Developmentally plausible learning of word categories from distributional statistics

Daniel Freudenthal[1], Julian M. Pine[1], Gary Jones[2] and Fernand Gobet[1]
[1]Department of Psychological Sciences, University of Liverpool
[2]Division of Psychology, Nottingham Trent University

Abstract

In this paper we evaluate a mechanism for the learning of word categories from distributional information against criteria of psychological plausibility. We elaborate on the ideas developed by Redington et al. (1998) by embedding the mechanism in an existing model of language acquisition (MOSAIC) and gradually expanding the contexts it has access to in a developmentally plausible way. In line with child data, the mechanism shows early development of a noun category, and later development of a verb category. It is furthermore shown that the mechanism can maintain high performance at lower computational overhead by disregarding token frequency information, thus improving the plausibility of the mechanism as something that is used by language-learning children.

Keywords: Word class acquisition; Distributional analysis

Introduction

A key issue in understanding how children acquire language is how they build word class categories such as verb or noun. Recent computational work has shown that there is a great deal of information in the distributional properties of different languages that can be used to distinguish between instances of different word class categories in the input. For example, in English, words that are preceded by determiners such as *a* and *the* and followed by auxiliary verbs such as *can* and *will* tend to be nouns, whereas words that are preceded by nominative pronouns such as *I* and *You* and followed by determiners such as *a* and *the* tend to be verbs.

Several approaches to this problem have been proposed, but they tend to focus on building large word classes with high accuracy rather than developing mechanisms that can be plausibly applied by, and fit developmental data from language-learning children. Thus, mechanisms for distributional analysis routinely collect data from large corpora and entire utterances, and make limited contact with the (developmental) child data. In this paper, we explore a more developmentally plausible mechanism by embedding it in an existing model of language acquisition that learns to produce increasingly long utterances, and thus gradually expands the contexts over which statistics are computed in a way that is consistent with children's processing biases.

The two dominant approaches to distributional analysis are those of Mintz (2003) and Redington et al. (1998). Mintz introduces the notion of a 'frequent frame' - a combination of two words with one word intervening (*You* X *A*). Mintz identifies the most (typically 45) frequent frames for a given corpus, and finds that the words that co-occur within a frame tend to belong to the same grammatical category. The approach of Mintz seems intuitively appealing since its relative simplicity means it is well the capabilities of language-learning children. However, it has been argued that it works less well for languages with relatively free word order such as Dutch (Erkelens, 2009) and German (Stumper et al. 2011). It has also been argued that its all-or-none nature makes it overly sensitive to noise - the presence of a small number of items that do not fit the dominant word class for a frame has large effects on classification accuracy (Freudenthal et al. 2013).

Redington et al. (1998) express the contexts in which words occur as vectors containing counts of frequent context words in preceding and following position. Similarity between words is then expressed as the (rank order) correlation between these vectors, and the matrix of correlations is used as input to a cluster analysis. The probabilistic nature of Redington et al.'s approach makes it naturally resistant to noise, and thus less prone to error. However, Redington et al.'s approach requires children to track the frequency of large numbers of words and has been criticized for carrying a high computational overhead (St. Clair et al. 2011), thus making it less appealing as a mechanism that language-learning children might employ. The plausibility of this approach would thus be greatly increased if it could be shown that its computational overhead can be reduced significantly without affecting performance.

While both approaches are intended as mechanisms that language-learning children employ, neither considers the fact that word classes are likely to be gradually built up over time. Thus, they collect statistics from complete utterances and across large corpora, and focus on building large word classes with high accuracy. However, it is debatable whether young children represent statistics that reflect all of the input they have encountered. Children's early utterances are just one or two words long, and only gradually increase in length over a period of years. While children may well attend to longer utterances than they produce, it seems unlikely that they would process entire utterances from a young age. A mechanism that tracks complete utterances may therefore employ statistics that are not available to language-learning children in the early stages of development. A developmentally more plausible mechanism would build word classes gradually by slowly expanding the contexts from which statistics are collected.

This suggestion is further supported by experimental evidence which suggests that children's productive use of words develops at different speeds for different word classes. In particular, production studies suggest that children develop a category of noun earlier than they develop a category of verb (Akhtar & Tomasello, 1997; Olguin & Tomasello, 1999; Tomasello & Olguin, 1993).

The main aim of this paper is to investigate if such developmental effects can be simulated by gradually expanding the contexts from which statistics are collected. This will be done by embedding variants of Redington et al.'s mechanism in an existing computational model of language acquisition (MOSAIC; Freudenthal et al. 2007, 2010, 2015), that successfully simulates a number of key phenomena in language acquisition. Full details of MOSAIC and the way in which it is trained are provided in Freudenthal et al. (2015).

MOSAIC is a simple learning model that takes as input orthographically transcribed corpora of child-directed speech. Training in MOSAIC takes place by feeding the input corpus through the model multiple times. A key feature of MOSAIC is that it builds up its representation of the utterances to which it is exposed by starting at the right edge of the utterance and slowly working its way to the left. Thus, with each exposure to the input MOSAIC represents increasingly long utterance-final phrases that become increasingly adult-like. The utterance-final bias is MOSAIC's key mechanism for simulating cross-linguistic differences in children's early speech and is thus independently motivated. MOSAIC thus provides a natural framework for testing the notion that word classes develop gradually as the contexts from which statistics are collected are expanded. MOSAIC employs a mechanism for distributional analysis that borrows from the ideas of Redington et al., and thus allows us to investigate the performance of different implementations of the mechanism in a developmental setting.

A first analysis will investigate how well Redington et al.'s mechanism performs when statistics are collected from increasingly long utterance-final phrases. This will be done by training MOSAIC on corpora of child-directed speech and producing output at several stages of development. The output (of increasing average length) from MOSAIC will then be used to derive the counts of context words on which the mechanism is based, and thereby test the performance of the mechanism in a developmentally more plausible setting.

A second analysis will investigate how well a substantially simplified version of Redington et al.'s approach performs. For all words that are to be classified (the *target* words), Redington et al.'s mechanism collects counts for a number (typically the 150 most frequent words for a corpus) of *context* words in preceding and following position. The vectors of counts are then concatenated and similarity between words is expressed as the rank order correlation between concatenated vectors. The mechanism thus collects counts for large numbers of context words, but only uses a limited amount of this frequency information. Here, we investigate the performance of a variant of Redington et al.'s mechanism that disregards token frequency information altogether. Rather than collecting counts for target words in preceding and following position, the approach simply notes the identity of the words in preceding and following position. The context for a given target word is therefore expressed as a list of words (types) rather than a vector of counts (tokens). Similarity is then expressed as a measure formally known as the Jaccard distance: the length of the intersection of two contexts divided by the length of the union - for two words that have been preceded by {a, the} and {the, green} respectively, the similarity is 1/3.

Disregarding token frequency reduces the computational overhead associated with the mechanism considerably: there is no need to collect large numbers of counts, and computation of the Jaccard distance is a mathematically simpler operation than the computation of a rank order correlation. While the simplified mechanism uses less information than the original approach, this may only have limited effects on performance, since the rank order correlation used by Redington et al. only utilizes a limited amount of frequency information.

A final question concerns how well the mechanism is capable of capturing developmental effects in the building of word classes. If the mechanism could be shown to display early emergence of a noun class and late emergence of a verb class, this would increase its developmental plausibility. Such a developmental pattern may arise naturally from MOSAIC's utterance-final bias. Nouns frequently occur near the end of utterances, whereas verbs tend to occur in medial position. Contexts for nouns may therefore register earlier than for verbs, resulting in the emergence of a noun category before a verb category.

In summary, the main aim of this paper is to develop a psychologically plausible mechanism for learning word classes from distributional information by 1. assessing the performance of Redington et al.'s mechanism when gradually extending its access to input in a developmentally plausible way, 2. comparing the performance of the original mechanism with a greatly simplified one that disregards token frequencies, and 3. determining if the developmental variation results in a developmentally plausible pattern of noun and verb linkage. These questions are investigated by applying the two variants of Redington et al.'s mechanism to the increasingly long phrases encoded by MOSAIC models in different stages of development. The main dependent variables are the standard measures of accuracy (number of correct classifications), and completeness (number of classifications), both overall and for nouns and verbs separately. Additionally, a measure of noun richness is computed to track the emergence of noun and verb categories.

The simulations

The first set of analyses was aimed at determining the performance of Redington et al.'s approach in a developmental setting. To this end, MOSAIC was trained on the child-directed speech for each of the 12 children in the Manchester corpus (Theakston et al., 2001). Each model was exposed to the input a total of 50 times. The average length of the utterances that MOSAIC represents increases from zero to approximately five words over training. The model's ability to classify words was assessed at several points in the model's training. As in Redington et al.'s original formulation, target and context words were restricted to the 1000 and 150 most frequent words for each corpus (based on corpus-wide counts). Target and context words were fixed throughout development.

At each point in development, context vectors for the target words were generated by determining how often the context words occurred in the position directly before or after the target words. A complicating factor here is that MOSAIC does not represent duplicate utterances, and may thus underestimate how often a context and target word co-occur. This was remedied by taking each utterance in the input corpus, determining the largest utterance-final phrase from that utterance that was represented in the model, and adding this utterance-final phrase to the pool of utterances over which statistics were computed. Thus, if the input corpus contained three instances of "it's a dog", and two instances of "that's a dog", and the model only represents the utterance-final phrase "a dog", then 5 instances of the phrase "a dog" were added to the pool of utterances. The pool of utterances was then searched for the target words, and any occurrence of context words in preceding and following position noted. The rationale behind this procedure was to generate accurate counts for an input corpus based on the utterance-final fragments from that corpus that were represented in the model. As training proceeds, and MOSAIC represents more and longer utterances, the counts generated in this manner will become a closer approximation of the counts that would be generated from a corpus-wide analysis.

Vectors containing counts of content words in preceding and following position were concatenated and rank order correlations were computed for every pair of target words. For the current analysis, two words were considered of the same class if the rank order correlation exceeded a certain threshold. That is, the cluster analysis performed by Redington et al. was omitted for ease of interpretation. Results are reported for two levels of the threshold: 0.5 and 0.6. The accuracy of the resulting classification was scored against the (most common) grammatical class assigned to each word based on the morphology (MOR:) line of the transcripts. The main dependent variables were the overall accuracy of the classification and the number of items classed together (number of links). Within class accuracy for nouns and verbs was computed separately by dividing the number of noun-noun (or verb-verb) pairs over the number of pairs containing at least one noun (or verb).

Table 1 shows the results of these analyses, averaged over the 12 different models. The top rows present the data for a threshold of 0.5, and the lower rows a threshold of 0.6. Results are reported for 5 developmental stages, ranging from 36 to 50 exposures to the input. For completeness, the rows labeled 'all' provide data for a corpus-wide analysis, that includes all complete utterances. In line with the current practice in MOSAIC, only declarative utterances were included in the analysis. As can be seen in Table 1, the mechanism manages to maintain high overall accuracy throughout, which tends to be higher in the later stages. The number of links is higher for the threshold of 0.5 than for 0.6, though accuracy scores are not much different.

The main developmental effect in these simulations is that the number of linked items in the corpus-wide analysis is lower than it is for most earlier developmental stages, which link more items with lower accuracy. This pattern seems implausible. Children's early language use has been characterized as relatively rote and lexically specific, and is thought to become more productive with age – the opposite of the pattern in Table 1. It is also apparent from Table 1 that, while the mechanism manages relatively high accuracy for nouns, it is far less accurate in linking verbs, particularly for the early stages – verb accuracy exceeds 0.6 for just 4 out of the 12 cells in Table 1. Taken together, these findings suggest that the mechanism is not sufficiently constrained in the early stages of development.

Table 1: Performance of Redington et al.'s approach at thresholds of 0.5 and 0.6, averaged over 12 models.

Runs	# of links	Acc.	Noun Acc.	Verb Acc.
0.50				
36	1,954	0.68	0.74	0.37
38	4,580	0.73	0.77	0.41
40	5,219	0.72	0.75	0.43
44	3,258	0.79	0.80	0.57
50	2,454	0.81	0.81	0.71
All	2,382	0.82	0.81	0.73
0.60				
36	1,328	0.71	0.76	0.39
38	2,645	0.73	0.78	0.42
40	2,493	0.71	0.74	0.42
44	1,132	0.79	0.79	0.50
50	708	0.82	0.81	0.69
All	667	0.83	0.81	0.70

The reason for the mechanism's initial lack of constraint is that, early in development, contexts for the target words are derived from short, utterance-final phrases, that may only be two words long. This means not only that a limited number of contexts may be available for a given target word, but also that the available contexts may be biased towards preceding or following words. Since vectors for preceding and following contexts are concatenated (and zeros are effectively ignored), this means that the mechanism initially links items on the basis of just preceding or following contexts. With increasing exposure to the input, MOSAIC will represent longer phrases that extend further to the left of the utterance. As a result, the mechanism registers not only more contexts, but a better mix of preceding and following contexts. The mechanism thus becomes more constraining, and links (fewer) items, with higher accuracy – in particular for verbs.

These observations suggest that the practice of concatenating vectors for preceding and following contexts may be inappropriate from a developmental perspective. They also illustrate the value of embedding the mechanism in an existing developmental model of acquisition as the constraint provided by developmental data may lead to insights that remain hidden in corpus-wide analyses.

The data in Table 2 show that a more plausible developmental pattern results when preceding and following contexts are separated. For this analysis, the rank order correlation was computed separately for preceding and following contexts, and two items were considered of the same category only if *both* correlations were sufficiently high. Table 2 reports results for thresholds of 0.4 and 0.5[1]. The pattern of results in Table 2 differs starkly from that in Table 1. Rather than becoming more constraining with development, the mechanism shows a steady increase in the number of links over development, a pattern that is consistent with children's language use becoming more productive with age. The mechanism also achieves higher accuracy, in particular for verbs – verb accuracy is lower than 60% for just two out of twelve cells.

Table 2: Performance of Redington et al.'s approach with separated vectors at thresholds of 0.4 and 0.5.

Runs	# of links	Acc.	Noun Acc.	Verb Acc.
0.40				
36	50	0.65	0.64	0.33
38	254	0.81	0.85	0.61
40	756	0.81	0.83	0.60
44	1,426	0.84	0.84	0.77
50	1,658	0.86	0.84	0.83
All	1,691	0.86	0.85	0.84
0.50				
36	21	0.74	0.73	0.29
38	107	0.85	0.89	0.60
40	267	0.83	0.83	0.62
44	405	0.86	0.85	0.78
50	433	0.89	0.87	0.86
All	442	0.89	0.87	0.88

The data in Table 2 thus suggest that, when similarity is computed separately for preceding and following contexts, Redington et al.'s mechanism can be applied in a developmental setting, which increases its plausibility as a mechanism that could be employed by language learning children. However, separating preceding and following contexts does not alter the basic statistics over which correlations are computed: counts for the 150 most frequent context words. As was argued in the introduction, the need to collect these counts increases the complexity of the mechanism considerably. Thus, once a child has identified the most frequent context words, it needs to track their frequency before and after all target words. However, since the similarity between words is expressed using a non-parametric measure (i.e. a rank order correlation), much of the frequency information is discarded. The following section explores the mechanism's performance when frequency information is disregarded completely.

Disregarding Token Frequencies

Table 3 shows the results for 12 new MOSAIC models trained on the individual corpora for children in the Manchester corpus. Similarities between words were computed on the basis of the Jaccard distance or the rank order correlation. The rank order correlation was computed as in the previous analysis: based on counts for the 150 most frequent contexts words in (separated) preceding and following contexts. The Jaccard distance is defined as the length of the intersection divided over the length of the union of two sets. As with the rank order correlation, preceding and following contexts are considered separately. The Jaccard distance disregards token frequencies, and thus greatly reduces the computational complexity of the mechanism, as it is no longer necessary to collect counts for context words. Since context is represented as a simple list of word types, there is also no need to restrict context words to the most frequent words in the corpus. Thus, while in practice most context words will be contained in the 150 most frequent words, in principle, any word can act as a context word.

Table 3: Performance of Jaccard distance and rank order at thresholds of 0.2 and 0.45

Runs	# of links	Acc.	Noun Acc.	Verb Acc.
Jaccard				
36	27	0.78	0.83	0.50
38	140	0.83	0.84	0.55
40	370	0.87	0.88	0.68
44	648	0.89	0.86	0.82
50	717	0.91	0.88	0.87
All	738	0.91	0.88	0.87
Rank order				
36	35	0.68	0.76	0.46
38	181	0.8	0.83	0.53
40	488	0.82	0.83	0.63
44	846	0.85	0.84	0.77
50	913	0.87	0.85	0.84
All	925	0.87	0.85	0.85

As in previous analyses, the criterion for considering two words to be of the same category was fixed: 0.2 for the Jaccard distance and 0.45 for the rank order to allow for a meaningful comparison of accuracy and completeness. As can be seen in Table 3, results for the two measures are quite similar, although the rank order measure tends to have lower accuracy – so may be more appropriately set at a slightly higher threshold. Data from Table 2, however, suggests that this may cause completeness to fall below that for the Jaccard

[1] The value of these thresholds differs from those used for the first analysis, and was chosen to enable a meaningful overall comparison in terms of accuracy and completeness.

distance. Nevertheless, the conclusion that can be drawn from the data in Table 3 is that the Jaccard distance performs as well as (if not slightly better than) the rank order measure, and thus that the computational overhead of the mechanism can be reduced significantly without affecting performance.

Noun Richness

The previous analyses showed that Redington et al.'s approach can be applied in a developmental setting provided similarity is computed for preceding and following contexts separately. Its computational overhead can also be significantly reduced without affecting performance by disregarding frequency information. These changes increase the mechanism's plausibility as something that could be employed by language-learning children. This plausibility would be further enhanced if the relative emergence of word classes displayed by the mechanism corresponded to that found in language-learning children. Several authors have argued that children show evidence of a productive noun category before a productive verb category. The relative emergence of noun and verb classes was investigated using a measure of noun richness. Noun richness is a measure of the relative size of the noun and verb category and is computed by dividing the number of noun-noun links over the number of noun-noun plus verb-verb links. A model that shows early emergence of a noun category and late emergence of a verb category would thus show decreasing noun richness. Table 4 shows the development of noun richness for the two models reported in Table 3 – the Jaccard distance at a threshold of 0.2 and the rank order correlation at a threshold of 0.45. For completeness, noun richness for the rank order measure for concatenated contexts (at a threshold of 0.6, data from Table 1) are reported as well. Table 5 lists the number of noun-noun and verb-verb links) for the Jaccard distance.

Table 4: Noun richness scores for Jaccard distance, separated rank order and concatenated rank order.

Runs	Jaccard, 0.2	Rank Order, 0.45	Concatened Rank Order, 0.6
36	0.72	0.68	0.87
38	0.80	0.80	0.89
40	0.80	0.82	0.88
44	0.64	0.71	0.87
50	0.57	0.65	0.82
All	0.57	0.68	0.81

Table 5: Number of noun-noun and verb-verb links for Jaccard distance

Runs	Nouns	Verbs
36	13.75	3.58
38	93.33	16.92
40	254.83	53.67
44	353.5	191.25
50	357.75	256.92
All	364.42	266.92

As can be seen in Table 4, the measures show the highest noun richness score for the second or third developmental phase (run 38/40). Noun richness scores subsequently decrease for all three measures. This decrease is smallest for the concatenated rank order (0.07), intermediate for the separated rank order (0.17) and largest for the Jaccard distance (0.23). Table 5 lists the number of noun-noun and verb-verb links for the Jaccard distance and shows that both the number of noun and the number of verb links increase with development, but verb links are added at a greater rate in the later stages. Results of these analyses thus confirm that computing similarity separately for preceding and following contexts aids verb learning, and leads to a more plausible developmental pattern of noun and verb linkage. This pattern is most pronounced for the computationally simplest measure - the Jaccard distance.

Conclusions

The main conclusion to be drawn from the analyses reported here is that Redington et al.'s (1998) approach can be adapted in such a way that it provides a developmentally plausible account of how children build grammatical word classes on the basis of distributional information. Thus, the mechanism is able to yield developmentally plausible results when the length of the utterances over which it computes statistics is gradually increased. In addition, it was shown that, when statistics are computed separately for preceding and following contexts, the mechanism is able to reach high levels of accuracy for both nouns and verbs, even in early stages of development when few contexts may be available. When similarity is computed for joint (concatenated) preceding and following contexts, the mechanism is too liberal in the early stages, when it tends to link items based on just preceding or following contexts.

It was also shown that disregarding token frequencies of context words substantially reduces the computational overhead of the mechanism without affecting its performance. Thus, rather than representing contexts as vectors of counts for context words, the mechanism simply represents a list of word types that have appeared in preceding and following position. This latter finding is significant, as it has been suggested that the computational complexity of the original mechanism makes it implausible as a mechanism used by language-learning children.

Finally, it was shown that the mechanism can simulate a plausible pattern of noun and verb linkage, with a noun category emerging ahead of a verb category, as is evident in language-learning children. This pattern was least pronounced in the variant that employed concatenated vectors, and most pronounced for the simplest variant (Jaccard distance), thus providing additional evidence that computing similarity separately for preceding and following contexts is particularly helpful for classifying verbs.

Taken together, the analyses reported here suggest that distributional analysis can be a viable approach to the learning of word classes from the earliest stages of development. Thus, while Redington et al. applied their method to complete utterances and corpora containing millions of word tokens, current analyses suggest that it's possible to accurately classify words on the basis of just a few contexts, provided these are drawn from both the preceding and following position. That is: two words that have overlap in both the preceding and following contexts are very likely to be of the same category, particularly if these are the only known contexts for these words.

The current approach could be argued to combine the best features of Mintz (2003) and Redington et al.'s (1998) approaches. The great strength of Mintz's approach is its ability to classify (large numbers of) items on the basis of very little information (meaning it can be plausibly used by language-learning children), while the probabilistic nature of Redington et al.'s approach means it is naturally resistant to noise, and thus likely to be more accurate. The analyses reported here suggest that a system of intermediate complexity can quickly begin to classify items with high accuracy on the basis of a small number of contexts and gradually expand these contexts as they become available. This necessarily means that the number of items classified in the early stages will be considerably lower than that classified by Mintz's approach. However, while the great strength of Mintz's approach is its ability to classify large numbers of items on the basis of very little information, it is typically applied to corpus-wide statistics and not subject to the developmental manipulations that were studied here, and which provide support for the amended mechanism of Redington et al. Freudenthal et al. (2013) also suggest that the number of classified items may be increased by including utterance boundaries as framing elements.

While disregarding token frequencies for context words makes the mechanism considerably simpler (and hence more psychologically plausible), it retains many of the desirable properties of Redington et al.'s approach. Since similarity is expressed probabilistically, it is naturally resistant to noise and appears better placed to deal with languages whose word order is more flexible than English.

The results presented here also illustrate the value of evaluating distributional learning mechanisms against developmental criteria. Using an established model of language acquisition we show that gradually expanding the contexts over which similarity is computed can result in a developmentally plausible pattern of noun and verb linkage with high accuracy, even when frequency information is discarded. However, the fact that few contexts are available early in development means that distributional statistics can become overly liberal unless similarity in preceding and following contexts is considered separately, a finding, that is likely to remain hidden in analyses that compute statistics over complete utterances and large corpora.

Acknowledgements

Daniel Freudenthal, Julian Pine, and Fernand Gobet are members of the International Centre for Language and Communicative Development (LuCiD) at the University of Liverpool, for which support of the Economic and Social Research Council [ES/L008955/1] is gratefully acknowledged. This research was supported by ESRC Grant ES/J011436/1.

References

Akhtar, N., & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology, 33*, 952-965.

Erkelens, M. A. (2009). *Learning to categorize verbs and nouns.* Unpublished PhD Thesis, Universiteit van Amsterdam, Amsterdam.

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

Freudenthal, D., Pine, J. M. & Gobet, F. (2010). Explaining quantitative variation in the rate of Optional Infinitive errors across languages: A comparison of MOSAIC and the Variational Learning Model. *Journal of Child Language*, *37*, 643-669.

Freudenthal, D., Pine, J.M., Jones, G. & Gobet, F. (2013): Frequent frames, flexible frames and the noun-verb asymmetry. In: M. Knauf, M. Pauen, N. Sebanz E I. Wachsmuth (Eds.), *Proceedings of the 35th annual meeting of the Cognitive Science Society*. (pp. 2327-2332). Austin, TX: Cognitive Science Society.

Freudenthal, D., Pine, J.M., Jones, G. & Gobet. F. (2015). Simulating the cross-linguistic pattern of Optional Infinitive errors in children's declaratives and Wh-questions. *Cognition*, *143*, 61-76.

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

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

Olguin, R., & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development*, 8, 245-272.

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

St. Clair, M.C. Monaghan, P., & Christiansen, M.H. (2010). Learning grammatical categories from distributional cues. Flexible frames for language acquisition. *Cognition*, *116*, 341-360.

Stumper, B., Bannard, C., Lieven, E., & Tomasello, M. (2011). Frequent frames in German child-directed speech: A limited cue to grammatical categories. Cognitive Science, 35, 1190-1205.

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.

Tomasello, M., & Olguin, R. (1993). Twenty-three-month-old children have a grammatical category of noun. *Cognitive Development*, *8*, 451-464.