Which distributional cues help the most? Unsupervised context selection for lexical category acquisition

Reference:

Cassani, Grimm Robert, Daelemans Walter, Gillis Steven.- Which distributional cues help the most? Unsupervised context selection for lexical category acquisition

Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning / Berwick, Robert [edit.]; et al. - ISBN 978-1-941643-32-7 - Association for Computational Linguistics, 2015, p. 33-39

Which distributional cues help the most? Unsupervised contexts selection for lexical category acquisition

Giovanni Cassani **Robert Grimm** **Walter Daelemans** **Steven Gillis**
University of Antwerp, CLiPS
{name.surname}@uantwerpen.be

Abstract

Starting from the distributional bootstrapping hypothesis, we propose an unsupervised model that selects the most useful distributional information according to its salience in the input, incorporating psycholinguistic evidence. With a supervised Parts-of-Speech tagging experiment, we provide preliminary results suggesting that the distributional contexts extracted by our model yield similar performances as compared to current approaches from the literature, with a gain in psychological plausibility. We also introduce a more principled way to evaluate the effectiveness of distributional contexts in helping learners to group words in syntactic categories.

1 Introduction and related work

The psycholinguistic research about language acquisition has long been concerned with how children crack the linguistic input to infer the underlying structures. In this respect, bootstrapping (Gillis and Ravid, 2009) has been an important concept, which generated a number of hypotheses. After semantic bootstrapping, introduced by Pinker (1984), other proposals were put forward, each strengthening one aspect as the starting level that informs the others (syntactic bootstrapping (Gleitman, 1990; Gleitman and Gillette, 1995), prosodic bootstrapping (Christophe et al., 2008), distributional bootstrapping (Maratsos and Chalkley, 1980; Mintz, 2003)). This debate is tightly interwoven with the more general controversy between a nativist (Chomsky, 1965) and an emergentist account (Bates and MacWhinney, 1987; MacWhinney, 1998; Tomasello, 2000): our work was set up to explore the possibility of learning useful linguistic information from the Primary Linguistic Data (PLD), only using general-purpose learning mechanisms. Thus, we look at language acquisition from an emergentist perspective, exploring the fruitfulness of the distributional bootstrapping hypothesis.

Starting with Cartwright and Brent (1997), a variety of models for Parts-of-Speech (PoS) induction has been proposed (Clark, 2000; Mintz et al., 2002; Mintz, 2003; Parisien et al., 2008; Leibbrandt, 2009; Chrupała and Alishahi, 2010; St. Clair et al., 2010), showing that PLD are rich enough in distributional cues to provide the child with enough information to group words according to their syntactic category. Among such models, two major approaches can be identified: i) a frame-based one which starts by selecting the relevant cues and then evaluate how these help categorization, and ii) a probabilistic approach that considers all possible contexts in a left and right window whose size is set in advance, and determines the best category for each word based on a probabilistic match between the context of each new word and the previously encountered contexts for all words. While the first approach has been more concerned with finding the right cues or the most useful type of context (Monaghan and Christiansen, 2004), usually by focusing on certain distributional patterns and assessing their effectiveness in inducing lexical categories, the second one has tackled the problem from a more global perspective, inducing categories – not necessarily syntactic – and evaluating them using other linguistic tasks (Frank et al., 2008).

The first approach has been more influential in the acquisition literature, and is the topic of active behavioral research with both adults (Reeder et al., 2013; Mintz et al., 2014) and infants (Zhang et al., 2015). The second approach has been more distinctive of the computational psycholinguistic literature, but has been largely neglected by the acquisition literature. In this short paper, we try

to suggest that the approach and the methods used in the second stream of research can be applied to the first, not only to induce plausible categories, but also a set of cues, without focusing on a specific kind of distributional pattern which is set in advance using linguistic knowledge. In this respect, we will review some of the major problems of the frame-based approach before suggesting a first way of tackling them.

In his seminal paper, Mintz (2003) suggested that the 45 most frequent A_x_B frames, defined as two words flanking a variable slot, are a plausible and accurate type of information – see also Wang and Mintz (2007) for an incremental model. This hypothesis was further tested on French by Chemla et al. (2009) with good success; however, its cross-linguistic validity was challenged by Erkelens (2009) for Dutch and Stumper et al. (2011) for German[1]. More recently, the *frequent frames* hypothesis was challenged by St. Clair et al. (2010), who proposed to use *flexible frames*, i.e. left and right bi-grams defined through the 45 most frequent words in a corpus, that can be combined on the fly to provide tri-gram level information – but see Mintz et al. (2014).

The main problem we see in both frequent and flexible frames, is the arbitrariness in deciding which contexts are important (Leibbrandt, 2009). While frequency drives the decision, what makes A_x_B (or A_x + x_B) frames so special that the child commits to them to infer lexical categories?

Moreover, restricting to token frequency can lead to retain contexts that do not help categorization, since they only occur with one word (like the frequent frame *have_X_look*), which in turn causes the model to not scale well to unseen data. Where the goal is explicitly to deal with reduced computational capacities, such behavior is far from desirable since it stores information that does not help to group words in more abstract categories.

A further problem of frequent frames, at least with English, is a strong verb bias: such cues provide information about a greater number of verbs, while the PLD typically contain many more nouns than verbs. This bias is a by-product of the definition of frames as fully lexical contexts: the shortest sentence from which a frame can be derived consists of three words, where the medial slot is usually taken up by a verb.

[1]However, better results were obtained with frames defined at the morpheme level, rather than at the word level (Wang et al., 2011).

At the same time, flexible frames suffer from other problems. Behavioral evidence suggests that children and adults store longer sequences as units (Bannard and Matthews, 2008; Arnon and Clark, 2011)[2], and arbitrarily excluding them does not seem a good strategy. Moreover, they were evaluated using a feed-forward neural network that was trained and tested *on the same data* (St. Clair et al., 2010). Since the utility of a set of distributional contexts cannot be restricted to its accuracy, the extent to which it scales to new, unseen words also needs to be taken into account.

Some of these problems have been addressed by Leibbrandt (2009), although his models are not incremental and rely heavily on arbitrary thresholds to remove very infrequent elements: while some sort of threshold seems to be unavoidable in a fully unsupervised model, a multitude of thresholds make it arbitrary and difficult to evaluate.

We will now introduce our model and then discuss the experiment that was set up to assess its effectiveness. We finally highlight the limitations of this work, sketch some ways to improve on it and draw the conclusions.

2 Model

We propose a model as a solution to the problems we highlighted in the previous section: it is entirely data-driven (reducing arbitrariness in the choice of the relevant dimensions) and more consistent with psycholinguistic evidence.

Three different pieces of information concerning a distributional context can be useful to the task at hand: i) its token frequency, i.e. how many times it occurs in the input; ii) its type frequency, i.e. the number of different words it occurs with; iii) the strength to which a context is predicted by a word, averaging across all the words it occurs with. Since it is hard to think to frequency without a comparison threshold, we divide token and type frequencies of a context by the average token and type frequencies across all contexts stored in memory at each sentence in the input.

These pieces of information can be combined in the following way:

$$score = token\_F \cdot type\_f \cdot p \qquad (1)$$

where each context is represented by a score resulting from the product of three pieces of infor-

[2]Although, see Baayen et al (2011) for an account in which n-grams effects are explained in a different way.

mation, defined as follows:

$$token\_F = \frac{log_2(count(c_i))}{avg(log_2(count(c))} \quad (2)$$

$$type\_f = \frac{log_2(\|W_{ci}\|)}{avg(log_2(\|W_c\|))} \quad (3)$$

$$p = \frac{1}{\|W_{ci}\|} \sum_{j=1}^{\|W_{ci}\|} \frac{log_2(count(w_j, c_i))}{log_2(count(w_j))} \quad (4)$$

In these formulas, c_i represents a distributional cue, W_{ci} is the set of words the cue occurs with; w_j represents a word and $count(w_j, c_i)$ the number of times a cue occurs with a specific word.

Raw counts are transformed with a base-2 logarithm to account for the fact that, as frequency grows, the contribution of every new occurrence to the total frequency is less and less important (Keuleers et al., 2010). Moreover, since the goal of this model is to discover structure, we assume that an item is only considered when it occurs more than once (items whose log is 0 are not considered). The formula in (4) closely resemble an average conditional probability – which children are likely to use to infer structure in language (Saffran et al., 1996) –, but differs from it since counts are again log-transformed for consistency with (2) and (3).

Salience can be thought of as the importance that a context might play in grouping words into categories, and the score we propose serves the purpose of selecting the most salient contexts. In this work, any context whose score is > 1 is considered to be salient, since 1 is the theoretical upper boundary of the *p* term, that can be increased or decreased by the following terms.

The formula in (1) is plugged into an incremental model that computes averages for token and type frequencies at every sentence *s*, and updates scores for contexts encountered in *s*. Contexts are harvested in a 2-word left/right window, looking at 2 bi-grams (*A_x; x_B*) and 3 tri-grams (*A_B_x, A_x_B* and *x_A_B*). A window cannot exceed a sentence boundary. At sentence initial and final positions, two dummy words were inserted, since sentence boundary information has been shown to be a useful distributional cue (Freudenthal et al., 2006; Freudenthal et al., 2008).

3 Experiment

3.1 Data

The experiment was carried out on the Aran section of the Manchester corpus (Theakston et al., 2001) from the CHILDES database (MacWhinney, 2000). In order to evaluate our model on unseen data, we divided the corpus chronologically in two sections: the first is used to select the distributional cues, the second for the evaluation phase.

We only considered sentences uttered by the mother, obtaining a corpus of 35K sentences. Our section for context selection (*selection set* henceforth) contains roughly 20K sentences, the section for the evaluation phase 15K. The corpus was not lemmatized. We removed false starts, onomatopoeia and other words based on their MOR PoS tags [3].

3.2 Setup

Different models - where each term from (1) is knocked out separately to assess its importance - were run on the selection set using only bi-grams, only tri-grams or both as contexts. The salient contexts at the end of this process were used as features in a supervised PoS experiment over types (not tokens) to evaluate their usefulness. As one reviewer pointed out, this evaluation is problematic for a number of reasons (Frank et al., 2008): however, we decided to use such approach because it is easy to interpret and provide a first indication about the potential effectiveness of the selected cues, serving as a first proof of concept.

In the selection set, only surface forms are considered[4]. We used the TiMBL package for memory-based learning (Daelemans et al., 2009), selecting the IB1 algorithm (Aha et al., 1991), weighted overlap as a distance metric with no feature weighting, and 1 nearest neighbor. In order to perform the experiment, the second part of the corpus was divided into a training and a test set (10K and 5K sentences, respectively), and two vector spaces were constructed, containing information about ho many times a word occurred with each cue.

The salient contexts harvested on the selection set were used as columns and the words occur-

[3] This is the list of MOR tags that were removed: *neo, on, chi, wplay, meta, fam, sing, L2, none*. Words without a tag were also removed, like errors, marked by a *0* before the tag, as in *0aux*.

[4] *Dog* and *dogs* are two different types, the modal *can* and the noun *can* are not.

Model	*# contexts*	*Useless*	*Missed words (%)*	*Hits*	*Acc.*
frequent frames	45	3 (6.7%)	83.7	290	**.83**
flexible frames	90	**0**	16.6	1405	.66
$p \cdot token\_F$					
2grams_bound	75	**0**	10.2	1559	.671
3grams_bound	348	13 (3.7%)	37.3	1073	.681
all_bound	490	11 (2.2%)	3.8	**1669**	.664
$p \cdot type\_f$					
2grams_bound	21	**0**	19.5	1377	.674
3grams_bound	42	**0**	56.7	788	.756
all_bound	97	**0**	8.7	1611	.679
$p \cdot token\_F \cdot type\_f$					
2grams_bound	211	**0**	2.6	1624	.641
3grams_bound	659	7 (1%)	25.5	1249	.653
all_bound	964	8 (0.8%)	**1.2**	1562	.609

Table 1: Evaluation of several sets of distributional cues, with baselines at the top and our models grouped according to the information included. Column 2 shows the number of salient contexts; column 3 shows how many of them could not be used for categorization. Column 4 provides the percentage of words from the training set (total = 3191) that could not be categorized by the contexts. Columns 5 and 6 raw number of hits (test set = 2600 words) and accuracy on supervised PoS tagging.

ring with at least one such context as rows. Words that never occurred with any of the salient contexts were not categorized. In the training and test sections, homographs were disambiguated when they were tagged differently: thus, the list of target words may well include *dog_noun*, *dogs_noun*, *can_verb* and *can_noun*.

Performances were evaluated on a tag-set consisting of 5 categories: nouns (including pronouns), verbs (including auxiliaries), adjectives, adverbs and function words, since we were mainly interested in content words, which make up the productive part of the lexicon. Performance is evaluated along 5 aspects: i) the number of salient contexts; ii) the percentage of salient contexts that could not be used in the training section, either because they were absent or because they only occurred with one word; iii) the proportion of words that were missed on the training set; iv) number of hits on the PoS-tagging experiment, and v) accuracy.

3.3 Results and discussion

Table 1 shows performances of all models on the five dimensions we introduced in (§3.2). Best scores on each dimensions are highlighted in bold. Intuitively, a model is good when it (i) selects a limited set of contexts, reducing the dimensionality of the vector space in which similar words are searched; (ii) minimizes the number of selected contexts that do not scale to new data; (iii) ensures high coverage on new data; (iv) allows to correctly categorize a high number of words; and (v) achieves a high accuracy, resulting in a reliable categorization.

While *frequent frames* achieve the highest accuracy, they also have the worst coverage and lowest number of hits. Plus, it is interesting that 3 contexts out of 45 are useless for categorization. When we turn to *flexible frames*, we see that they scale perfectly and achieve rather good accuracy, but do not ensure wide coverage and many hits.

A first global trend involves accuracy, which is inversely correlated with the number of selected contexts (Pearson r = -0.68), suggesting that distributional information is noisy and it is vital to focus on certain cues and discard the majority of them[5] to achieve reliable categorization. Finally, conflating bi-grams and tri-grams - which is closer to the psycholinguistic evidence we have - does not harm the model.

Turning to model-specific features[6], $p \cdot token\_F$ results in a rather large set of contexts, some of them being useless. Coverage is generally high, as the number of hits. When all three terms are in-

[5] A further analysis, not reported, was conducted by retaining all contexts and showed that both accuracy and number of hits were worse than most of the models evaluated here.

[6] The $token\_F \cdot type\_f$ models performed much worse than the others, thus results are not reported.

cluded, we still have large sets of contexts, few of which don't scale to new data. Coverage is high as the raw number of hits, but each model here is less accurate than its twin models. The reason for this behavior could be that *type_ f* strongly correlates with *token_ F* (the first cannot exceed the latter), and when they are both considered their contribution is inflated, resulting in more contexts and noise.

The $p \cdot type\_f$ models result in the smallest set of contexts, with perfect scalability and high accuracy. The downsides pertain coverage, and number of hits. Overall, no model performs high across all dimensions. However, the model combining *p* and *type_ f* displays parsimony, scalability, coverage and accuracy, although it is not the best on any dimension (it is also similar to flexible frames, but with better coverage, hits and accuracy). As we noted earlier, $token\_F$ and $type\_f$ are strongly positively correlated: this result suggests that the latter can be more useful to categories induction, since high type frequency ensures that a cue is systematic. We also evaluated contexts' token frequencies because of the well-attested frequency effects in language acquisition (Bybee, 1995), but the results suggest its effect in category formation can be better accounted for by contexts' type frequency. Nevertheless, further evidence is needed to confirm this hypothesis.

4 Limitations and future work

As one reviewer pointed out, this approach should be extended to be fully incremental and categorize tokens instead of types and evaluated with external linguistic task (see §3.2). However, unlike the probabilistic approach to category induction (§1), the focus of this paper was on the cues rather that on the categories: our goal was to show that it is possible to explicitly select the most informative distributional cues that infants are likely to rely on using a principled metric that does not simply rely on token frequency and predetermined distributional patterns. At the same time, if the presented model is indeed relevant can be only determined by directly evaluating categories of tokens induced in an unsupervised way on several linguistic task and looking at the time-course of learning, which was not discussed here.

A further limitation of the current work is that it arbitrarily focuses on words, neglecting morphological information, which is crucial in languages such as German, Turkish, Finnish and alike. A full model for distributional bootstrapping should automatically decide which are the relevant cues to categories, with no a priori restrictions on which units to focus on – see Hammerström and Borin (2011) for a review on unsupervised learning of morphology. This work only suggests a first way of moving away from pre-defined distributional patterns, since it can be equally applied to morphemes but it needs a presegmented input. A possible solution would be that of combining segmentation and category formation, looking at which cues are given more importance by the model and how useful they are to grouping words. Again, this falls outside of the scope of this paper and will be addressed in the future.

Finally, our model can be degraded in a variety of ways to introduce more plausible cognitive constraints in the form of free parameters that can reproduce attention and memory limitations. Such degraded versions would constitute a further and more informative test for this model, but are left for future work.

5 Conclusions

While no strong conclusion can be drawn without more data from typologically different languages, we think the goal of the paper was matched: we showed that the limitations of current frame-based approaches to distributional bootstrapping can be tackled with a simple model that incorporates evidence from psycholinguistic experiments and takes the number of different words a cues occurs with into account to decide whether the cue is informative. Furthermore, we showed that a model should be evaluated on different levels, since it is hard to achieve globally good performances.

The work by Mintz (2003) was crucial in showing that the PLD were rich enough to support an emergentist account of language learning. However, we contend that it is better to evaluate a process and its output, rather than a pre-selected set of cues, since it will more likely shed light on how certain cues but not others become important. It appears clear that focusing on fewer contexts is better: the central issue in a frame-based account of distributional bootstrapping should be to devise a model that identifies which cues give the best information.

Acknowledgments

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

References

David W. Aha, Dennis Kibler, and Marc K. Albert. 1991. Instance-based learning algorithms. *Machine Learning*, 6:37-66.

Inbal Arnon and Eve V. Clark. 2011. Why *Brush your teeth* is better than *Teeth*. Children's word production is facilitated in familiar sentence-frames. *Language Learning and Development*, 7:107-129.

Harald R. Baayen, Petar Milin, Dusica F. Durdević, Peter Hendrix, and Marco Marelli. 2011. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological Review*. 118(3):438.

Elizabeth Bates and Brian J. MacWhinney. 1987. Competition, variation, and language learning. in Brian MacWhinney (Ed.), *Mechanisms of Language Acquisition*. Hillsdale, NJ:Lawrence Erlbaum.

Joan L. Bybee. 1995. Regular morphology and the lexicon. *Language and Cognitive Processes*, 10:425-455.

Colin Bannard and Danielle Matthews. 2008. Stored word sequences in language learning. The effect of familiarity on children's repetition of four word combinations. *Psychological Science*, 3:241-248.

Timothy A. Cartwright and Michael R. Brent. 1997. Syntactic categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition*, 63:121-170.

Emmanuel Chemla, Toben H. Mintz, Savita Bernal, and Anne Christophe. 2009. Categorizing words using "frequent frames": What cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science*, 12:396-406.

Noam Chomsky. 1965. *Aspects of the theory of syntax*. MIT Press.

Anne Christophe, Séverine Millotte, Savita Bernal, and Jeffrey Lidz. 2008. Bootstrapping lexical and syntactic acquisition. *Language and Speech*, 51:61-75.

Grzegorz Chrupała and Afra Alishahi. 2010. Online entropy-based model of lexical category acquisition. *Proceedings of the 14th Conference on Computational Natural Language Learning*, ACL:Stroudsburg, PA. 182:191

Alexander Clark. 2000. Inducing syntactic categories by context distribution clustering. *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational Natural Language Learning, Vol 7. ACL:Stroudsburg*.

Walter Daelemans, Jakub Zavrel, Ko van der Sloot, and Antal van den Bosch. 2009. *TiMBL: Tilburg Memory Based Learner, version 6.3, Reference Guide*. ILK Technical Report 10-01

Marian A. Erkelens. 2009. Restrictions of frequent frames as cues to categories: the case of Dutch. *Supplement to the Proceedings of the 32nd Boston University Conference on Language Development (BUCLD 32)*. Boston, MA.

Stella Frank, Sharon Goldwater, and Frank Keller. 2009. Evaluating models of syntactic category acquisition without using a gold standard. *Proceedings of the 31st Annual Conference of the Cognitive Science Society*.

Daniel Freudenthal, Julian M. Pine, and Fernand Gobet. 2006. Modeling the development of children's use of optional infinitives in Dutch and English using MOSAIC. *Cognitive Science*, 30:277-310.

Daniel Freudenthal, Julian M. Pine, and Fernand Gobet. 2008. On the utility of conjoint and compositional frames and utterance boundaries as predictors of word categories. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*. Washington, DC.

Steven Gillis and Dorit Ravid. 2009. Language Acquisition. In Dominik Sandra, Jan-Ola Östman and Jef Verschueren (Ed.), *Cognition and pragmatics*. Amsterdam: Benjamin.

Lila R. Gleitman. 1990. The structural sources of verb meaning. *Language Acquisition*, 1:3-55.

Lila R. Gleitman and Jane Gillette. 1995. The role of syntax in verb-learning. In Paul Fletcher & Brian MacWhinney (Ed.), *The Handbook Of Child Language*. Oxford: Blackwell. 413-427.

Harald Hammarström and Kars Borin. 2011. Unsupervised learning of morphology. *Computational Linguistics*, 37(2):309-350.

Emmanuel Keuleers, Kevin Diependaele, and Marc Brysbaert. 2010. Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono-and disyllabic words and nonwords. *Frontiers in Psychology*, 1:174.

Richard E. Leibbrandt. 2009. Part-of-speech Bootstrapping using Lexically-specific Frames (PhD). *Flinders University, School of Computer Science, Engineering and Mathematics*.

Brian J. MacWhinney 1998. Models of the emergence of language. *Annual Review of Psychology*, 49:199-227.

Brian J. MacWhinney. 2000. *The CHILDES Project: Tools for analyzing talk. Third Edition*. Mahwah, NJ:Lawrence Erlbaum Associates.

Michael P. Maratsos and Mary A. Chalkley. 1980. The Internal Language of Children Syntax. In K. Nelson (Ed.) *Children's language*, Hillsdale, NJ: Erlbaum. 2:127-213.

Toben H. Mintz, Elissa L. Newport, and Thomas G. Bever. 2002. The distributional structure of grammatical categories in speech to young children. *Cognitive Science*, 26:393-424.

Toben H. Mintz. 2003. Frequent frames as a cue for grammatical categories in Child-Directed Speech. *Cognition*, 90(1):91-117.

Toben H. Mintz, Felix Hao Wang, and Jia Li. 2014. Word categorization from distributional information: Frames confer more than the sum of their (bigram) parts. *Cognitive Psychology*, 75:1-27.

Padraic Monaghan and Morten H. Christiansen. 2004. What distributional information is useful and usable for language acquisition? *Proceedings of the 26th annual conference of the Cognitive Science Society*.

Christopher Parisien, Afsaneh Fazly, and Suzanne Stevenson. 2008. An incremental Bayesian model for learning syntactic categories. *Proceedings of the twelfth conference on Computational Natural Language Learning, ACL:Stroudsburg*.

Steven Pinker. 1984. *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Patricia A. Reeder, Elissa L. Newport, and Richard N. Aslin. 2013. From shared contexts to syntactic categories: The role of distributional information in learning linguistic form-classes. *Cognitive Psychology*, 66:30-54.

Jenny R. Saffran, Elissa L. Newport, and Richard N. Aslin. 1996. Word segmentation: The role of distributional cues. *Journal of Memory and Language*, 35:606-621.

Michelle C. St. Clair, Padraic Monaghan, and Morten H. Christiansen. 2010. Learning grammatical categories from distributional cues: Flexible frames for language acquisition. *Cognition*, 116:341-360.

Barbara Stumper, Colin Bannard, Elena V. M. Lieven, and Michael Tomasello. 2011. "Frequent Frames" in German Child-Directed Speech: A limited cue to grammatical categories. *Cognitive Science*, 35:1190-1205.

Anne L. Theakston, Elena V. M. Lieven, Julian M. Pine, and Caroline F. Rowland. 2001. The role of performance limitations in the acquisition of "mixed" verb-argument structure at stage I. *Journal of Child Language*, 28:127-152.

Michael Tomasello. 2000. Do young children have adult like syntactic competence?. *Cognition*, 74:209-253.

Hao Wang and Toben H. Mintz. 2007. A dynamic learning model for categorizing words using frames. *Proceedings of the 32nd Annual Boston University Conference on Language Development (BUCLD 32)*. Boston, MA.

Hao Wang, Barbara Höhle, F. Nihan Ketrez, Aylin C. Küntay, and Toben H. Mintz. 2011. Cross-linguistic distributional analyses with frequent frames: the cases of german and turkish.. *Proceedings of 35th Annual Boston University Conference on Language Development (BUCLD 35)*. Boston, MA.

Zhao Zhang, Rushen Shi, and Aijun Li. 2015. Grammatical categorization in Mandarin-Chinese-learning infants. *Language Acquisition*, 22:104-115.