160,000 utterances. For Spanish, the corpus of Juan was used (Aguado-Orea, 2004). This corpus consists of two 30-minute recordings per week in the home environment between the ages of 1;11 and 2;6 and includes approximately 25,000 utterances.

4 Results

4.1 Results for the English and Dutch simulations

The version of MOSAIC used for the present simulations differs in various respects from the earlier implementation. The changes made to the model all have the potential to affect the previous good fit between the model and the children. One of the aims of this paper was therefore to investigate whether the new model provides as good a fit to the data as the previous version did. Figures 3 and 4 show this to be the case. Both with the children and the model, non-finites generally decrease with increasing MLU, and compound finites increase. The absolute proportions, which differ between English and Dutch, are also well captured by the model. Thus, despite several changes to the model (implementation of an utterance-final bias in learning, relaxation of a constraint on generativity, implementation of unlearning and the addition of a novel chunking mechanism), the model still provides a good fit to the data from English and Dutch.

Squared correlation coefficients, which reflect the degree of correspondence in the developmental pattern between the children and models are .98 for Anne, .98 for Becky, .88 for Matthijs and .86 for Peter (correlation coefficients were computed on the proportion of non-finites). Root Mean-Square Errors (RMSEs) which reflect the degree of correspondence between the absolute numbers (computed over all categories) are .06, .08, and .09 for the three MLU points for Anne, and .16, .05 and .04 for Becky. For the Dutch simulations, the RMSEs are .08, .01, .21 and .12 for Matthijs, and .03, .06, .25, and .11 for Peter.

---------- Insert Fig. 3 about here ----------

---------- Insert Fig. 4 about here ----------

The model's fit to the data from English and Dutch has changed little from the earlier version of the model. In this earlier version, the utterance-final bias was implemented as a filter in production rather than a constraint in learning. The finding that the model maintains a good fit provides support for the notion that the English and Dutch data can be successfully simulated in terms of an utterance-final bias in learning. However, while the changes to the model have not impacted on the fit to the data, the relaxation of the constraint in generativity and the addition of the chunking mechanism may have impacted on the quality of the output. This possibility was investigated by assessing how many 'ill-formed' utterances or errors the output contained.

4.1.1 Error rates

The quality of MOSAIC's output was assessed by coding samples of the output for the occurrence of errors of commission. For each simulated child a sample of 500 utterances was drawn from MOSAIC's output at an MLU of approximately 3.0. Utterances where the substitution of a word or phrase in a novel context had resulted in an ungrammatical utterance were classified as errors. Coding of the Dutch samples was carried out by the first author who is a native speaker of Dutch. The English samples were coded independently by the first and second author. Error rates were generally low, 7% for Anne, 6% for Becky, and 5% for Matthijs and Peter. Inter-rater reliability (for the English samples) was high at 97% (Kappa = .79). The relaxation of the constraints in generativity thus has not given rise to large amounts of error.

The error rates compare favourably with those obtained with the earlier simulations of Dutch and English (error rates in those simulations varied from 14 to 19%). Two factors contribute to this decreased error rate. First, the proportion of novel utterances that the model generates is slightly lower than it was in the earlier simulations, as the increase in the overlap parameter offsets the increased generativity due to the relaxation of the constraints in generativity. A consequence of increasing the overlap parameter is that the distributional similarity between words needs to be larger for these words to be linked. Thus, an increase in the overlap parameter will generally result in higher quality generative links that give rise to fewer errors.

A second factor in reducing the error rates is the addition of the chunking mechanism, which was designed specifically to avoid the substitution of words in inappropriate

contexts. A more detailed description of why the chunking mechanism is successful in reducing error rates can be found in Freudenthal et al. (2005).

4.1.2 The utterance-final bias revisited

It is worth recalling at this point how MOSAIC's utterance-final bias interacts with the structure of the language to produce OIs at rates that closely match those displayed by the children. The model produces utterances such as *He go* and *Hij eten* (He eat-INF) by producing the final phrases of compound finites such as *Can he go* and *Wil hij eten* (Wants he eat-INF), where finite modals and auxiliaries precede the non-finite verb forms. MOSAIC simulates the basic developmental patterning as a result of it producing increasingly long utterances. OIs are slowly absorbed into compound finites as the length of the utterances MOSAIC produces increases and auxiliaries and modals start appearing. This leads to a gradual decrease in the rates of OI errors. On the face of it, however, the rates at which compound finites occur are not sufficient to explain the high rates of OI errors that children display early in development (particularly for Dutch). Compound finites only make up approximately 30% of the input, yet Dutch children's early verb use is virtually exclusively non-finite. However, the structure of Dutch is such that this limited (30%) amount of compound finites actually results in the large majority of verbs in utterance-final position being non-finite. In Dutch main clauses, finite verbs take second position and are followed by complements such as objects, whereas non-finite verbs take sentence-final position and are preceded by their complements. This is illustrated in the following examples:

Hij trapt de bal (He kicks-FIN the ball)

Hij wil de bal trappen (He wants-FIN the ball kick-INF/ He wants to kick the ball)

Compound finites thus pattern in such a way that they give rise to non-finite verb forms in sentence-final position. Finite utterances, however, do not give rise to many finite verbs in utterance-final position. While second position may overlap with utterance-final position in intransitive constructions such as *Hij loopt* (He walks-FIN), objects (in main clauses) occur after finite verbs, leading to low levels of utterance-final finite verbs. In fact, an analysis of the Dutch input shows that the large majority (~85%) of verbs in utterance-final position is non-finite. MOSAIC's output at low MLUs will of course closely mimic these numbers as the early output consists predominantly of utterance-final phrases that are one or two words long.

Rates of Optional Infinitive errors in English are generally lower than they are in Dutch (though the restriction to third singular context -and accompanying higher MLU values for the first data point- make a direct comparison of the numbers difficult). However, given that English has SVO word order, the rates for English may actually be considered relatively high as non-finite verbs do not appear to occur in utterance-final position very frequently. As it turns out, however, non-finite verb forms in utterance-final position do outnumber finite verb forms in utterance-final position: approximately 75% of all verb forms that occur in utterance-final position (in utterances containing a third singular) are non-finite. An indication of why the ratio of non-finite to finite verbs in utterance-final position is not much lower than it is in Dutch is given in table 1: compound finites in English are far more frequent in English (~ 70%) than in Dutch

(~30%). As a result of this, non-finite verb forms are actually relatively frequent in English.

4.2 Results for the German simulation

4.2.1 Child data and simulations

Having established that MOSAIC still successfully simulates English and Dutch, we can now turn to the question of how well MOSAIC simulates the German data. As was mentioned earlier, German, like Dutch is an SOV/V2 language. Thus, verb placement follows the same rules as in Dutch: finite verbs take second position while non-finite verbs take sentence-final position. On the basis of this similarity one would expect German children to produce OIs at rates that are comparable to Dutch children. However, as was pointed out earlier, some subtle differences exist between German and Dutch that may result in compound finites being less frequent in German child-directed speech. German, for example, does not allow the use of *go (*Dutch: *gaan*, German: *gehen)* plus infinitive to express future intentions, and lacks a progressive construction, which includes a sentence-final infinitive in Dutch. These differences may result in non-finite verb forms being less frequent in sentence-final position, leading to lower rates of OI errors in early child German. Fig. 5 shows that this is indeed the case for the German child.

---------- Insert Fig. 5 about here ----------

Whereas the Dutch children produce OIs at rates close to 80% during the early stages, the German child only produces 60% OIs at a comparable MLU value. MOSAIC closely simulates the data from the German child (RMSEs: .05, .05, .02, and .04, r^2= .98).

The finding that MOSAIC simulates this difference suggests that it will be mirrored by a difference in the proportion of non-finite verbs in utterance-final position in the input. An analysis of the maternal speech directed at the different children shows that this is indeed the case. The proportion of non-finite verbs in utterance-final position (relative to all verbs in utterance-final position) in the input files was .90 for Matthijs, .87 for Peter, and .66 for Leo. The comparison across these three children thus suggests that the rates at which children produce OIs closely reflect the proportion of non-finites that occur in utterance-final position in the input they hear.

4.2.2 An analysis of additional corpora

While the differences between the Dutch and German children, their input and the models suggest a genuine cross-linguistic difference, care should be taken in drawing conclusions on the basis of one German and two Dutch children, as the differences found in these corpora may reflect individual rather than cross-linguistic differences. This appears to be at least partly true of Leo's corpus. It was suggested by the researcher responsible for the collection of Leo's corpus that Leo's mother uses many complex constructions containing subordinate clauses, which, in German, have finite verb forms in sentence-final position[5] (Behrens, personal communication).

[5] The same is true of Dutch, though Dutch word order in subordinate clauses is slightly less constrained in that both finite and non-finite verbs can take final position in subordinate clauses with double verb constructions.

In order to discount the role of individual differences, it was therefore decided to analyse all the children and the relevant child-directed speech from the Groningen corpus (Bol, 1995) (of which Matthijs and Peter are part) and the control children from the Szagun (2001) corpus (which consists of data sets from German-speaking children with Cochlear implants as well normal-hearing controls). For all children, all the relevant child-directed speech, and a sample of child speech at an MLU of roughly 1.5 was analyzed in the same way as was done for the other children and simulations. This resulted in data from 7 Dutch and 6 German children being available (including Leo, Matthijs and Peter). For the Dutch input, an average proportion of .85 of the verbs in utterance-final position was non-finite. For the German input the average proportion was .77. The overall proportion of non-finites in utterance-final position in German is thus higher than for Leo, suggesting Leo's maternal speech does differ somewhat from that for the average German child. However, the difference between Dutch and German was statistically significant ($t(11) = 2.63$, $p < .05$), suggesting that overall compound finites are more frequent in Dutch.

The higher rate of Dutch utterance-final non-finites was also reflected in the children's use of OIs. At an MLU of approximately 1.5,[6] an average proportion of .74 of the Dutch children's utterances containing verbs were OIs. For the German children this proportion was .60. Again, this difference was statistically significant ($t(11) = 2.33$, $p < .05$). Finally, the level of non-finites in utterance-final position in the input was also predictive of the levels of OIs that the children produced: the Spearman's rank order

[6] The average MLU for the Dutch children was 1.53. For the German children this was 1.59. This difference was not significant ($t(11) = .64$, $p > .50$).

correlation between the proportion of non-finites in utterance-final position in the input and OIs produced by the children is .70 ($p < .01$).

The comparison of Dutch and German thus shows that the initial use of non-finite verb forms in German children is less pronounced than it is in Dutch children. This difference is mirrored in the input files for several children. In the Dutch input, non-finite verbs make up a higher proportion of the verbs in sentence-final position. This finding suggests that subtle differences in the distributional characteristics of languages that display the OI phenomenon may determine the rates at which children produce OI errors. It thus provides strong support for the role of input-driven learning and the utterance-final bias as determinants of children's early multi-word speech.

4.2.3 A comparison of the Dutch and German input

As was mentioned earlier, there are several subtle differences between German and Dutch grammar that might account for the lower levels of non-finites in utterance-final position in German. Thus, German lacks a progressive, and appears to use auxiliary/modal plus infinitive constructions for future events less frequently. In order to ascertain whether these differences could account for the lower levels of non-finites in utterance-final position in German, the rate at which these constructions occurred in the child-directed speech for Leo, Peter and Matthijs was examined.

4.2.3a Progressives in Dutch and German

German, unlike Dutch, does not have a progressive construction. Thus, in German, an enquiry into what a person is doing would be phrased as a (finite) present tense (e.g. *Was*

machst-FIN du? (What do you?/ What are you doing?)). Dutch does have a progressive, which (unlike in English) is formed using auxiliary *be*, the infinitive and the phrase *aan het* (e.g. *Wat ben je aan het doen?* (Wat are you 'on it' do-INF?/What are you doing?)). Furthermore, Dutch uses verbs like *zitten* (sit), *lopen* (walk) and *staan* (stand) in combination with the infinitive to indicate ongoing events in phrases like *Zit je te spelen?* (Sit you to play-INF?/ Are you (sitting and) playing?). As ongoing events will presumably feature relatively frequently in child-directed speech, this difference between Dutch and German may lead to higher rates of infinitives in Dutch child-directed speech. An analysis of the input files for Peter and Matthijs showed roughly 90 occurrences of the phrase '*aan het*', which (when preceding an infinitive) unambiguously identifies the progressive in each file. A search for utterances containing forms of the verbs *sit, walk*, and *stand* as well as the infinitival marker *te* (to) revealed roughly 50 instances in both Peter's and Matthijs' input files. While these absolute numbers may not seem particularly large, these progressive constructions make up almost 2 percent of all utterances containing a verb in the maternal speech directed at Peter and Matthijs. The Dutch use of the progressive therefore accounts for some of the difference in utterance-final non-finites in Dutch and German.

4.2.3b Future tense in German and Dutch

Dutch and German also differ in the manner in which future tense is expressed. In both languages, future events and intentions can be expressed using a modal + infinitive construction, as in the English *I will go*. These constructions are relatively rare, and the present tense is often used to refer to future events, particularly when the verb or other

words already indicate that the event is to take place in the future (*Wir kommen morgen* (We come tomorrow/We will be coming tomorrow)). Dutch, however, has a third means for expressing future events/intentions: the use of auxiliary *go* + infinitive. Thus, a phrase like *Ik ga spelen* (I go play/ I am going to play) can be used to express an intent to play. Such constructions (especially in interrogative contexts) are quite frequent in child-directed speech where mothers enquire about children's actions and intentions. An analysis of the maternal speech in Matthijs's and Peter's corpus showed that both corpora contained about 850 (present tense) instances of the verb *go*. This amounts to approximately 10% of all utterances containing a verb. An analysis of a random sample of 100 utterances containing a present tense form of *go* from Matthijs's maternal speech showed that in 75% of these utterances the verb *go* was used as an auxiliary combined with an infinitive form. Thus, modal *go* makes up around 7-8% of all the child-directed speech containing a verb for Matthijs. In contrast, present tense forms of *go* occur in less than 2% of the German maternal speech. In a sample of 100 of these utterances *go* was used as an auxiliary only 3 times. Thus, while modal *go* makes up a significant portion of Dutch maternal speech, it is virtually non-existent in German.

A final possibility is that German speakers tend to use the regular future tense (Dutch *zullen*, German *werden*) where Dutch speakers use modal *go.* If this were the case, it would tend to work against the effect described above. However, our data do not provide any support for this idea. Leo's input corpus contained 129 instances of the regular future tense, Peter's corpus contained 106, and Matthijs' contained 194. While these absolute numbers do not differ very much, it should be borne in mind that the German corpus is

2.5 times larger than the Dutch corpora. The relative frequency of regular future tense is thus considerably higher in Dutch than in German.

In conclusion, in the three input corpora analysed, utterance-final non-finites are more frequent in Dutch by about 20 percentage points. Roughly half of this difference is accounted for by the Dutch use of the progressive (2%), modal *go* (8%) and the regular future tense (2%).

4.2.4 Summary of the Dutch and German comparison

In summary, the group analysis shows that, early in development, Dutch OIs outnumber German OIs by about 15 percentage points. This difference is also apparent in the proportion of utterance-final non-finites in the input across 13 Dutch and German children, with Dutch utterance-final non-finites being more frequent by about 8 percentage points. Utterance-final non-finites are also predictive of the children's rates of OIs with a rank-order correlation of .70. A detailed analysis of three individual corpora finally showed that the higher rates of utterance-final non-finites in Dutch closely maps onto the use of modal constructions such as modal *go*, and the progressive (which are absent from German) as well the regular future tense. Given that Dutch and German compound and finite (main) clauses pattern identically, these results strongly suggest that the distributional statistics of the input directly affect the rates at which children produce OI errors.

4.3 Results for the Spanish simulation

--------- Insert Fig. 6 about here ----------

Fig. 6 shows the results for the Spanish simulation. For the sake of simplicity and ease of comparison with earlier analyses, bare infinitives and other non-finites are classed together in this graph. A more detailed analysis of the rates of infinitives and other non-finites will be conducted later. As Fig. 6 shows, OIs are rare in Spanish compared to the other languages (in particular German and Dutch). This pattern is clearly reflected in the simulations as well (RMSEs are .05, .05 and .11 for the three different MLU points; r^2=.88). Thus, despite compound finites in Spanish occurring at rates that are similar to German and Dutch, MOSAIC successfully simulates the low levels of OIs in Spanish.

Having established that MOSAIC successfully simulates the low levels of OIs in Spanish, we can now turn to an analysis of the input to investigate the reasons why MOSAIC simulates this result. As in OI languages, compound finites are the potential sources from which MOSAIC could produce Spanish OIs. In Spanish compound finites, the auxiliary or modal precedes the infinitive, as it does in Dutch and German. Thus, the compound finite *Quiero beber café* ((I) want drink-INF coffee) would give rise to the OI *beber café* (drink-INF coffee) if the modal were omitted. Why then does MOSAIC produce so few Spanish OIs when compound finites occur in Juan's input file at rates that are roughly comparable to OI languages (.25, .35 and .22 for the Spanish, Dutch and German input files used for the MOSAIC simulations, respectively)? The answer is that

despite compound finites being equally frequent across the languages, the rates of non-finites in utterance-final position are very different. Thus, only 26% of the Spanish verbs in sentence-final position are non-finite, compared to 65% for German and 85% for Dutch. MOSAIC's utterance-final bias is thus the key factor in transforming roughly equal rates of compound finites into radically different rates of OIs. Underlying this radical transformation of the numbers when read through an utterance-final bias are differences between the respective grammars. Spanish, German and Dutch share the feature that, in main clauses, finite verbs occur early in the sentence and normally precede potential complements such as (prepositional) objects. Object-Verb order for non-finite verbs differs across the languages, however. In Dutch and German, (direct) objects[7] precede the non-finite verb form, which takes sentence-final position, whereas in Spanish objects are normally placed after the non-finite verb form. This is illustrated in Table 1 and the following Dutch and Spanish examples:

Ik wil koffie drinken (I want coffee drink-INF)

Quiero beber café ((I) want drink-INF coffee)

Ik ga in het park wandelen (I go in the park walk-INF)

Voy a pasear en el parque ((I) go to walk-INF in the park).

Thus, while non-finites occur later in the sentence than finites in compound constructions in both Spanish and OI languages, they are more likely to occur in sentence-final position in Dutch and German.

[7] Indirect objects such as prepositional phrases are occasionally placed after the non-finite for reasons of stress, but normally precede the non-finite verb form.

There is an additional feature of Spanish that may lead to finites being more frequent in utterance-final position. Dutch, German and Spanish share the feature that intransitive finite verbs take sentence-final position (as this overlaps with V2) in simple Subject-Verb utterances (e.g. *Hij loopt* (He walks)). For transitive verbs, a lexical object will appear after the verb (e.g. *Hij trapt de bal* (He kicks the ball)). Unlike German and Dutch, however, Spanish allows the placing of pronominal objects before the finite verb (e.g. *(Yo) Lo quiero* ((I) it want)), leading to an utterance-final finite verb in a construction that does not give rise to an utterance-final verb in Dutch or German.

Thus, while the relative proportions of simple and compound finites in Spanish, Dutch and German are roughly equal, the differing rules of verb and object placement lead to differing proportions of utterance-final non-finites. The rates at which children produce OIs closely reflect these proportions of utterance-final non-finites. MOSAIC in turn closely matches the rates of OIs that children produce as a result of its utterance-final bias.

As indicated in Table 1, Spanish and English follow similar rules of object and verb placement. However, OI errors in English are more frequent than they are in Spanish. One reason for this difference is that Spanish allows preverbal pronominal objects (i.e. clitics) resulting in utterance-final finites. More importantly, however, compound finites make up a much larger proportion of the input for English[8] (.70), than it does for Spanish (.25). There are two main reasons for this difference. First, usage of the progressive (e.g *he is eating*) is far more frequent in English that it is in Spanish. An analysis of the input for Anne and Becky shows that this construction alone makes up approximately 25% of

[8] For English, the input refers to utterances containing a third singular subject.

the utterances in the input that contain a third singular and a main verb. Thus, the English progressive alone makes up a portion of the input that is equivalent to all combined compound finites for Spanish. The progressive in Spanish is relatively infrequent: an analysis of Juan's input shows that only 3% of his input (utterances containing verbs) consists of progressives. A second reason for the high proportion of compound finites in the English input is the use of the English dummy modal *do* in negation and question formation. In English, a simple finite (e.g. *He wants a cookie*), is transformed into a compound finite through negation (*He doesn't want a cookie*) or question formation (*Does he want a cookie?*). In Spanish, these processes do not result in a compound finite. Thus, the negated form of (*Juan) Quiere una galleta* ((Juan) Wants a cookie) is the simple finite (Juan) No *quiere una galleta* ((Juan) Not wants a cookie). Likewise, this utterance is transformed into a (finite) question by either changing the intonation contour, or through main verb inversion *¿Quiere (Juan) una galleta?* (Wants (Juan) a cookie?). Like the English progressive, the use of dummy modal *do* makes up a significant proportion of the English input: 15% of all utterances containing a third singular and a main verb. Thus, the relatively high level of OI errors in English is explained by the fact that compound finites make up 75% of the English input as opposed to 25% of the Spanish input. Almost 40 percentage points of this difference are explained by the extensive use of progressives (~25%) and the use of dummy modal *do* (~15%). A further difference between Spanish and English is (as mentioned earlier) that Spanish allows preverbal pronominal objects (e.g. (*él) lo* quiere ((he) it wants)). Such constructions result in an utterance-final finite, where the English equivalent contains an utterance-final object.

4.3.1 Rates of bare infinitives and other non-finites in Spanish

Wexler (1998) predicts that bare infinitives and other non-finites occur at differential rates in Spanish. According to Wexler, children are subject to a Unique Checking Constraint that impacts on their ability to check the D-feature of the subject DP against more than one D-feature. In languages like English and Dutch this results in the child producing an infinitive in a finite context, as the finite main verb requires the checking of both Tense and Agreement. In Spanish, only the D-feature of Tense needs to be checked in these (simple) finite contexts. As a result, these contexts rarely give rise to bare infinitives. Constructions containing an auxiliary and a progressive or perfect participle however require checking of the D-feature for both Tense and Agreement on the auxiliary. This results in omission of the Auxiliary, and consequently, a bare progressive or past participle. The rates of bare infinitives (relative to bare infinitives plus simple finites) and other non-finites (relative to bare non-finites plus auxiliary + progressive/perfect constructions) for Juan and the simulation are shown in Table 2.

---------- Insert Table 2 about here ----------

Table 2 shows a clear disparity between the rates of bare infinitives and other non-finites for both the child and the simulation. The reason MOSAIC simulates this result is because OIs in MOSAIC are not infinitives produced in a (simple) finite context (as assumed by Wexler), but rather compound finites with a missing modal or auxiliary. In line with Wexler's analysis the rate of bare infinitives is expressed relative to simple

finites. Since simple finites are quite frequent, the rate of bare infinitives will be low. The rate of bare other non-finites however is expressed relative to auxiliary plus perfect/progressive constructions (which are assumed to give rise to bare perfect/progressive participles). The denominator for this comparison will be low since auxiliary plus perfect/progressive constructions are relatively infrequent. As a result, the rate of bare participles is relatively high.

4.3.2 Omission of Auxiliary estar (be) and haber (have)

A further distinction that can be made in the Spanish data relates to the different distribution of bare progressive and perfect participles. Aguado-Orea (2004) analysed the data for Juan and a second Spanish child and found that these children omitted auxiliary *estar (be)* in auxiliary plus progressive constructions more often than they omitted auxiliary *haber (have)* in auxiliary plus perfect constructions. This finding is problematic for Wexler, as it is unclear how such a distinction between progressive and perfect constructions would be accounted for in his theory. In order to test whether MOSAIC simulates this finding, the ratio of bare progressives to bare plus compound progressives was compared to the ratio of bare perfects to bare plus compound perfects. The resulting rates for Juan and the simulation are shown for the three different MLU points in Table 3. While MOSAIC underestimates the omission rates for progressives and overestimates omission rates for perfects, it does capture the effect.

---------- Insert Table 3 about here ----------

An identical analysis of the input file suggests two reasons for this effect. First, as pointed out by Aguado-Orea (2004), bare perfects and progressives (which are acceptable as elliptical answers) occur at different rates in the input. Ten percent of the progressives occur in a bare form, whereas this is the case for only 3% of the perfects. A second reason for the lower omission rates for perfects is that perfects are more frequent than progressives. Thus, the input file for Juan contains approximately 400 progressive and 950 perfect participles. Since perfect participles are more frequent, they will be learned more quickly by MOSAIC, and will subsequently be absorbed into compound constructions more quickly.

4.4 Summary of results

MOSAIC has been applied to the simulation of the OI phenomenon in four languages: English, Dutch, German and Spanish. OIs occur at different rates in these languages, and MOSAIC provides a good fit to all languages without fitting any free parameters. MOSAIC produces OIs in all languages from the same constructions: compound finites. In all four languages, finite verbs precede non-finite verbs in compound finites. Omission of sentence-initial words from compound finites therefore results in OIs in all languages. Despite compound finites occurring at roughly equal rates across three of the languages, the rates with which children produce OIs differ across these languages. The analyses presented here indicate that the levels of OIs produced in early child speech map closely

onto the proportions of non-finites that occur in utterance-final position. Table 4 summarizes this finding[9].

---------- Insert Table 4 about here ----------

Two main reasons have been identified for the differing rates of non-finites in utterance-final position. First, the comparison of Dutch and German showed that compound finites are actually more frequent in Dutch. This effect is largely carried by the Dutch use of modal *go*. Second, differences in the way in which (compound) finites pattern account for the differences between German and Dutch on the one hand and Spanish on the other hand. Spanish compound finites are less likely to include the infinitive in sentence-final position, as this position may be occupied by an object argument. In simple finites, the finite verb form is more likely to occur in sentence-final position in Spanish, as pronominal objects may be placed before the verb. English differs from all other languages in that it has very high levels of compound finites, which results in high levels of OI errors compared to Spanish, despite the fact that the two languages have the same word order.

The strong cross-linguistic relation between the proportion of non-finites in utterance-final position and the rates of OI errors produced by the children thus provides strong support for the view that the cross-linguistic patterning of finiteness marking can be

[9] While English is included in the table, it should be noted that a direct comparison between English and the other three languages is not possible as the English measures are necessarily based only on data from third singular contexts.

explained in terms of the interaction between an utterance-final bias in learning and the distributional properties of the input.

5. Discussion

This paper set out to investigate whether the interaction between MOSAIC's utterance-final bias and the distributional statistics of the input is sufficient to explain the variation in the occurrence of the OI phenomenon across four languages: English, Dutch, German and Spanish. MOSAIC's utterance-final bias results in the production of OI errors through the omission of utterance-initial phrases from compound finites. Thus, in MOSAIC, OIs are viewed as incomplete compound finites. As the length of the utterances MOSAIC produces increases, OIs are slowly absorbed into the compound finites from which they were learned.

An earlier version of MOSAIC has already successfully simulated the OI phenomenon in English and Dutch, but suffered some implementational weaknesses that resulted in the model not accurately reflecting the theory underlying it. For the present simulations, the implementation of MOSAIC was therefore brought more in line with the underlying theoretical model. To this end, MOSAIC's learning mechanism was altered so that the model builds up its representation of the input to which it is exposed from the right edge of the utterance. Furthermore, an unrealistic constraint in the generation of novel utterances was removed, and a mechanism for unlearning was developed. In addition, a new chunking mechanism, which treats frequent phrases as single units, was implemented. Despite these implementational changes, which all had the potential to affect MOSAIC's fit to the data, MOSAIC maintains a good fit to the developmental data

from English and Dutch. This finding suggests that MOSAIC's basic mechanism for simulating the OI phenomenon (production of increasingly long utterance-final phrases) does so in a fairly robust way that is not too dependent on implementational details. This robustness is further underscored by the fact that the simulations for English and Dutch as well as the novel simulations for Spanish and German were conducted using one, identical model. Thus, no free parameters were fitted and no changes were made to the model across the languages. In fact, the only difference between the simulations for the different languages was that child-directed speech from these different languages was used as input.

A second aim of the paper was to extend the range of languages that MOSAIC simulates. The two languages modelled with the earlier version of MOSAIC constituted a fairly restricted sub-set of the range of languages that display the OI phenomenon. It was argued that even within OI languages the OI phenomenon may be subject to cross-linguistic variation related to the distributional characteristics of the input, and the simulation of a third OI language might highlight such differences. German was chosen as a third OI language because it shares many relevant grammatical features with Dutch, yet it differs in relatively subtle ways with respect to the use of some compound finite constructions.

A greater challenge however, was to investigate whether MOSAIC was capable of simulating a non-OI language such as Spanish. The simulation of Spanish is a challenge for MOSAIC because compound finites, the source of OI errors in MOSAIC, do not occur at significantly lower rates in Spanish than in OI languages such as Dutch and German. However, OI errors are rare in Spanish child speech. On the face of it, the

finding that OI errors are rare in Spanish while compound finites are not appears to count against the notion that OIs are incomplete compound finites.

MOSAIC has no free parameters that can be manipulated to produce differential levels of OIs from comparable rates of compound finites in different languages. The only source of variation that can result in the differing levels of OIs is therefore the distributional statistics of the input. Since these do not differ in terms of the rates at which compound finites occur, MOSAIC is dependent on compound finites patterning differently across languages. The mechanism in MOSAIC that is sensitive to such differences is its learning mechanism, in particular its utterance-final bias. Thus, the interaction between the distributional statistics of the target language and MOSAIC's utterance-final bias is crucial for MOSAIC's successful simulation of the difference between OI and non-OI languages.

MOSAIC clearly provides a good fit to the data from all four languages, suggesting there is sufficient variation in the input sets to explain the cross-linguistic variation in the child data. This finding strongly suggests that differences in the rates at which children produce OI errors are graded, quantitative differences which reflect quantitative differences in the distributional statistics of the input, rather than qualitative, structural differences between the languages. Thus, when it comes to the OI phenomenon, languages are better classified on a continuum rather than a dichotomy.

The suggestion that the difference between OI and non-OI languages is quantitative rather than qualitative is not a new one. Phillips (1995) provides a scatterplot depicting rates of OIs from 27 children learning nine different languages. The scatterplot shows considerable variation across languages and MLU. Phillips notes that rates of OIs are

very low in Null-Subject languages, but concludes that 'the rates are by no means zero' (Phillips, 1995 p. 264). In fact, Phillips reports that for one Italian child, OIs make up 13% of the root verbs at a very young age. Phillips concludes that 'these findings raise the possibility that what causes optional infinitives is not absent in Italian, but rather that it drops away at an extremely early age…' (Phillips, 1995; p. 265). The analyses reported in this paper suggest that what causes OIs is the occurrence of non-finite verb forms in utterance-final position in the input. In OI languages, these occur with high frequency. In non-OI languages they are infrequent.

Two sources of variation have been identified that affect these rates of non-finites in utterance-final position. The comparison between Dutch and German showed that Dutch mothers use compound finites with utterance-final non-finites more when addressing their children than German mothers do. In line with this higher rate of compound finites, Dutch children produce more OIs (early in development) than German children. In fact, across seven Dutch and six German children, the rank order correlation between the proportion of utterance-final non-finites and the proportion of OIs produced by the children was .70. This strong relation between the proportion of compound finites and number of OI errors in two languages that are very closely related is also borne out in the larger difference between English and Spanish. Thus, the high rate of OI errors in English compared to Spanish is accounted for by the high rates of compound finites in English, in particular progressives and constructions including dummy modal *do*.

A second source of variation results from the way in which different languages pattern. The input for the Spanish child contains slightly more compound finites than the input for the German child. OIs in the Spanish child and simulations, however, occur at

considerably lower rates than in the German child. The levels of OIs displayed by the children again closely mirror the proportion of non-finites in utterance-final position. This is a result of the fact that compound finites pattern differently in the two languages. In German (and Dutch), objects precede non-finite verb forms, which take utterance-final position. In Spanish, non-finite verb forms may take sentence-final position, but they are often followed by an object argument, leading to lower rates of non-finites in utterance-final position.

Two main conclusions can be drawn from how these sources of variation affect the rates of OI errors that children produce. First, the finding that the proportion of compound finites predicts the rates of OI errors across two closely related languages (Dutch and German) strongly suggests that the distributional statistics of the input directly affect the rates with which children produce OIs. Thus, within structurally equivalent languages, there is input-dependent variation in terms of the rates with which children produce OIs. The implication is that greater consideration should be given to the role of the input in theories of the OI phenomenon. It is also worth noting that this input-dependent variation has only become apparent as a result of the relatively detailed quantitative analyses performed here. Since this variation is most apparent at low MLU values, it is likely to remain hidden in more coarse-grained analyses that do not distinguish between different MLU points. The findings of the present study therefore invite a greater focus on detailed quantitative analyses of the rate at which OI errors occur in different languages at different points in development.

The second main conclusion to be drawn is that differences in how languages pattern can render cross-linguistic comparisons in terms of the frequency of sentence types that

can give rise to OI errors meaningless. Thus, while compound finites are equally frequent in Spanish and German, compound finites pattern differently in the two languages, giving rise to very different rates of non-finites in utterance-final position. Such an interaction may easily obscure the relevance of certain sentence types for phenomena in child speech. It therefore suggests that a thorough understanding of both the statistics and the structure of a language is required if one is to make meaningful statements about the potential sources of errors in child speech.

Envisaging how the statistics and structure of a language interact is of course no trivial task. Indeed, the fact that the frequency of compound finites in English leads to drastically higher rates of non-finites in utterance-final position than in Spanish, a language that in principle has the same word order, is not something we expect many researchers would have predicted. However, the apparent intractability of such interactions does highlight one of the strengths of our approach. By implementing a theory as a computational model, and subjecting the model to child-directed speech that has a realistic frequency distribution, such interactions can be quantitatively investigated while constraints on the learning mechanism are independently manipulated. These constraints on the learning mechanisms need not be overly complex. In fact, the analyses reported here show that what look like complicated cross-linguistic patterns can be understood in terms of the interaction between differences in the surface properties of different languages and a relatively simple utterance-final bias in learning. As such, they illustrate the dangers of assuming that complex cross-linguistic patterns in the developmental data require a complex formal explanation, and the potential value of an

approach that uses cross-linguistic variation in the developmental data to identify processing mechanisms and constraints that are common to all children.

Acknowledgements:

The authors would like to thank the Max Planck Institute for Evolutionary Anthropology, Leipzig for providing access to the Leo Corpus and Heike Behrens for her valuable help and advice when analysing the Dutch and German data. This research was funded by the Economic and Social Research Council under grant number R000223954.

References

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

Behrens, H. (2006). The input-output relationship in first language acquisition. *Language and Cognitive Processes*, *21*, 2-24.

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: The early stages*. Cambridge, MA: Harvard University Press.

Chase, W.G. & Simon, H.A. (1973). Perception in chess. *Cognitive Psychology*, *4*, 55-81.

Freudenthal, D., Pine, J.M. & Gobet, F. (2006). Modelling the development of children's use of Optional Infinitives in Dutch and English using MOSAIC. *Cognitive Science*, *30*, 277-310.

Freudenthal, D., Pine, J.M. & Gobet, F. (2005). On the resolution of ambiguities in the extraction of syntactic categories through chunking. *Cognitive Systems Research*, 6, 17-25.

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences, 5*, 236-243.

Gobet, F., & Simon, H. A. (1998). Expert chess memory: Revisiting the chunking hypothesis. *Memory, 6*, 225-255.

Harris, T. & Wexler, K. (1996). The optional-infinitive stage in child English: Evidence from negation. In H. Clahsen (Ed.), *Generative perspectives in language acquisition* (pp. 1-42). Philadelphia: John Benjamins.

Ingram, D. & Thompson, P. (1996). Early syntactic acquisition in German: evidence for the modal hypothesis. *Language*, *72*, 97-120.

Jordens, P. (1990). The acquisition of verb placement in Dutch and German. *Linguistics, 28,* 1407-1448.

MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk (3rd Edition).* Mahwah, NJ: Erlbaum.

Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child-directed speech. *Cognition,90,* 91-117.

Phillips, C. (1995). Syntax at age two: Cross-linguistic differences. In C. Schütze, J. Ganger & K. Broihier (eds), *Papers on Language Processing and Acquisition*. MITWPL #26, 225-282.

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

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

Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language*, 21, 109-141.

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* (pp. 305-350). 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). Bare infinitives in Dutch early child language: an effect of input? *Journal of Child Language*, *28*, 629-660.

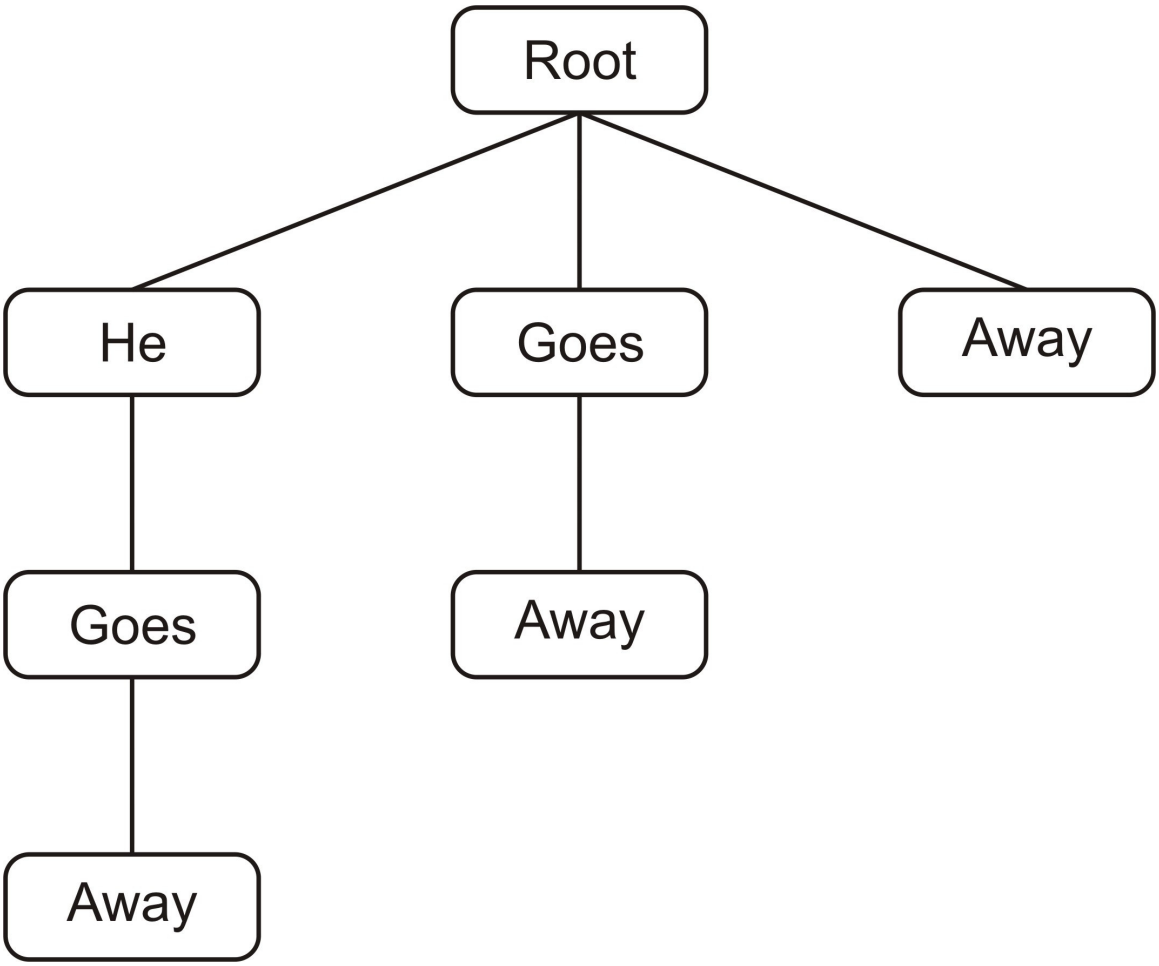

Fig. 1: A MOSAIC network after it has seen the phrase *He goes away* five times.

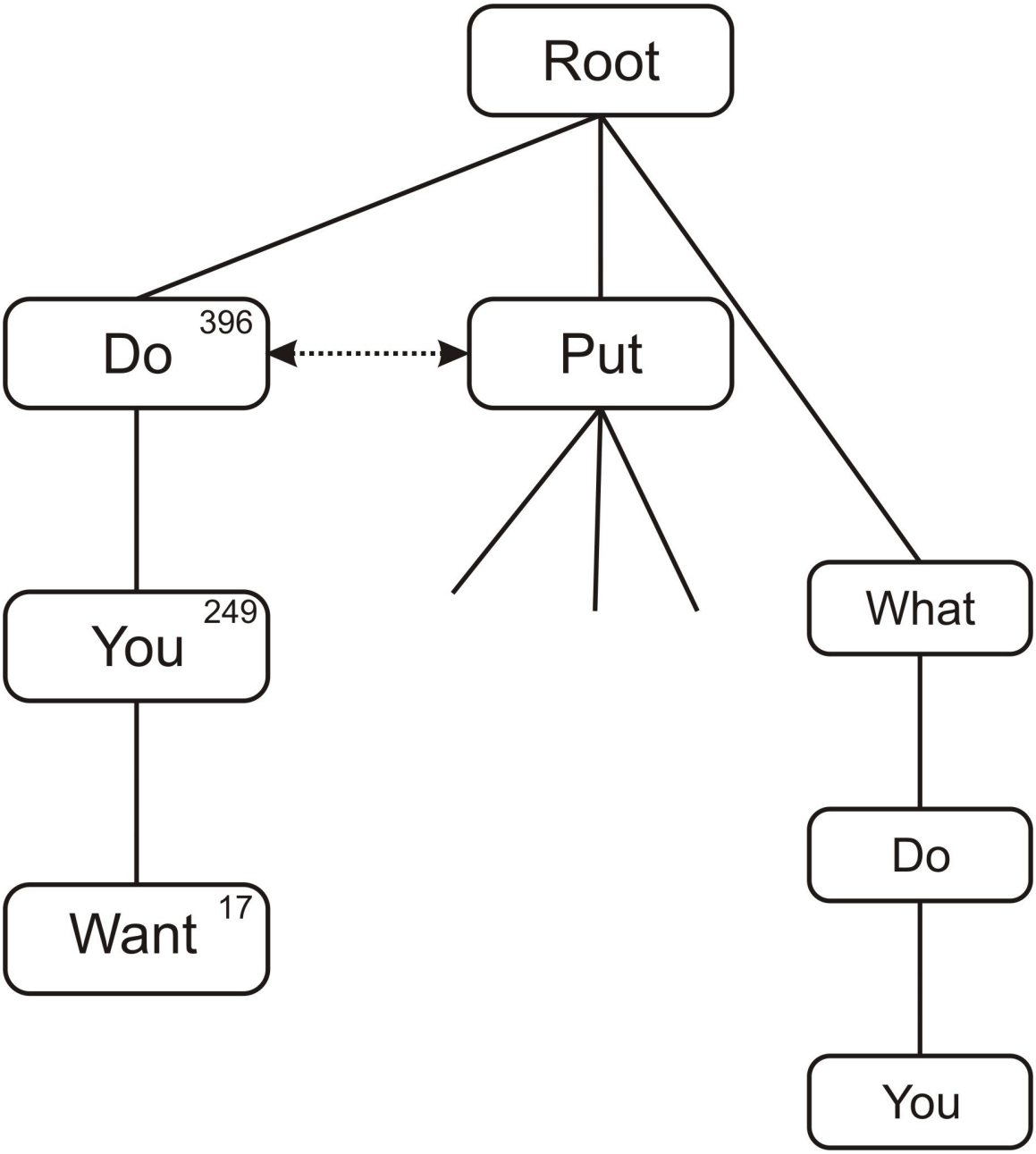

Fig. 2a: A sample MOSAIC network before the phrase *do you* has been chunked.

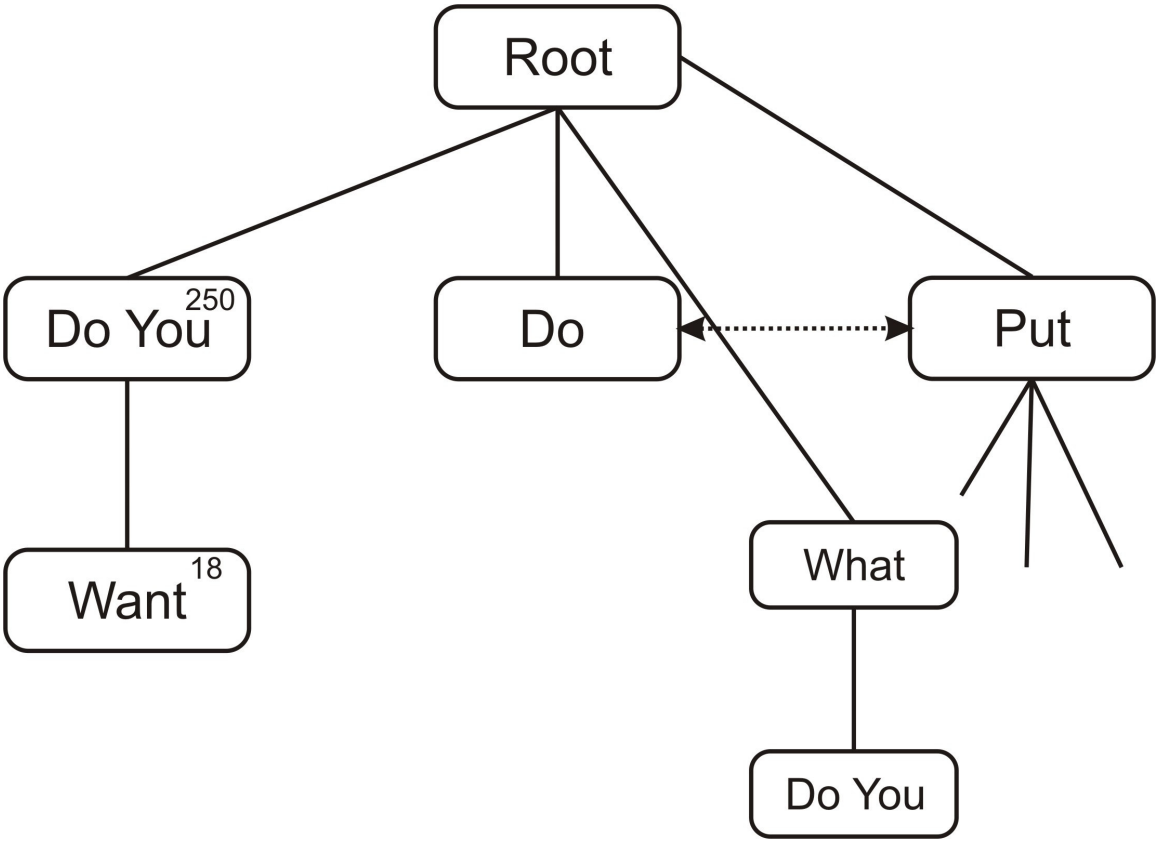

Fig. 2b: The same network after the phrase *do you* has been chunked.

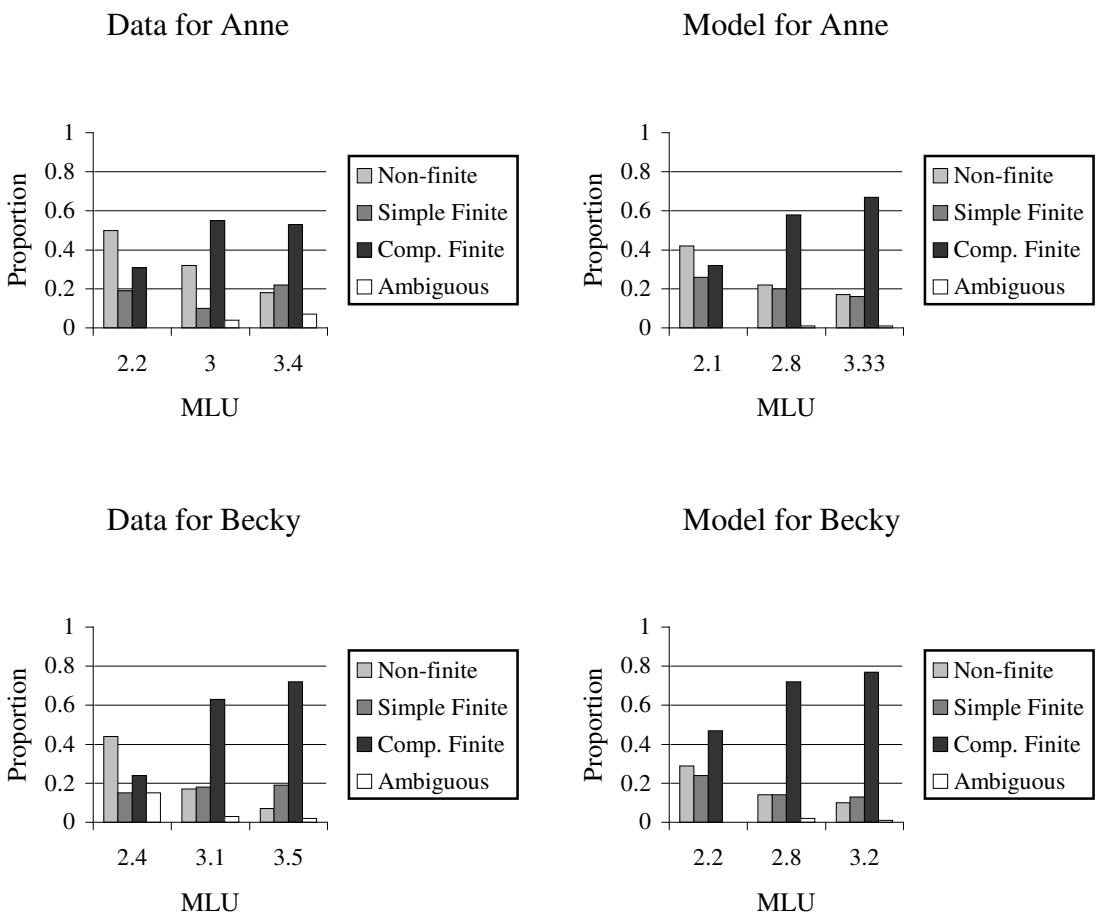

Fig. 3: Data and simulations for English. The number of utterances contributing to the analysis increases (across MLU points) from 26 to 74 for Anne and from 33 to 103 for Becky. For the models, the increase is from 57 to 1730 (Anne), and 59 to 1127 (Becky).

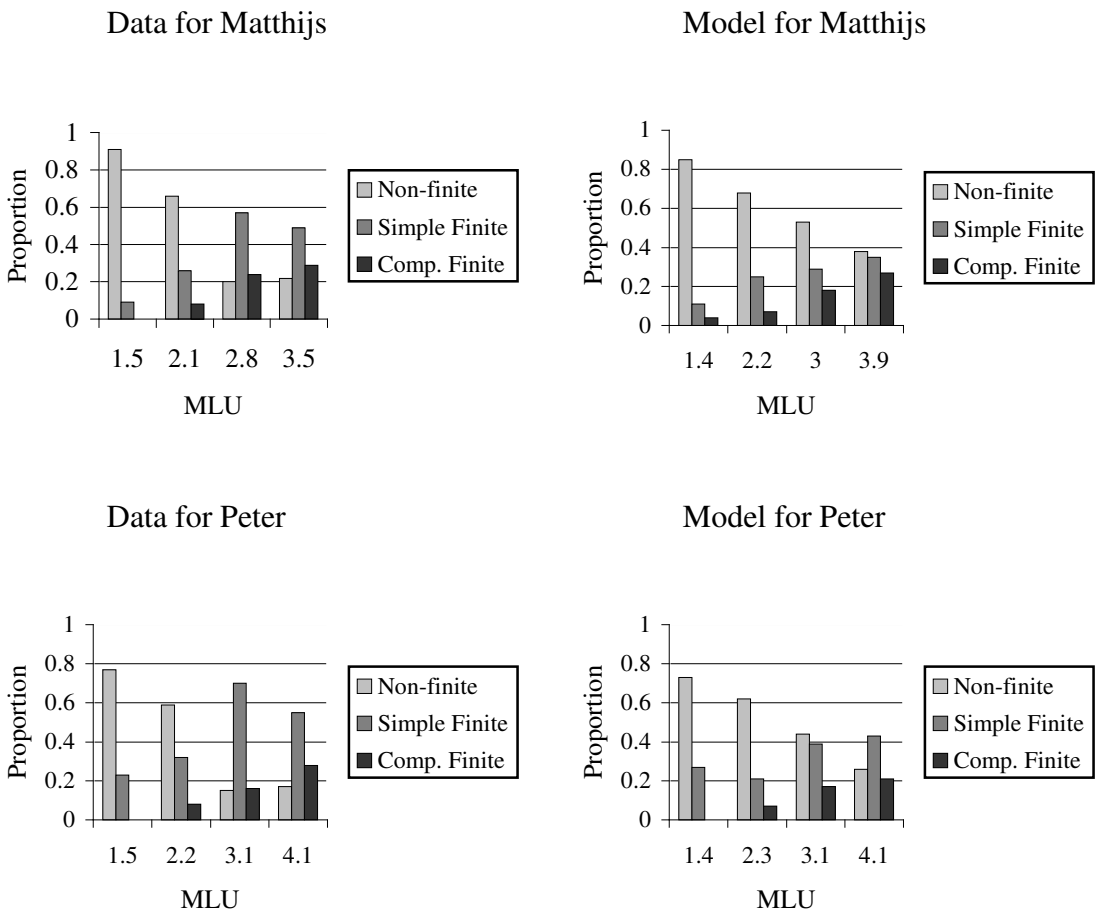

Fig. 4: Data and Simulations for Dutch. The number of utterances contributing to the analysis increases (across MLU points) from 98 to 1459 for Matthijs and from 65 to 676 for Peter. For the models, the increase is from 54 to 7117 (Matthijs), and 88 to 6857 (Peter).

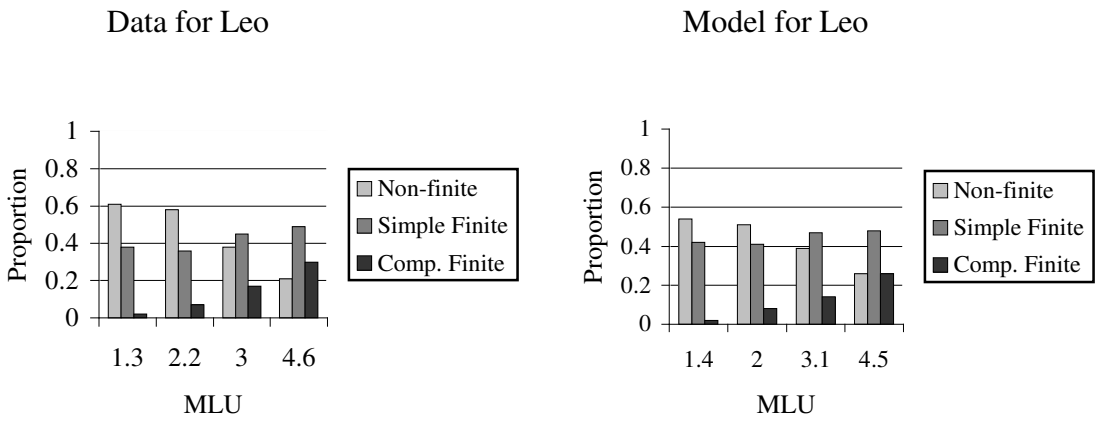

Fig. 5: Data and Simulations for German. The number of utterances contributing to the analysis increases (across MLU points) from 345 to 4696 for Leo, and from 197 to 18774 for his model.

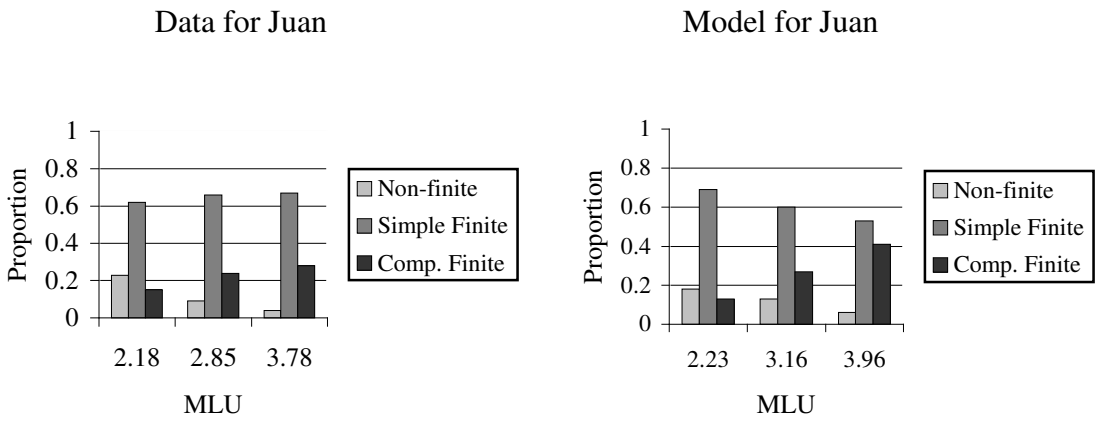

Fig. 6: Data and Simulations for Spanish. The number of utterances contributing to the analysis increases (across MLU points) from 429 to 1995 for Juan, and from 1722 to 19455 for his model.

Table 1: Summary descriptions and examples of simple and compound finite constructions for English, Spanish, Dutch and German.

English	**Spanish**	**Dutch**	**German**
Canonical word order: SVO	Canonical word order: SVO	Canonical word order: SOV/V2 (SVO for simple finites)	Canonical word order: SOV/V2 (SVO for simple finites)
Proportion of compound finites in input: .70	Proportion of compound finites in input: .25	Proportion of compound finites in input: .34	Proportion of compound finites in input: .29

Simple Finites	**Simple Finites**	**Simple Finites**	**Simple Finites**
He jumps	(él) salt-a ((He) jump-3S)	Hij spring-t (He jump-3S)	Er spring-t (He jump-3S)
He eats ice cream	(él) com-e helado ((He) eat-3S ice cream)	Hij eet ijs (He eat-3S ice cream)	Er iss-t Eis (He eat-3S ice cream)
He goes to the park	(él) va al parque ((He) go-3S to the park)	Hij gaat naar het park (He go-3S to the park)	Er geh-t in den Park (He go-3S to the park)
He puts the book on the table	(él) pon-e el libro en la mesa ((He) put-3S the book on the table	Hij leg-t het boek op de tafel (He put-3S the book on the table)	Er leg-t das Buch auf den Tisch (He put-3S the book on the table)
He wants it	(él) lo quier-e	Hij wil-t het	Er will es

	((He) it want-3S)	(He want-3S it)	(He want-3S it)
Compound Finites	**Compound Finites**	**Compound Finites**	**Compound Finites**
He can jump	(él) pued-e salt-ar ((He) can-3S jump-INF)	Hij kan spring-en (He can-3S jump-INF)	Er kann spring-en (He can-3S jump-INF)
He can go to the park	(él) pued-e ir al parque ((He) can-3S go-INF to the park)	Hij kan naar het park gaan (He can-3S to the park go-INF)	Er kann in den Park geh-en (He can-3S to the park go-INF)
He can put the book on the table	(él) pued-e pon-er el libro en la mesa ((He) can-3S put-INF the book on the table)	Hij kan het boek op de tafel leg(g)-en (He can-3S the book on the table put-INF)	Er kann das Buch auf den Tisch leg-en (He can-3S the book on the table put-INF)

He has put the book on the table	(él) ha puesto el libro en la mesa ((He) have-3S put-PERF the book on the table)	Hij heeft het boek op de tafel ge-leg-d (He have-3S the book on the table put-PERF)	Er hat das Buch auf den Tisch ge-leg-t (He have-3S the book on the table put-PERF)
He is eating	(él) está com-iendo ((He) be-3S eat-PROG	Hij is aan het eten (He be-3S 'on it' eat-INF)	N.A.
He is going to walk	(él) va a andar ((He) go-3S to walk-INF)	Hij gaat lopen (He go-3S walk-INF)	N.A.

Table 2: Rates of bare infinitives and other non-finites for Juan and his simulation for the three different MLU points shown in figure 6.

	Juan Data		Juan Model	
	Infinitives	Other Non-Finites	Infinitives	Other Non-Finites
MLU1	.20	.41	.14	.48
MLU2	.09	.23	.13	.26
MLU3	.04	.11	.08	.11

Table 3. Omission rates for auxiliary *estar* (*be)* and *haber* (*have)* in progressive and perfect participle constructions for Juan and his simulation at three MLU points.

		MLU1	MLU2	MLU3
Juan	Progressives	.72	.57	.57
	Perfects	.36	.14	.04
Simulation	Progressives	.57	.41	.29
	Perfects	.51	.26	.11

Table 4: Proportion of compound finites and utterance-final non-finites in the input and proportions of OIs produced by a German, a Dutch, a Spanish and an English child.

	Proportion of compound finites in input	Proportion of Utterance-final Non-finites in input	Proportion of OIs produced by children
Dutch (Peter)	.34	.87	.74
German (Leo)	.29	.66	.61
Spanish (Juan)	.25	.26	.22
English (Anne)	.70	.76	.50