Freudenthal, D., Pine, J. M., & Gobet, F. (in press). Simulating the referential properties of Dutch, German and English Root Infinitives in MOSAIC. *Language, Learning and Development.*

Simulating the referential properties of Dutch, German and English Root Infinitives in MOSAIC

Daniel Freudenthal, Julian M. Pine

School of Psychology, University of Liverpool

Fernand Gobet

School of Social Sciences, Brunel University

ABSTRACT

Children learning many languages go through an Optional Infinitive stage in which they produce non-finite verb forms in contexts in which a finite verb form is required (e.g. ‘That go there’ instead of ‘That goes there’). MOSAIC (Model of Syntax Acquisition in Children) is a computational model of language learning that successfully simulates the developmental patterning of the Optional Infinitive (OI) phenomenon in English, Dutch, German and Spanish (Freudenthal, Pine, Aguado-Orea & Gobet, 2007). In the present study, MOSAIC is applied to the simulation of certain subtle but theoretically important phenomena in the cross-linguistic patterning of the OI phenomenon that are typically assumed to require a more complex formal analysis. MOSAIC is shown to successfully simulate 1) The Modal Reference Effect: the finding that Dutch and German children tend to use Root Infinitives in modal contexts, 2) The Eventivity constraint: the finding that Dutch and German Root Infinitives refer predominantly to actions rather than static situations, and 3) The absence or reduced size of these effects in English. These results provide strong support for input-driven explanations of the Modal Reference Effect as well as MOSAIC’s mechanism for producing Root Infinitives, and the wider claim that it is possible to explain key aspects of children’s early multi-word speech in terms of the interaction between a resource-limited distributional learning mechanism and the surface properties of the language to which children are exposed.

1. Introduction

A central task facing models of language acquisition is to explain cross-linguistic variation in children's early language. Within the framework of generative linguistics such variation is typically explained in terms of the interaction between Principles of Universal Grammar and differences in the syntactic properties of the language being acquired. However, an alternative approach is to explain cross-linguistic variation in terms of the interaction between processing mechanisms that are common to all children and differences in the semantic-distributional properties of the language to which different children are exposed. In our previous work, we have developed a computational model of children's early language (MOSAIC) that is able to simulate cross-linguistic variation with respect to the Optional Infinitive (OI) phenomenon — a phenomenon that has been the subject of considerable generativist theorising. This work shows that it is possible to simulate differences in the speech of children learning OI and non-OI languages as a function of the interaction between psychologically motivated constraints on learning and differences in the distributional properties of OI and non-OI languages. In the present study, we investigate whether the same mechanism that is able to capture differences in the speech of children learning OI and non-OI languages can also account for subtle differences in the patterning of errors across OI languages. The aim is to show that the kind of input-driven model of the OI phenomenon implemented in MOSAIC can explain both the qualitative differences between OI and non-OI languages predicted by generativist models, and certain quantitative differences between OI languages that are more difficult to explain in generativist terms.

1.1. The Optional Infinitive Phenomenon

A central feature of children's early multi-word speech is that, in many languages, children go through a stage in which they produce utterances with verbs that appear to lack Tense or Agreement markers that are obligatory in the adult grammar. For example, English-speaking children often produce utterances like 1a instead of the correct 1b or 1c.

* 1a. He go to school.

1b. He goes to school.

1c. He went to school.

Errors like 1a have traditionally been interpreted as reflecting the omission of inflectional morphemes through lack of knowledge (Brown, 1973) or performance limitations in production (Bloom, 1990; Valian, 1991). However, data from languages such as German and Dutch suggest that these errors may actually reflect the use of infinitive verb forms in contexts in which a finite verb form is required. Thus, Dutch children produce utterances like 2a instead of the correct 2b and German children produce utterances like 3a instead of the correct 3b.

* 2a. Hij ijs eten.

He ice-cream eat-INF.

2b. Hij eet ijs.

He eat-FIN ice-cream.

*	3a. Er Kaffee trinken.

He coffee drink-INF

3b. Er trinkt Kaffee

He drink-FIN coffee

These errors involve the use of infinitive verb forms (marked with the infinitival morpheme: -en and correctly ordered with respect to their complements), in contexts in which a finite verb form is obligatory, and have come to be known in the literature as 'Optional Infinitive (OI)' errors (Wexler, 1994) or 'Root infinitive (RI)' errors (Rizzi, 1994). In fact, children in the Optional Infinitive stage typically make errors with non-finite participles (e.g. *He going to the shops; *He gone to work) as well as with infinitives. As a result, the term 'Optional Infinitive' is often used rather loosely, particularly in the literature on English, to refer to both 'Root Infinitive' and 'Root Participle' errors. Previous work with MOSAIC simulates the patterning of both types of error. However, the phenomena discussed in this paper (i.e. the Modal Reference Effect and the Eventivity Constraint) apply specifically to Root Infinitive errors. For this reason we will use the term 'Optional Infinitive' when discussing the Optional Infinitive phenomenon in general and 'Root Infinitive' when discussing the Modal Reference Effect and the Eventivity Constraint in particular.

A number of theories have been proposed to account for the occurrence of OI errors in children's speech. For example, Hyams (1996) argues that children can leave functional heads such as I (Inflection) and D (Determiner) underspecified in the underlying representation of the sentence, which results in a lack of finiteness in the verbal domain and a lack of specificity in the nominal domain; Rizzi (1994) argues that, rather than projecting a full CP (Complementizer Phrase) structure, children have the option of truncating lower down the clause, with a structure truncated below TP (Tense Phrase) resulting in a non-finite clause; and Phillips (1995) argues that OI errors are fully represented finite clauses, in which merger of the verb with inflection has been delayed and so Tense and Agreement markers have not attached to the lexical verb.

One particularly influential account of OI errors is that of Wexler (1994, 1998; Schütze & Wexler, 1996). Wexler's account is designed to explain the occurrence of OI errors in obligatory subject languages such as English, Dutch, and German (Wexler, 1994), and the absence of such errors in optional subject languages such as Spanish and Italian (Wexler, 1998). According to Wexler (1998) children have correctly set all the inflectional and clause structure parameters of their language from a very early age, but are subject to a 'Unique Checking Constraint' (UCC), which prevents them specifying both Tense and Agreement in the underlying representation of the sentence. Since the under-specification of either Tense or Agreement has the effect of blocking the production of agreeing tensed forms, this results in the production of an infinitive verb form where a finite verb form is required. Optional subject languages such as Spanish require the checking of only one D-feature (Tense) on finite (main) verbs. The UCC therefore does not tend to result in under-specification of Tense or Agreement in these languages and OI errors are rare. In fact, Wexler's (1998) theory predicts that Spanish and Italian children will make Root Participle but not Root Infinitive errors. The data on Spanish and Italian are broadly consistent with this prediction, with Spanish and Italian children producing Root Participle errors at reasonably high rates, but rarely producing Root Infinitive errors (Phillips, 1995; Aguado-Orea, 2004).

The great strength of Wexler's theory is that it provides a unified account of the cross-linguistic pattern of occurrence and non-occurrence of OI errors. At the same time, it explains the low frequency of other types of errors, such as word order errors or the provision of finite verb forms that fail to agree with the subject of the sentence (e.g. *I walks* instead of *I walk*) (Harris & Wexler, 1996). However, the theory fails to provide an explanation of certain subtle effects in the patterning of RI errors across languages, in particular the Modal Reference Effect and the Eventivity Constraint (Hoekstra & Hyams, 1998). The Modal Reference Effect refers to the fact that, in contrast to English, where RI errors and correct finite forms appear to occur in free variation, RI errors in German and Dutch tend to have a modal reading (i.e. to express wishes, desires or intentions rather than to describe ongoing events). The Eventivity Constraint refers to the fact that, in contrast to English, where RI errors are not restricted to a certain subset of verbs, RI errors in German and Dutch occur almost exclusively on verbs that denote actions rather than static situations (Wijnen, 1998; Ingram & Thompson, 1996; Hoekstra & Hyams, 1998).

The aim of the present paper is to investigate whether it is possible to simulate the Modal Reference Effect and the Eventivity Constraint using MOSAIC — a computational model of language learning that has already been used to simulate cross-linguistic variation in the patterning of OI errors. Freudenthal, Pine, Aguado-Orea, and Gobet (2007) have recently shown that it is possible to simulate the pattern of developmental change in OI errors across four languages: English, Dutch, German, and

Spanish, using the same version of MOSAIC. The present study will investigate whether this version of the model can also simulate the Modal Reference Effect and the Eventivity Constraint. Since MOSAIC provides an account of variation in the patterning of OI errors in English, Dutch and German, the success of the model will be evaluated both in terms of its ability to simulate the Modal Reference Effect and the Eventivity Constraint in Dutch and German, and in terms of its ability to simulate the absence (or reduced size) of these effects in English.

1.2. The Modal Reference Effect and the Eventivity Constraint

As Hyams (2001) points out, one problem with optional rule accounts like that of Wexler (1998) is that they predict that root infinitives and finite clauses will be in essentially free variation (i.e. will occur in a similar range of semantic contexts to correct finite forms). However, as indicated above, several authors have noted that when Dutch and German children produce RI errors, they do so predominantly to refer to wishes, intentions, and unrealized events (Ingram & Thompson, 1996; Wijnen, 1998). One possible explanation of this phenomenon is that, rather than being non-finite structures, RIs are finite modal structures (e.g. ‘He can kick the ball’) from which the modal verb (in this case ‘can’) has been omitted. For example, Boser, Lust, Santelmann and Whitman (1992) argue that German children’s apparently non-finite utterances are structurally identical to finite adult utterances, but contain a phonologically null modal in the underlying structure. The Null Modal Hypothesis provides a straightforward explanation of the modal reading of RIs in Dutch and German. Moreover, if it is assumed that modal verbs select eventive predicates, it can also be extended to explain why RIs tend to include eventive rather than stative verbs (Ferdinand, 1996). However, as Hoekstra and Hyams (1998) point out, the Null Modal Hypothesis provides no means of explaining why some languages exhibit an OI stage and others do not.

Hoekstra and Hyams (1998) offer an alternative account of the Modal Reference Effect. On the basis of data collected by Deen (1997), they conclude that the Modal Reference Effect does not hold in English. Deen found that only 13% of the English RIs carry a modal meaning. This stands in marked contrast to data from Dutch and German. Wijnen (1998) found that, across 4 Dutch children, 86% of the RIs were modal. For German, Ingram and Thompson (1996) report a rate of 52% modal RIs using a strict criterion and 77% using a lenient criterion[1]. Hoekstra and Hyams (1998) explain this difference with reference to qualitative differences between the Dutch/German and the English infinitive. They argue that unlike the Dutch and German infinitive (which carries a morphological marker: *-en*), the English infinitive is not a true infinitive, but rather a *bare form*. According to Hoekstra and Hyams, the Dutch/German infinitival morpheme carries an *irrealis* feature, which is responsible for the modal reading of the Dutch/German infinitive. This, they argue, is evident from an analysis of the following utterances:

[1] In fact, the figures reported by Ingram and Thompson are 55% and 79%, but these figures are for the proportion of infinitives (as opposed to RIs) in modal contexts (including infinitives used correctly in combination with finite modals). The figures reported here have therefore been recalculated after excluding infinitives in modal constructions from the relevant counts.

*	4. I see John cross the street.

5. I saw John cross the street.

6. I see John crossing the street.

Utterance (4) is ungrammatical in English, because the English bare form denotes 'not only the processual part of the event, but includes the completion of that event' (Hoekstra & Hyams, 1998, p. 105). A correct description of an ongoing event in English would therefore require the use of the past tense as in (5), or the progressive as in (6). Utterance 7 makes it clear that this constraint does not operate in Dutch: an ongoing event may be described using a present tense construction and an infinitive. Apparently, the Dutch infinitive does not signal completion of the event.

7. Ik zie/zag Jan de straat oversteken.
I see/saw John the street cross-INF.
I see/saw John cross the street.

The differential status of the English bare form and the Dutch/German infinitive assumed by Hoekstra and Hyams can also explain a further difference between RI errors in these languages. In Dutch and German, the verbs that feature in RI errors are almost exclusively eventive (i.e. verbs denoting an action rather than a static situation), a feature which Hoekstra and Hyams dub the Eventivity Constraint. Hoekstra and Hyams argue that the Eventivity Constraint is less pronounced in English. Thus, whereas Wijnen

(1998) found that as many as 95% of the 1883 RIs in his Dutch data were eventive, Deen (1997) found that only 75% of the 264 RI errors in his English data featured eventive verbs (see Table 1). Hoekstra and Hyams explain this difference through the modal reading of the Dutch and German infinitive: it is the modal reading of the infinitive that forces the selection of an eventive verb. Since English RIs are not exclusively modal they can occur with stative as well as eventive verbs.

Hoekstra and Hyams' (1998) account, later developed into the Semantic Opposition Hypothesis (Hyams, 2001), has the advantage of providing an integrated explanation both of the difference between OI and non-OI languages and of differences in the referential properties of RIs across different OI languages. However, it suffers from one major empirical weakness. Since it relies on the assumption that the infinitival morpheme carries an *irrealis* feature, it predicts that RIs in languages with an infinitival morpheme should be restricted to modal contexts (i.e. that the proportion of RIs with a modal reading should be close to 1.00). This prediction is not borne out by the data on Dutch and German. Thus, although Wijnen (1998) reports an average proportion of 0.86 modal RIs in 4 Dutch children's speech (Range = 0.74 to 0.95), Blom (2003) reports an average proportion of only 0.74 (Range = 0.64 to 0.80) in 6 Dutch children; Ingram and Thompson (1996) report an average proportion of 0.52 (Range = 0.21 to 0.84) using a strict criterion and 0.77 (Range = 0.48 to 1.00) using a lenient criterion in 4 German children (with the actual proportion presumably falling somewhere in between); and Lasser (1997) reports proportions of only 0.69 and 0.73 for two further German children.

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

When taken together, these results suggest that the rate at which Dutch and German children produce RIs with modal reference is closer to 0.70 than it is to 1.00. However, the situation is complicated by the fact that, in all of the above studies, researchers were forced to exclude a relatively high proportion of uninterpretable RIs, which makes it difficult to estimate the proportion of RIs with modal reference accurately. In order to overcome this difficulty, Blom, Krikhaar, and Wijnen (2001) adopted an experimental approach to the problem in which they elicited descriptions of modal and non-modal events from Dutch and English children. In this experiment, the majority (68%) of the Dutch children's RIs were used to describe modal events (a figure that is significantly greater than 50%). For the English children, the relevant figure was 44% (a figure that is not significantly lower than 50%, but is significantly lower than the figure for Dutch). These results provide converging evidence that there is a modal preference in Dutch, and a significant difference between the proportion of modal RIs in Dutch and English children's speech (though see Blom (2007 for a counter-argument). However, they also confirm that the use of RIs in Dutch is not restricted to modal contexts. The implication is that both the Modal Reference Effect and the difference between Dutch and English are graded quantitative rather than qualitative effects. The graded nature of these effects is difficult to explain in terms of the presence or absence of infinitival morphology, and is hence a problem for Hoekstra and Hyam's (1998) account.

1.3. An alternative account of Optional Infinitive errors.

The vast majority of work on OI errors has been conducted within the framework of generative linguistics. However, Freudenthal, Pine, and Gobet (2006) and Freudenthal et al. (2007) have recently shown that the cross-linguistic pattern of OI errors can be understood in terms of the interaction between a relatively simple learning mechanism (MOSAIC) and cross-linguistic differences in the distributional statistics of the input that children receive. MOSAIC is a computational model of language learning, with no built-in knowledge of syntactic categories or rules, which takes as input corpora of orthographically transcribed child-directed speech and learns to produce as output 'child-like' utterances that become progressively longer as learning proceeds. As a result of these characteristics, MOSAIC can be used to generate corpora of utterances at different stages of development, and hence to model the behaviour of children learning different languages across a range of Mean Length of Utterance (MLU) values.

MOSAIC simulates OI errors because it has a strong utterance-final bias in learning. This bias results in the production of partial utterances that were present as utterance-final phrases in the input to which the model was exposed. The utterances in the input that give rise to OI errors are *compound finites*: utterances that contain a (finite) modal or auxiliary and a non-finite main verb. Thus, MOSAIC learns to produce utterances resembling English OIs such as 'Go home' and 'He go home' as truncated versions of utterances such as '(He can) go home' and '(Can) he go home?'. Similarly, MOSAIC learns to produce utterances resembling Dutch OIs such as 'IJs eten' and 'Hij ijs eten' as truncated versions of utterances such as '(Hij wil) ijs eten' (He wants to eat ice cream) and 'Wil hij ijs eten?' (Does he want to eat ice cream?).

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 high rates, and produce fewer OIs as the length of their utterances increases. MOSAIC simulates this phenomenon because of the way that compound finites pattern in OI languages. 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.

In an initial study, Freudenthal et al. (2006) showed that MOSAIC was able to simulate the developmental patterning of the OI phenomenon in two languages: English and Dutch. More specifically, they showed that one identical version of MOSAIC was able to provide 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. The model was also able to simulate the tendency for Dutch children to correctly order finite and non-finite verb forms with respect to their complements. This phenomenon is simulated very readily within MOSAIC because the model learns finite + complement strings such as 'Eet ijs' ('Eat-FIN ice cream') from simple finite utterances such as 'Hij eet ijs' ('He eat-FIN ice cream') and complement + infinitive strings such as 'IJs eten' ('Ice cream eat-INF) from compound finite utterances such as 'Hij wil ijs eten' ('He want-FIN ice cream eat-INF).

In a more recent study, Freudenthal et al. (2007) showed that a modified version of MOSAIC was able to simulate the developmental patterning of finiteness marking in four languages including English and Dutch, a third OI language (German) and a non-OI language (Spanish). Again, it was possible to achieve a close quantitative fit to the developmental data in terms of the interaction between one identical learning mechanisms and cross-linguistic variation in the distributional properties of the input language.

It is perhaps worth noting at this point that, although a useful cross-linguistic model of the OI stage, MOSAIC does not provide a complete model of the language acquisition process. On the contrary, MOSAIC is a relatively simple distributional analyser with no access to semantic information, which is currently not powerful enough to acquire many aspects of adult syntax. Nevertheless, because of its ability to produce child-like utterances across a range of different languages, MOSAIC does provide a powerful means of testing hypotheses about the relation between cross-linguistic variation in children's early language and cross-linguistic differences in the language to which they are exposed. It can thus be used to investigate the extent to which such variation can be explained in terms of the interaction between processing mechanisms that are common to all children and differences in the distributional properties of the input language.

When viewed in this context, Freudenthal et al.'s results are interesting for a number of reasons. First, they show that building a simple utterance-final constraint into the learning mechanism has a profound effect on the proportion of finite and non-finite forms that are learned during the early stages. Thus, MOSAIC learns OI errors (including bare participles) from utterances that include a finite modal or auxiliary and a non-finite main verb. Although such utterances constitute around 70% of the utterances including verbs in English child-directed speech, they constitute only around 30% of the utterances including verbs in Dutch child-directed speech. However, Freudenthal et al. (2006) show that, because the verb forms that occur in utterance-final position in Dutch are much more likely to be non-finite than finite, MOSAIC's utterance-final bias results in high rates of OI errors in both English and Dutch during the early stages.

Second, they show that the interaction between MOSAIC's utterance-final processing constraint and the distributional characteristics of the input can result in striking cross-linguistic differences in the developmental patterning of OI errors. Thus, although Dutch and Spanish both have compound finite constructions, which occur at similar rates in the input, the same version of MOSAIC is able to simulate both the high rate of OI errors in Dutch and the low rate of OI errors in Spanish. Freudenthal et al. (2007) show that this is because of the way that compound finites pattern in the two languages. In Dutch (which is an SOV/V2 language), non-finite verb forms occur in sentence-final position, after their complements (see sentence 8).

8. Ik ga in het park wandelen.

   I go-FIN in the park walk-INF.

9. Voy andar en el parque.

   (I) go walk-INF in the park.

10. Lo Quiero.

    (I) it want-FIN.

11. Ik wil het.

    I want-FIN it.

In Spanish (which is an SVO language), non-finite verb forms occur before their complements (see sentence 9). Coupled with the fact that Spanish allows finite verb forms in utterance-final position in some constructions in which they are not allowed in Dutch (see sentences 10 (Spanish) and 11 (Dutch)), this results in the proportion of utterance-final verb forms that are non-finite being high in Dutch (.87) and low in Spanish (.26). Since MOSAIC learns from the right edge of the utterance, the model initially produces a high proportion of OI errors when exposed to Dutch input, and a low proportion of OI errors when exposed to Spanish input.

Finally, Freudenthal et al. show that, in addition to simulating the qualitative difference between an OI language like Dutch and a non-OI language like Spanish, MOSAIC can also simulate fine-grained quantitative differences between OI languages. Thus, Freudenthal et al. (2007) show that despite the fact that Dutch and German are structurally very similar languages, Dutch children produce OIs at significantly higher rates than German children during the early stages. They also show that MOSAIC is able to simulate this difference as a result of a difference in the proportion of utterance-final verbs that are non-finite (0.87 in Dutch versus 0.66 in German). This difference can be traced back to the higher proportion of compound finites in Dutch child-directed speech. Further evidence that this factor is critical in determining the level of OI errors in early Dutch and German is provided by an analysis of the relation between 7 Dutch and 6 German children's early language and the distributional properties of their input. Freudenthal et al. report a correlation of 0.70 between the proportion of OIs in these 13 children's early speech and the proportion of non-finite utterance-final verbs in their input.

When taken together, these results suggest that it is possible to simulate the developmental patterning of OI errors across languages surprisingly well in terms of the interaction between an utterance-final bias in learning and differences in the distributional properties of compound finites in the input. However, the central role of compound finites in the simulation of OI errors in MOSAIC also raises the possibility that MOSAIC may be able to provide a unified explanation both of cross-linguistic variation in the rate at which OI errors occur, and of those subtleties in the patterning of RI errors identified by Hoekstra and Hyams (1998). Thus, as Ingram and Thompson (1996) have argued, one possible explanation of the Modal Reference Effect is that Dutch and German children come to associate infinitives with modal contexts because infinitives tend to occur in modal + infinitive compounds in these languages, and one possible explanation of the Eventivity Constraint is that only eventive (and not stative) verbs tend to occur in such modal + infinitive constructions. MOSAIC's mechanism for simulating OI errors makes it particularly well suited to exploring these input-driven explanations and the extent to which they can account for the differences between Dutch/German and English identified by Hoekstra and Hyams (1998). Note that the view that OI errors are learned from modal + infinitive constructions in the input is similar in some respects to a class of generativist models (e.g. Boser et al. 1992; Ferdinand, 1996), that treats OI errors as finite clauses that contain a null modal. However, the null modal hypothesis provides no explanation of why OI errors occur so much more frequently in early Dutch and German than modal constructions in the input, or why OI errors are so rare in languages like Spanish and Italian, which also have modal + infinitive constructions. The learning mechanism implemented in MOSAIC provides a simple and elegant explanation of both of these phenomena, which provides a good fit to quantitative data on the rate at which children produce OI errors at different MLU levels in English, Dutch, German and Spanish. This mechanism is not necessarily incompatible with generativist models (in the sense that it could be adapted to operate at a more abstract level). However, it does suggest the need for such models to take processing limitations more seriously and to incorporate some kind of probabilistic element into the learning mechanism.

In view of these considerations, the aim of the present paper is to assess the extent to which MOSAIC is able to simulate the Modal Reference Effect and the Eventivity Constraint and the way in which the modal reference of RIs varies across languages. Showing that MOSAIC is able to simulate the Modal Reference Effect and the Eventivity Constraint in Dutch and German would provide further evidence in favour of the idea that OI errors are learned from compound finites in the input. Showing that MOSAIC is able to simulate cross-linguistic variation in the modal reference and eventivity of RI errors would constitute a strong test both of MOSAIC's mechanism for generating OI errors, and of the idea that the cross-linguistic patterning of OI errors reflects the interaction between an utterance-final bias in learning and the distributional characteristics of the input language.

2. MOSAIC

2.1. The MOSAIC network

MOSAIC is an unsupervised learning mechanism consisting 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, which has no contents. Nodes immediately underneath the root node are called primitive nodes, and are used to encode single words[2]. Nodes at deeper levels in the network are used to store sequences of words that have been encoded at the primitive level. A MOSAIC network is slowly built up from exposure to the input it receives. Early in learning it will contain just a few nodes. As it sees more input, it will encode more words at the primitive level, and more and longer sequences of words at deeper levels in the network.

MOSAIC learns from orthographically transcribed child-directed speech with whole words being the unit of analysis. Thus, MOSAIC assumes that the speech stream has been segmented into words. Learning in MOSAIC is anchored at the right edge of the utterance: a word or phrase will only be encoded in the network if everything following that word or phrase in the utterance has already been encoded in the network. MOSAIC thus has a strong utterance-final bias in learning. The processing of an utterance in MOSAIC can be likened to a moving window or buffer, the size of which is determined by how much of the utterance has already been encoded by the model. MOSAIC processes an utterance in a left-to-right fashion, depositing the words it encounters into the buffer. Whenever the model encounters a word it has not yet encoded, the buffer is emptied, and the new word or sequence is deposited in it. The new word or sequence will only remain in the buffer (thus making it eligible for encoding) if everything that follows it in the utterance has already been encoded in the network. MOSAIC thus processes the utterance in a left-to-right fashion, but builds up its representation of the utterance by

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

starting at the end, and slowly working its way to the beginning of the utterance. In terms of a child attending to the speech stream this can be likened to the occurrence of an unknown word effectively clearing the contents of the speech stream encountered so far. This leaves the new word and the rest of the utterance for analysis. MOSAIC thus implements a view of language learning that has the child strongly biased towards the most recent (final) elements in the speech stream. Such an utterance-final bias (or recency effect) is psychologically plausible and is consistent with the literature on auditory processing in both adults (Penney, 1989; Cowan, Saults & Brown, 2004; Conway & Christiansen, 2005) and children (Naigles & Hoff-Ginsberg, 1998; Shady & Gerken, 1999; Wijnen, Kempen, & Gillis, 2001)[3].

As an example of how a MOSAIC network is built up, consider an empty network that is shown the utterance ‘he goes away’. The model will first deposit the word ‘he’

[3] It might be argued that MOSAIC’s utterance-final processing bias should be complemented by an utterance-initial processing bias (in order to simulate both recency and primacy effects). In fact, this issue is not as straightforward as it might at first appear, since recency effects are generally taken to reflect capacity limitations in short-term memory (which *are* likely to be a factor in language learning), whereas primacy effects are often taken to reflect active elaboration processes such as rehearsal (which *are not* likely to be a factor, at least in younger children). However, this issue can be safely ignored since Freudenthal, Pine and Gobet (2005a) have shown that building some sensitivity to utterance-initial position into MOSAIC does not have any significant effect on the model’s ability to simulate the developmental patterning of OI errors in English, Dutch, German and Spanish.

into the buffer, but will replace this with the word ‘goes’ as this has not yet been encoded. The word ‘goes’ will itself be replaced by the word ‘away’, which will be in the buffer when the model reaches the end of the utterance (as signalled by an *end marker*). At this point, the model will encode the word ‘away’ as a primitive node. After a second presentation of the same utterance, the buffer will contain the phrase ‘goes away’ when the end of the utterance is reached. The model will attempt to encode this phrase. However, as the word ‘goes’ has not yet been encoded in the model, it will create this node first. A third presentation will result in the phrase ‘goes away’ being encoded in the network through the creation of an ‘away’ node under the ‘goes’ node. A fourth presentation will result in a primitive node for ‘he’. Finally, a fifth presentation will result in the ‘goes away’ branch being copied underneath the ‘he’ node. Figure 1 shows the network after these five presentations.

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

2.2. Node Creation Probability

In the example given, we have assumed that a node is created whenever the opportunity arises. In practice, however, node creation is probabilistic, and learning is much slower. The probability of creating a node 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.

This formula results in a basic sigmoid curve (when plotted as a function of number of utterances seen). Early in learning, when the model has seen few utterances, the probability of creating a node is low (i.e. learning is slow). As the model sees more and more utterances the node creation probability increases and learning speeds up. The learning rate thus increases as the amount of knowledge encoded in the model grows. This is consistent with empirical data showing that children learn new words more readily as their vocabulary size increases (Bates & Carnavale, 1993). The formula also includes the size of the corpus used. The reason for this is that the corpora used for simulations with MOSAIC are quite varied in size. Including the size of the corpus in the formula for node creation probability ensures that after *n* presentations of a corpus the node creation probability is identical for corpora of different sizes. Finally, the base number in the formula is raised to the power of the square root of *d* (the distance to the end of the utterance). This ensures that the probability of encoding material that occurs near the end of the utterance is higher than the probability of encoding material near the beginning of the utterance. Note that utterance position in MOSAIC is defined in terms of distance in number of words (or chunks) from the right edge of the utterance, regardless of the structural properties of the utterance. For example, MOSAIC does not distinguish between utterances consisting of a single clause (e.g. 'He eats cookies') and utterances ending in an embedded clause (e.g. 'These are the kind of cookies he eats'). This feature of the model means that MOSAIC could, and occasionally does, learn simple finite sequences such as 'he eats' from utterances ending in embedded clauses. It also means that, in German, where the order of the finite modal and non-finite main verb is reversed in embedded clauses, MOSAIC could, and occasionally does, learn compound non-finite + finite-modal sequences such as 'essen kann' (eat can) from utterances ending in an embedded clause. In practice, however, embedded clauses are not sufficiently frequent in the input to have a significant effect on the proportion of utterance-final verb forms that are finite as opposed to non-finite

When building a MOSAIC network, an input corpus is fed through the model several times. With successive presentations of the input, MOSAIC will encode more and longer (utterance-final) phrases and the learning rate will increase. Output can be generated from the model after every presentation of the input. With every presentation of the input the amount and average length of output will increase. This allows for the simulation of developmental variation. Thus, output files can be generated after each presentation of the input and selected for comparison with child data at different points in development on the basis of their MLU.

2.3. Generating output from MOSAIC

MOSAIC employs two mechanisms for generating output. The first mechanism involves traversing all the branches to their terminal nodes, and outputting the (utterance-final) phrases they encode. The output produced through this mechanism is *rote* output; all the phrases produced in this manner were present as (utterance-final) phrases in the input.

This mechanism is complemented by a second mechanism that allows for the substitution of distributionally similar words in the phrases MOSAIC encodes. For all words that are encoded in the network, MOSAIC stores the context (preceding and following words) in which they have occurred. Words that have occurred in similar contexts (are preceded and followed by the same words) are connected by a *generative link*, which allows them to be substituted in production. This mechanism allows MOSAIC to produce utterances that were not present in the input it received. The rationale for allowing substitution on the basis of co-occurrence statistics is that it has been shown that words that occur in similar contexts tend to be of the same word class (Redington, Chater, & Finch, 1998; Mintz, 2003). For the present simulations, two words are connected by a generative link when there is an overlap of 20% or more in both the words preceding and following the target words. This value is the same as that used by Freudenthal et al. (2007), where it was chosen because it prevented output files at later MLU stages from becoming unmanageably large. In practice, neither increasing nor decreasing this value has very much effect on the rate at which MOSAIC produces OI errors in the different languages simulated to date. However, it does affect the quality of the generative links and hence the rate at which the model produces incorrect substitutions. Freudenthal et al. show that when the overlap parameter is set to 20%, utterances that include incorrect substitutions occur in MOSAIC's output at rates of between 5% and 7%.

2.4. Chunking

A final feature of MOSAIC is that it employs a chunking mechanism, which results in frequent phrases being treated as one unit by the substitution 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 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 (2005b), 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 (also see Gleitman & Wanner, 1982). The verbs *do* and *get,* for example, may share considerable context due to their occurrence as main verbs, 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 *get* for *do* in this context will result in anomalous utterances such as *Get 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 *get* 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.

3. Simulations

Simulations are run in MOSAIC by exposing the model to corpora of orthographically transcribed child-directed speech. Given that the majority of publicly available child-directed speech corpora are orthographically rather than phonetically transcribed, MOSAIC's ability to accept such corpora as input obviously has certain advantages. However, it is clearly also an important simplification and means that MOSAIC is insensitive to information that is not included in this format, such as information about intonation and relative stress. As a result, MOSAIC is unable to simulate aspects of the data that depend on such factors. For example, MOSAIC is insensitive to the difference between stressed and unstressed morphemes and will learn sequences including unstressed function words (e.g. 'want a cookie') as readily as sequences of stressed content words (e.g. 'Jack likes milk').

When building a MOSAIC network, input corpora are fed through the model several times, and an output file is created after each cycle through the input corpus. This file consists of all the utterances (both rote-learned and generated) that the model is able to produce at that particular point in development. Because the network grows as a function of the amount of input to which it has been exposed, the average length of the model's output increases with every cycle through the input corpus. Output files can therefore be selected for analysis on the basis of their MLU and compared with data from children at the same stage of development (i.e. with similar MLUs). Since MOSAIC's output files consist of sets of utterance types rather than utterance tokens, corpora of child utterances are also reduced to sets of utterance types. In practice, however, the difference between child measures based on utterance types and utterance tokens is much smaller than one might expect. This is because, although there is substantial variation in the frequency with which children produce particular lexical items, there is much less variation in the frequency with which children produce particular multi-word utterances, with the vast majority of multi-word utterances in children's output occurring only once.

3.1. Input Corpora

The corpora used as input for the present simulations were the Manchester corpus (Theakston, Lieven, Pine, & Rowland, 2001) for English, and the Groningen corpus (Bol, 1995) for Dutch. The Manchester corpus consists of recordings of caregiver/child interactions for 12 different children. The Groningen corpus contains recordings of caregiver/child interactions for 7 different children. Both of these corpora are available in the CHILDES database (MacWhinney, 2000). In addition, two dense datasets were used. These were the Thomas corpus (Dabrowska & Lieven, 2005) for English, and the Leo corpus (Behrens, 2006) for German. These datasets consist of particularly rich corpora of adult and child speech since recordings were made up to 5 times per week during the period in which the child was between 2 and 3 years of age.

3.2. Preparation of the Input Corpora

Basic preparation of the input was performed in the same way as described in Freudenthal et al. (2006, 2007). Briefly, for each child, this involved extracting all maternal speech from the CLAN files for that child and aggregating this into one file (resulting in several aggregate files containing the maternal speech directed at the individual children). A limited amount of (automated) filtering was subsequently carried out on the maternal files. Filler material, such as *hmm*, *ah*, *oh*, was deleted, as were retracings, duplicated material, vocatives and tags occurring at the end of utterances, and utterances where one or more words were unintelligible to the transcriber. However, no attempt was made to separate utterances into clauses. The resulting input corpora differed considerably in size. The average size of the 12 input sets from the Manchester corpus was approximately 25,000 utterances. The 7 Dutch input sets consisted of approximately 10,000 utterances on average. The German input set (Leo) contained 80,000 utterances, while the speech directed at Thomas consisted of 240,000 utterances. All of these corpora contained a wide range of utterances including simple sentences, sentences with embedded clauses, single-word utterances, and sentence fragments (where these occurred as complete utterances in the original transcripts).

In addition to the initial filtering of the input, some preparation was needed that was specific to the simulation of the Modal Reference Effect. In order to distinguish between infinitives produced in a modal and non-modal context, it was necessary to mark infinitives learned from a modal or non-modal context separately. This was done in the following manner: First, the maternal speech files were (automatically) searched for utterances containing words that denote a modal or not-realised context. These words consisted of the standard English modals (can, will, may, shall, must) as well as the

English semi-auxiliaries (want to/wanna, going to/gonna, need to/needta, ought to/oughta) and their Dutch and German equivalents. Next, the utterances identified as constituting a modal context were searched for words that matched the infinitive. Where these were found, they were marked for being part of a modal context by adding the tag '+mod' to the infinitive form. Note that, because of the way MOSAIC represents words (as character strings), this procedure means that separate entries are created in the model for the same word produced in a modal and a non-modal context. To MOSAIC *walk* and *walk+mod* are different words that are represented in different primitive nodes. Similarly, *walk home* and *walk+mod home* are different phrases. Distinguishing between modal and non-modal RIs in this way is not intended as a realistic way of representing children's knowledge. It is simply an implementational device that allows us to use MOSAIC to investigate the extent to which the Modal Reference Effect and the Eventivity constraint can be explained in terms of the semantic distributional properties of the input. However, since the use of this device does have the potential to affect the fit between MOSAIC's output and the child data, simulations were run to ensure that the model was still able to simulate the developmental patterning of the OI phenomenon in English, Dutch and German as reported in Freudenthal et al. (2007) (see below)[4].

[4] Note that this procedure inevitably mis-categorises some utterances in the input. For example, it treats adult RIs (which are allowed in all three languages in certain contexts, some of which are modal) as non-modal contexts. It also treats all go + infinitive constructions as modal contexts, although it is possible to use go + infinitive constructions in Dutch with a non-modal present tense reading. However, neither of these utterance types are frequent enough to have a serious impact on the results of the present

3.3 Running the simulations

Three preliminary simulations (one for English, one for Dutch and one for German) were carried out to assess whether the practice of marking infinitive forms for modal contexts (i.e. creating separate model entries for an infinitive uttered in a modal and non-modal context) affected the fit of the simulations to the basic OI phenomenon. These three simulations constitute a replication of the simulations performed by Freudenthal et al. (2007) using the input files containing modal coding. (Note that these files are identical to those used by Freudenthal et al., 2007, except for the modal coding contained in the new files.) These simulations involved running the model up to an MLU of approximately 4. A number of output files of increasing MLU were then selected and analyzed with respect to the proportion of utterances that were non-finite, simple finite and compound finite. The output files selected were compared to (MLU matched) child output at different stages in development, as well as the results of the earlier simulations.

Simulations of the referential properties of RIs were carried out by running separate models on the child-directed speech for each of the 12 English children in the Manchester corpus and each of the 7 Dutch children in the Groningen corpus. In addition, one simulation was run on a German dense dataset: the Leo corpus, and one simulation was run on an English dense dataset: the Thomas corpus. Thus, a total of 21 models were run, each on the child-directed speech of one of 21 different children. The simulations were run in an identical manner for all of the corpora. For each simulation the corpus of

study. Thus, an analysis of the Dutch input data revealed that only 5.1% of utterances containing a verb were modal RIs and only 1.5% were go + infinitive constructions with a present tense reading.

speech directed at the child was fed through the model several times. Output (of increasing average length) was generated from the model after each exposure to the input and the output file with an MLU closest to 2.5 was identified. This MLU value was chosen to ensure comparability with Blom et al.'s (2001) data, where the average MLU of her Dutch subjects was 2.62. Files identified in this way were then analysed with respect to the proportion of modal and non-modal RIs and the proportion of eventive and stative RIs.

3.4. Coding and data analysis

3.4.1. Preliminary Simulations

In order to ensure comparability with Freudenthal et al. (2007), the preliminary simulations were coded as follows.

In the case of Dutch and German, each file was searched for utterances that contained at least one verb form other than the copula. Each of the utterances identified in this way was then classified as a simple-finite, a compound-finite or a non-finite utterance.

Simple-finite utterances were defined as utterances that only included an unambiguously finite verb form (e.g. utterances containing first person singular, second person singular or third person singular present tense verb forms).

Compound-finite utterances were defined as utterances containing both an unambiguously finite verb form and a verb form that was not unambiguously finite (e.g. utterances containing a singular present tense verb form and an infinitive).

Non-finite utterances were defined as utterances that did not include an unambiguously finite verb form (e.g. utterances containing an infinitive or a plural present tense morpheme).

Note that an important feature of this coding scheme is that it treats all ambiguous verb forms as if they were non-finite verb forms. This feature of the coding scheme is necessary because there are some finite verb forms in Dutch and German that are indistinguishable from infinitival verb forms. Thus, although there is strong evidence that Dutch and German children do produce infinitival verb forms in contexts in which a finite verb form is required, it is actually impossible to be sure whether the verb form included in any particular utterance is an infinitival verb form as opposed to a finite plural present tense verb form.

An obvious disadvantage of coding the data in this way is that the resulting measures are always likely to underestimate to some degree the model's ability to produce correct finite forms (particularly later in development). This should obviously be borne in mind when interpreting the absolute values reported in the simulations. However, it is important to realise that they do not affect the validity of any analyses of the closeness of fit between model and child, because model and child data are analysed using exactly the same (automated) procedure. Indeed, we would argue that they illustrate one of the strengths of our modelling approach, which is that it allows us to measure the closeness of fit between model and child in a way that is independent of any assumptions about the knowledge underlying the child's use of particular utterances at particular points in development.

Nevertheless, it could be argued that, because of its impoverished verb morphology, the level of ambiguity in English is so great that treating all zero-marked forms as non-finite in both the model and the child is likely to result in a trivially good fit between model and child. In order to deal with this problem, analysis of the English simulations is restricted to utterances that contain a third person singular (pronominal) subject (e.g. *He go(es)*), as the provision of a zero-marked form in such contexts is clearly incorrect. Analysis is restricted to pronominal third person singular subjects as this allows an automated lexical search and hence an automated analysis. However, even when restricting the analysis to third person singular contexts, a certain level of ambiguity remains due to English regular past tense forms being indistinguishable from non-finite perfect 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 are therefore classified as ambiguous and counted separately.

3.4.2. Simulations of the Referential Properties of RIs

The simulations of the Referential Properties of RIs were coded in two ways. In a first set of analyses, all utterances in which the only verb forms in the utterance matched an infinitive (or bare form) were selected. The proportion of these utterances that had been learned from a modal context (i.e. contained a +mod tag) was then calculated. This number constituted the proportion of RIs produced in a modal context. Note that, since not all bare stems in English are infinitives, and not all verb forms marked with –en are infinitives in Dutch and German, this analyses is relatively crude. However, it does have the advantage of allowing us to conduct a strong test of MOSAIC's ability to simulate the Modal Reference Effect (i.e. the preference for modal over non-modal RIs in Dutch and German) across the full range of the model's output. This is because, while some of the non-modal RIs produced by the model may be finite forms that are indistinguishable from infinitives, all of the modal RIs produced by the model are necessarily RIs since they have been learned from infinitives in the input. The measures generated by this analysis can therefore only err on the side of underestimating the proportion of modal RIs in the model's output.

In a second set of analyses, utterances were only coded if, in addition to containing verb forms that matched the infinitive (or bare form), they also contained a third person singular subject. This analysis has the disadvantage that it forces one to exclude a large proportion of the model's output. However, because it controls for differences in the ambiguity of verb forms matching the infinitive across the different languages, it provides a much stronger test of MOSAIC's ability to simulate cross-linguistic differences between Dutch/German and English. The data selected in this way were therefore also used to assess MOSAIC's ability to simulate quantitative differences in the eventivity of RIs in Dutch/German and English.

4. Results

4.1. The basic Optional Infinitive phenomenon

In this section, we report the results of simulations designed to investigate whether distinguishing between infinitives occurring in modal and non-modal contexts in the input adversely affects MOSAIC's ability to simulate the basic OI phenomenon. This question is addressed by computing Root Mean Square Error statistics (RMSEs), which measure the closeness of fit between model and child data (the smaller the value, the better the fit), and comparing those obtained for the new simulations with those obtained for the simulations reported in Freudenthal et al. (2007).

Fig. 2 shows the results for the new English simulation (c), as well as the simulations (b) and child data (a) reported by Freudenthal et al. (2007) for Anne, an English child. As can be seen, the new simulations display the same pattern as the child and the earlier simulations. RMSEs for the old simulation are .06, .07, and .08 for the three MLU points. For the new simulations RMSEs are .11, .11 and .10.

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

Fig. 3 shows the new and old simulations and child data for Matthijs, a Dutch child. Although neither the old nor the new simulation capture all of the fine detail of the child data (particularly the pronounced drop in the proportion of RIs between the second and third MLU points), the new simulations again show the same pattern as the old simulations. RMSEs for the old simulations are .04, .01, .25 and .12. For the new simulations these are .01, .02, .28 and .12.

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

Fig. 4 shows the simulations and data for Leo, a German child. RMSEs for the old simulations are .06, .05, .02, and .04. For the new simulations they are .04, .05, .02 and .07.

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

It is evident from these simulations that, apart from a slight reduction in the fit to the English data, distinguishing between infinitives occurring in modal and non-modal contexts in the input has little effect on MOSAIC's ability to simulate the basic OI phenomenon, with all of the new simulations displaying the same pattern of results as the old simulations. These results indicate that the mechanism MOSAIC uses to produce OI errors simulates the developmental data in a relatively robust way that is not greatly affected by minor differences in the way the input is represented. They thus validate the practice of distinguishing between infinitives occurring in modal and non-modal contexts in the input as a way of assessing whether this mechanism is able to capture cross-linguistic differences in the referential properties of Dutch/German and English RIs.

4.2 Simulating the Modal Reference Effect

In this section we report the results of simulations designed to investigate whether MOSAIC captures the Modal Reference Effect (i.e. the preference for modal over non-modal RIs in Dutch and German) and the differences in the proportion of modal RIs in Dutch/German and English. Two sets of analyses are reported in each case. The first focuses on all utterances that could be interpreted as RIs. The second focuses only on utterances that could be interpreted as third person singular RIs.

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

The results of the first set of analyses are reported in Table 2, which shows the (average) proportions of modal RIs for the English, Dutch, and German simulations. The (average) number of utterances contributing to this analysis was 1441 for English, 706 for Dutch, 6512 for Leo, and 3928 for Thomas. As can be seen from Table 2, even though this analysis is likely to underestimate the proportion of modal RIs in MOSAICs output, there is still a small preference for modal RIs in both Dutch and German. This preference was analysed by comparing the observed figure to chance using a single sample t-test for the Dutch simulations and a binomial test for the German simulation. In both cases, the preference was found to be statistically significant (t(6) = 1.97, p < .05, one-tailed for Dutch and p < 0.0001 for German).

It can also be seen from Table 2 that the proportion of Modal RIs in English is substantially lower than it is for both Dutch and German. This difference was analysed by comparing the data from the English (Manchester) simulations with the data from the Dutch (Groningen) simulations using an independent t-test and the data from the English (Thomas) simulation with the data from the German (Leo) simulation using Chi-square. The difference between the English (Manchester) and Dutch (Groningen) simulations is statistically significant ($t(17) = 7.18$, $p < .0001$). The difference between the English

(Thomas) and the German (Leo) simulations is also significant: ($\chi^2(1) = 392.63$, $p <$ .0001).

These results suggest that MOSAIC does simulate the preference for modal RIs in Dutch and German and the difference in the modal reference of Dutch/German and English. However, they ignore the complication that, particularly in English, it is often difficult to unambiguously identify the infinitive form. Thus, in the English present tense, the third person singular form is the only form that can be distinguished from the infinitive. A similar, but less pronounced problem occurs in Dutch and German where the present tense plural forms cannot be distinguished from the infinitive. As noted above, this complication actually works against MOSAIC's ability to simulate the modal preference in Dutch and German. However, since the level of ambiguity is greater in English than it is in Dutch and German, it has the potential to exaggerate any differences that exist between Dutch/German and English. In order to deal with this problem, a second analysis was carried out on the subset of utterances that contained only verb forms matching the infinitive plus a third person singular subject. The rationale behind this restriction is that it creates a level playing field on which to compare Dutch/German and English since the use of infinitives (or bare stems) are clearly incorrect in such contexts in all three languages. The results of the modal analysis on this restricted data set are shown in Table 3.

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

It should be noted that the restriction to third person singular subjects reduced the number of utterances contributing to the analysis considerably. Thus, the average number of utterances included in the restricted analysis was 52 for the English (Manchester) simulations and 31 for the Dutch (Groningen) simulations; the German (Leo) analysis was conducted on 95 utterances, and the English (Thomas) analysis was carried out on 70 utterances. Nevertheless, the rate of modal reference was significantly greater than 0.50 in both the Dutch and the German data ($t(6) = 3.31$, $p < 0.05$, two-tailed for Dutch, $p < 0.0001$ by a Binomial test for German). Moreover, the difference between the English (Manchester) and Dutch (Groningen) simulations remained statistically significant ($t(17) = 8.24$, $p < .0001$), as did the difference between the German (Leo) and English (Thomas) simulations: ($\chi^2(1) = 68.19$, $p < .001$). These results show that differences in the modal reference of MOSAIC's Dutch/German and English output cannot be explained in terms of the greater level of ambiguity in English as opposed to Dutch and German, and hence confirm that MOSAIC is able to simulate differences in the modal reference of Dutch/German and English RIs.

4.3. Simulating the Eventivity Constraint

Having established that MOSAIC successfully simulates the finding that RIs in Dutch and German carry a modal reading more often than in English, we now turn to the related finding that, compared to English, the verbs that feature in Dutch/German RIs have a higher likelihood of being drawn from the pool of eventive verbs (i.e. verbs denoting actions, rather than static situations). Wijnen (1998) found that 95% of Dutch children's

RIs are eventive. Hoekstra and Hyams (1998) cite a paper by Deen (1997) who found that 75% of English RIs contained an eventive verb.

The eventive/stative nature of the RIs produced by MOSAIC was established by hand coding the verbs in the third person singular RIs for their eventivity. As can be seen in Table 4, the (average) proportion of eventive RIs in the Dutch and German data was over 90% in both cases. Moreover, it was significantly higher in the Dutch (Groningen) simulations than in the English (Manchester) simulations ($t(17) = 3.39$, $p < .01$) and in the German (Leo) simulation than the English (Thomas) simulation ($\chi^2(1) = 5.68$, $p < .05$). MOSAIC therefore simulates both the Eventivity Constraint (the predominance of eventive RIs in Dutch and German) and the reduced size of this effect in English.

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

4.4. What causes the Modal Reference Effect?

The finding that MOSAIC simulates the differences between Dutch/German and English for both the Modal Reference Effect and the Eventivity Constraint suggests that these differences are related to differences in the surface characteristics of the input sets for these languages rather than the deep structural differences between the Dutch/German infinitives and the English bare form posited by Hoekstra and Hyams (1998). MOSAIC uses no built-in syntactic knowledge, and the same model was used for the simulation of all of the three languages. In fact, the only difference between the simulations for the different languages was the fact that input from the respective languages was used. Therefore, differences in the model’s output for the different languages must necessarily reflect differences in the distributional or surface characteristics of the input. In order to understand what these surface characteristics are, it is useful to examine more closely the contexts in which third person singular subjects can precede a verb form matching the infinitive in the three languages. In both English and Dutch/German, the majority of these contexts are likely to be questions (e.g. *Can he go?*; *Kan hij gaan?* (Can he go-INF?)). There is, however, an important difference between English and Dutch/German in the way in which questions are formed. Both English and Dutch/German use subject-modal/auxiliary inversion to create an interrogative form of a compound finite construction (e.g. *He can go* → *Can he go?*; *Hij kan gaan* → *Kan hij gaan?*). However, Dutch and German use subject-main-verb inversion to construct a question from a simple finite utterance. Thus, an interrogative version of the simple finite utterance *Hij gaat* (He goes) is constructed by placing the main verb before the subject (*Gaat hij?*/ Goes he?). In English, such questions are formed by constructing a compound finite using dummy modal *do* (*He goes* → *Does he go?*). Importantly, English dummy modal *do* patterns like a modal, but does not assign a modal meaning to the utterance. Thus, constructing a question from a simple finite (third person singular) utterance results in a third person singular subject followed by an infinitive *in a non-modal context* in English, but not in Dutch/German. As a result, third person singular subject plus infinitive constructions occur in modal contexts relatively less frequently in English than in Dutch/German. Further sources of third person singular subject plus infinitive constructions are adult RIs, which can occur as both full clauses and elliptical answer to questions in all three languages (Lasser, 2002), double verb constructions (e.g. *I see John walk*/ *Ik zie Jan lopen*), and, in Dutch, progressive constructions (e.g. Is Jan aan het lopen?/ Is John on the walk-INF?/ Is John walking? and Zit Jan te spelen?/ Sit-FIN John to play-INF? Is John sitting and playing?).

In summary, third person singular subject plus infinitive constructions occur in different contexts in English and Dutch/German. Due to the non-modal nature of English dummy modal *do*, a larger proportion of these contexts in English are non-modal. If a child learns RIs from the input, she is likely to produce them in the type of context in which they are most frequently encountered. One way of directly testing such an input-driven account of RIs is to search the input for occasions where a third person singular subject is followed by an infinitive form, and noting whether the context is modal or not. Table 5 provides the results of such an analysis. For English, the analysis was carried out on the child-directed speech contained in the Thomas corpus. For German, the Leo corpus was used. Since the individual input files for the Dutch children are relatively small (~10,000 utterances) compared to the dense datasets of Leo and Thomas, it was decided to aggregate the speech directed towards all 7 children in the Groningen corpus into one file of approximately 70,000 utterances. This large file was used for the input analysis.

--------------- Insert Table 5 about here -----------------

It can be seen from Table 5 that the proportion of modal contexts is very different for English and Dutch/German, and that all three proportions are very similar to the (average) proportion of modal RIs in the relevant model’s output. Also of interest is the question of what the non-modal contexts are in which these constructions occur. For

English, 52% of these turn out to be ‘do-questions’ (the other main categories being RIs in the input (e.g. *Mummy do it*), and double verb constructions (e.g. *Watch Daddy do it*). Thus, while third person singular subject plus infinitive constructions can occur both in modal and non-modal contexts, non-modal *do* alone accounts for 44% of all third person singular contexts in the speech directed at Thomas. In the Dutch and German input, non-modal contexts are limited in number, and largely confined to adult RIs (20% of all third person singular contexts in Dutch and 6% in German), double verb constructions (4% of all third person singular contexts in Dutch and 5% in German) and progressive constructions (9% of all third person singular contexts in Dutch). The majority of non-modal RIs therefore appear to be learned off dummy modal do-constructions in English and off RIs in Dutch and German, though there is also a role for double-verb constructions and a role for progressive constructions that is peculiar to Dutch.

Turning to the eventive-stative distinction, it now also becomes apparent why RIs have a higher likelihood of containing eventive verbs in Dutch/German. In English, stative verbs (preceded by a third person singular subject) frequently occur in questions like *Does he want it?*. In Dutch such questions do not carry the inflection on the dummy modal, but on the inverted main verb (*Wil hij dat?; Wants he that?*). As a result, stative verbs only occur as infinitives in Dutch in double verb constructions (e.g. *Dat zou hij willen/That would he want*) and in compound (modal) questions (e.g. *Wil hij slapen?*/Wants he sleep?). Both of these construction types are relatively infrequent compared to those that give rise to stative RIs in English. Moreover, compound questions, tend to include only a subset of stative verbs (e.g. *see* and *sleep*, but not *want* and *need)* because of their modal semantics. These observations are borne out by an analysis of the types of stative verbs that feature in the Dutch and English models' RIs. Table 6 presents the most frequent[5] stative verbs that occurred in the RIs of the models trained on the Manchester and Groningen corpora. The first thing to note about Table 6 is that the relative frequency of the English statives is higher than it is for Dutch. It is also clear that, while there is some overlap between the verbs that occur in the two languages (e.g. sit, have, sleep, see, wait), there are a few high frequency English statives that do not occur in the Dutch list: Fit, Like, Want and Need. These verbs frequently feature in English *do*-questions (e.g. *Does it fit?*, *Does he want it?*). Compound questions that would give rise to a third person singular subject followed by an infinitive form of these verbs are rare in Dutch.

The higher proportion of statives in English RIs can therefore also be explained in terms of the use of dummy modal *do* in the input. The use of Dummy modal *do* gives rise to RIs with verbs that are infrequently used in Dutch compounds. It should be noted that English children readily use these verbs in conjunction with third person singular subjects. An analysis of the speech of the children in the Manchester corpus showed that 11 of the 12 children produced *need* as an RI with a third person singular subject and all 12 of the children produced *want*, *like* and *fit* as RIs with third person singular subjects.

---------- Insert Table 6 about here -----------------

[5] A total of 20 different stative verbs occurred in the output of the English models. Only 10 stative verbs occurred in the Dutch models' output.

5. Discussion

This paper set out to establish whether MOSAIC, a computational model that has already been shown to successfully simulate cross-linguistic differences in the developmental patterning of the OI phenomenon, is capable of simulating certain subtle effects in the patterning of RI errors across languages, in particular the Modal Reference Effect and the Eventivity Constraint in Dutch and German, and the absence (or reduced size) of these effects in English. This question is of particular interest because the cross-linguistic patterning of these phenomena is difficult to explain in terms of current generativist accounts of the OI stage (e.g. Wexler, 1998; Hoekstra & Hyams, 1998). On the other hand, it appears consistent with the assumption (instantiated within MOSAIC) that RIs are learned from *Compound Finites* (Aux/Modal plus infinitive constructions) in the language to which children are exposed (e.g. Ingram & Thompson, 1996).

MOSAIC clearly simulates both the Modal Reference Effect in Dutch and German and the differential reading of RIs in Dutch and German compared to English. Thus, in the Dutch simulations, an average of 68% of the RIs with third person singular subjects in MOSAIC's output had been learned from modal contexts and, in the German simulation, 78% of the RIs with third person singular subjects had been learned from modal contexts. However, in the English simulations, an average of only 28% of the RIs with third person singular subjects had been learned from modal contexts for the Manchester corpus and only 13% of the RIs with third person singular subjects had been learned from modal contexts for the Thomas corpus. MOSAIC also successfully simulates the Eventivity Constraint in Dutch and German and differences in the levels of eventive RIs in Dutch and German compared to English. Thus, in the Dutch simulations, an average of 91% of the RIs in MOSAIC's output were eventive and, in the German simulation, 92% of the RIs in MOSAIC's output were eventive. However, in the English simulations, an average of only 76% of the RIs in MOSAIC's output were eventive for the Manchester corpus, and only 79% of the RIs in MOSAIC's output were eventive for the Thomas corpus. As would be predicted by an input-driven account, these differences in the output for the different languages could be traced back to quantitative differences in the distributional characteristics of English and Dutch/German child-directed speech. Third person singular subject plus infinitive constructions occurred in modal contexts far less often in English (16%) than in Dutch (68%) and German (88%). An analysis of the non-modal contexts in which these constructions occurred showed that the main source of non-modal occurrences of third person singular subject plus infinitive in English is the use of dummy modal *do*, which patterns like a modal but does not assign a modal meaning to the utterance.

The present results are important for a number of reasons. First, they show that, in addition to simulating cross-linguistic variation in patterns of finiteness marking, the mechanism used by MOSAIC to simulate OI errors can capture subtle features of children's use of RIs in Dutch, German and English that are difficult to explain in terms of current generativist models of the OI stage (Wexler, 1998; Hoekstra & Hyams, 1998). Thus, according to Wexler, RIs are not truncated modal utterances, but rather attempts at finite utterances in which the finite verb form fails to surface, as a result of the failure to check either Tense or Agreement, and an infinitive verb form is produced instead. Wexler's theory therefore provides no explanation for the predominantly modal reference of Dutch and German RIs, or for the fact that the probability with which RIs occur in modal contexts is different in Dutch/German and English. On the other hand, according to Hoekstra and Hyams (1998), the modal reference of Dutch and German RIs (and differences in the reference of Dutch/German and English RIs) reflect the fact that the infinitival morpheme (which is present in Dutch and German and absent in English) assigns a modal reading to the Dutch/German infinitive. Hoekstra and Hyams' account therefore predicts that all Dutch/German RIs will have a modal reading and is unable to explain the fact that the proportion of RIs in Dutch and German with modal reference is typically closer to 0.70 than it is to 1.00. According to MOSAIC, RIs are incomplete compound finites, whose semantics reflects the semantics of the utterances from which they have been learned. The simulations presented here show that this kind of account can explain not only the predominantly modal reference of Dutch and German RIs, but also the pattern of differences in the modal reference of Dutch/German and English RIs, and the fact that these differences appear to be quantitative rather than qualitative in nature. These findings provide further support both for the claim that OI errors are incomplete compound finites, and for the idea that cross-linguistic variation in the modal reference and eventivity of RIs can be explained in terms of quantitative differences in the surface properties of different languages (i.e. in the extent to which infinitives occur in modal contexts in the input).

Of course, this conclusion is not necessarily incompatible with generativist accounts of the OI stage. As was noted earlier, the way in which MOSAIC simulates the OI stage shares some similarities with accounts that view OI errors as finite clauses that contain a null modal (Boser et al., 1992; Ferdinand, 1996). Nor does it rule out the possibility that other generativist accounts, such as those of Wexler and Hoekstra and Hyams, could be extended to provide a better fit to the data on modal reference (see Blom (2007) and Hyams (2007) for developments along these lines). However, the fact that quantitative differences in the referential properties of RIs in different languages appear to be so directly related to quantitative differences in the surface properties of those languages does suggest the need to modify such accounts by making them more sensitive to quantitative variation in the distributional properties of the input language. It is difficult to see how this could be done without incorporating some kind of probabilistic element into the learning mechanism (e.g. Legate & Yang, 2007).

Second, the present results illustrate the potential power of cross-linguistic modelling as a way of investigating the relation between variation in children's early multi-word speech and variation in the distributional properties of the language to which children are exposed. Thus, the use of one identical model to simulate cross-linguistic differences in the modal reference of RIs not only allows us to conduct a strong test of Ingram and Thompson's (1996) input-driven accounts of the Modal Reference effect in Dutch and German; it also allows us to identify critical differences in the distributional properties of Dutch/German and English child-directed speech that have the potential to explain differences in the modal reference of Dutch/German and English children's RIs. The simulations reported in the present paper suggest that it is possible to explain differences in the modal reference of Dutch/German and English children's RIs in terms of a relatively subtle difference between Dutch/German and English: the fact that English has a dummy modal (i.e. a form that patterns like a modal auxiliary, but does not assign a modal meaning to the utterance). It is worth noting that this result would have been difficult to predict a priori. Indeed, it would probably not have been found even in the simulations reported here if, as is often the case in computational modelling research, artificially created input sets had been used. This is because, although utterances with dummy modal *do* account for approximately 50% of all third person singular plus infinitive constructions in real child-directed speech, dummy *do* is only one of a large number of different English modals. It is therefore unlikely that, without prior knowledge of the importance of *do* to the success of the simulations*,* a researcher constructing artificial input sets would include constructions with *do* at a sufficient rate to capture the relevant effects. This example underlines both the potential value of using computational modelling techniques as a tool for investigating the relation between children's speech and the characteristics of the input language, and the importance of using realistic input in which the relative frequencies of different items and constructions are properly represented (cf. Christiansen & Chater, 2001).

Finally, our results show that, provided realistic input data is used, it is possible to simulate both cross-linguistic differences in the rate at which OI errors occur at different points in development and cross-linguistic variation in the modal reference and eventivity of RI errors using one identical learning mechanism. Thus, in the present study, we used a version of MOSAIC that simulates the pattern of developmental change in OI errors across English, Dutch, German, and Spanish to simulate cross-linguistic variation in the modal reference of Dutch/German and English children's RIs. This version of MOSAIC simulates the cross-linguistic pattern of finiteness marking in terms of the interaction between an utterance-final bias in learning and the distributional characteristics of the speech to which children learning different languages are exposed. The fact that the same version of MOSAIC can also simulate quite subtle differences in the characteristics of RIs in Dutch/German and English suggests that this kind of model is capable of explaining cross-linguistic variation in the OI phenomenon at a surprising level of detail. It also provides further support for the view that it is possible to explain key features of children's multi-word speech in terms of the interaction between a resource-limited distributional learning mechanism and the surface properties of the language to which children are exposed.

Acknowledgements

The authors would like to thank the Max Planck Institute for Evolutionary Anthropology, Leipzig for providing access to the Leo and Thomas corpora. This research was funded by the Economic and Social Research Council under grant number RES000230211.

References

Aguado-Orea, J. (2004). *The acquisition of morpho-syntax in Spanish: Implications for current theories of development*. Doctoral Dissertation, University of Nottingham, United Kingdom.

Bates, E. & Carnavale, G. F. (1993). New directions in research on child development. *Developmental Review, 13*, 436-470.

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

Blom, E. (2003). *From root infinitive to finite sentence: The acquisition of verbal inflections and auxiliaries*. Doctoral dissertation, University of Utrecht, The Netherlands.

Blom, E. (2007). Modality, infinitives and finite bare verbs in Dutch and English child language. *Language Acquisition, 14*, 75-113.

Blom, E., Krikhaar, E. & Wijnen, F. (2001). Nonfinite clauses in Dutch and English child language: An experimental approach. In: A. H-J. Do, L. Domínquez & A. Johansen (Eds.): *Proceedings of the 25th annual Boston University Conference on Language Development* (pp. 133-144). Somerville, MA: Cascadilla Press.

Bloom, P. (1990). Subjectless sentences in child language. *Linguistic Inquiry*, *21*, 491-504.

Bol, G.W. (1995). Implicational scaling in child language acquisition: The order of production of Dutch verb constructions. In M. Verrips & F. Wijnen, (Eds.), *Papers from the Dutch-German Colloquium on Language Acquisition*, Amsterdam Series in Child Language Development, 3, Amsterdam: Institute for General Linguistics.

Boser, K., Lust, B., Santelmann, L. & Whitman, J. (1992). The syntax of CP and V2 in early child German: The strong continuity hypothesis. *Proceedings of NELS, 22*, 51-65.

Brown, R. (1973). *A first language: The early stages*. Cambridge, MA: Harvard University Press.

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

Conway, C. M. & Christiansen, M. H. (2005). Modality-constrained statistical learning of tactile, visual and auditory sequences. *Journal of Experimental Psychology: Learning, Memory and Cognition, 31*, 24-39.

Cowan, N., Saults, J. S. & Brown, G. D. A. (2004). On the auditory modality superiority effect in serial recall: separating input and output factors. *Journal of Experimental Psychology: Learning Memory and Cognition, 30*, 639-644.

Dabrowska, E. & Lieven, E. (2005). Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics*, *16*, 437-474.

Deen, K. U. (1997). *The interpretation of root infinitives in English: Is eventivity a factor?* Unpublished Manuscript, UCLA, California.

Ferdinand, A. (1996). *The development of functional categories: The acquisition of the subject in French*. Doctoral dissertation. University of Leiden, The Netherlands.

Freudenthal, D., Pine, J. M. & Gobet, F. (2005a). Simulating the Cross-Linguistic Development of Optional Infinitive Errors in MOSAIC. In B. G. Bara, L. Barsalou & M. Buchiarelli (Eds.), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 702-707). Mahwah, NJ: Erlbaum.

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

Freudenthal, D., Pine, J. M. & Gobet, F. (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., Aguado-Orea, J. & Gobet, F. (2007). Modelling the developmental patterning of finiteness marking in English, Dutch, German and Spanish using MOSAIC. *Cognitive Science*, *31*, 311-341.

Gleitman, L. & Wanner, E. (1982). Language acquisition: The state of the art. In E. Wanner and L. Gleitman (Eds.), *Language acquisition: The state of the art* (pp. 3-48). Cambridge: Cambridge University Press.

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.

Hoekstra, T. & Hyams, N. (1998). Aspects of root infinitives. *Lingua*, *106*, 81-112.

Hyams, N. (1996). The underspecification of functional categories in early grammar. In H. Clahsen (Ed.), *Generative perspectives in language acquisition* (pp. 1-42). Philadelphia: John Benjamins.

Hyams, N. (2001). Now you hear it, now you don't: The nature of optionality in child language. In A. H-J. Do, L. Domínquez & A. Johansen (Eds.), *Proceedings of the 25th annual Boston University Conference on Language Development* (pp. 34-58). Somerville, MA: Cascadilla Press.

Hyams, N. (2007). Aspectual effects on interpretation in early grammar. *Language Acquisition, 14*, 231-268.

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

Lasser, I. (1997). *Finiteness in adult and child German*. Doctoral dissertation, Max Planck Institute, Nijmegen, The Netherlands.

Lasser, I. (2002). The roots of root infinitives: Remarks on infinitival main clauses in adult and child language. *Linguistics, 40*, 767-796.

Legate, J. A. & Yang, C. (2007). Morphosyntactic learning and the develoment of tense. *Language Acquisition, 14*, 315-344.

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.

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

Penney, C. G. (1989). Modality effects and the structure of short-term verbal memory. Memory and Cognition, 17, 398-422.

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

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

Rizzi, L. (1994). Some notes on linguistic theory and language development: The case of root infinitives. *Language Acquisition*, 3, 371-393.

Schütze, C. T. & Wexler, K. (1996). Subject case licensing and English root infinitives. In A. Stringfellow, D. Cahma-Amitay, E. Hughes & A. Zukowski (Eds.), *Proceedings of the 20th Annual Boston University Conference on Language Development*. Somerville, MA: Cascadilla Press.

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

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. (1998). The temporal interpretation of Dutch children's root infinitivals: The effect of eventivity. *First Language*, *18*, 379-402.

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

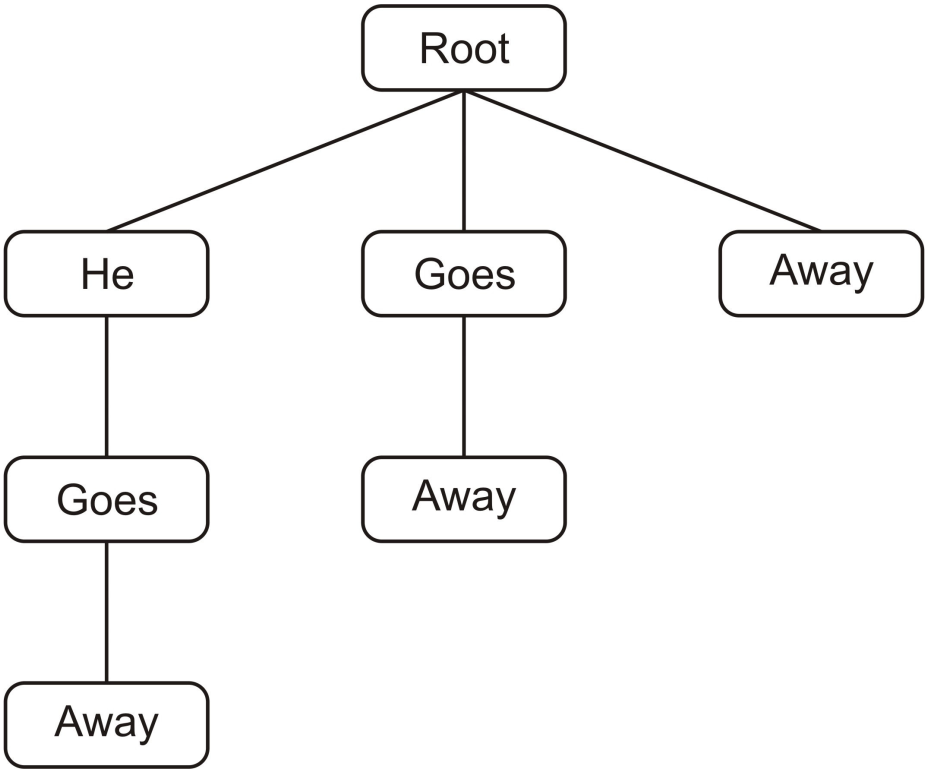

Fig. 1: A MOSAIC network after it has seen the phrase *He goes away* five times.

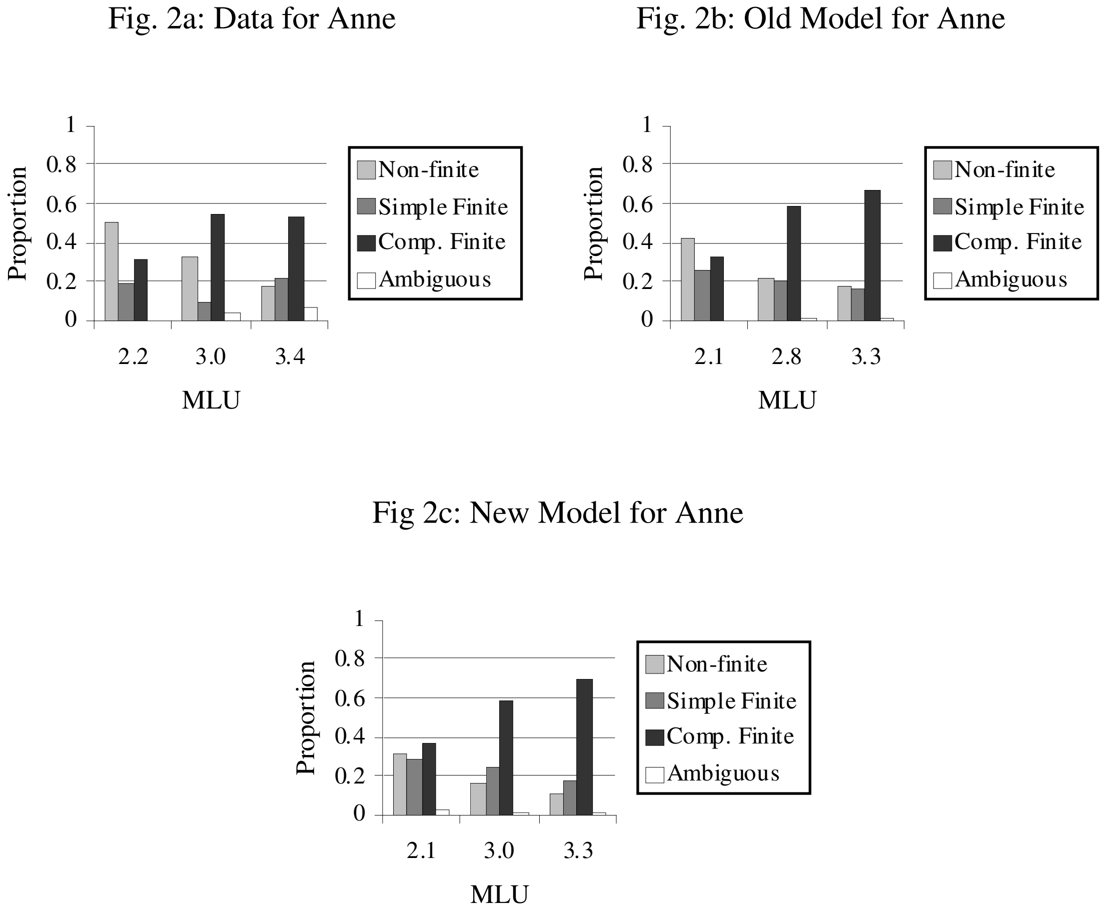
Fig. 2a: Data for Anne

Fig. 2b: Old Model for Anne

Fig 2c: New Model for Anne

Fig. 2: Proportions of non-finite, simple, and compound finites (a) for an English child, (b) with the earlier MOSAIC simulations, and (c) with the new MOSAIC simulations.

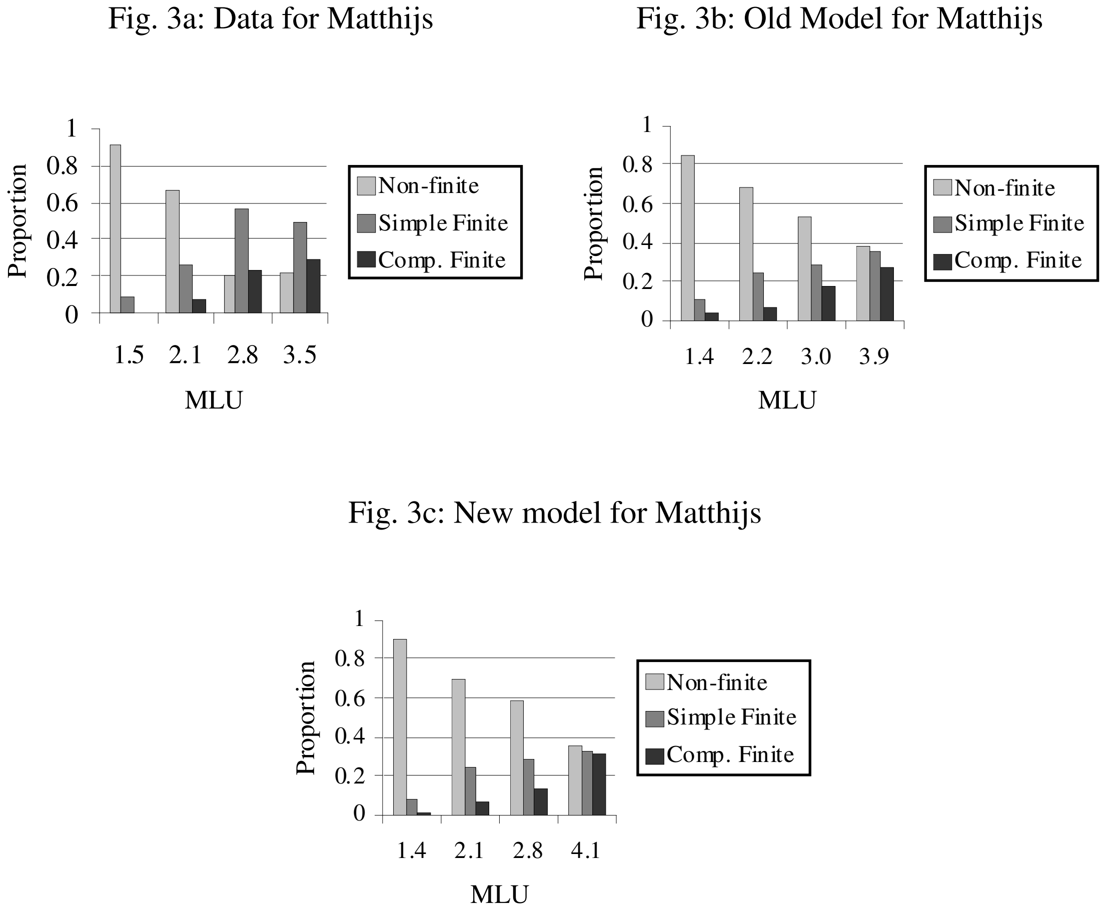

Fig. 3a: Data for Matthijs

Fig. 3b: Old Model for Matthijs

Fig. 3c: New model for Matthijs

Fig. 3: Proportions of non-finite, simple, and compound finites (a) for a Dutch child, (b) with the earlier MOSAIC simulations, and (c) with the new MOSAIC simulations.

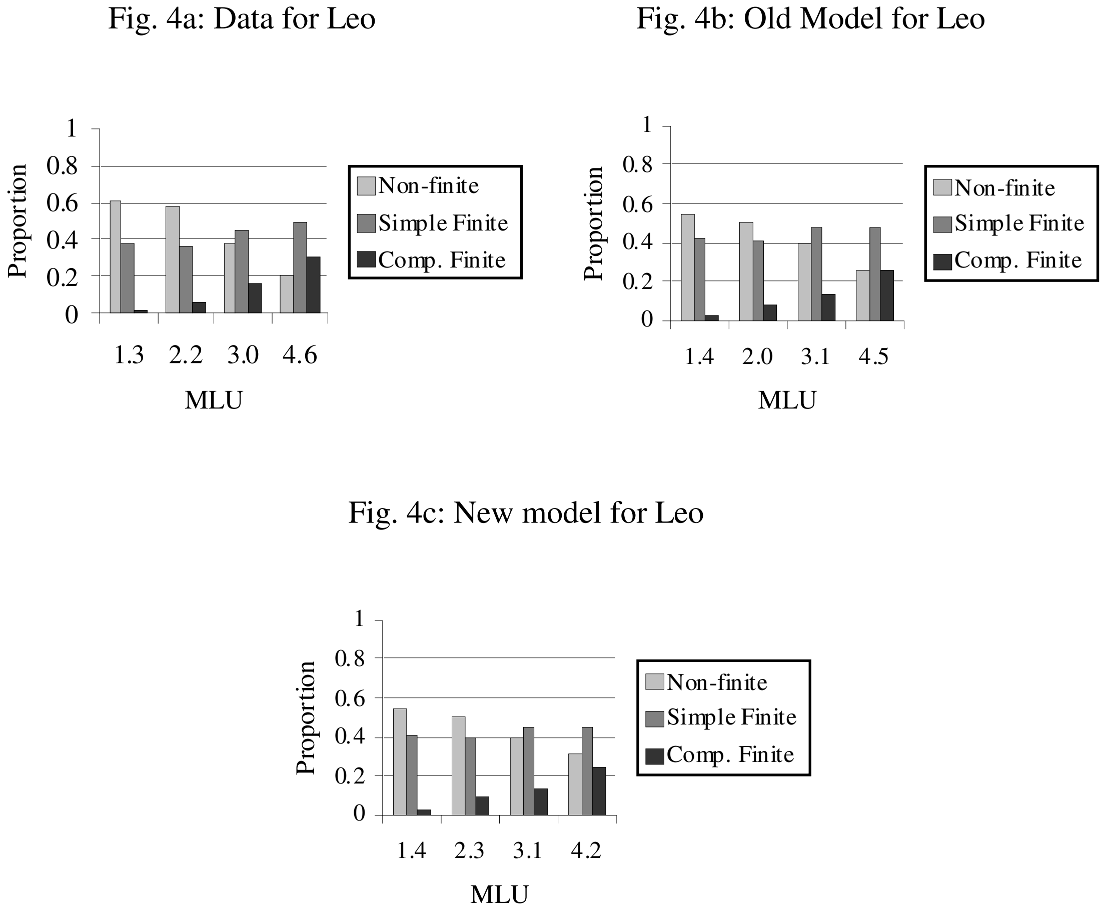

Fig. 4a: Data for Leo

Fig. 4b: Old Model for Leo

Fig. 4c: New model for Leo

Fig. 4: Proportions of non-finite, simple, and compound finites (a) for a German child, (b) with the earlier MOSAIC simulations, and (c) with the new MOSAIC simulations[6].

[6] In order to allow a proper comparison with the earlier simulations, this simulation was run on the same random sample of 30,000 utterances from the Leo corpus that was used in the earlier simulations. Other simulations reported in this paper were conducted using the full Leo corpus as input.

Table 1: (Average) proportions of modal RIs reported in the previous literature for English, Dutch, and German

	Prop. Modal RIs	Range
English		
Deen (1997)	.13	N.A.
Blom, Krikhaar & Wijnen (2001)	.44	N.A.
Blom (2007)	.36	N.A.
Dutch		
Wijnen (1998)	.86	.74 to .95
Blom, Krikhaar & Wijnen (2001)	.68	N.A.
Blom (2003)	.74	.64 to .80
Blom (2007)	.61	N.A.
German		
Ingram & Thompson (1996)		
Lenient	.77	.48 to 1.00
Strict	.52	.21 to .84
Lasser (1997)	.71	.69 to .72

Table 2: (Average) MLU and proportions of modal RIs for the English, Dutch, and German simulations (standard deviations in parentheses).

	MLU	Prop. Modal RIs
English (Manchester, N=12)	2.66 (.28)	.35 (.07)
Dutch (Groningen, N=7)	2.65 (.22)	.56 (.08)
German (Leo)	2.61 (N.A.)	.56 (N.A.)
English (Thomas)	2.62 (N.A.)	.36 (N.A.)

Table 3: (Average) proportions of modal RIs for the English, Dutch and German simulations restricted to third person singular contexts (standard deviations in parentheses).

	Prop. Modal RIs
English (Manchester, N=12)	.28 (.07)
Dutch (Groningen, N=7)	.68 (.14)
German (Leo)	.78 (N.A.)
English (Thomas)	.13 (N.A.)

Table 4: (Average) proportions of eventive RIs for the English, Dutch, and German simulations restricted to third person singular contexts (standard deviations in parentheses).

	Prop. Eventive RIs
English (Manchester, N=12)	.76 (.11)
Dutch (Groningen, N=7)	.91 (.08)
German (Leo)	.92 (N.A.)
English (Thomas)	.79 (N.A.)

Table 5: Proportion of contexts that are modal for (third) singular subjects followed by an infinitive form for English, Dutch, and German (Number of contexts in parentheses).

	Proportion of modal contexts
English	.16 (3078)
Dutch	.68 (980)
German	.88 (254)

Table 6: Most frequent stative verbs that occurred in the RIs of the models trained on the Manchester and Groningen corpus (frequency per 100 RIs in parentheses).

English	Dutch
Fit (4.33)	Zitten (sit) (2.92)
Sit (4.00)	Slapen (sleep) (1.30)
Like (3.37)	Hebben (have) (1.30)
Want (2.40)	Blijven (stay) (.97)
Have (1.44)	Liggen (lie) (.97)
Sleep (1.24)	Zien (see) (.65)
Need (1.12)	Horen (hear) (.65)
See (1.12)	Vergeten (forget) (.32)
Wait (.96)	Wachten (wait) (.32)