In Proceedings of the Sixth International Conference on Cognitive Modeling, 94-99.
Mahwah, NJ: Lawrence Earlbaum.

Resolving Ambiguities in the Extraction of Syntactic Categories through Chunking

Daniel Freudenthal (DF@Psychology.Nottingham.Ac.Uk
Julian Pine (JP@Psychology.Nottingham.Ac.Uk
School of Psychology, University of Nottingham
University Park, Nottingham, NG7 2RD, UK

Fernand Gobet (Fernand.Gobet@Brunel.Ac.Uk)
Department of Human Sciences, Brunel University
Uxbridge, Middlesex, UB8 3PH, UK

Abstract

In recent years, several authors have investigated how co-occurrence statistics in natural language can act as a cue that children may use to extract syntactic categories for the language they are learning. While some authors have reported encouraging results, it is difficult to evaluate the quality of the syntactic categories derived. It is argued in this paper that traditional measures of accuracy are inherently flawed. A valid evaluation metric needs to consider the well-formedness of utterances generated through a production end. This paper attempts to evaluate the quality of the categories derived from co-occurrence statistics through the use of MOSAIC, a computational model of syntax acquisition that has already been used to simulate several phenomena in child language. It will be shown that derived syntactic categories which may appear to be of high quality will quickly give rise to errors which are not typical of child speech. A solution to this problem is suggested in the form of a chunking mechanism which serves to differentiate between alternative grammatical functions of identical word forms. Results are evaluated in terms of the error rates in utterances produced by the system as well as the quantitative fit to the phenomenon of subject omission.

Introduction

In recent years, several authors have argued that co-occurrence statistics can provide powerful cues that may aid children in extracting syntactic categories for the language they are learning. Redington, Chater and Finch (1998) analysed large corpora of child directed speech and performed a cluster analysis on vectors describing the lexical context in which words occurred. They found that words that occurred in linguistically similar contexts (tended to be preceded and followed by the same words) had a high likelihood of belonging to the same syntactic class.

Mintz (2003) expanded on the work of Redington et al. Rather than analysing vectors describing lexical context, Mintz's unit of analysis was a frame: two jointly occurring words with one word in between. Mintz restricted his analysis to the 45 most frequent frames that occurred in a large corpus.

While both Redington et al. and Mintz showed that their procedure resulted in apparently good syntactic categories, there is an inherent difficulty with the use of co-occurrence statistics to derive syntactic categories. As Pinker (1987) points out, words that occur in similar contexts may not be of the same category. Pinker argues that a distributional learning mechanism faced with utterances 1a, b and c, would produce an ungrammatical utterance like 1d.

1a. John ate fish
1b. John ate rabbits
1c. John can fish
1d. *John can rabbits

Mintz (2003) claims that 'in children's actual input, these problems do not significantly undermine the informativeness of distributional patterns' (p. 92). He also suggests that 'although problematic environments may exist, there is nonetheless enough "signal" in the distributional patterns compared to the noise created by the problematic environments that categorization from distributional patterns is not intractable' (p. 93).

There is, however, an inherent difficulty with the approach taken by Mintz and Redington et al., which may obscure the extent of the problem identified by Pinker. Mintz and Redington et al. evaluated the quality of the extracted categories using criteria of accuracy and completeness. Accuracy was computed by classifying every word-pair within a category as a hit (same syntactic class), or miss (different syntactic class). Where the grammatical class of a word was unclear, the corpus was consulted to disambiguate and label the word. Mintz used two types of labeling. In standard labeling, all nouns and pronouns were classed as nouns, and all verbs (lexical verbs, auxiliaries and the copula) were classed as verbs. In expanded labelling, nouns and pronouns were labeled as distinct categories, as were lexical verbs, auxiliaries and the

copula. While Mintz achieved high levels of accuracy with both types of labelling, closer inspection of his categories reveals that they may not be as accurate as his analyses suggest. One of Mintz's verb categories contains verbs in present tense and past tense as well as progressive particles, verbs that can and cannot be used in an imperative frame, and verbs such as *do* and *have* that can be used both as a main verb and as an auxiliary.

This heterogeneity of the derived word classes may not appear problematic since neither Mintz nor Redington et al. concern themselves with production (Mintz views the process of extracting distributional categories as a precondition for a (relatively unspecified) process of bootstrapping into a parametrized universal grammar). When one considers how the extracted categories might be used in production, however, it quickly becomes apparent that heterogeneous word classes will result in utterances that deviate considerably from child speech. The simplest way in which a child producing speech could use the categories arrived at through a distributional analysis of the input is by considering the members of a category as equivalent. That is, if words *a* and *b* occur in the same category, the child may simply substitute *a* for *b* in a context where it knows *b* has occurred. Taking the words *do*, *have* and *put* (which were classed together in Mintz's analysis) as an example, such a substitution mechanism will result in (clearly incorrect) utterances such as *Do you got an ice-cream* and *Put you want a drink*.

It is argued here that when syntactic categories derived from co-occurrence statistics are used to generate speech, more subtle problems emerge that are not apparent with the use of an evaluation metric based on a researcher's intuition about a word's syntactic class. These problems become especially apparent in detailed quantitative simulations of child data, where seemingly correct substitutions may drastically affect the fit to actual child data. This became clear when Freudenthal, Pine & Gobet (2002a) used MOSAIC, a computational model of syntax acquisition which utilizes co-occurrence statistics to substitute phrases that occurred in similar contexts, to simulate the phenomenon of subject omission and the associated verb phrase length effect (Bloom, 1990). This phenomenon revolves around the fact that there is a stage in development where children produce subjectless utterances such as *Want a cookie*. While the model simulated the general pattern of results, it tended to overestimate the levels of subject omission. One of the reasons for this was that, in order to identify ungrammatical subjectless utterances, the analyses were restricted to utterances containing 'non-imperative verbs'. Since MOSAIC tended to substitute non-imperative verbs for imperative verbs it generated a relatively high number of subjectless utterances. The reason why these verbs were substituted was that both verb types were linked because they both occur in non-imperative frames. While their substitution in imperative frames did result in child-like utterances, the substitution rate was too high to allow a good quantitative fit to the data. This type of problem is not apparent in an approach that simply extracts syntactic categories and does not use a production end to generate utterances.

Thus, the main cause of problematic substitutions is that a substitution that is correct in one context is incorrect in another context. This paper aims to show that one possible solution to this problem is to compute co-occurrence statistics over longer units. Redington et al. considered longer contexts (two or three words preceding and following the target word), and found this did not improve the quality of their syntactic categories. This paper investigates a different approach, inspired by the well-established chunking theory (Chase & Simon, 1973; Gobet et al., 2001).

A new version of MOSAIC has been developed which incorporates a novel chunking mechanism[1] which results in frequent phrases being treated as one unit. One consequence of this is that single words that have been chunked up will no longer be substituted unless when substituted as part of a chunk. It will be shown that this mechanism decreases the amount of unwanted as well as incorrect substitutions, resulting in a decreased overall error rate as well as a better fit to the phenomenon of subject omission.

The remainder of this paper is organized as follows. Firstly, MOSAIC and its chunking mechanism will be described. MOSAIC will be trained on corpora of child directed speech while the parameter governing chunking frequency is manipulated. In order to provide an evaluation of the quality of the output, a sample of generated utterances is judged against criteria of 'well-formedness'. The output is also compared to actual child speech, which is analysed with respect to the phenomenon of subject omission.

Simulating Language Acquisition in MOSAIC

MOSAIC has already been used to simulate several phenomena in child speech. Earlier versions have been used to simulate the Verb-Island phenomenon (Jones, Gobet & Pine, 2000; negation errors (Croker, Pine & Gobet, 2003) the Optional Infinitive phenomenon in Dutch, Spanish and English (Freudenthal, Pine & Gobet, 2002a, 2003, in preparation), as well as phenomena related to subject omission in English Freudenthal, Pine & Gobet 2002b). Whilst the version

[1] Earlier versions of MOSAIC employed a chunking mechanism as well. The novel chunking mechanism differs from this in that chunks can now occur at the primitive level.

used for the simulations discussed here has changed from the earlier simulations, the main theoretical underpinning of the model remains the same. The basic tenet of the model is that the learning of language is a performance-limited process which is heavily weighted towards the most recent elements in the speech stream (i.e., which has an utterance final bias). Several authors have argued that children are better at learning material that occurs towards the end of the utterance (Shady & Gerken, 1999; Wijnen et al., 2001).

MOSAIC learns from orthographically coded input, with whole words being the unit of analysis. The model is a simple discrimination net (an n-ary tree) which is headed by a root node. At the start of learning the discrimination net consists of just the root node. More nodes (encoding words or phrases) are added as the model is shown more utterances. An important requirement for nodes to be added is that whatever follows the word to be encoded in the input must already have been encoded in the model. That is, the model will only learn a new word when it has already encoded the rest of the utterance. Thus, while the model processes utterances from left to right, it builds up its representation of the utterances it receives by starting at the end of the utterance, and slowly working its way to the beginning[2]. The probability of creating a node in MOSAIC is given by the following formula:

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

where: NCP = Node Creation Probability

m = a constant, set to 20 for these simulations.
c = corpus size.
u = total number of utterances seen.
d = distance to the end of the utterance.

The formula results in a basic sigmoid curve (when plotted as a function of the number of utterances the model has seen). The formula contains the size of the corpus and the total number of utterances seen. The size of the corpus is included because the size of the available input corpora differs considerably. The use of the term (m – u/c) ensures that after n presentations of the complete input corpus the Node Creation Probability is identical for corpora of different sizes. The 'distance to the end of the utterance' in the exponent causes material that occurs near the beginning of the utterance to have a lower likelihood of being encoded than material that occurs near the end. This effect decreases as the model sees more input. Since learning in MOSAIC is slow, the input corpus is fed through the model several times, so that output of increasing average length can be generated after consecutive exposures to the input corpus.

[2] Earlier versions of MOSAIC simulated such an utterance final bias by restricting production to utterances that had appeared in sentence final position.

Production of Novel Utterances

Utterance production in MOSAIC involves outputting all the utterances the model has encoded. However, the output that MOSAIC produces consists of more than the input it has seen. MOSAIC has a mechanism for linking words or phrases that have occurred in similar contexts. All nodes being traversed when processsing input are deposited into a buffer of limited size reflecting the most active/recently encountered input. The nodes in the buffer are then compared with respect to their preceding and following context. When the overlap between two nodes is sufficiently high (more than 20% of both the context that preceded and followed the target node are the same), a *generative link* is created between them. The contents of nodes that are linked can be substituted for each other when the model produces output. This mechanism allows MOSAIC to produce utterances that were not present in the input.

Chunking in MOSAIC

MOSAIC employs a chunking mechanism, which results in frequent multi-word phrases being treated as one unit. Nodes in the network contain a frequency slot, the value of which is increased every time that node is traversed when the net sorts an input utterance. The frequency of a node at one of the lower levels (non-primitive nodes[3]) in the tree encodes the frequency of the entire phrase leading up to that node. Thus, if a node for *you* occurs underneath *do*, the frequency of that node encodes the number of times the phrase *do you* has been encountered. When the frequency of a non-primitive node exceeds a pre-determined value, the node is chunked up with the node above it: the two nodes are merged into one node at the primitive level. Thus, in the above example the two nodes encoding the phrase *do you* will be merged into one node at the primitive level. The chunk is then propagated through the network; all occurrences of the phrase *do you* are chunked up. Nodes encoding chunks can be linked to other nodes encoding chunks (or words) in the same way that nodes encoding individual words are linked. When two nodes are chunked, it is no longer possible to substitute words for the individual words making up the chunk. Thus, the chunk *Do I* may be substituted for the

[3] The distinction between nodes directly underneath the root node (primitive nodes) and those at lower levels (non-primitive nodes) is an important one. Due to the structure of the discrimination net, primitive nodes encode all the context the word or phrase has been seen in (the 'global context'). Non-primitive nodes encode 'local context'.

chunk *Do you*, when they share sufficient context. However, should the words *I* and *you* be linked, they can be substituted in unchunked contexts, but not in chunks. In this way, chunking serves to differentiate different grammatical functions of the same word form: if the dummy modal *Do* is chunked with the subject *you*, it will no longer be substituted by verbs that are linked to *Do* by virtue of its occurrence as a main verb.

Chunking affects the substitution of words in two ways. Firstly, chunks themselves are deposited into the buffer, making phrases the target for the creation of generative links. Secondly, the context preceding and following a target node may be chunked up. Thus if the phrase *He goes into the house* contains the chunks *he goes* and *the house*, the context for the word *into* will be *he goes* and *the house* rather than *goes* and *the*. Chunking thus serves to increase the length of items considered for a generative link, as well as increase the context considered in the creation of a generative link.

The chunking mechanism also affects learning. If the model receives a novel input utterance containing a phrase that has been chunked up earlier, it will treat the phrase as a unit, rather than attempting to encode its constituent words separately.

The Simulations

Simulations were run using two corpora of English (maternal) child-directed speech (those of Anne and Becky) taken from the Manchester corpus (Theakston, Lieven, Pine & Rowland, 2001) available through the CHILDES data-base (MacWhinney, 2000). The size of the input sets is approximately 33,000 and 24,000 utterances. Simulations were run using different levels of chunking. The models' output was analysed with respect to error rates and levels of subject omission.

Error rates

For the first simulation, models were trained with and without chunking for both children. For the chunked model, the chunking threshold (frequency required for a node to be chunked up) was set at 1/4 times the square root of the number of nodes in the net. The chunking threshold was expressed relative to the square root of the nodes in the net to ensure that the chunking rate was relatively constant over the development of the model. For all simulations, an output file was selected at an MLU (Mean Length of Utterance) of approximately 3.5

Next, a sample of 500 utterances from each of the output files was coded by two independent raters for the presence or absence of syntactic errors. Syntactic errors were defined as cases in which one or more of the substitutions made by the model resulted in an utterance that was grammatically incorrect (e.g. *Pegs find fallen down* from *Pegs have fallen down*). Note that this definition of syntactic errors is designed to exclude cases in which the model substituted a grammatically correct word into a sentence fragment (e.g. *My toys out* v *Your toys out*) and cases in which the substitutions made by the model were semantically but not syntactically anomalous (e.g. *Shall I cut them with the puzzle?* v *Shall we cut them with the knife?*). The vast majority of the errors identified in this way fell into one of the following categories: word-class errors (e.g. *To vest on his tummy* v *To lie on his tummy*); subject-verb agreement errors (e.g. *They am sitting* v *I am sitting);* missing argument errors (e.g. *Putting the story* v *Reading the book*); errors involving the use of a verb with the wrong particle (e.g. *Shall we use her t-shirt off* v *Shall we take his dungarees off*); errors involving the use of a verb form with the wrong auxiliary (e.g. *I've just finish that off* v *You've just taken that off*) and errors involving the use of a particular type of noun with the wrong determiner (e.g. *Put it on a sand* v *Put it on the sand*). Interestingly, virtually all of the errors falling into these categories seemed to involve either the substitution of a word from the wrong syntactic category for a word that is a member of two or more syntactic categories (e.g. the use of *vest* as a verb instead of *lie* which can be both a noun and a verb) or the substitution of a word from the correct syntactic category into a context in which that particular instance of the category is not permitted to occur (e.g. the use of the indefinite article *a* with a mass noun instead of the definite article *the* which can be used with both mass and count nouns in English). Note that these are precisely the kinds of errors that are likely to be hidden by the kind of evaluation metrics used in previous research using distributional learning mechanisms.

Agreement between the raters was high, at .93 (Kappa = .74). The results are shown in Table 1. Error rates are lower for the chunking models ($X^2 = 40.70$, $p < .001$ for Anne, and $X^2 = 5.42$, $p < .05$ for Becky).

Table 1: Syntactic error rates for Anne and Becky's simulations at two levels of chunking.

	No chunking	Chunking
Anne	.21	.07
Becky	.24	.18

However, a potentially confounding factor is that the unchunked models simply generate more novel utterances. It could therefore be argued that any mechanism that restricts the generativity of the model will reduce the error rate. In order to test this possibility, the generativity of the unchunked models was reduced by increasing the overlap parameter governing the creation of generative links to .25. This resulted in the proportion of novel utterances being similar to that in the chunked models. The error rate for Anne's model was reduced to .16, less than for the high generativity model $X^2 = 4.15$, $p < .05$, but more than for the chunked version $X^2 = 19.90$, $p < .001$. For Becky's model, the error rate was .23, not significantly different from the high generativity version $X^2 < 1$, $p < 1$, and still higher than the chunked version $X^2 = 4.44$, $p < .05$.

Thus, the reduced error rates in the chunking version are not just a result of chunking reducing the proportion of generated utterances, but rather of chunking reducing unwanted substitutions. The reason why error rates remain higher than in the chunked version is that even at a high overlap percentage, some links may remain which give rise to errors. As the overall rate of generativity decreases, these undesirable links may even gain weight, and could conceivably even increase the error rate.

Subject omission

A second analysis assessed whether chunking can decrease the levels of subject omission. For these simulations, the chunking threshold was set to three levels: 4, 1, and .25 times the square root of the number of nodes in the net. These simulations are referred to as low, medium and high chunking, respectively.

In order to match the models' output to child speech, models were trained iteratively to match the MLU of the children at two points in time. The models' output was then compared against children's output with respect to the phenomenon of subject omission.

The analysis of the levels of subject omission was performed in the same way as in Bloom (1990), and Freudenthal, Pine and Gobet (2002b). Utterances were limited to those that Bloom identified as non-imperative (though the verb *see* was excluded from this list as it was not considered non-imperative). The analysis was restricted to declaratives. Double verb constructions and utterances containing the words *don't*, *no*, or *not* were excluded from the analysis.

The remaining utterances were scored with respect to the inclusion of a subject. Figure 1 shows the results for the children and the six simulations at two different MLUs. Model MLUs were matched as closely as possible to the children's MLUs.

Figure 1 shows that the fits for the simulations increase as the chunking rate increases. While the overall fits are not particularly good, the chunking mechanism appears to have been successful in avoiding unwanted substitutions, with the high chunking model providing the best fit at high MLU (particularly for Becky). It may be worth stressing that only one parameter in the chunking mechanism has been manipulated. Future work may suggest manipulations that result in a better fit.

Generative Chunks

While it could be argued that the main effect of the chunking mechanism is to reduce error by avoiding incorrect substitutions, it is worth pointing out that the chunking mechanism leads to different types of substitutions (and errors) as well. The reason for this is that chunks themselves can be linked (both to single words and other chunks). While a full analysis of the role of linked chunks is beyond the scope of this paper, some interesting examples can be given. In some of the high chunking models phrases like *I can* were linked to *Can I*, thus allowing the model to generate declaratives off questions and the other way round. Similarly, in one of the models the phrase *I wouldn't* was linked to *I don't want to*. The chunking mechanism thus aids in linking phrases as well as words that fullfil a similar grammatical role. The chunking mechanism resulted in some interesting errors as well. One of the simulations substituted *don't want to* for *want to*. While this resulted in some grammatical utterances, it also resulted in phrases like *Do you don't want to*. This is clearly a syntactic error. However, it is a type of error that children do occasionally make.

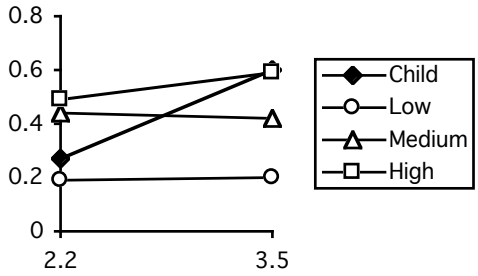

Figure 1a: Levels of subject provision for Anne and simulations

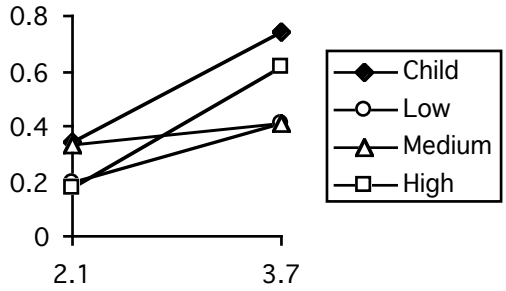

Figure 1b: Levels of subject provision for Becky and simulations.

Conclusions

Several conclusions can be drawn from the simulations reported here. Firstly, in a global analysis of generated utterances, clear word class errors do occur, but not at very high rates. The problem identified by Pinker (1987) therefore does not appear to be particularly significant. However, when the analysis is restricted to a subset of the data (such as utterances containing non-imperative verbs), it becomes apparent that the fact that a simple distributional analysis does not pick up subtle differences between different verb classes can greatly affect the fit to child data. It was shown that the chunking mechanism was able to reduce the overall error rates as well as prevent the substitution of similar words in incorrect contexts.

It should be stressed that chunking does not simply cut generativity in all contexts (as increasing the overlap parameter does). Rather, chunking restricts the contexts in which two words may be substituted. Thus, two single words that share a generative link may be substituted in unchunked contexts, but not in contexts where the word is chunked up (unless of course the chunk itself has a generative link). The chunking mechanism is thus able to cut generativity selectively. Besides diminishing unwanted generativity, chunking also adds to generativity by substituting phrases rather than words.

It is important to bear in mind that the only parameter manipulated in these simulations is the chunking threshold. There is clearly a range of parameters that can be manipulated in conjunction with chunking threshold. At present, the chunking threshold is a function of the square root of the number of nodes in the net. Variations of this formula may affect the chunking rate differentially for different stages of development, thus affecting more detailed fits to child data. We are not committed to the fits and specific implementation used in these simulations, but rather stress the fact that chunking can be a powerful tool in resolving ambiguities in the extraction of syntactic categories.

On a more general level, these analyses illustrate two strengths of MOSAIC as an approach to modelling language acquisition: the use of realistic child-directed speech, and the production of utterances that can be compared with child speech. The use of child-directed speech is important because it ensures a realistic frequency distribution. As all distributional analyses are frequency sensitive, a realistic frequency distribution in the input is crucial for obtaining good fits to detailed phenomena in child language.

The use of a production end has shown that traditional measures of accuracy are insufficient to evaluate the quality of syntactic categories derived from co-occurrence statistics, as accuracy not only depends on the researcher's intuitions regarding a word's syntactic class, but also on the context in which the word is used. Researchers should therefore be careful about relying on measures of accuracy to evaluate the quality of syntactic categories as the addition of a production end may show the accuracy to be considerably lower than it appeared.

Acknowledgments

This research was funded by the ESRC under grant number RES000230211

References

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

Croker, S., Pine, J. M., & Gobet, F. (2003). Modelling children's negation errors using probabilistic learning in MOSAIC. In F. Detje, D. Dörner & H. Schaub (Eds.) *Proceedings of the Fifth International Conference on Cognitive Modeling* (pp. 69-74). Bamberg: Universitäts-Verlag.

Freudenthal, D., Pine, J. & Gobet, F. (2002a). Modelling the development of Dutch optional infinitives in MOSAIC. In W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Meeting of the Cognitive Science Society* (pp. 322-327). Mahwah, NJ: LEA.

Freudenthal, D., Pine, J. & Gobet, F. (2002b). Subject omission in children's language: The case for performance limitations in learning. In W. Gray & C. Schunn (Eds.), *Proceedings of the 24th Annual Meeting of the Cognitive Science Society* (pp. 328-333). Mahwah, NJ: LEA.

Freudenthal, D. Pine, J. & Gobet, F. (2003). Modelling syntax acquisition in MOSAIC. Paper presented at the 8th International Cognitive Linguistics Conference, Logroño, Spain.

Freudenthal, D., Pine, J. & Gobet, F. (in preparation). Modelling the development of children's use of optional infinitives in English and Dutch using MOSAIC.

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

Jones, G., Gobet, F. & Pine, J. M. (2000). A process model of children's early verb use. *Proceedings of the Twenty Second Annual Meeting of the Cognitive Science Society* (pp. 723-729). Mahwah, NJ: LEA.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. Third Edition.* Mahwah, NJ: LEA.

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

Pinker, S. (1987). The bootstrapping problem in language acquisition. In B. MacWhinney (Ed.). *Mechanisms of language acquisition*. Hillsdale, NJ: LEA.

Redington, M., Chater, N., & Finch S. (1998). Distributional Information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*, 435-469.

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

Theakston, A.L., Lieven, E.V.M., Pine, J.M. & Rowland, C.F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, *28*, 127-152.

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