Distributional learning and lexical category acquisition: What makes words easy to categorize?

Giovanni Cassani, Robert Grimm, Steven Gillis, and **Walter Daelemans**
Computational Linguistics and Psycholinguistics (CLiPS) Research Center
Department of Linguistics, University of Antwerp, 13 Prinsstraat
B-2000 Antwerpen, Belgium
{name.surname}@uantwerpen.be

Abstract

In this study, results of computational simulations on English child-directed speech are presented to uncover what distributional properties of words make it easier to group them into lexical categories. This analysis provides evidence that words are easier to categorize when (i) they are hard to predict given the contexts they occur in; (ii) they occur in few different contexts; and (iii) their contextual distributions have a low entropy, meaning that they tend to occur more often in one of the contexts they occur in. This profile fits that of content words, especially nouns and verbs, which is consistent with developmental evidence showing that children learning English start by forming a noun and a verb category. These results further characterize the role of distributional information in lexical category acquisition and confirm that it is a robust, reliable, and developmentally plausible source to learn lexical categories.

Keywords: Distributional bootstrapping; Lexical category acquisition; Statistical learning; Computational psycholinguistics; Language acquisition

Introduction

Distributional bootstrapping (Maratsos & Chalkley, 1980) is an influential account of how children start breaking into language, and specifically of how they start grouping words into lexical categories such as nouns and verbs. More specifically, it claims that children use patterns of co-occurrences across linguistic units, such as words and morphemes, to group words that share similar contexts. Several computational simulations have shown that distributional information is a rich, useful, and usable source of knowledge about lexical categories (Mintz, 2003; Redington, Chater, & Finch, 1998; St. Clair, Monaghan, & Christiansen, 2010). Moreover, a number of behavioral experiments have confirmed that children use this information to group words together (Mintz, Wang, & Li, 2014; Reeder, Newport, & Aslin, 2013).

Research on distributional bootstrapping has mostly focused on investigating which contexts constitute the best cues for the acquisition of lexical categories. Several proposals that have been put forward share the approach of grouping together those words that share similar contexts of occurrence, but differ in the starting assumptions and the types of contexts they evaluate. For example, Mintz (2003) suggested that *frequent frames*, i.e. trigrams consisting of two words flanking an empty slot (*a_X_b*), are a psychologically plausible and highly effective type of context for acquiring lexical categories. St. Clair et al. (2010), on the contrary, provided evidence that better categorization can be achieved by using bigrams (*a_X* + *X_b*) that can be readily combined to obtain trigram level information.

This paper aims to explore distributional bootstrapping further and uses computational simulations to answer the following research question: what distributional properties of words make it easier to categorize them on the basis of the contexts they co-occur with? The relation between distributional properties of words and the extent to which these can be easily categorized in terms of lexical categories has been largely neglected in previous research, but characterizing it is important for two main reasons. Firstly, it generates predictions about the effect that several distributional properties of words have on lexical category acquisition in English speaking children: testing them can shed further light on the plausibility of distributional learning as an underlying mechanism for lexical category acquisition. Importantly, it is not enough that a model behaves like humans: a statistical analysis of what drives the model's behavior is necessary to assess whether it is driven by the same factors that affect human behavior. Secondly, it can help to constrain the development of psychologically motivated models of lexical category acquisition, by showing what information children are sensitive to when solving the task of grouping words into lexical categories.

In this work, computational simulations are used to carry out a categorization experiment whose outcome is used as the dependent variable in a regression analysis aimed to uncover the effect of several distributional properties of words on categorization accuracy. Results shed light and generate predictions on the mechanisms underlying distributional learning of lexical categories, and ultimately provide information to guide and constrain the development of psychologically motivated models of bootstrapping in language acquisition.

Methods

Corpora and pre-processing

In order to perform the computational simulations, transcribed interactions involving children and caretakers available in the CHILDES database (MacWhinney, 2000) were used. More specifically, the Manchester corpus (Theakston, Lieven, & Pine, 2001) from the British English part, and the Suppes corpus (Suppes, 1974) from the American English part were selected, since they have both been widely used in previous research on distributional bootstrapping of lexical categories. The Suppes corpus consists of transcripts of one child, Nina, recorded from 1;11 to 3;3, while the Manchester corpus contains data of 12 children, recorded for varying periods within the age range 1;8 to 3;0. Both come with an au-

tomatic categorization in terms of Part-of-Speech (PoS) tags, which can be accessed on the MOR tier of the CHILDES annotation scheme. The child-directed speech from the corpora was pre-processed to deal with some aspects of the transcriptions. Two dummy symbols, *#start* and *#end*, were inserted at the beginning and end of each utterance. This manipulation is motivated by evidence that sentence boundaries provide useful distributional information (Freudenthal, Pine, & Gobet, 2008). It also allows us to exploit every utterance from the corpus, including single word utterances: words occurring in isolation are considered to be occurring in the bigrams *#start_X* and *X_#end*, and in the trigram *#start_X_#end*.

Corpora from individual children were processed separately using a sliding window approach: starting from the first lexical element of the utterance, each word was considered as target, and all bigrams and trigrams occurring next to it were collected. These types of contexts were chosen given that they have been widely explored in previous research (Mintz, 2003; Monaghan & Christiansen, 2008; St. Clair et al., 2010). As an example, consider the following utterance from the Manchester corpus: *#start are~v you~n going~v to~funct put~v that~adv one~n inside~adv? #end*. The first target word, *are*, occurs in two bigrams, *#start_X* and *X_you~n*, and two trigrams, *#start_X_you~n* and *X_you~n_going~v*. For words in the middle of the utterance, three trigrams are available. The tags after the tilde indicate the lexical category to which each word belongs according to the automatic categorization. The original categories were collapsed to a coarser set, consisting of five categories: nouns (*n*), including pronouns; verbs (*v*), including auxiliaries, copulas, and non-finite forms[1]; adjectives (*adj*), adverbs (*adv*), and function words (*funct*). The idea is to zoom in on the open classes, conflating the closed class words in a single category given that function words are categorized later in development. No lemmatization is performed, and all information about lexical categories is preserved[2], although it is only used to evaluate whether categorization has been successful.

In order to minimize both the number of assumptions and that of possible decisions in the design of the experiment, all bigrams and trigrams are considered: some will turn out to be more informative to the categorization task than others, but the analysis of this aspect of the problem falls outside of the scope of this study. Larger *n*-grams are not considered due to the limited size of the corpora: they would be too infrequent to affect categorization.

Experimental setting

A categorization experiment was carried out, in which words were clustered together based on the similarity of the contexts in which they occurred in corpora of English child-directed speech (Redington et al., 1998). Words that tend to occur in the same contexts are considered to be more similar and clustered together: target words are categorized correctly if they are assigned the correct lexical category by the computational simulation. The experiment was performed using Memory-Based Learning (MBL, (Daelemans & van den Bosch, 2005)), a class of machine learning algorithms which implements an exemplar-based strategy and categorizes new items using retrieval of or similarity to items stored in memory, with no explicit abstraction.

The categorization experiment consists of two main phases, which are referred to as *training* and *testing* in the paper. During training, co-occurrence counts between target words and contexts are collected on a portion of the input data and stored in memory. Each word is represented as a vector of counts, with each count indicating the co-occurrence frequency of the corresponding word and context. During testing, a new portion of the input is considered and the same procedure is applied. At the end of this second stage, the learner has created two matrices of co-occurrence counts. Each word from the test matrix is categorized by comparing its vector of co-occurrences with all the vectors from the training matrix, looking for the most similar one; the two are then clustered together. During learning, the model has no access to the correct lexical categories of the words and only groups them together based on their co-occurrence patterns, in an unsupervised way. At the end of the process, the category of two words that were clustered together is inspected: if they share the same lexical category, the word from the test set has been categorized correctly. In this framework, the only factor driving clustering is similarity, which is a well-documented cognitive mechanism in categorization (Sloutsky, 2003).

In order to divide each individual corpus into a training and a test set, utterances of child-directed speech were ordered chronologically and split in two parts: (i) the first 70% of the utterances were allocated for training; and (ii) the last 30% of the utterances were used as test set. To evaluate how different distributional properties interact with time, operationalized as a larger exposure to the input language, an incremental training approach was implemented. In detail, training started on the first 40% of all the utterances, then proceeded on the first 45%, always increasing by 5 percentage points, up to the full training set (70% of the total utterances). The test set was kept constant to make sure that any change in performance came from the knowledge inferred from the training set and not by differences in the test set.

The TiMBL package (Daelemans, Zavrel, van der Sloot, & van den Bosch, 2009) was used to carry out the simulation, using the default IB1 algorithm (Aha, Kibler, & Albert, 1991) and cosine as a distance metric, because of its robustness to different frequencies in the co-occurrence vectors, and setting the number of nearest neighbors to 1. Moreover, no feature weighting based on co-occurrence statistics from the training corpus was applied during the categorization experiment: this allows us to perform the categorization experiment without weighting contexts according to their informativity, avoiding

[1] Results from Mintz (2003) show that merging pronouns with nouns, and auxiliaries, copulas, and non-finite forms with verbs does not bias categorization results.

[2] *X_dog~n* and *X_dogs~n* are different contexts, just as *light~n_X*, *light~v_X*, and *light~adj_X*

the effect of supervision on classification, which would be psychologically questionable and bias the results.

Importantly, no claim is put forward that children actually keep track of all available bigrams and trigrams, or that they implement an analogue of the IB1 algorithm with the chosen parameter setting. The interest of the current analysis is purely in the information that supports learning and in the analysis of the effects that distributional properties of words have on categorization, as operationalized using MBL.

Statistical analysis

Four pieces of distributional information were computed for each word on the test set (last 30% of utterances of each corpus) and used as predictors in a regression model:

Token frequency: the log-transformed frequency count of each token. The transformation is motivated by evidence from Keuleers, Diependaele, and Brysbaert (2010) that lexical frequency effects are better captured by log-transformed frequency counts. A positive effect of frequency is expected (Ambridge, Kidd, Rowland, & Theakston, 2015), since more frequent items are typically learned better than less frequent ones.

Contextual diversity: the log-transformed count of how many different contexts a word occurs in. A negative effect for contextual diversity is predicted: if a word occurs in many different contexts, its co-occurrence vector is noisy and it is harder to reliably group it with other words. This is the case, e.g., of function words, like conjunctions and determiners: they occur in all sort of contexts, making it hard to group them with similar words.

Average conditional probability: the average conditional probability of a word given all the contexts it occurs with. Consider a toy example where the context *the_X* occurs 100 times, 15 of which with the word *cat*: $p(cat|the\_X)$, is thus 0.15. Assume also that the word *cat* occurs 40 times in the context *a_X*, which in turn occurs 200 times: $p(cat|a\_X)$ is 0.2. In order to obtain the average conditional probability for the word *cat*, $p(cat|the\_X)$ and $p(cat|a\_X)$ are averaged, yielding 0.175. This independent variable is predicted to have a negative effect on categorization: high conditional probability means that the contexts in which a target word occurs do not occur with other words, making it hard to find shared contexts of occurrence between the target and other words.

Entropy: the entropy of the co-occurrence vector of a word (Shannon, 1948), normalized by the number of contexts it occurs with, so that entropy lies between 0 and 1. The entropy of a word is low when it occurs in the same context the majority of the times, while the more even the distribution of co-occurrences for a word, the higher its entropy. Entropy relates to diversity and its effect should go in the same direction: the more a word occurs equally frequently in the contexts it co-occurs in, the noisier its co-occurrence vector and the harder it is to correctly group it with similar words. Importantly, normalized entropy provides a related but different piece of information than contextual diversity: the normalization ensures that the number of different contexts a word occurs in does not affect entropy.

A further independent variable was considered for both words and contexts, i.e. *time*, operationalized as the amount of training input on which the computational simulations were trained: time goes from 0 (i.e. 40% of all utterances in the corpus used as training set) to 6 (70% of all utterances in the corpus used as training set). Time should have a positive effect, since exposing the model to more input language should provide more reliable and robust information about co-occurrence patterns.

The analysis was restricted on words that appeared in all 13 individual corpora (12 from the Manchester corpus and 1 from the Suppes corpus), to reduce the effect of idiosyncrasies and focus on general patterns. All words with a token frequency of 1 were also excluded from the analysis, because when this is the case, contextual diversity and entropy are fully determined. If a word occurred only once, then it also occurred in only one context (diversity of 1), and its entropy is 0, because the full probability mass is on the only context the word occurred in.

In order to analyze how easy it is to categorize a word, logistic mixed-effects models (Baayen, Davidson, & Bates, 2008) were fitted using the "lme4" package in R (D. Bates, Maechler, Bolker, & Walker, 2015). Random intercepts for corpus (13 levels) and word (456 levels, i.e. the single words that survived the filtering steps just detailed) were included. The categorization outcome of each word was used as a binary dependent variable, with each correctly categorized word coded as 1. Covariates were included in a step-wise fashion, according to the improvement in fit measured by the Akaike Information Criterion (AIC, (Akaike, 1973)).

Results

The best converging logistic mixed-effects model included main effects for average conditional probability, entropy, time, and contextual diversity. Adding a main effect for token frequency resulted in the model not converging. Two-way interactions between time and conditional probability, entropy, and lexical diversity were tested; however, when these were entered, the model did not converge. Table 1 provides the βs estimated for this model, expressed on the log-odds scale, while Figure 1 represents the effects graphically, with accuracy expressed as proportion. The final model resulted in a marginal R^2 of 0.055 and in a conditional R^2 of 0.913, suggesting that while the effect of predictors is significant, they do not explain much variance in the data. This is further addressed in the discussion.

As predicted, the average conditional probability of a word given the contexts in which it occurs has a strong negative effect on the estimated accuracy ($\beta = -12.17, t = -11.56, p < 0.001$), and the same is true for the entropy of the distribu-

Table 1: Mixed-effects model fitted to analyze what distributional properties make words easier to categorize. Estimates (Est.) and standard errors (Std. Err.) are provided on the log-odds scale. (*Cond. Prob.*: average conditional probability of words given contexts; *Cont. Div.*: contextual diversity.

Ind. Vars.	Est.	Std. Err.	z	p val.
(Intercept)	14.185	1.298	10.928	$< .001$
Cond. Prob.	-12.170	1.053	-11.560	$< .001$
Entropy	-11.027	1.215	-9.077	$< .001$
Time	0.078	0.011	6.838	$< .001$
Cont. Div.	-0.893	0.255	-3.509	$< .001$

tion of co-occurrence counts of a word over all the contexts it occurs in ($\beta = -11.027, t = -9.077, p < 0.001$). Time has a significantly positive effect ($\beta = 0.078, t = 6.838, p < 0.001$), showing that the clustering algorithm is actually exploiting the larger amount of input language to better group similar words together. Finally, contextual diversity has a significant negative effect ($\beta = -0.893, t = -3.509, p < 0.001$), suggesting that words are easier to categorize when they occur in fewer contexts, matching the initial hypothesis. As it was reported, adding frequency resulted in convergence issues: this is most likely due to the filtering step. It is possible that surviving words had similar frequency counts, making it impossible for the model to find sufficient variation to estimate the effect of token frequency on categorization accuracy, once contextual diversity already entered the model (since it improved the fit more than token frequency).

Discussion

The results that have been presented point to a relation between distributional properties of words and the degree to which it is easy to categorize them into lexical category. The easiest words appear to (i) be on average hard to predict given the contexts in which they occur; (ii) have a very skewed distribution of co-occurrence counts with the contexts they occur in, meaning that they tend to occur most often in one or few contexts; and (iii) tend to generally occur in few contexts.

First, being able to predict a word given the contexts it occurs in is detrimental to categorization. This entails that effective categorization depends on some uncertainty in the co-occurrence patterns of words and contexts. Since categorization works on similarity (Sloutsky, 2003), two words can only be grouped together if they occur in the same context, i.e. they have something in common. The negative effect of conditional probability of words given contexts also points to a feature that contexts should have in order to be useful and usable, namely that they need to occur with more than one word. As a matter of fact, the conditional probability of words given contexts is computed by dividing the co-occurrence count of the word and the context by the frequency count of the context itself. For the average conditional probability of word given context to be low, each context must occur with other words

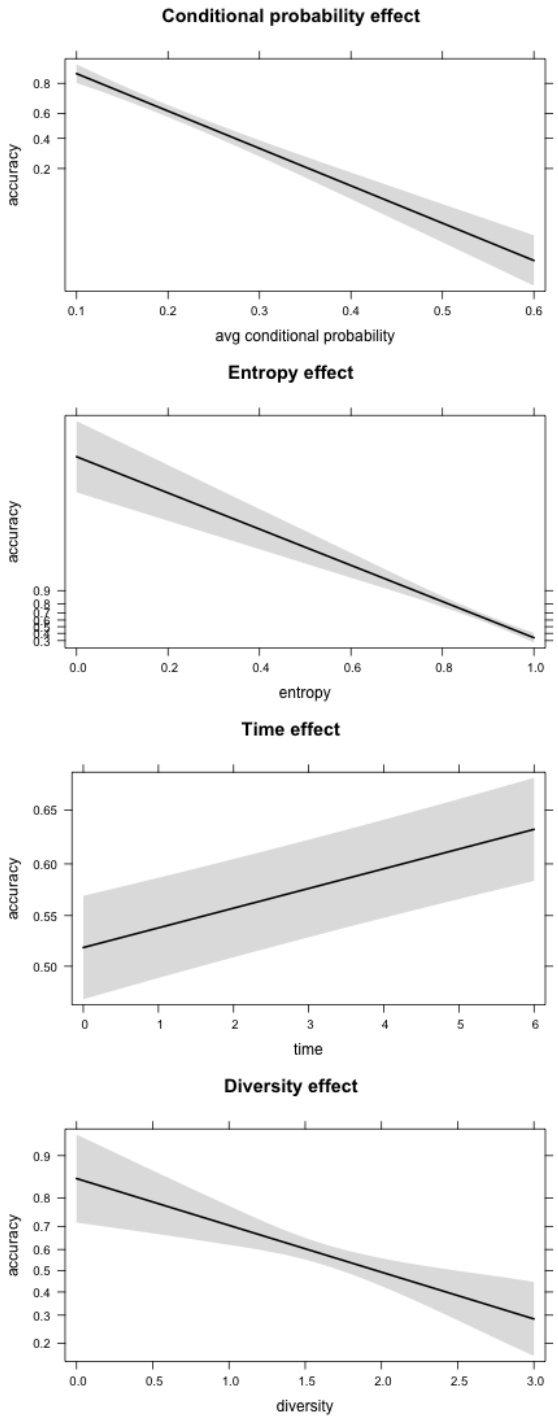

Figure 1: Main effects, with confidence bands, of average conditional probability, entropy, time, and contextual diversity on how easy it is to categorize a word in terms of lexical categories. The order, from top to bottom, reflects the improvement in fit brought by each predictor. The y-axis represents probabilities estimated from the log-odds reported in Table 1. Each axis is automatically scaled to provide a clear depiction of the effect. The plots were obtained using the *effects* package in R (Fox, 2003).

a substantial amount of times. This hypothesis fits evidence provided by Matthews and Bannard (2010) that children find it easier to group words together when these occur in contexts that, in turn, occur with several different words.

The negative coefficients of entropy and contextual diversity complement the negative effect of average conditional probability: the latter indicates that words are easier to categorize when they tend to occur in contexts only a fraction of the times the contexts themselves occur. βs for normalized entropy and contextual diversity, on the contrary, tell that words are easier to categorize when they tend to occur most often in one or few contexts. The ideal situation is thus that of a word that always and only occurs in a single context, which however occurs with many other words, which also only occur in that context (to reduce noise). The effects of entropy and contextual diversity indicate that uncertainty in word-context co-occurrence patterns is necessary at the context level but detrimental at the word level: words need to occur in few contexts for effective categorization. This is likely due to the fact that when contextual diversity and entropy are high, the co-occurrence pattern of a word can be very noisy.

The distributional properties that make a word easier to categorize are rather distinctive of content words, especially nouns: knowing a context, e.g. a determiner, it is hard to predict exactly which noun will appear next to it, because many different nouns (and some adjectives) are possible, which translates into a low conditional probability of words given contexts. Moreover, it is likely that a noun occurs with one of the few determiners or possessive pronouns of the English language, thus scoring low on contextual diversity, and that most of the times it occurs with just a couple of specific determiners or possessive pronouns, scoring low on entropy. In order to get a grasp of which lexical categories easier words belonged to, those words that were categorized correctly for at least 80% of the 13 individual corpora at the last stage of training were selected. This analysis highlighted 127 such words: 2 function words, 101 nouns, and 24 verbs. This shows that the distributional properties of words that make them easier to categorize strongly correlate with lexical categories, and that the same features are a possible candidate to explain why certain lexical categories are formed earlier than others[3]. Furthermore, the majority of the 51 words that are never categorized correctly predominantly consists of function words (26) and adverbs (18), the categories that are learned later in development (E. Bates, Dale, & Thal, 1995). The observation that nouns are categorized best also relates to the observation that children form a productive noun category earlier than any other category (Tomasello, 2000). The reported evidence lends support to the hypothesis that the so-called noun bias can be traced back to the distributional properties of words belonging to different lexical categories (Cassani, Grimm, Daelemans, & Gillis, submitted), showing that regardless of the fact that the set of target words contained an equal number of nouns and verbs, noun categorization is more effective.

The reported evidence also parallels and complements results about word learning, which suggest children find it easier to learn words (particularly nouns) when they occur in a variety of different contexts (Hills, Maouene, Riordan, & Smith, 2010). While a comprehensive experiment is still lacking that explicitly contrasts the effect of contextual diversity on word learning and categorization, it emerges that this factor impacts both phenomena, although in opposite directions. While a higher contextual diversity is beneficial for word learning, it is detrimental to word categorization, as appears from the statistical analysis reported here. Further research about the interplay between different frequency effects (Ambridge et al., 2015) is needed to clarify to what extent distributional learning drives and explains language acquisition in its many different aspects and sub-tasks.

Lastly, this study investigated a fully distributional explanation of the developmental pattern of lexical category acquisition. However, the low R^2 shows that the distributional properties we investigated leave a substantial portion of variance unexplained, calling for further research on which properties affected the machine learner and whether these also influence children during lexical category acquisition. Moreover, current research has highlighted the importance of other sources of information during lexical category acquisition and word learning (Roy, Frank, DeCamp, Miller, & Roy, 2015), including morphology, phonetics, semantics and prosody (Monaghan & Christiansen, 2008). The influence of these sources of information should be further analyzed to complement research on distributional bootstrapping.

Summarizing, this study provided evidence about the effect of different distributional properties of words on the acquisition of lexical categories from distributional information. Conditional probability, entropy, and contextual diversity have a negative effect on categorization accuracy. Words with these features tend to be content words, mostly nouns, which also appear to be the words children start grouping earlier and most effectively. Future studies should assess the cross-linguistic validity of these findings, to understand whether the same distributional properties have similar effects in typologically different languages. Moreover, a similar approach — performing statistical analysis on the outcome of computational simulations — could be used to investigate what distributional properties make contexts more useful. Finally, other computational models should be tested, to compare their outcome to developmental data and shed light on which architectures are closer to what children actually do.

Conclusion

The evidence presented in this study shows that specific distributional properties of words determine how easy it is to cluster them together based on the similarity of their co-occurrence patterns. In detail, words are easier to categorize

[3]The bias towards nouns and verbs in categorization does not result from an imbalance in the set of target words, consisting of 40 adjectives, 47 adverbs, 76 function words, 145 nouns, and 148 verbs.

(i) when they are hard to predict given the contexts they occur in, (ii) when they generally occur in few contexts, and (iii) when they tend to occur more often in one context, having low entropy. This study extends previous research on distributional bootstrapping by providing evidence that distributional properties also affect which words are categorized more easily and which lexical categories are formed earlier.

Acknowledgments

This research was supported by a BOF/TOP grant (ID 29072) of the Research Council of the University of Antwerp.

References

Aha, D. W., Kibler, D., & Albert, M. K. (1991). Instance-based learning algorithms. *Mach Learn*, *6*(1), 37-66.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov & F. Csáki (Eds.), *2nd International Symposium on Information Theory, Tsahkadsor, Armenia, USSR*. Budapest, Hungary: Akadémiai Kiadó.

Ambridge, B., Kidd, E., Rowland, C. F., & Theakston, A. L. (2015). The ubiquity of frequency effects in first language acquisition. *J Child Lang*, *42*(2), 239-273.

Baayen, H. R., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *J Mem Lang*, *59*(4), 390-412.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J Stat Softw*, *67*(1), 1-48.

Bates, E., Dale, P. S., & Thal, D. (1995). Individual differences and their implications for theories of language development. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language* (p. 96-151). Oxford, UK: Blackwell.

Cassani, G., Grimm, R., Daelemans, W., & Gillis, S. (submitted). Distributional bootstrapping and the noun bias: Zooming in on developmental plausibility.

Daelemans, W., & van den Bosch, A. (2005). *Memory-based language processing*. Cambridge, UK: Cambridge University Press.

Daelemans, W., Zavrel, J., van der Sloot, K., & van den Bosch, A. (2009, 31 May 2010). *Timbl: Tilburg Memory Based Learner, version 6.3. reference guide.* (Tech. Rep.). Tilburg University.

Fox, J. (2003). Effect displays in r for generalised linear models. *J Stat Softw*, *8*(15).

Freudenthal, D., Pine, J. M., & Gobet, F. (2008). On the utility of conjoint and compositional frames and utterance boundaries as predictors of word categories. In V. Sloutsky, B. Love, & K. McRae (Eds.), *Proceedings of the 30th annual meeting of the Cognitive Science Society* (p. 1947-1952). Austin, TX: Cognitive Science Society.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *J Mem Lang*, *63*(3), 259-273.

Keuleers, E., Diependaele, K., & Brysbaert, M. (2010). Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords. *Front Psychol*, *1*, 174.

MacWhinney, B. J. (2000). *The childes project: Tools for analyzing talk. the database.* (3rd ed., Vol. 2). Mahwah, NJ: Lawrence Erlbaum Associates.

Maratsos, M. P., & Chalkley, M. A. (1980). The internal language of children syntax: The nature and ontogenesis of syntactic categories. In K. E. Nelson (Ed.), *Children's language* (Vol. 2, chap. 2). New York, NY: Gardner Press.

Matthews, D., & Bannard, C. (2010). Children's production of unfamiliar word sequences is predicted by positional variability and latent classes in a large sample of child-directed speech. *Cognitive science*, *34*(3), 465-488.

Mintz, T. H. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*(1), 91-117.

Mintz, T. H., Wang, F. H., & Li, J. (2014). Word categorization from distributional information: Frames confer more than the sum of their (bigram) parts. *Cognitive psychology*, *75*, 1-27.

Monaghan, P., & Christiansen, M. H. (2008). Integration of multiple probabilistic cues in syntax acquisition. In H. Behrens (Ed.), *Corpora in language acquisition research: History, methods, perspectives* (Vol. 6, p. 139-164). Amsterdam: John Benjamins Publishing.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive science*, *22*(4), 425-469.

Reeder, P. A., Newport, E. L., & Aslin, R. N. (2013). From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive psychology*, *66*(1), 30-54.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. K. (2015). Predicting the birth of a spoken word. *Proc Natl Acad Sci U S A*, *112*(41), 12663-8.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*, 379-423.

Sloutsky, V. M. (2003). The role of similarity in the development of categorization. *Trends in cognitive sciences*, *7*(6), 246-251.

St. Clair, M. C., Monaghan, P., & Christiansen, M. H. (2010). Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, *116*(3), 341-360.

Suppes, P. (1974). The semantics of children's language. *Am Psychol*, *29*(2), 103-114.

Theakston, A. L., Lieven, E. V. M., & Pine, J. M. (2001). The role of performance limitations in the acquisition of "mixed" verb-argument structure at stage i. *J Child Lang*, *28*(1), 127-152.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, *74*(3), 209-253.