J. Child Lang., Page 1 of 27. © Cambridge University Press 2010
doi:10.1017/S0305000909990523

Explaining quantitative variation in the rate of Optional Infinitive errors across languages: A comparison of MOSAIC and the Variational Learning Model*

DANIEL FREUDENTHAL AND JULIAN PINE
University of Liverpool

AND

FERNAND GOBET
Brunel University

(*Received* 19 *December* 2008 – *Revised* 14 *August* 2009 – *Accepted* 5 *December* 2009)

ABSTRACT

In this study, we use corpus analysis and computational modelling techniques to compare two recent accounts of the OI stage: Legate & Yang's (2007) Variational Learning Model and Freudenthal, Pine & Gobet's (2006) Model of Syntax Acquisition in Children. We first assess the extent to which each of these accounts can explain the level of OI errors across five different languages (English, Dutch, German, French and Spanish). We then differentiate between the two accounts by testing their predictions about the relation between children's OI errors and the distribution of infinitival verb forms in the input language. We conclude that, although both accounts fit the cross-linguistic patterning of OI errors reasonably well, only MOSAIC is able to explain why verbs that occur more frequently as infinitives than as finite verb forms in the input also occur more frequently as OI errors than as correct finite verb forms in the children's output.

INTRODUCTION

An important feature of language development research over the past forty years has been the collection of rich corpora of early child language and child-directed speech from a variety of different languages (MacWhinney,

[*] We would like to thank the Max Planck Institute for Evolutionary Anthropology, Leipzig for allowing us access to the Leo Corpus. This research was funded by the ESRC under Grant Number RES-062-23-1348. Address for correspondence: Daniel Freudenthal, School of Psychology, University of Liverpool, Bedford Street South Liverpool, L69 7ZA, United Kingdom. tel: 0151 794 1108; e-mail: D.Freudenthal@liverpool.ac.uk

2000). The availability of these corpora has led to the identification of some key cross-linguistic phenomena in children's early multiword speech, and facilitated the development of models that seek to integrate across data from a wide range of different languages. However, these models are often not sufficiently well specified to generate quantitative predictions about the cross-linguistic data and, as a result, can be rather difficult to test.

In the present paper, we show how using corpus analysis and computational modelling techniques to test models of cross-linguistic phenomena can both allow us to identify the weaknesses of particular accounts, and provide us with important insights into the relation between children's early language and the distributional properties of the language to which they are exposed. We focus on one particular cross-linguistic phenomenon: the Optional Infinitive (OI) Stage, and two models of this phenomenon: Freudenthal, Pine & Gobet's (2006) Model of Syntax Acquisition in Children (MOSAIC) and Legate & Yang's (2007) Variational Learning Model (VLM). We first describe the OI phenomenon and the two models of the OI stage that are the focus of our investigation. We then evaluate these models in terms of their ability to explain the data from five different languages (English, Dutch, German, French and Spanish), before concluding with a discussion of the implications of our results for the field as a whole.

The Optional Infinitive phenomenon

In many languages, children go through a stage in which they produce non-finite verb forms in contexts in which a finite verb form is obligatory. For example, English-speaking children produce utterances like (1a) instead of the correct (1b); Dutch children produce utterances like (2a) instead of the correct (2b); and French children produce utterances like (3a) instead of the correct (3b):

(1a) Mummy go to work
(1b) Mummy goes to work
(2a) Ik ijs eten
I ice cream eat-INF
(2b) Ik eet ijs
I eat-FIN ice cream
(3a) La poupée dormir
The doll sleep-INF
(3b) La poupée dort
The doll sleep-FIN

These errors involve the use of an infinitival verb form (zero-marked in English, but marked with the infinitival morphemes *-en* in Dutch and, in

our examples, *-ir* in French) in contexts in which a finite verb form is obligatory. Since they tend to occur at a stage when the child is also producing correctly marked finite forms, they have come to be known in the literature as Optional Infinitive (OI) errors (Wexler, 1994).

A number of theories have been proposed to explain the occurrence of OI errors in children's speech (e.g. Rizzi, 1994; Hyams, 1996; Hoekstra & Hyams, 1998; Wexler, 1998). These accounts can explain why children tend to make OI errors in some languages and not in others. For example, Wexler's (1998) account predicts that children will make OI errors in obligatory subject languages such as English, Dutch and French, but not in optional subject languages such as Spanish and Italian. However, they are unable to explain the wide range of variation that exists in the rate at which OI errors occur across languages. For example, Phillips (1995) reviews data from children learning nine different languages (including five OI languages and four non-OI languages) and concludes that rates of OI errors vary along a continuum from high in English and Swedish through moderate in Dutch, French and German to low (but by no means zero) in Catalan, Hebrew, Italian and Spanish.

Two recent models of the OI stage that take this quantitative variation more seriously are Legate and Yang's VLM (Legate & Yang, 2007), and Freudenthal *et al*.'s MOSAIC (Freudenthal *et al*., 2006; 2009; Freudenthal, Pine, Jones & Gobet, submitted; Freudenthal, Pine, Aguado-Orea & Gobet, 2007). Although these models differ considerably in their underlying assumptions, both are explicitly designed to address the graded nature of the OI phenomenon by making quantitative predictions about the developmental data. The aim of the present study is to use corpus-based and computational modelling techniques to compare these two accounts and evaluate their ability to explain quantitative variation in the rate and patterning of OI errors across languages.

The Variational Learning Model

Legate & Yang's (2007) VLM is one of a class of models explicitly designed to account for quantitative aspects of the developmental data within a generativist framework. According to the variational learning approach (Yang, 2002; 2004), the child's grammar at any particular point in development can be modelled as a population of innately derived hypotheses whose composition changes during the course of learning. The child is seen as entertaining a number of possible grammars (or parameter settings) each of which is associated with a particular probability. However, the distribution of these probabilities is assumed to change adaptively in response to linguistic data in the environment. Thus, when a particular grammar (or parameter setting) is used to parse linguistic data, it is rewarded

by utterances that are consistent with it, and punished by utterances that are not consistent with it. The child is assumed to converge on the correct grammar of the language by gradually abandoning hypotheses that are not consistent with the input data. However, the probabilistic nature of the learning process means that there may be a long period during which the child continues to entertain two or more competing hypotheses.

According to the VLM, OI errors reflect the fact that children learning tense-marking languages (i.e. languages with a [+Tense] grammar) initially entertain the hypothesis that they are learning a language such as Mandarin Chinese, which does not manifest tense marking (i.e. a language with a [−Tense] grammar). This hypothesis is gradually abandoned in response to utterances in the input language that reward the [+Tense] grammar. However, the time taken to abandon the [−Tense] grammar varies as a function of the amount of morphological evidence for tense marking in the input. Thus, children learning a morphologically rich language such as Spanish emerge from the OI stage relatively early because a large proportion of the utterances in their input reward the [+Tense] grammar, whereas children learning a morphologically impoverished language such as English emerge from the OI stage relatively late because only a small proportion of the utterances in their input reward the [+Tense] grammar.

The great strength of the VLM is that, because it incorporates a probabilistic learning mechanism, it can be used to derive predictions about the order in which children learning different languages will emerge from the OI stage, and hence about variation in the rate of OI errors in different languages at particular points in development. Indeed, it is possible to use the model to compute precise quantitative measures of the extent to which the input in any particular language rewards the [+Tense] grammar, and hence to derive clear quantitative predictions about the rate at which OI errors will occur across the full range of tense-marking languages.

Legate & Yang (2007) test the VLM by deriving corpus-based measures of the extent to which English, French and Spanish input reward the [+Tense] grammar. Their results show that Spanish input rewards the [+Tense] grammar more than French input, and that French input rewards the [+Tense] grammar more than English input. Since rates of OI errors tend to by very high in English, moderately high in French and very low in Spanish, these results provide some support for the VLM. However, given that the model is intended to capture variation across the full range of tense-marking languages, evaluating its success on languages with such different rates of OI errors might be regarded as rather a weak test of its predictions.

In view of this potential criticism, one of the aims of the present study is to conduct a stronger test of the VLM by assessing its ability to explain more subtle differences in the rate of OI errors across languages. This will be done by deriving corpus-based measures of the extent to which input from five different languages rewards the [+Tense] grammar, and examining the relation between these measures and the rate of OI errors in the speech of children learning these languages. The languages investigated include the three languages examined by Legate and Yang (English, French and Spanish), and two further languages (Dutch and German). Since both Dutch and German are described by Phillips (1995) as languages with moderately high rates of OI errors, assessing the model's ability to explain variation across these five languages (and particularly across Dutch, German and French, the languages in the middle range of the distribution) constitutes a much stronger test of the VLM than that conducted by Legate and Yang.

MOSAIC

MOSAIC is a constructivist model of language learning, with no built-in knowledge of syntactic categories or rules, which is implemented as a working computational model. MOSAIC takes as input corpora of 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 in different languages across a range of Mean Length of Utterance (MLU) values, and hence to model cross-linguistic variation in the rate at which particular kinds of errors occur at particular points in development.

The basis of MOSAIC is an n-ary discrimination net consisting of nodes and arcs that connect these nodes. At the top of the network is an empty root node. Nodes directly below the root node are called primitive nodes and store the words that MOSAIC has encoded. Nodes at deeper levels in the network store sequences of words or phrases encoded by the model. MOSAIC learns from orthographically transcribed input with whole words being the unit of analysis. As the model sees more input, it creates more nodes encoding the words that it has encountered; it also creates nodes at deeper levels in the network, representing progressively longer phrases. The amount and average length of the output that MOSAIC can produce thus increases as a function of the amount of input to which the model is exposed.

A major constraint on the way that MOSAIC networks are built up is that the model is subject to an utterance-final bias in learning. That is, MOSAIC does not encode a word or phrase unless everything that follows

that phrase has already been encoded in the network. MOSAIC thus builds up its representation of an utterance by starting at the end of the utterance and slowly working its way to the beginning. The version of the model used for the present simulations complements MOSAIC's utterance-final bias or right-edge learning mechanism with a (smaller) utterance-initial bias or left-edge learning mechanism. This version of the model was first described in Freudenthal, Pine & Gobet (2005) and has been developed further in Freudenthal *et al*. (submitted). The utterance-initial bias enables MOSAIC to associate utterance-initial words and short (frequent) phrases with (longer) utterance-final phrases. As a result, MOSAIC's output now consists of a mixture of utterance-final phrases and concatenations of utterance-final and utterance-initial phrases. The main reason for developing this new version of the model was that earlier versions of MOSAIC were unable to produce strings with missing sentence-internal elements, and the model was thus unable to simulate certain types of errors that are relatively frequent in the output of young language-learning children. For example, English-speaking children often produce errors such as *Where he go?*, which MOSAIC was unable to simulate because strings such as *Where he go?* do not occur in utterance-final position in the input. A further reason for implementing an utterance-initial bias was that MOSAIC was dependent on questions as the source of OI errors with third person singular subjects (e.g. *He jump*). Previous versions of MOSAIC produced such utterances by learning utterance-final phrases from compound questions such as *Can he jump?* The utterance-initial bias allows MOSAIC to produce such forms by learning to associate the utterance-initial subject and the utterance-final verb in declarative utterances such as *He can jump*. Thus, MOSAIC learns the *He* in *He jump* using the new utterance-initial learning mechanism and the *jump* in *He jump* using the old utterance-final learning mechanism. It then learns to associate the two words on the basis of their co-occurrence as utterance-initial and utterance-final words in utterances in the input. This mechanism has the potential to generate non-child-like concatenations such as *He was jump* from utterances such as *He was going to jump*. Such errors are avoided by making the utterance-final bias stronger than the utterance-initial bias and making the probability of associating utterance-initial and utterance-final sequences dependent on the distance between these sequences in the relevant input utterance.

MOSAIC now represents declaratives and questions separately, and the analyses reported in the present paper focus on declarative output from MOSAIC that was learned from declarative input to the model. Figure 1 shows an example of a network that has learned an utterance-initial and utterance-final phrase.

Learning in MOSAIC is a probabilistic process that is relatively slow. Input corpora are fed through the model multiple times, and output

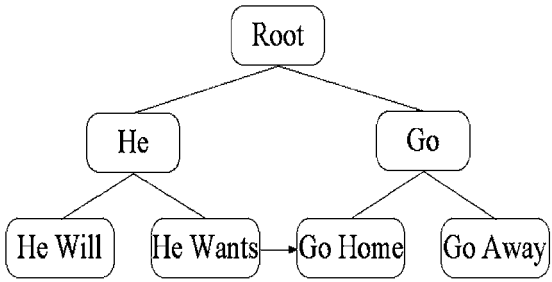

Fig. 1. A sample MOSAIC network that has learned an utterance-initial and utterance-final phrase.

(of increasing average length) can be generated after every exposure to the input. The probability of creating a node encoding a word or phrase in MOSAIC is governed by the following formula:

$$NCP = \left(\frac{1}{1 + e^{(m-u/c)/3}}\right)^{\sqrt{d}}$$

where: NCP = Node Creation Probability.

m = a constant, set to 40 for these simulations.
c = corpus size (number of utterances).
u = total number of utterances seen.
d = distance to the edge of the utterance.

This formula is designed to ensure that the model displays the following features. First, the base number in the formula is a sigmoid bounded between 0 and 1. The formula includes the term (m–u/c), which results in the learning rate speeding up with increasing exposures to the input corpus. The term u/c (or number of times the input corpus has been seen) is subtracted from the constant m. This ensures that after n exposures to the input corpus (i.e. 40) the learning rate is identical for corpora of different sizes (i.e. 0·5). Second, the basic learning rate in the formula is raised to the power of the square root of the distance to the edge of the utterance. As the basic learning rate is bounded between 0 and 1, this means that the probability of encoding a longer phrase is lower than the probability of creating a short phrase. This advantage for shorter phrases decreases with training as the base probability approaches 1. For left-edge learning, the distance to the edge of the utterance is increased by 2, making it slower than right-edge learning.

MOSAIC has two ways of generating output. First, output is produced from MOSAIC by traversing all the branches in the network. When a

terminal node (or end-of-utterance marker) is encountered, the phrase encoded in that branch is produced. This mechanism results in the production of all the utterance-final phrases that the model has encoded. When an utterance-final phrase has been associated with an utterance-initial phrase, the concatenation of the utterance-final and utterance-initial phrase is also produced. Thus, the sample model in Figure 1 is capable of producing the utterance-final phrases *go away* and *go home*, as well as the concatenation *he wants go home*. This first mode of generating output can only result in the production of (partial) utterances that were present in the input. Second, MOSAIC employs a generativity mechanism that allows it to substitute distributionally similar words in novel contexts. This leads to the production of output that was not present in MOSAIC's input. MOSAIC's generativity mechanism is described in more detail in Freudenthal *et al.* (2007).

It is worth emphasising at this point that MOSAIC is a relatively simple distributional analyzer with no access to semantic information, which is clearly not powerful enough to acquire many aspects of adult syntax. MOSAIC is therefore best viewed, not as a realistic model of the language acquisition process itself, but as one of many possible ways of implementing an utterance-final (and in the current version of the model, utterance-initial) bias in learning.

MOSAIC simulates OI errors by learning them from COMPOUND FINITES: utterances that contain a (finite) modal or auxiliary and a non-finite main verb (e.g. *He can go home*). As noted above, MOSAIC learns from the right and left edges of the utterance, and links together (short) utterance-initial and (longer) utterance-final phrases. This learning procedure results in the production of truncated utterances (i.e. utterances that occur as utterance-final phrases in the input) and utterances with missing sentence-internal material (i.e. utterances that reflect the concatenation of an utterance-initial and an utterance-final phrase). Thus, MOSAIC learns English OI errors such as *Go home* and *He go home* from compound finite utterances such as *He can go home*. Similarly MOSAIC learns Dutch OI errors such as *Ijs eten* 'Ice cream eat-INF' and *Hij ijs eten* 'He ice cream eat-INF' from compound finite utterances such as *Hij wil ijs eten* 'He wants ice cream eat-INF'.

MOSAIC simulates the developmental patterning of OI errors because it learns to produce progressively longer utterances as a function of the amount of input to which it is exposed. Children start out producing OI errors at high rates, and produce fewer OI errors 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 most of what MOSAIC learns is learned from the right edge of the utterance, the early

(short) utterances produced by the model tend to contain only non-finite verb forms. As the utterances MOSAIC produces become longer, finite modals and auxiliaries start to appear, and OI errors are slowly replaced by compound finites.

Previous work with MOSAIC has shown that a model that learns compound finites from the right edge of the utterance is surprisingly good at simulating cross-linguistic variation in the rate at which OI errors occur. Thus, in an initial study, Freudenthal, Pine & Gobet (2006) showed that MOSAIC was able to simulate developmental changes in the rate of OI errors in two languages: English and Dutch. In a more recent study, Freudenthal, Pine, Aguado-Orea & Gobet (2007) showed that a modified version of MOSAIC was able to simulate the developmental patterning of OI errors in four languages, including English and Dutch, a third OI language (German) and a non-OI language (Spanish). Freudenthal *et al.* (2007) also showed that MOSAIC's success in simulating the differences between Dutch, German and Spanish could be explained in terms of the interaction between the model's utterance-final bias in learning and the relative frequency of non-finite and finite verb forms in utterance-final position in the input (0·87 for Dutch, 0·66 for German and 0·26 for Spanish).

Like the VLM, MOSAIC provides a powerful means of predicting cross-linguistic variation in the rate at which OI errors occur. However, MOSAIC currently only simulates two of the three languages studied by Legate and Yang (English and Spanish). Moreover, the way in which MOSAIC's output has been analyzed in previous simulations of English makes it difficult to conduct a fair comparison of MOSAIC and the VLM. Thus, because it is only possible to distinguish between correct finites and OI errors in third person singular contexts in English, previous simulations of English have focused only on utterances with an explicit third person singular subject (e.g. *He go there* versus *He goes there*, but not *Go there* versus *Goes there*). These simulations have shown that MOSAIC provides a good fit to the English data. However, they cannot be directly compared with MOSAIC simulations in Dutch, German and Spanish.

In view of these considerations, a second aim of the present study is to extend on previous work with MOSAIC so that the model's predictions can be directly compared with those of the VLM. This will be done, first by supplementing previous simulations of English, Dutch, German and Spanish with simulations of a further OI language (French), and second by using a new method to simulate English that will allow us to focus on OI errors with and without explicit subjects. These developments will allow a direct comparison of the ability of MOSAIC and the VLM to predict cross-linguistic variation in the rate of OI errors at a particular point in development.

Explicitly distinguishing between MOSAIC and the VLM

Comparing MOSAIC's and the VLM's ability to explain quantitative variation in rates of OI errors across languages is clearly a potentially powerful way of distinguishing between the two accounts. However, it is possible that the predictions of both accounts will fit the cross-linguistic data reasonably well. It is therefore also worth trying to distinguish more explicitly between the two models by identifying areas of the data where they can be shown to make radically different predictions.

One area of the data where MOSAIC and the VLM do appear to make very different predictions is the relation between the errors made by children during the OI stage and the distributional properties of the input to which they have been exposed. Thus, according to MOSAIC, OI errors are compound finite constructions with missing modals or auxiliaries, learned directly from compound finite constructions in the input. One prediction that follows from this view is that verbs that tend to occur as the non-finite form (e.g. *kick*) in compound finite constructions (e.g. *He can kick the ball*) will be more likely to occur as OI errors than verbs that tend to occur as the finite form (e.g. *wants*) in simple finite constructions (e.g. *He wants a drink*). That is to say, MOSAIC predicts input-driven lexical effects on the distribution of OI errors in children's speech.

According to the VLM, on the other hand, OI errors are NOT learned directly from the input, but reflect the current state of the child's underlying grammatical system (i.e. the fact that the child has yet to reject the hypothesis that she is learning a language with a [−Tense] grammar). Thus, although the speed with which the child emerges from the OI stage, and hence the rate at which OI errors occur, reflects the amount of morphological evidence for tense marking in the input, the distribution of non-finite verb forms in the input language has no role to play in determining the distribution of OI errors in the child's speech. That is to say, the VLM does not predict input-driven lexical effects on the distribution of OI errors in children's speech. Indeed, since OI errors are assumed to reflect the probabilistic use of a [−Tense] grammar, the VLM would seem to predict that OI errors will occur at more or less the same rate across different verbs (i.e. at a rate determined by the probability associated with the [−Tense] grammar).

This difference between MOSAIC and the VLM suggests that it is possible to distinguish between the two models by looking for lexical effects on the distribution of OI errors in children's speech in each of the languages under investigation. A third aim of the present study is therefore to explicitly distinguish between MOSAIC and the VLM by investigating the relation between the relative frequency with which particular verbs occur as OI errors as opposed to correctly marked finite forms in children's speech,

and the relative frequency with which they occur as finite forms in simple finite structures and non-finite forms in compound finite structures in the children's input.

METHOD

The present study consists of four sets of analyses aimed at: (1) determining the rate of OI errors in children's speech in each of the five languages; (2) evaluating the VLM by determining the proportion of clauses rewarding a [+Tense] grammar in child-directed speech in each of the five languages; (3) evaluating MOSAIC by determining the proportion of OI errors produced by MOSAIC in each of the five languages; and (4) explicitly distinguishing between MOSAIC and the VLM by looking for lexically specific effects on the patterning of OI errors.

Determining the rate of OI errors in children's speech

Measures of the rate of OI errors in children's speech were derived by analyzing data from fourteen children: Anne, Aran, Becky, Dominic, Gail and John (English), Tim and Anais (French), Juan and Lucía (Spanish), Matthijs and Peter (Dutch), and Leo and Rah (German). The data for the English children are part of the Manchester corpus (Theakston, Lieven, Pine & Rowland, 2001); the data for the French children are part of the Lyon corpus (Demuth & Tremblay, 2008); the data for the Spanish children make up the Madrid corpus (Aguado-Orea, 2004) and the data for the Dutch children are part of the Groningen corpus (Bol, 1996). The data for the German children come from two different sources. Rah's data form part of the Szagun corpus (Szagun, 2001) and Leo's data constitute a dense corpus (consisting of over 140,000 parental utterances and over 50,000 child utterances) made available to us by the Max Planck Institute for Evolutionary Anthropology in Leipzig (Behrens, 2006). All of these corpora are available or, in the case of Juan, Lucía and Leo, soon to be available, in the CHILDES database (MacWhinney, 2000).

Rates of OI errors were computed by selecting blocks of consecutive transcripts at an MLU of approximately 2·0 words, extracting all utterances including a verb other than the copula and coding them in the manner set out below. Note that, since both the VLM and MOSAIC make predictions about the rate at which children will use non-finite forms in finite contexts, rather than just about the rate at which children will produce infinitives in contexts in which a finite main verb is required, these measures do not distinguish between bare infinitive errors (e.g. *That go there*) and bare participle errors (e.g. *That going there* or *That gone there*). Both types of error are treated as OI errors in the analysis. Note also that, since

MOSAIC's output consists of a corpus of utterance types rather than tokens, the child measures are also based on utterance types rather than tokens (where, for example, *Go there*, *That go there* and *That goes there* are treated as three different utterance types, but multiple instances of *Go there*, *That go there* and *That goes there* are only counted once). In practice, coding utterances in this way makes very little difference to the results since, although multiple uses of the same word are common in children's speech, the combinatorial properties of language are such that multiple uses of the same multiword utterance are relatively rare.

Coding procedure

All utterances including a verb form other than the copula were coded as either (incorrect) non-finite utterances or (correct) finite utterances. The first category included all utterances that contained one or more non-finite verb forms but no finite verb forms (i.e. utterances including bare infinitives and bare past and progressive participles). The second category included utterances that contained at least one finite verb form (i.e. simple finite utterances and utterances including a finite auxiliary or modal and a non-finite verb form). In French and Spanish, coding the data in this way is relatively straightforward, since finite forms are readily distinguishable from non-finite forms in these languages. However, in Dutch and German there is the complication that plural present tense forms are indistinguishable from the infinitive. Since plural present tense forms are relatively infrequent, this problem was dealt with in the present study by treating all forms that matched the infinitive as non-finite. Note that coding the data in this way means that the measures reported for Dutch and German are likely to overestimate the rate of OI errors in Dutch and German to some extent. However, given the low frequency with which plural present tense forms occur in the input data, it is unlikely to have a major impact on the results. Moreover, it has the advantage of allowing a direct comparison between the child data and MOSAIC's output, where plural present tense verb forms are also indistinguishable from infinitives. For English, the problem of distinguishing between finite and non-finite verb forms is much more serious than in Dutch and German, since infinitives are indistinguishable from all present tense verb forms other than the third person singular. Analysis was therefore restricted to third person singular contexts in English since, in this case, the provision of an uninflected verb form in a third person singular context is, by definition, an OI error. This analysis was conducted by hand-coding the English data for third person singular contexts, including contexts with an explicit third person singular subject and contexts in which it was possible to infer an implicit third person singular subject from the surrounding discourse.

Evaluating the VLM

The VLM was evaluated by determining the proportion of utterances that rewarded a [+Tense] grammar in the input of one child from each of the five languages. The selected children were: Anne for English, Matthijs for Dutch, Leo for German, Tim for French and Juan for Spanish. In the case of English, French and Spanish, this analysis simply involved following the procedures laid out in Legate & Yang (2007). In the case of Dutch and German, past tense forms and second and third person singular present tense forms all have overt tense or tense-dependent morphology and so were assumed to reward the [+Tense] grammar, whereas plural present tense forms match the infinitive, and were thus assumed not to reward the [+Tense] grammar. The status of the first person singular present tense form is somewhat ambiguous. In Dutch, this form consists of the stem, and in German the suffix is a single schwa, which is often not pronounced (and consequently often not transcribed in corpora of child-directed speech). In Dutch (but not in German), this ambiguity extends to the second singular present tense, where the *-t* suffix is dropped in instances of subject–main-verb inversion (e.g. in question formation), resulting in a verb form that matches the stem. Thus, the interrogative equivalent of the phrase *jij loopt* 'you walk' is *loop jij?* 'walk you?'. These forms are clearly different from the infinitive. However, they do not carry overt tense or tense-dependent morphology. They were therefore assumed not to reward the [+Tense] grammar. Note, however, that, in line with Legate and Yang's analysis, zero-suffixed forms involving a stem change were treated as rewarding the [+Tense] grammar.

Coding procedure

All corpora were analyzed in an automated fashion. That is, lists of verb forms with tense or tense-dependent morphology were drawn up, and the utterances in the input corpus where searched for these words. Utterances that contained forms with tense or tense-dependent morphology (e.g. *He goes, They went*, *He is going*) were counted as rewarding the [+Tense] grammar; utterances that only included verb forms with no tense or tense-dependent morphology (e.g. *we go*, *he can go*) were counted as not rewarding the [+Tense] grammar. Note that there is a slight complication in the case of English, Dutch and German in that some past tense forms (which, in line with Legate and Yang's analysis reward a [+Tense] grammar) are indistinguishable from past participles, which do not occur in tensed position, and are hence irrelevant to Legate and Yang's analysis. This ambiguity was dealt with by treating such forms as past tense forms in the absence of auxiliary HAVE (e.g. *I walked*), and past participles in the presence of auxiliary HAVE (e.g. *I have walked*).

Evaluating MOSAIC

MOSAIC was evaluated by training the model on the same five input files used to evaluate the VLM (i.e. Anne for English, Matthijs for Dutch, Leo for German, Tim for French and Juan for Spanish). No changes were made to the model or its parameters when simulating the different languages. Thus, the only difference between the simulations in each of the five languages was the use of a different input file. Further details regarding preparation of the input files can be found in Freudenthal *et al.* (2007).

Recall that the main aim of these simulations was to extend previous work with MOSAIC so that the model's predictions could be directly compared with those of the VLM. One way in which the present simulations extend previous work is by focusing on a fifth language (French), analyzed by Legate and Yang, but not yet simulated in MOSAIC. Another is by using a new method to simulate the English data. Thus, as noted earlier, because it is only possible to identify OI errors in third person singular contexts in English, previous simulations of English have focused exclusively on utterances with explicit third person singular subjects (e.g. *He go there*). This approach does not allow a direct comparison of the rates of OI errors in English and Dutch, German and Spanish, since the English rates are necessarily based only on utterances with subjects, whereas the Dutch, German and Spanish rates are based on utterances with and without subjects.

In order to solve this problem a new method was used to simulate the English data in the present study. This method did not involve making any changes to the model itself. It simply involved distinguishing all of the verb forms in the English input file on the basis of whether or not they occurred in a third person singular context. This was done by hand coding the English input file, and explicitly marking all finite and non-finite verbs that occurred in a third person singular context by adding the tag +3SG to the relevant verb form. Third person singular contexts included contexts with an explicit third person singular subject as well as contexts where an implicit third person singular subject could be inferred from the surrounding discourse. Note that, because of the way that MOSAIC represents words (as character strings), exposing the model to input coded in this way results in separate entries being created for the same word encountered in third person singular and non-third person singular contexts. Thus, as far as MOSAIC is concerned, *went* and *went*+3SG are different words that are represented in different primitive nodes. This means that it is possible to distinguish between infinitives learned from third person singular contexts (e.g. *go*-3SG learned from *That can go*-3SG *there*) and finite forms learned from non-third person singular contexts (e.g. *go* learned from *I go shopping*), and hence to identify subjectless third person singular OI errors in the

model's output (e.g. *go*-3SG *there*). Coding the English input thus allows a meaningful comparison of the rate of OI errors in simulations of English and the rate of OI errors in simulations of languages where OI errors can be readily identified in the absence of an explicit subject.[1]

The present simulations were run by repeatedly feeding input corpora through the model and creating an output file after each cycle through the input corpus in each of the five languages. This was done until the MLU of the model's output in each of the languages was greater than 2·0, at which point the output file with the MLU closest to 2·0 was selected for analysis. This output file was then searched for utterances that included a verb form, or, in the case of English, utterances that included a verb form learned from a third person singular context. The utterances identified in this way were then analyzed in the same way for all languages. That is, utterances were divided into (correct) finite and (incorrect) non-finite utterances (i.e. OI errors). Non-finite utterances were utterances that only contained verb forms matching a non-finite form, whereas finite utterances were utterances that contained at least one finite form. Note that this method of analysis is identical to the method used to analyze the child data.

Lexical analysis

The analysis of possible lexical effects in the data was carried out in the following manner. First, a sample of child speech was selected from each of the five children used to evaluate the VLM and MOSAIC. Then the proportion of incorrect infinitive versus correct finite uses of each verb was determined on a verb-by-verb basis. Although modals can be used as main verbs in Dutch and German, they were excluded from the analysis to avoid the possibility that their inclusion might artificially inflate the correlations. For English, the sample was also restricted to verbs that were used in a third person singular context. Where possible (in English, Dutch and German), the sample was selected from a period in which correct finite utterances and OI errors occurred at approximately equal rates. This was not possible for Spanish and French. For these languages a sample of speech at an MLU of approximately 2·0 was used (i.e. the same samples on which the rates of OI errors reported in Table 1 are based). Note that, if a relation does exist between the number of times particular verbs are

[1] Note that, in principle, a similar strategy could have been used to disambiguate infinitival and plural present tense forms in Dutch and German. However, in practice, the size of the German corpus and the absence of a %mor coding line in the Dutch transcripts made such an analysis prohibitively expensive. Distinguishing between infinitival and plural present tense forms will be considerably easier when %mor coding lines are available for all of the transcripts in CHILDES.

TABLE 1. *Proportion of OI errors in 6 English, 2 Dutch, 2 German, 2 French and 2 Spanish children's speech at an MLU of approximately 2·0 words*

	MLU	Proportion OI errors (Total number of utterances)
English		
Anne	2·16	0·87 (109)
Aran	2·01	0·85 (130)
Becky	2·17	0·97 (98)
Dominic	2·05	0·82 (95)
Gail	1·99	0·87 (90)
John	2·17	0·87 (31)
Dutch		
Matthijs	2·06	0·77 (347)
Peter	2·05	0·74 (290)
German		
Leo	2·08	0·58 (3,967)
Rah	1·94	0·58 (178)
French		
Anais	2·14	0·41 (203)
Tim	1·96	0·32 (250)
Spanish		
Juan	2·17	0·20 (305)
Lucía	2·31	0·05 (62)

encountered in non-finite as opposed to finite form in the input and the rate at which children produce those verbs as OI errors, it is most likely to be found at a point at which the child is producing roughly equal numbers of correct finites and OI errors. The fact that Spanish and French children do not produce many OI errors even at relatively low MLU points therefore works against finding lexical effects in these languages. Finding a significant correlation even in these unfavourable circumstances would thus provide particularly strong evidence for lexical effects in the data.

Once the child data had been analyzed, the input was searched for finite and infinitive forms of the verbs used by the relevant child, and the proportion of infinitive uses in the input was determined. Since the analysis of lexical effects was aimed at determining the plausibility of MOSAIC's account of OI errors being learned from compound constructions in the input, this analysis of the input only considered infinitives that occurred in compound constructions. Thus, instances where infinitives occurred as bare forms in the input were ignored. For finite verb forms, only instances where a verb was used as a main verb were counted. Thus, the verb form *has* counted as a finite form in the utterance *This dog has a very loud bark*. On the other hand, in the utterance *This dog has gone to sleep*, *has* is used as an auxiliary rather than a main verb, and this use of *has* was not

TABLE 2. *Proportion of clauses that reward the [+Tense] grammar in English, Dutch, German, French and Spanish input*

	Total number of clauses	Proportion of clauses rewarding the [+Tense] grammar
English	20,548	0·57
Dutch	8,176	0·49
German	18,413	0·62
French	14,169	0·67
Spanish	19,044	0·81

counted. Plural present tense forms, which match the infinitive in Dutch and German, were counted as finite forms.

Finally, the correlation between the proportion of OI errors for each verb in the child data and the proportion of occurrences as infinitives in compound finites in the input was computed across the verbs used by each child. The VLM predicts that there will be no relation between the rate of OI errors for each verb in the child's speech and the proportion of times that that verb occurs as an infinitive in compound structures in the input. Thus, the VLM predicts that the correlation will not deviate statistically from zero. A significant (positive) correlation on the other hand, implies lexical effects in the data, and provides support for MOSAIC's account of OI errors.

RESULTS

Measures of the proportion of OI errors in the fourteen children's speech at MLU $\approx$ 2·0 are presented in Table 1. These measures are consistent with Phillips' (1995) conclusion that rates of OI errors vary along a continuum from very high in English and Swedish to very low in Spanish, Italian and Hebrew, with Dutch, German and French falling somewhere in between. However, they also suggest that there is systematic cross-linguistic variation in the middle range of the distribution, with rates for French closer to Spanish than to English and rates for Dutch closer to English than to Spanish.

Measures of the proportion of clauses that reward the [+Tense] grammar in the five different languages are provided in Table 2. It is clear from a comparison of Tables 1 and 2 that, as predicted by the VLM, there is a strong negative correlation between the proportion of clauses that reward the [+Tense] grammar and the rate of OI errors across the five languages (ρ ($N=5$) $=-0{\cdot}90$, $p=0{\cdot}04$, two-tailed). If we focus on the results for English, French and Spanish, it is clear that the Spanish input rewards

the [+Tense] grammar more than the French input and the French input rewards the [+Tense] grammar more than the English input. These results replicate those of Legate & Yang (2007). Indeed the rates reported in Table 2 are remarkably similar to those reported in the earlier study (0·81 vs. 0·80 for Spanish, 0·67 vs. 0·70 for French, and 0·57 vs. 0·53 for English). They thus provide some validation for the coding procedures used in the present study, and confirm that the VLM correctly predicts the pattern of differences in the rate of OI errors across these three languages (i.e. English > French > Spanish).

However, if we expand the focus to include Dutch and German, the pattern of results becomes a little more complex. Thus, although rates of OI errors tend to be lower in both Dutch and German than they are in English, the German input rewards the [+Tense] grammar more than the English input (0·62 vs. 0·57), whereas the Dutch input rewards the [+Tense] grammar less than the English input (0·49 vs. 0·57). There is thus a discrepancy between the pattern of differences in the child data (i.e. English > Dutch > German > French > Spanish) and the pattern of differences predicted by the VLM (i.e. Dutch > English > German > French > Spanish), with the VLM incorrectly predicting higher rates of OI errors in Dutch than in English, when the English rates are actually considerably higher than the Dutch rates (between 0·82 and 0·97 for English versus 0·74 and 0·77 for Dutch).

These results suggest that the predictions of the VLM provide a relatively good fit to the cross-linguistic patterning of OI errors (particularly in the middle range of the distribution, where the model correctly predicts the relative ordering of Dutch > German > French). However, they also suggest that the VLM may have a particular problem explaining the very high level of OI errors in early child English. Thus, while the results for Spanish, French and English replicate those of Legate and Yang very closely, the results for Dutch and German suggest that the figure of 0·57 for English (0·53 in Legate and Yang) is not extreme enough to explain why OI errors are more common in English than in any of the other four languages.

One possible explanation of the VLM's problem in accounting for the English data is that, although English has a very impoverished morphological system for lexical verbs, there is actually quite a lot of tense and tense-dependent marking on copulas and auxiliaries. This, together with the fact that auxiliary structures are more common in English than in Dutch, may be the reason why there is actually more evidence for the [+Tense] grammar in English than in Dutch.

In order to investigate this possibility, separate measures were computed for the extent to which lexical verbs and copulas and auxiliary verbs rewarded the [+Tense] in English and Dutch. This analysis revealed that, in both languages, the rate for copulas and auxiliaries was much higher than

TABLE 3. *Proportion of OI errors in MOSAIC's output across five languages at an MLU of approximately 2·0 words*

	MLU	Proportion OI errors (Total number of utterances)
English	1·94	0·63 (150)
Dutch	1·95	0·65 (561)
German	1·96	0·49 (1,508)
French	1·95	0·32 (510)
Spanish	2·08	0·15 (1,514)

the rate for lexical verbs (0·70 versus 0·16 in English and 0·66 versus 0·27 in Dutch). This pattern combined with the higher proportion of copulas and auxiliaries in English than in Dutch (0·80 versus 0·56) to result in an overall figure that was closer to the figure for lexical verbs in Dutch than it was in English, and hence lower for Dutch than it was for English. The implication is that the VLM's failure to predict the correct pattern for Dutch and English reflects the fact that its learning mechanism does not differentiate between evidence for the [+Tense] grammar derived from lexical verbs (of which there is less in English than in Dutch) and evidence for the [+Tense] grammar derived from copulas and auxiliaries (of which there is considerably more in English than in Dutch).

Measures of the proportion of OI errors in the five MOSAIC simulations at MLU $\approx$ 2·0 are presented in Table 3. It is clear from a comparison of Tables 1 and 3 that there is a strong positive correlation between the rate of OI errors in MOSAIC's output and the rate of OI errors in children's speech across the five languages (ρ ($N=5$) = 0·90, $p=0{\cdot}04$, two-tailed). If we focus on the pattern of differences between Spanish, French, German and Dutch, it is clear that MOSAIC simulates this pattern of cross-linguistic variation remarkably well. Thus, as reported in previous work, MOSAIC simulates the low level of OI errors in Spanish, the much higher rates of OI errors in German and Dutch and the more subtle quantitative difference in the rate of OI errors between Dutch and German. MOSAIC also simulates the relatively low level of OI errors in French. However, MOSAIC appears to have a similar problem to the VLM in accounting for the very high level of OI errors in early child English. Thus, the model actually produces more OI errors in Dutch than in English and hence fails to differentiate appropriately between these two languages.

This pattern of results was investigated further by running an analysis of the proportion of non-finite verbs in utterance-final position in the input in each of the five languages. Previous work with MOSAIC has used this kind of analysis to demonstrate that MOSAIC's utterance-final bias in

learning is the key factor in allowing the model to simulate the cross-linguistic data. The results showed that there was a close relationship between the level of OI errors produced by the model and the proportion of non-finite verb forms in utterance-final position across the five languages (0·15 and 0·21 for Spanish; 0·32 and 0·40 for French; 0·49 and 0·69 for German, 0·63 and 0·78 for English; and 0·65 and 0·87 for Dutch). Thus, although the level of OI errors was always lower than the proportion of utterance-final non-finites (because it is based on output of MLU $\approx$ 2·0 whereas the input analysis is effectively based on one-word strings), the measures show the same rank order across languages, with higher rates of utterance-final non-finites in Dutch than in English (ρ ($N = 5$) = 1·00, $p < 0{\cdot}01$, two-tailed).

These results confirm that MOSAIC's success in simulating differences in the rate of OI errors across Spanish, French, German and Dutch can be explained in terms of the model's utterance-final bias in learning. On the other hand, they also suggest that learning from the right edge of the utterance is unlikely to result in the very high rate of OI errors found in early child English, and that some additional mechanism may be required to explain this phenomenon.

To summarize, both MOSAIC and the VLM appear to predict a similar pattern of differences in the rate of OI errors across languages, with both models providing a good fit to the data on Spanish, French, German and Dutch, and both models unable to explain the very high levels of OI errors in early child English. However, one area of the data about which MOSAIC and the VLM make very different predictions is the relation between the errors made by children during the OI stage and the distributional properties of the input to which they have been exposed. Thus, MOSAIC predicts input-driven lexical effects on the distribution of OI errors in children's speech, whereas the VLM predicts that OI errors will occur at more or less the same rate across different verbs.

In order to distinguish between the two models, these predictions were tested by computing the rate of OI errors for each of the verbs produced by the child and correlating these rates with the rate at which those same verbs occurred as non-finite forms in compound structures versus finite forms in simple finite structures in the input language. The results of this analysis are reported in Table 4. Two correlations are listed for each language: one computed over all of the verbs used by the child, and one computed over a restricted set of verbs (i.e. all of those verbs that were used by the child at least three times). As can be seen from Table 4, both sets of correlations provide strong evidence for lexical effects in the data. Thus nine of the ten correlations are statistically significant at $p < 0{\cdot}05$, two-tailed, and even the remaining correlation (on the restricted set of verbs in Spanish) is marginally significant ($p = 0{\cdot}06$, two-tailed).

TABLE 4. *Correlations between the rate of OI errors for each verb used by the child and the proportion of occurrences of that verb as an infinitive in a compound structure in the input. Numbers of contributing verbs are listed in brackets* (+ = p < 0·10, * = p < 0·05, ** = p < 0·01)

	Full set (all verbs)	Restricted set (verbs used 3 or more times)
English	0·35* (43)	0·55* (15)
Dutch	0·71** (102)	0·83** (59)
German	0·48** (143)	0·68** (69)
French	0·45** (75)	0·57** (37)
Spanish	0·40** (69)	0·29+ (43)

These results suggest that the relative frequency with which children produce OI errors with particular verbs is related to the relative frequency with which those verbs occur as infinitives in compound finite constructions in the language to which the children are exposed. They thus provide strong support for the idea that OI errors are learned from compound finite constructions in the input.

DISCUSSION

The aim of the present study was to compare two recent accounts of the OI stage (one generativist and one constructivist), both of which have been explicitly designed to address the graded nature of the OI phenomenon. According to the VLM, cross-linguistic variation in rates of OI errors reflects differences in the speed with which children establish that they are acquiring a [+Tense] grammar as a function of differences in the amount of morphological evidence for tense marking in the input. According to MOSAIC, cross-linguistic variation in rates of OI errors reflects the interaction between an utterance-final bias in learning and the distributional patterning of finite and non-finite verb forms in the input language.

In a first set of analyses, we assessed the extent to which each of these accounts could explain the level of OI errors across five different languages (English, Dutch, German, French and Spanish). In a second set of analyses we attempted to differentiate between the two accounts by testing their predictions about the relation between children's OI errors and the distribution of infinitival verb forms in the input language.

If we focus on the results of the first set of analyses, it is clear that both the VLM and MOSAIC do a relatively good job of predicting the cross-linguistic data. Thus, both models provide a good approximation to the

pattern of differences between Spanish, French, German and Dutch at MLU ≈ 2·0. In the case of the VLM, the key factor is the proportion of bare stem forms in the input language. This factor is critical to the VLM in differentiating between a non-OI language (Spanish) and two languages with relatively high rates of OI errors (German and Dutch). In the case of MOSAIC, the key factor is the proportion of utterance-final verb forms that are non-finite. This factor interacts with MOSAIC's utterance-final bias in learning to result in very different rates of OI errors in languages in which compound finites occur at roughly equivalent rates (Spanish, German and Dutch).

A particularly interesting feature of these results is the extent to which both MOSAIC and the VLM are able to differentiate between languages in the middle range of the distribution. Thus, both models are not only able to explain the difference between OI and non-OI languages, but also to simulate differences between languages with relatively similar rates of OI errors (French, German and Dutch). These results provide strong support for the view that there is systematic quantitative variation in the rate at which OI errors occur across different languages (e.g. Phillips, 1995). They also show that there are two important correlates of this variation in the distributional properties of different tense-marking languages: the proportion of bare stems in the input and the proportion of utterance-final verbs that are non-finite. They thus illustrate how building and testing models that are sufficiently well-specified to make quantitative predictions about child language data can provide us with important insights about the relation between children's early language and the distributional properties of the language to which they are exposed.

Of course, because both models fit the cross-linguistic data relatively well, this first set of analyses fails to distinguish very clearly between MOSAIC and the VLM. However, if we focus on the results of the second set of analyses, it is clear that there are important lexical effects on the distribution of OI errors in children's speech that are difficult for the VLM to explain. Thus in all five of the languages investigated there is a significant correlation between the extent to which particular verbs occur as OI errors in the child's speech and the extent to which those same verbs occur as infinitives in compound structures in the input. These correlations are obviously consistent with the idea, implemented in MOSAIC, that OI errors are learned from compound finite structures in the input. However, they are not predicted by the VLM since, according to the VLM, OI errors are not learned from the input, but simply reflect the child's continued use of the incorrect [−Tense] grammar. Of course, it might be tempting to think that because the VLM is sensitive to quantitative variation in the input, it is also sensitive to the distribution of particular lexical items in the input. In fact, however, this is not the case. Thus, an important feature of

the VLM is that its learning mechanism focuses not on the distribution of non-finite forms in the input, but on the extent to which finite forms in the input reward the [+Tense] grammar. This feature can be seen as both a strength and a weakness of the model. On the positive side, it allows the model to predict high rates of bare infinitives in Dutch and German children's speech as a result of lack of evidence for tense marking in the shape of finite stem forms in Dutch and German input. On the negative side, however, it means that the model has no mechanism for explaining the relation between the distributional patterning of OI errors in the child's speech and the distributional patterning of infinitival forms in the input. The lexical effects reported in the current paper thus not only provide support for the idea that OI errors are learned from compound structures in the input, but also count against the idea that such errors reflect the probabilistic use of a [−Tense] grammar.

A final interesting feature of the present results is the extent to which both MOSAIC and the VLM struggle to explain the very high level of OI errors in early child English. In the case of the VLM, this problem would seem to be another symptom of the VLM's lack of sensitivity to lexical patterning in the data. Thus, the reason why the VLM is unable to explain the very high level of OI errors in early child English is that it does not discriminate between evidence for the [+Tense] grammar in the form of inflected copula and auxiliary verb forms and evidence for the [+Tense] grammar in the form of inflected main verbs. Although English lexical verbs provide very little evidence for the [+Tense] grammar, and considerably less evidence than Dutch lexical verbs, English copulas and auxiliaries actually provide a great deal of evidence for the [+Tense] grammar. This, together with the fact that copulas and auxiliaries make up a much greater proportion of the English child's input, means that the VLM actually predicts lower levels of OI errors in English than in Dutch. A more lexically oriented input-driven account could probably deal with this problem relatively easily by simply distinguishing between what the child is learning about copulas and auxiliaries and what the child is learning about lexical verbs, and predicting high levels of OI errors on lexical verbs and lower levels of OI errors on copulas and auxiliaries. Interestingly, this is exactly the pattern of results reported in two recent lexically oriented analyses of early child English (Wilson, 2003; Pine, Conti-Ramsden, Joseph, Lieven & Serratrice, 2008). The VLM, however, is not a lexically oriented account since it assumes that the child is not learning how to inflect particular lexical forms but learning to reject a particular grammar or parameter setting. The VLM is thus unable to explain why English-speaking children entertain the [−Tense] grammar for so long when there is so much evidence for the [+Tense] grammar in the form of tensed and agreeing copulas and auxiliaries in the input.

MOSAIC's problems simulating the high level of OI errors in early child English would seem to reflect the fact that the proportion of utterance-final verbs that are non-finite in English is simply not high enough to explain the almost exclusive production of OI errors during the early stages. One obvious reason why this might be the case is that English differs from the other four languages in that for lexical verbs the infinitive is indistinguishable from the bare stem. Since the only present tense form that is not a bare stem in English is the third person singular, this means that a much higher proportion of lexical verb forms in the input are either infinitives or forms that are indistinguishable from the infinitive. This fact is likely to slow down the process of paradigm building in English and result in default effects where the child produces a bare stem/infinitive in the absence of knowledge of the relevant third person singular or past tense form. Since MOSAIC is insensitive to the morphological structure of the verbs that it encodes, it is clearly unable to simulate this kind of default effect, and hence predicts fewer OI errors than children actually produce.

Of course, if it is necessary to supplement MOSAIC's account of OI errors with some kind of paradigm-building account in order to explain the data on early child English, one might wonder whether it is possible to simply replace MOSAIC with a paradigm-building account (e.g. MacWhinney, 1978; Pinker, 1984). We would argue that there are at least three reasons for seeing MOSAIC and paradigm building as complementary rather than competing accounts. First, although a paradigm-building account provides a very natural way of explaining OI errors in English, it fares less well as an account of OI errors in other languages. This is because, in languages other than English, infinitives tend to carry morphology that distinguishes them from the most frequent and/or least marked form in the present tense paradigm. It is therefore difficult to see why children learning these languages would default to the infinitive rather than to some other less marked form.[2] Indeed, there is some evidence that children learning languages other than English produce both OI errors and non-infinitive default errors during the early stages. For example, Aguado-Orea (2004) reports that Juan (the Spanish child whose data are simulated in the present paper) produces errors that involve defaulting to the third person singular (the most common and the least marked form in the present tense paradigm) alongside OI errors during the early stages. This result is consistent with the idea that learning OI errors from compound finites and defaulting to the most common and/or the least marked form in the

[2] This is particularly true for Spanish and French, where forms that match the infinitive are less frequent than less marked forms. However, it is also true for Dutch and German where forms that match the infinitive are more frequent than bare stems (40% vs. 30% in Dutch and 33% vs. 22% in German), but do not dominate the paradigm in the way that they do in English (where they account for 61% of all verb forms).

language are both processes that occur early in development. These processes happen to produce the same effect in early child English: OI errors/bare stems in the child's output. However, this is not necessarily the case in languages other than English.

Second, a paradigm-building account of OI errors would seem to predict that defaulting to the infinitive would reflect some more general confusion on the part of the child between finite and non-finite forms. In fact, however, a key feature of the OI stage is that, although children produce infinitives in root clauses, their use of these forms is highly sensitive to differences in the distributional properties of finite and non-finite forms in the input language (Wexler, 1994). For example, in French, where finite forms precede and non-finite forms follow the negative particle *pas*, children correctly place finite forms before the negative participle and the infinitives in OI errors after the negative particle (Pierce, 1992; Joseph & Pine, 2002). Similarly, in Dutch and German, where finite verbs precede and infinitives follow their complements, children correctly place finite forms before their complements and the infinitives in OI errors after their complements (Jordens, 1990; Poeppel & Wexler, 1993). This feature of the OI stage, which is often taken by generativists as evidence for very early parameter setting (Wexler, 1998), is difficult to explain in paradigm-building terms. However, it fits naturally with the idea that OI errors are truncated compound finites, which retain the properties of the compound structures from which they have been learned.

Finally, a paradigm-building account offers no obvious explanation of the lexical effects found in the present study. These effects, which can be seen in all five of the languages under investigation, suggest that those verbs that occur as OI errors in children's speech also tend to occur as infinitives in adult compound finites. They thus provide strong support for the idea that OI errors are learned from compound structures in the input. Interestingly, this idea can also explain the fact that Dutch and German OI errors are more likely to have a modal reading than English OI errors, and that Dutch and German children are more likely to make OI errors with eventive than stative verbs. (Freudenthal *et al.*, 2009). For all of these reasons, we would argue that MOSAIC and paradigm-building accounts are best seen as providing complementary rather than competing explanations of the pattern of OI errors in children's speech.

To conclude, in the present study we have used corpus-based and computational modelling techniques to test two alternative accounts (MOSAIC and the VLM) of the cross-linguistic patterning of OI errors in five different languages. Our results suggest that although both of these accounts fit the cross-linguistic patterning of OI errors reasonably well, both have similar problems in explaining the very high rate of OI errors in early child English. In the case of the VLM, this difficulty seems to reflect a problem

with the level of abstraction at which the learning mechanism operates. In the case of MOSAIC, this difficulty probably reflects the model's lack of sensitivity to the morphological structure of words. However, of the two models, only MOSAIC is able to explain why verbs that occur more frequently as infinitives than as finite verb forms in the input also occur more frequently as OI errors than as correct finite verb forms in the children's output. The implication is that the OI phenomenon reflects the learning of OI errors from compound structures in the input rather than the probabilistic use of a grammar that does not mark tense. The extent to which these results generalize to languages with different word order patterns and different verb paradigms is, of course, an empirical issue. Ongoing work with MOSAIC seeks to answer this question in two ways: first by using the current version of MOSAIC to simulate data from a wider range of languages; and second by developing a new version of the model that learns from syllabified input and can hence be used to simulate omission errors below the level of the word. This new version of the model is currently being used to simulate developmental data in K'iche' Mayan, an agglutinative language of Central America, in which children produce bare verb stems that do not occur as isolated words in the input language (Pye, 1983; Pye, Pfeiler, De León, Brown & Mateo, 2007). Preliminary results suggest that such errors can be simulated by a mechanism that builds verb forms syllable by syllable from the right edge of the word.

REFERENCES

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

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

Bol, G. W. (1996). Optional subjects in Dutch child language. In C. Koster & F. Wijnen (eds), *Proceedings of the Groningen Assembly on Language Acquisition*, 125–35.

Demuth, K. & Tremblay, A. (2008). Prosodically-conditioned variability in children's production of French determiners. *Journal of Child Language* **35**, 99–127.

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**, 311–41.

Freudenthal, D., Pine, J. M. & Gobet, F. (2005). Simulating Optional Infinitive errors in child speech through the omission of sentence-internal elements. In B. G. Bara, L. Barsalou & M. Buchiarelli (eds), *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, 708–713. Mahwah, NJ: Erlbaum.

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

Freudenthal, D., Pine, J. M. & Gobet, F. (2009). Simulating the referential properties of Dutch, German and English root infinitives in MOSAIC. *Language Learning and Development* **5**, 1–29.

Freudenthal, D., Pine, J. M., Jones, G. & Gobet, F. (submitted). Simulating the cross-linguistic pattern of finiteness marking in children's declaratives and Wh- questions in terms of edge effects in utterance learning.
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*, 91–128. Philadelphia: John Benjamins.
Jordens, P. (1990). The acquisition of verb placement in Dutch and German. *Linguistics* **28**, 1407–448.
Joseph, K. L. & Pine, J. M. (2002). Does error-free use of French negation constitute evidence for Very Early Parameter Setting? *Journal of Child Language* **29**, 71–86.
Legate, J. A. & Yang, C. (2007). Morphosyntactic learning and the develoment of tense. *Language Acquisition* **14**, 315–44.
MacWhinney, B. (1978). The acquisition of morphophonology. *Monographs of the Society for Research in Child Development* **43**, 1–123.
MacWhinney, B. (2000). *The CHILDES project: Tools for analysing talk*, 3rd edn. Mahwah, NJ: Erlbaum.
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**, 325–82.
Pierce, A. (1992). *Language acquisition and syntactic theory: A comparative analysis of French and English*. Kluwer: Dordrecht.
Pine, J. M., Conti-Ramsden, G., Joseph, K., Lieven, E. V. M. & Serratrice, L. (2008). Tense over time: Testing the Agreement/Tense Omission Model as an account of the pattern of tense-marking provision in early child English. *Journal of Child Language* **35**, 55–75.
Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.
Poeppel, D. & Wexler, K. (1993). The full competence hypothesis of clause structure in early German. *Language* **69**, 1–33.
Pye, C. (1983). Mayan telegraphese: Intonational determinants of inflectional development in Quiche Mayan. *Language* **59**, 583–604.
Pye, C., Pfeiler, B., De León, L., Brown, P. & Mateo, P. (2007). Roots or edges? Explaining variation in children's early verb forms across five Mayan languages. In B. Pfeiler (ed.), *Learning indigenous languages: Child language acquisition in Mesoamerica*, 15–46. Berlin: Mouton de Gruyter.
Rizzi, L. (1994). Some notes on linguistic theory and language development: The case of root infinitives. *Language Acquisition* **3**, 371–93.
Szagun, G. (2001). Learning different regularities: The acquisition of noun plurals by German-speaking children. *First Language* **21**, 109–141.
Theakston, A. L., Lieven, E. V. M., Pine, J. M. & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language* **28**, 127–52.
Wexler, K. (1994). Optional infinitives, head movement and the economy of derivation in child grammar. In N. Hornstein & D. Lightfoot (eds), *Verb movement*, 305–365. 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.
Wilson, S. (2003). Lexically specific constructions in the acquisition of inflection in English. *Journal of Child Language* **30**, 75–115.
Yang, C. (2002). *Knowledge and learning in natural language*. New York: Oxford University Press.
Yang, C. (2004). Universal Grammar, statistics or both? *Trends in Cognitive Sciences* **8**, 451–56.