Modelling Children's Case-Marking Errors with MOSAIC

Steve Croker (sfc@psychology.nottingham.ac.uk)
Julian M. Pine (jp@psychology.nottingham.ac.uk)
Fernand Gobet (frg@psychology.nottingham.ac.uk)
ESRC Centre for Research in Development, Instruction and Training
School of Psychology, University of Nottingham,
University Park, Nottingham, NG7 2RD

Abstract

We present a computational model of early grammatical development which simulates case-marking errors in children's early multi-word speech as a function of the interaction between a performance-limited distributional analyser and the statistical properties of the input. The model is presented with a corpus of maternal speech from which it constructs a network consisting of nodes which represent words or sequences of words present in the input. It is sensitive to the distributional properties of items occurring in the input and is able to create 'generative' links between words which occur frequently in similar contexts, building pseudo-categories. The only information received by the model is that present in the input corpus. After training, the model is able to produce child-like utterances, including case-marking errors, of which a proportion are rote-learned, but the majority are not present in the maternal corpus. The latter are generated by traversing the generative links formed between items in the network.

Case-Marking Errors

Children in the early stages of language development are known to make case-marking errors. These errors are utterances in which a nominative pronoun (e.g. 'he') has been replaced with a non-nominative pronoun, such as the accusative (e.g. 'him'), resulting in utterances such as 'him does it' instead of 'he does it' and 'her get it' instead of 'she gets it'.

There are a number of possible explanations for this phenomenon. According to Schütze and Wexler's (1996) ATOM (Agreement/Tense Omission Model) (see also Wexler, 1998, for an overview of this model), pronoun case-marking errors occur because the child produces the accusative form of the pronoun as a default when the abstract features of agreement (which are necessary for correct case assignment) are absent from the child's underlying representation of the sentence.

The ATOM predicts that once children have the relevant nominative and accusative forms in their productive lexical inventories, they will produce nominative subjects when agreement is present in the underlying representation of the sentence, and non-nominative subjects when agreement is absent. Since agreement can be present but hidden when tense is absent, nominative subjects are predicted to occur with both agreeing forms (e.g. 'he goes' and 'she's singing'), and non-agreeing forms (e.g. 'he go' and 'she singing'). In the latter, agreement is 'hidden' in the surface structure, but present in the underlying structure. However, since agreement assigns nominative case, non-nominative subjects are predicted to occur only with non-agreeing forms (e.g. 'him go' and 'her singing', and not 'him goes' and 'her is singing'). Since the ATOM makes no predictions about how often case and agreement will surface in children's speech, and allows for the occurrence of nominative subjects with agreement, and nominative and non-nominative subjects without agreement, the only real prediction that it makes is that one will never find children who make certain kinds of errors (i.e. 'him goes' and 'her is' type errors), or, more realistically, that one will never find children who make such errors at rates higher than would be consistent with the notion that they can be disregarded as noise (Schütze, 1999).

Rispoli (1999) shows that, in fact, accusative pronouns do occur with agreeing verbs. Another related phenomenon is presented by Rispoli (1998): overextension of 'her' for 'she' occurs with greater frequency than overextension of 'him' for 'he'. Rispoli explains this in terms of a 'double-cell' effect, whereby 'her' fills the 'slots' for both accusative and genitive pronouns. 'Her' is used in the same contexts as both 'him' and 'his', which could lead to 'her' appearing in more contexts than 'him'.

The model we present here is an attempt to simulate two basic effects found in child case-marking errors. First, a greater proportion of case-marking errors occur with feminine subjects than masculine subjects. Second, not only do agreeing verbs occur with case-marking errors, they occur more often in feminine contexts. Nina, a child presented by Schütze (1997), uses 'her' proportionally more often than 'him', both with and without an agreeing verb. However, she also produces 'she' less often than 'he'. As a result, the ratio of 'her' for 'she' is greater than that of 'him' for 'he' and any analysis in which masculine and feminine forms are analysed together will show a lower rate of

case-marking errors than an analysis of just feminine forms.

A Distributional Account

The ATOM derives much of its power from the abstractness of the categories with which its proponents are prepared to credit the language-learning child — and hence to analyse the multi-word speech data. Thus, analysing the child's correct performance in terms of abstract features such as tense and agreement, as opposed to the lexical items which instantiate these features, means that any correct use of third singular present or past tense verb forms, whether these forms are copulas, auxiliaries or lexical verbs, can be used to support the notion that knowledge of tense and agreement is available to the child from the outset. On the other hand, analysing the child's incorrect performance in terms of tense optionality ignores the possibility that, in the early stages at least, many of the verbs used by children in untensed form may never occur in tensed form in their speech, or at least only as non-finite verb forms in combination with tensed auxiliaries (e.g. 'she will go').

An alternative explanation for the phenomena described by Schütze and Wexler is that the formation of syntactic relationships in children's speech can be explained in terms of the input the child receives from external sources, in particular parental input. In the account we propose here, the child's use of lexical forms is a result of the distribution of these forms in the input. Most of the predicted speech patterns have models in the speech of adults which act as the linguistic input to the child. For example, a child may produce the utterance 'her go' which is accounted for in the ATOM as the use of an untensed form in a position where tense is required. Obviously, a child should not hear her mother saying 'her go'. However, she will hear utterances such as 'did you see her go?'. Likewise, a child may hear 'where did she go?' and 'she goes out'. Thus, all three of the combinations predicted to occur by the ATOM can be rote-learned. Our main concern is to show the ability of a simple distributional analyser to produce utterances with errors of the same types as those made by children (see Croker, Pine & Gobet, 2000 for a preliminary analysis).

MOSAIC

MOSAIC (Model Of Syntax Acquisition In Children; Gobet & Pine, 1997) is a computational model based on the CHREST architecture (Gobet 1993, 1998; De Groot & Gobet, 1996). CHREST is, in turn, a member of the EPAM family (Feigenbaum & Simon, 1984). Variants of CHREST have been used to model a number of areas of human cognition including the acquisition of multiple representations in physics (Lane, Cheng & Gobet, 1999) and the acquisition of vocabulary (Jones, Gobet & Pine, 2000). See also Gobet et al. (in press) for an overview.

Network Formation

Knowledge is modelled in CHREST as a discrimination network, which is a hierarchically structured network consisting of nodes and vertical links between layers of nodes. Each node has an 'image', which contains the information available at this node (in MOSAIC, it consists of information regarding the links traversed to arrive at that node). Nodes and links each consist of one or more words.

When an utterance is presented, each word in the utterance is considered in turn, which allows the utterance to be sorted to a given node. If the word currently considered has not previously been seen by the model, the process of discrimination is used to create a new node corresponding to that word. The new node is created at the first layer of the network, just below the root node. This first layer may be seen as the layer where the 'primitives' of the network (i.e., the individual words that have been seen by the model) are learned and stored.

In cases where nodes only consist of one word, the image of the node matches the test link (described below) immediately above it, as that is the only link to have been traversed. However, at deeper levels, the image will contain more information relating to the sequence of tests. As noted above, at their first presentation, all words are encoded as primitives at the first layer of the network; a particular word must be 'seen' again in order for it to occupy a second location in the network. Subsequent words in an utterance are represented as nodes below the primitive, as long as they are already encoded as primitives themselves. Test links above nodes refer to the 'test' (one word or a

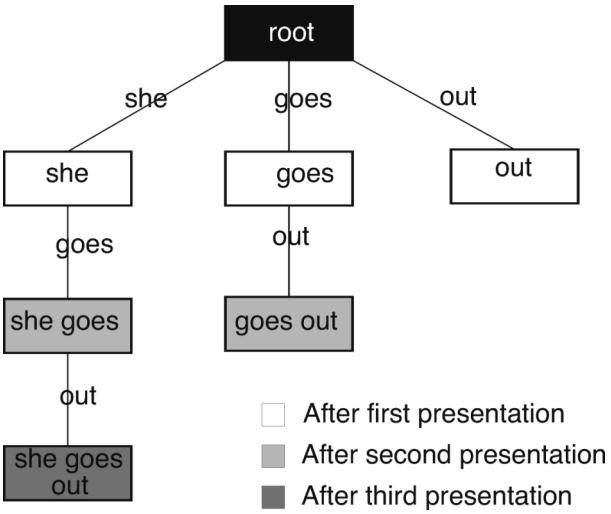

Figure 1: Network formed after the utterance 'she goes out' is presented 3 times to the model.

sequence of words) that has to be passed to travel down that link to the node, and are represented as the final element in the image of that node. These are created during discrimination, at the same time as a new node.

Figure 1 shows a small network created by presenting the utterance 'she goes out' to the model 3 times. On the first presentation, the primitives (white nodes) are created. When the model sees the utterance again, the network can be extended as the primitives have already been learnt (light grey nodes). The dark grey level 3 'she goes out' node is created on the third presentation.

When an utterance starts with a word already seen by the model, the *image* of the matching node is compared to the utterance. The utterance is then compared at the next level down to see if the second word of the utterance is already in the network below the primitive. The network is followed down as far as possible until one of two possibilities occurs: 1) The entire utterance is already accessible by traversing the network; 2) A point is reached where the utterance can not be traced down the network any further. In this case, discrimination takes place and a new node is created.

In contrast to earlier versions of the model (Croker, Pine & Gobet, 2000; Jones, Gobet & Pine, 2000), MOSAIC learns both from left to right and right to left. We feel this provides a more plausible account of learning as it allows sensitivity to the context of a word in terms of both the preceding and succeeding items (see Figure 2).

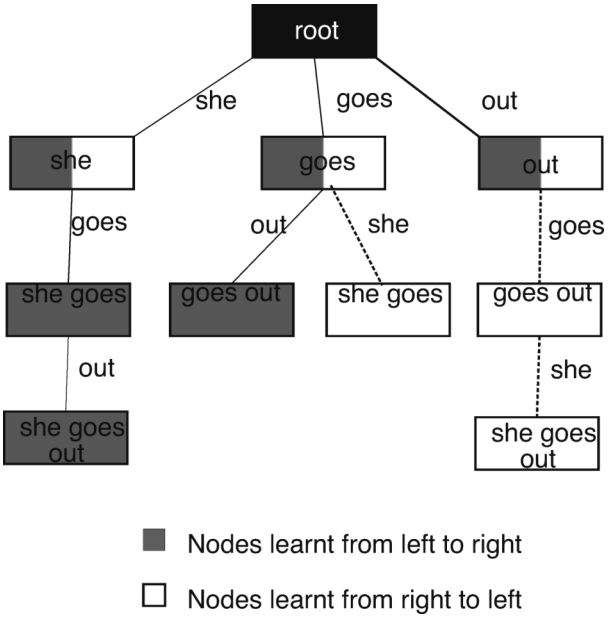

Figure 2: Right to left learning.

Generative Links

As well as learning utterances by rote, MOSAIC is able to generate novel utterances using generative links, an important feature of the model. Generative links are 'lateral' links between nodes which have contextual similarities. If two words occur frequently in similar contexts, then a generative link can be made between these items. These two nodes do not have be on the same level – a level 2 node can be linked to a level 3 node, for instance. The similarity measure is the degree of overlap between items that precede and succeed any two nodes. This is calculated by taking all the children of any two nodes and assessing whether the proportion of children shared by both nodes exceeds a certain threshold with respect to the total number of child nodes. For the purposes of this study, that parameter was set to 4% in each direction.

This, again, is in contrast to earlier versions of MOSAIC in which an absolute value was used in creating generative links (e.g. 15 succeeding items in common). It is our hope that this change, coupled with bidirectional learning, will enable the model to capture more subtle effects found in children's speech.

Production of Utterances

Once a network has been created, it can be used to produce utterances in two ways: by recognition and by generation. Utterances produced by recognition are essentially rote-learned (i.e. they are utterances or portions of utterances presented to the model in the

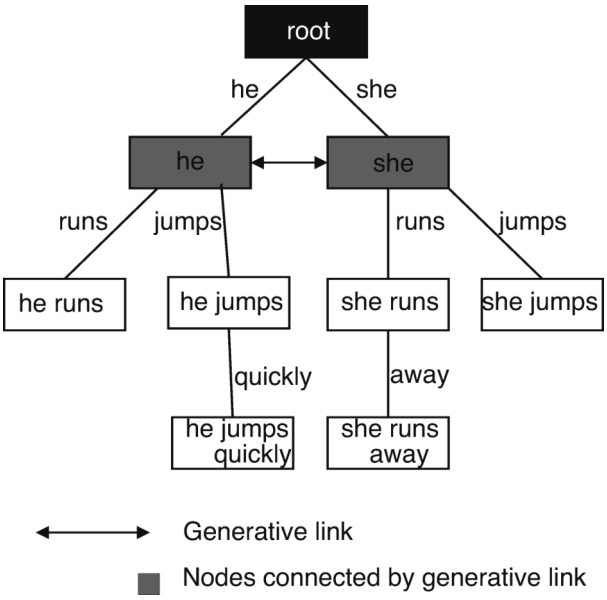

Figure 3: Generative link formation: 'he' and 'she' are linked by virtue of possessing child nodes in common. (Nodes learnt from right to left have been omitted in order to preserve clarity.)

input corpus). These are produced by starting at each node in turn, and following the left-to-right test links down the network. For example, from the fragment of a network shown in Figure 3, utterances such as 'she runs away' and 'he jumps' could be produced by recognition. Production by generation utilises the generative links to create utterances not seen in the input. This occurs in a similar way to production by recognition, the difference being that lateral generative links can be traversed as well as vertical test links, although only one generative link can be followed per generated utterance – this is simply to limit the number of generated utterances produced by the model. Thus, utterances such as 'he runs away' and 'she jumps quickly' could be produced by generation.

Methods

In this paper, we present data obtained from MOSAIC, trained on maternal input to one child between the ages of 1;10 and 2;9, which was taken from the Manchester corpus (Theakston, Lieven, Pine & Rowland, 2000) of the CHILDES database (MacWhinney & Snow, 1990). This corpus consists of transcripts of audio recordings made twice every three weeks for a period of 12 months. There are two half-hour recordings for each session, one made during free play and the other made during structured play. The model was trained on 15,000 utterances of maternal speech. We also present data from three children from this corpus, Anne, Becky and Gail.

There are two important points regarding data produced by MOSAIC. First, a word used as a verb is often used in other syntactic categories. An analysis was made of the frequencies with which words were used as verbs by the mother. A word was classified as a verb for the purpose of this research if it occurred as a verb in 90% or more of its instances in the mother's speech corpus. Second, the data used in analysing the performance of the model consists of types, not tokens. Much of the research in children's speech is based on analysis using tokens as the entire corpus is considered. MOSAIC, however, does not produce multiple instances of utterances in the same way that a child does. The model produces all the utterances it is capable of producing, whereas a child produces speech in response to the context in which the child is situated. We analysed only utterances which started with the pronouns 'he', 'him', 'she' and 'her' which contained a verb of which that pronoun was the subject. The restriction to third-person-singular forms is necessary as this is the only case in which agreement can be reliably distinguished in English. The same analysis was made of the utterances produced by the children.

Comparison With Human Data

Case-marking errors with an agreeing main verb are predicted not to occur within the ATOM. The literature in the field (see Rispoli, 1998) provides evidence of these error types occurring in children's speech. As we shall see, MOSAIC can also produce these errors. Although the model embodies a very simple learning mechanism, it captures aspects of the data which, at best, the ATOM cannot explain and which at worst count directly against it.

Anne, Becky and Gail (Tables 1 - 3) all produce a number of case-marking errors with both masculine and feminine subjects. It is immediately apparent that the rate of overextension of 'her' for 'she' is greater than that of 'him' for 'he'. What these data also show is that there is no consistent error rate across children. As a result, we can not model 'typical' performance, as there is no typical error rate for children. The 'him' for

Table 1: Case-marking errors (Anne)

	Subject	
Case	He	She
Nominative	263	36
Accusative	3	7
% Accusative	1.13%	16.28%

Table 2: Case-marking errors (Becky)

	Subject	
Case	He	She
Nominative	398	95
Accusative	5	13
% Accusative	1.24%	12.04%

Table 3: Case-marking errors (Gail)

	Subject	
Case	He	She
Nominative	176	17
Accusative	10	19
% Accusative	5.38%	52.78%

Table 4: Case-marking errors (MOSAIC)

	Subject	
Case	He	She
Nominative	783	631
Accusative	34	52
% Accusative	4.16%	7.61%

$\chi^2 = 8.201$, $p=0.004$

Table 5: C/M errors with agreeing verbs (Anne)

	Subject	
Case	He	She
Nominative	183	17
Accusative	1	4
% Accusative	0.54%	19.05%

Table 6: C/M errors with agreeing verbs (Becky)

	Subject	
Case	He	She
Nominative	296	62
Accusative	3	13
% Accusative	1.00%	17.33%

Table 7: C/M errors with agreeing verbs (Gail)

	Subject	
Case	He	She
Nominative	124	14
Accusative	4	9
% Accusative	3.13%	39.13%

Table 8: C/M errors with agreeing verbs (MOSAIC)

	Subject	
Case	He	She
Nominative	523	409
Accusative	21	30
% Accusative	3.86%	6.83%

$\chi^2 = 4.367$, p=0.037

'he' error rate varies from 1.13% (Anne) to 5.38% (Gail) and the 'her' for 'she' error rate varies from 12.4% (Becky) to 52.78% (Gail). The results from MOSAIC (Table 4) are consistent with these findings – the masculine and feminine error rates are significantly different. In addition, it can be observed that case-marking errors often occur with verbs that carry agreement, in direct opposition to the predictions of the ATOM. Tables 5-8 show the error rates for utterances which contain an agreement-inflected verb-form. Once more, it is apparent that 'her' for 'she' overextensions outnumber 'him' for 'he' overextensions. Again, we make no attempt to model the absolute frequency of such errors, as they are variable across children (Tables 5-7). All that we can claim is that there is a 'gender bias' and that MOSAIC (Table 8) can capture this effect.

Discussion

This study utilises a computational model to present a distributional account of case-marking errors in children's speech. The output of the model was compared to phenomena found in child data. The results show that, after training, MOSAIC was able to produce case-marking errors with both masculine and feminine subjects, and with verb forms in which agreement can be present or absent. Some of these errors were produced by virtue of being present in the input corpus. For example 'did you see her sit?' gives the child the potential to produce 'her sit'. Others were produced by traversing generative links between lexical items linked by virtue of similarity. The reason that MOSAIC was able to reproduce the differences between masculine and feminine forms is due to the distributional difference between 'her' and 'him'. This difference can be explained in terms of a 'double-cell' effect in which 'her' is the feminine equivalent not only of 'him', but also of 'his', and therefore appears in more contexts than 'him'. The ability of MOSAIC to produce these errors suggests that an account of these phenomena in which the child is attributed with abstract categories such as 'tense' and 'agreement' as parts of underlying representation of an utterance is unnecessarily abstract. Once one abandons the assumption that children are operating with adult-like syntactic categories from the outset, their apparently sophisticated use of tense and agreement and the absence of particular kinds of errors in their speech can be explained in much more limited-scope terms.

The results presented here suggest that children's variable use of verb forms with respect to case-marking errors can be explained in terms of the learning of different verb forms from different positions in the surface structure in the mothers' speech from which they have been extracted. The implication is that children's early knowledge can be characterised as a vocabulary of unanalysed verb forms, or unanalysed sequences including verb forms, and a set of limited-scope formulae which specify how these verb forms pattern with respect to other items in their vocabularies (e.g. Braine, 1976, 1987).

Verbs with and without agreement marking are likely to come from different populations with the exception of high-frequency verbs, for which the child may have learned several morphological markers (Brown, 1973; Pine, Lieven & Rowland, 1998). As a result, errors which involve verbs with third-person-singular inflection (e.g. 'goes') are less likely to occur if the child has not learned many third-person-singular verb forms, not because children understand agreement. This explains why case-marking errors can occur with agreeing verb forms relatively infrequently. What is not considered in the ATOM is the likelihood of

particular errors being made. An argument based on the low frequency of a particular error type needs to take into account the expected error rate (Lieven, Pine & Baldwin, 1997; Rubino & Pine, 1998). This means that the low error rates which are either dismissed as performance errors or predicted not to occur at all, as in the case of the accusative+agreement errors discussed in this paper, are not necessarily due to an understanding of the rules of grammar by the child, but more simply due to the low probability of two lexical items being used in conjunction, given the statistical distribution of such items in a language, in this case English.

It is clear that although MOSAIC is perfectly capable of capturing the phenomena found in child speech, it is not fully capturing the degree to which masculine and feminine forms behave differently. The children whose data we report show very sparing use of the word 'she', which results in a higher proportion of 'her' for 'she' errors than would be the case if 'she' were produced in similar quantities to 'he'. MOSAIC does show greater production of 'he' than 'she', but still produces 'she' and 'her' disproportionately, when compared with the frequencies at which children produce these items. We hope that by making MOSAIC sensitive to the frequency of individual words in future, we may be able to model the proportions in which children produce different pronouns, which should, in turn, bring our results in line with the child data.

References

Braine, M.D.S. (1976). Children's first word combinations. *Monographs of the Society for Research in Child Development*, *41* (1, Serial No. 164).

Braine, M.D.S. (1987). What is learned in acquiring word classes - A step towards an acquisition theory. In B. MacWhinney (Ed.), *Mechanisms of language acquisition* (pp. 65-113). Hillsdale, NJ: Erlbaum.

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

Croker, S., Pine, J.M. & Gobet, F. (2000) Modelling optional infinitive phenomena. In *Proceedings of the Third International Conference on Cognitive Modeling* (pp.78-85). Veenendaal: Universal Press

De Groot, A. & Gobet, F. (1996) *Perception and memory in chess*. Assen: Van Gorcum.

Feigenbaum, E.A. & Simon, H.A. (1984) EPAM-like models of recognition and learning. *Cognitive Science, 8,* 305-336.

Gobet, F. (1993) A computer model of chess memory. *Proceedings of the 15th Annual Meeting of the Cognitive Science Society* (pp. 463-468). Hillsdale, NJ: Erlbaum.

Gobet, F. (1998) Memory for the meaningless: How chunks help. *Proceedings of the 20th Meeting of the Cognitive Science Society* (pp. 398-403). Mahwah, NJ: Erlbaum.

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

Gobet, F. & Pine, J.M. (1997) Modelling the acquisition of syntactic categories. *Proceedings of the Nineteenth Annual Meeting of the Cognitive Science Society* (pp. 265-270). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jones, G., Gobet, F., & Pine, J.M. (2000) Learning novel sound patterns. In *Proceedings of the Third International Conference on Cognitive Modeling* (pp. 169-176). Veenendaal: Universal Press.

Lane, P.C.R., Cheng, P.C-H., & Gobet, F. (1999). Learning perceptual schemas to avoid the utility problem. In M. Bramer, A. Macintosh and F. Coenen (Eds.) *Research and Development in Intelligent Systems XVI*, Cambridge, UK, pp. 72-82 (Springer-Verlag).

Lieven, E.V.M., Pine, J.M. & Baldwin, G. (1997) Lexically-based learning and early grammatical development. *Journal of Child Language*, 24, 187-219.

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

Pine, J., Lieven, E. & Rowland, C. (1998) Comparing different models of the development of the English verb category. *Linguistics,* 36, 807-830.

Rispoli, M. (1998) Patterns of pronoun case error. *Journal of Child Language,* 25, 533-554.

Rispoli, M. (1999) Case and agreement in English language development. *Journal of Child Language,* 26, 357-372.

Rubino, R.B. & Pine, J.M. (1998) Subject-verb agreement in Brazilian Portuguese: What low error rates hide. *Journal of Child Language,* 25, 35-59.

Schütze, C. (1997) INFL in child and adult language: Agreement, case and licensing. Ph.D. Thesis. MIT.

Schütze, C. (1999) Different rates of pronoun case error: Comments on Rispoli (1998). *Journal of Child Language*, 26, 749-755.

Schütze, C. & 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* (pp. 670-681). Somerville, MA: Cascadilla Press.

Theakston, A.L., Lieven, E.V.M., Pine, J.M., & Rowland, C.F. (2000) The role of performance limitations in the acquisition of 'mixed' verb-argument structure at stage 1. In M. Perkins & S. Howard (Eds.) *New directions in language development and disorders*. New York: Plenum

Wexler, K. (1998) Very early parameter setting and the unique checking constraint: A new explanation of the optional infinitive stage. *Lingua,* 106, 23-79.