Wang, H., Höhle, B., Ketrez, N. F., Küntay, A. C., & Mintz, T. H. (2011). Cross-linguistic Distributional Analyses with Frequent Frames: The Cases of German and Turkish. In N. Danis, K. Mesh, & H. Sung (Eds.), Proceedings of the 35th annual Boston University Conference on Language Development (pp. 628-640). Somerville, MA: Cascadilla Press.

BUCLD 35 Proceedings
To be published in 2011 by Cascadilla Press
Rights forms signed by all authors

Cross-Linguistic Distributional Analyses with Frequent Frames: the Cases of German and Turkish

Hao Wang[1], Barbara Höhle[2], F. Nihan Ketrez[3], Aylin C. Küntay[4], and Toben H. Mintz[1]

1. Introduction

Syntactic categories (e.g., nouns and verbs) are the basic building blocks of grammar. Hence, establishing the syntactic categories of words is an essential step in the acquisition of syntax. As syntactic categories are not directly labeled in the speech directed to children (and also to adults), an important question is how children acquire knowledge about this crucial aspect of their language.

While there are no unique markers of syntactic categories, various kinds of correlational cues have been shown to be informative about word categories, such as phonological properties of words (Cassidy & Kelly, 1991; Kelly, 1992; Monaghan, Christiansen, & Chater, 2007; Shi, Morgan, & Allopenna, 1998) and semantic commonalities across words that belong to the same syntactic category (Braine, 1976; Grimshaw, 1981; Pinker, 1984; Schlesinger, 1971). Maratsos and Chalkley (1980) proposed that distributional information in the input—e.g., word co-occurrence patterns—could be a cue for categorizing words. The idea stemmed from the observation that words occurring in the same linguistic context often belong to the same grammatical category (Bloomfield, 1933; Harris, 1951). For example, the fact that *cat* and *mat* in (1) are both preceded by the same word would be evidence for a distributional learner that *cat* and *mat* belong to the same category.

(1) the cat is on the mat

* [1]University of Southern California, [2]University of Potsdam, [3]İstanbul Bilgi University, [4]Koç University. We would like to thank Dilara Koçbaş for transcribing and coding the Turkish data, Frauke Berger for labeling German data and the Developmental Brown bag at the University of Southern California for the feedback on an early version of this study. This research was supported by Turkish Academy of Sciences, in the framework of the Young Scientist Award Program granted to Aylin C. Küntay (EA-TÜBA-GEBİP/2001-2-13) and by a grant from the National Science Foundation (BCS-0721328) to Toben H. Mintz. Correspondence should be addressed to Hao Wang, Department of Psychology, SGM 501, University of Southern California, Los Angeles, CA 90089-1061, USA. E-mail: haowang@usc.edu

Researchers have examined a variety of types of distributional contexts ranging from *bigrams*—two word sequences in which one word provides a categorization context for the other, as in the example above—to whole sentences. They have also analyzed different kinds of computational mechanisms for deriving categories from these contexts (Cartwright & Brent, 1997; Mintz, 2003; Mintz, Newport, & Bever, 2002; Redington, Chater, & Finch, 1998). One distributional environment, *frequent frames*, has been shown to be particularly successful at reliably categorizing words in English (Mintz, 2003), and is the distributional context investigated in this paper.

A frame is defined as two jointly occurring words with one word intervening. The frames occurring most often in a corpus are selected as *frequent frames*. For example, *you__it* is a frequent frame in many English corpora of child-directed speech. In this frame, *you* and *it* jointly serve as contexts, and words occurring between *you* and *it* are treated as a single frame-based category. Mintz (2003) analyzed the frequent frames in speech directed to six children under the age of 2;6. Overall, the analyses showed that frequent frames are a robust cue to word categories in English. For example, Mintz reported that in one corpus, the *you__it* frame contained 433 tokens comprised of 93 unique words (types), all of which were verbs. He concluded that frequent frames could provide information for bootstrapping into the syntactic categories of a language by providing an initial basis for categorization.

1.1. Cross-linguistic investigations of frequent frames

Frequent frames have been proven to be very accurate categorization contexts in English, and there is behavioral evidence that adults and infants can track frame contexts and categorize words based on them (Gomez & Maye, 2005; Mintz, 2002, 2006). However, an important question is whether frequent frames are informative in typologically different languages. Several cross-linguistic studies of frequent frames have been conducted so far, including Dutch (Erkelens, 2009), French (Chemla, Mintz, Bernal, & Christophe, 2009), German (Stumper & Lieven, 2009), Spanish (Weisleder & Waxman, 2010) and Chinese (Cai, 2006; Xiao, Cai, & Lee, 2006). There are several features in these languages that could be detrimental to the account of frequent frames. For example, Romance languages, like French and Spanish, show a significant amount of homophony among function words, such as clitic object pronouns and determiners in French. In English, the most frequent frames generally consist of closed-class items (Mintz, 2003), so similar patterns in Romance languages could result in frequent frames that are heterogeneous with respect to the category of words that occurs within them. A further influence that could diminish the informativeness of frequent frames is the existence of different gender categories, as in the case for French, Spanish and German; for instance, gender marking on determiners leads to a higher variation in the forms of determiners compared to English. In addition, in pro-drop languages like Spanish and Chinese, verb frames may be more variable than in English.

However, despite these complicating issues, frequent frames turned out to be remarkably reliable in these languages as well.

1.2. Current analyses: German and Turkish

All studies on frequent frames done so far considered *words* as units forming the left and right element of a frame. In our study we asked whether in languages with a rich morphological system frames based on morphemes would provide better categorization than frames based on words, assuming that morphological richness leads to a higher amount of form variation at the word level. To this end, we compared word based and morpheme based frequent frame analyses for German and Turkish – two morphologically complex languages with typologically interesting differences in their morphological and syntactic systems.

German is a highly inflected language. Nominal elements are inflected for case, gender and number with many suppletive forms, especially in the case of function words. For example, there are six different forms of the definite article. Determiners can be used as strong pronouns and have the same form as relative pronouns which may lead to confusion for frequent frames (2).

(2) der Mann (the man)
der läuft (he runs)
der Mann, der gestern lief (the man who yesterday ran)

Word order in German is also more variable than English. As it is typical for V2 languages, a variety of constituents can occur before the finite verb (subjects, objects, adverbials).

Turkish is an agglutinative language that often uses inflectional suffixes to form new words and phrases. Turkish has a large number of word forms compared to English, in the sense that a given root corresponds to many more inflected forms in Turkish compared to English. One could expect, then, that the types of distributional patterns that occur are very different across the two languages. A complementary distributional fact is that function words are less prevalent in Turkish compared to English. Many English frequent frames rely on function words and pronouns as framing elements (Mintz, 2003), so their relative scarcity in Turkish could impact the informativeness of frequent frames. Moreover, Turkish is known for its free word order. The canonical order is SOV, but all other five orders (SVO, OVS, OSV, VSO, VOS) are possible, further disrupting the kinds of regular distributional word-order patterns that exist in, say, English, French, and Spanish. In addition, Turkish is a pro-drop language, so the environments in which verbs occur are more variable compared to languages without pro-drop (however, pro-drop did not appear to have a substantial negative effect on frequent frame categories in Spanish (Weisleder & Waxman, 2010)).

In summary, languages like German and Turkish potentially present challenges for frequent frames as a universal cue to the initial bootstrapping of lexical categories. Some of their features appear detrimental to word-level frequent frames, such as free word order, a higher variation in the word forms due to a frequent occurrence of bound morphemes or suppletive forms or the absence of function words. On the other hand, in languages with rich morphology, suffixes are often very informative about the category membership of a word, making morphemes good candidates as predictors of category membership. In order to investigate these issues, we tested how successfully frequent frames categorized corpora of German and Turkish child-directed speech, specifically comparing the performance of analyses at the word level (as in prior analyses of English, French, etc.) and at the morpheme level. In morpheme-level analyses, bound suffixes were segmented from stems, and the sequences of suffixes and bare stems were treated just as sequences of words in the traditional frequent frames method.

2. Frequent Frame Analysis of German and Turkish

2.1. Data

The data analyzed for German is the speech directed to one middle-class German-learning child, Simone (1;10.22-2;5.19) (Miller, 1976) taken from the CHILDES data base (MacWhinney, 2000), which yields 5685 utterances. The data were tape-recorded for at least six hours in a six-week interval. Some shorter intermediate recordings were also made at intervals of about ten days. During the recordings, the father and/or the mother of the child were present – often in addition to other family members or friends - and interacted with the child in various everyday life situations.

The data analyzed for Turkish is the speech directed to two Turkish-learning children by their regular caregivers: Elif (0;9.10-1;9.28) and Irmak (0;9.0-2;0.16) (Ural, Yuret, Ketrez, Koçbaş, & Küntay, 2009). The numbers of utterances analyzed are 21741 and 16024 for each child, respectively. The child-caregiver interactions were video-recorded at the homes of the children for one hour every two weeks. Some differences on the frequency distribution of verbs between the two corpora and on the socio-educational backgrounds of the families were reported by Ural et al. (2009).

2.2. Method

The frequent frame analysis as described in Mintz (2003, Experiment 1) was carried out on both datasets at the word level. The method was then modified (in a slightly different way for each language, as explained below) to analyze frequent frames at the morpheme level. To conduct the word-level frequent frame analysis, each utterance was segmented into frames. All frames were tallied through the entire corpus and ranked by their frequency. The most frequent 45 frames were then selected as frequent frames. Each frequent frame

defined a category that consisted of all the word tokens that occurred within the frame in the corpus. To evaluate the success of the resulting categorization, a native speaker assigned a lexical category label to each categorized word, consulting the context within the corpus if there was any ambiguity.

The category labels used for the German data were adjective, adverb, conjunction, determiner, interjection, noun, preposition, pronoun, particle, verb and wh-word. The category labels used for the Turkish data were adjective, adverb, determiner, communicative (phrases such as *yes*, *no*, and greetings), conjunctions, existentials (*var* and *yok)*, interjection, noun, negation, numerical, postposition, pronoun, verb, and wh-word.

To further evaluate the success of the frame-based categorizations, for each corpus analysis we calculated a 'control categorization' by randomly selecting two frame-based categories and exchanging a randomly selected token from one category with a randomly selected token from the other, repeating this process for one million trials (see Stumper & Lieven, 2009, for a similar control procedure). This differed from the procedure used in Mintz (2003), in which control categorization was based on analyzing a scrambled corpus. The current method was simpler to implement, but may be biased to yield more accurate 'control categories,' since there is more homogeneity in the linguistic categories pre-selected by frequent frames than in an entire corpus. This control method thus provides a stringent comparison for testing the informativeness of frequent frames.

2.2.1. Further details of German analyses

In the morpheme-level analysis, grammatical morphemes were manually identified and separated from their roots. They were prefixed with a hyphen to indicate their status as a suffix (3). The roots appeared in each frequent frame were labeled with their actual grammatical categories and grouped together to form a frame-based category.

(3) soll ich mal gucken, was die machen
soll ich mal guck -en was die mach -en

Two additional distributional environments were also tested on the German corpus, in both word- and morpheme-level analyses. In French, Chemla at al. (2009) tested categorization with frequent frames, but also trigrams in which the joint co-occurrence of the first and second words was the categorizing environment for the third, and in which the joint co-occurrence of the second and third words was the categorization context for the first. The question was whether there was something special about the particular configuration of the context words—i.e., the frame—in frequent frames that was critical for good categorization, or whether any trigram with similar joint-occurrence constraints on two context positions was sufficient. Chemla at al. found that the frame trigram resulted in superior categories compared to the other two related trigram

configurations. We were interested in the same question here, so in addition to analyzing frequent frames, we analyzed trigrams by treating the first and second words (morphemes, in the morpheme-level analysis) as categorizing contexts for the third word. We use the notation F_1F_2__ for such trigrams. In addition, we analyzed trigrams by treating the second and third words (or morphemes) as categorizing contexts for the first word (__F_1F_2). (Frames are schematized as F_1__F_2 in our notation.)

2.2.2. Further details of Turkish analyses

The morpheme-level analysis in Turkish differs slightly from the German analysis due to the morphological coding used in the Turkish corpora. First, the grammatical morphemes were manually identified in each word. Each grammatical morpheme was then labeled with its grammatical function, such as plural, possessive or accusative (see (4), for an example). In Turkish there is usually one-to-one correspondence between phonological form and the morpheme, therefore, we glossed over the predictable variations of the same morpheme due to vowel harmony and consonant assimilation, which Turkish children master very early on. At last, the roots and morpheme labels were separated and treated just like individual units in morpheme-level analysis. The morpheme labels were in all-capital case.

(4) sen hep ayak-lar-ın-ı sok-uyo(r)-sun.
sen hep ayak-PL-POSS&2S-ACC sok-IPFV-2S[1]
'you always put in your feet'

Because inflections are suffixes in Turkish, a stem that occurs in a frequent morpheme frame will thus be framed by an inflection to the right that is bound to the stem, and an inflection to the left that is bound to the preceding stem. To the degree that stems are categorized accurately in morpheme frames, it could be the result of the bound affix, which is highly informative of the category of the stem/word. In that case, the leftmost framing element may not play a significant role in categorization. To test for this possibility, in addition to the frequent frame analyses, we ran morpheme-level bigram analyses in which frequently occurring morphemes provided categorization contexts to an immediately neighboring morpheme. We ran two versions of these analyses: In one, the frequent morpheme categorized the word immediately to its right (F_1__), in the other it categorized the word immediately to its left (__F_1). The latter bigram pattern includes stems that are categorized by an immediately adjacent bound morpheme.

[1] 2S=2nd person singular, ACC=Accusative case, IPFV=Imperfective (progressive marker -*Iyor*), PL=Plural, POSS&2S=Possessive and 2nd person singular fused

2.2.3. Quantitative evaluation

The categorization outcome is evaluated with standard measures in the literature—*accuracy* and *completeness*, which are similar to the metrics *precision* and *recall* in signal processing. To compute these measures, categorized words are labeled with their actual grammatical categories according to the contexts they appear. The categories of every pair of word tokens are compared to each other. Accuracy is penalized when words belonging to different grammatical categories are grouped together. Accuracy ranges from 0 to 1, where a score of 1 indicates that each distributional category contains items from only one grammatical category. Completeness is penalized when words belonging to the same grammatical category occur in different distributional categories. Completeness also ranges from 0 to 1, where a score of 1 indicates that words of the same grammatical category were classified together. (See Mintz (2003) section 2.1.3 for the details of the computation.) For a bootstrapping device, forming categories that are linguistically accurate is arguably more important than forming comprehensive categories. Thus, accuracy is a more relevant measure of categorization success for these analyses.

2.3. Results

Both word-level and morpheme-level frequent frames achieved relatively good accuracies in German (Table 1). Frequent frames have a slightly higher accuracy in morpheme-level analysis compared to the word-level analysis. However, the morpheme-level categorization included more than twice as many tokens as the word-level frequent frames. The completeness for frequent frames at both levels is comparable to that in previous studies.

The accuracies for $\_\_F_1F_2$ and $F_1F_2\_\_$ at both levels are substantially lower than the accuracies of frequent frames at those levels, although $\_\_F_1F_2$ at the morpheme-level has an exceptionally high accuracy among the two environments themselves, and approaches the accuracy of the two frame-based analyses. The accuracies in all conditions are substantially higher than the accuracies of the random controls.

Some of the most frequent morpheme frames are listed in Table 2. The first column is the frames. The second and third columns are the number of word types and word tokens categorized by each frame. The fourth column is the majority category of each frame, which is the category with the largest number of tokens than any other categories in that frame. The last column is the proportion of the tokens of the majority category to the total number of tokens. Values in this column are positively correlated with the overall accuracy of the analysis.

Table 1
German accuracies and completeness (FF is frequent frame)

	Actual Analysis			Randomized Categories	
	Accuracy	Completeness	Tokens	Accuracy	Completeness
FF Word	0.86	0.07	884	0.37	0.03
FF Morpheme	0.88	0.05	1857	0.51	0.03
__F_1F_2 Word	0.47	0.04	1216	0.31	0.03
F_1F_2__ Word	0.32	0.05	1462	0.17	0.03
__F_1F_2 Morpheme	0.78	0.10	2742	0.32	0.04
F_1F_2__ Morpheme	0.30	0.07	2672	0.16	0.03

Table 2
Some German frequent morpheme frames

Frame	Token	Type	Majority Category	Token % of Majority Cat.
was__-st	122	12	V	99%
Maxe__-t	107	32	V	100%
was__-t	91	18	V	100%
ge__-t	88	25	V	98%
-e__-e	65	32	Adj	38%
du__-st	65	26	V	98%
wir__-en	63	22	V	100%
n__-chen	59	3	Pro	91%
-e__-en	57	21	V	82%
pass__auf	54	2	Pt	100%
-en__mal	52	11	Pro	44%
das__-t	49	18	V	100%

Most of the frequent morpheme frames are perfect or near perfect (the token percentage of the majority category is close to 100%) verb frames. There are also a few good noun frames that are not shown in Table 2, such as die__-n and -m__-er. Table 2 also shows that several categories are less linguistically homogeneous.

Most of the frequent morpheme frames categorize a large number of word types and/or tokens. For example, one of the morpheme frames Maxe__-t categorized 107 tokens and 32 types, all of which are verb stems.

Turning to the Turkish analysis, the word-level frequent frame accuracies were quite low, whereas the morpheme-level frequent frames reached high

accuracies and categorized more tokens than word-level frequent frames (Table 3).

Some frequent morpheme frames are listed in Table 4. Many of the verb frames are perfect or near perfect. The most frequent morpheme frame is a very good noun frame. In fact, all the frames captured a large number of word types and tokens. As an example, all of the 203 tokens (71 types) appeared in the frame ACC_PAST are verb stems.

It is interesting to note that, for each corpus, the morpheme bigram analysis that includes stems categorized by their first sequential bound morpheme (__F_1) is not as accurate as the corresponding morpheme frame analysis (although it substantially exceeds the F_1__ morpheme bigram analysis). This suggests that in the frame analysis, the morpheme to the left of the stem plays an important role, even though it is part of the previous word. Thus, even in the morpheme-level analysis, word order regularities apparently are informative.

Table 3
Accuracies and completeness for Turkish

			Actual Analysis			Randomized Categories	
		Corpus	Accuracy	Comp.	Tokens	Accuracy	Comp.
FF Word		Elif	0.54	0.09	1269	0.19	0.03
		Irmak	0.40	0.11	1656	0.27	0.08
FF Morpheme		Elif	0.93	0.06	6102	0.49	0.03
		Irmak	0.88	0.06	2764	0.38	0.03
Bigram Morpheme	F_1__	Elif	0.31	0.05	33793	0.20	0.03
		Irmak	0.39	0.05	16678	0.26	0.04
	__F_1	Elif	0.66	0.10	41540	0.24	0.04
		Irmak	0.72	0.09	29017	0.29	0.03

Table 4
Some Turkish frequent morpheme frames (Elif)[2]

Frame	Token	Type	Majority Category	Token % of Majority Cat.
GEN__POSS&3S	538	163	N	96%
ne__IPFV	348	24	V	100%
ne__PAST	316	26	V	98%
QUE__PAST	260	77	V	98%
DAT__PAST	217	43	V	100%
QUE__IPFV	215	59	V	99%
DAT__IPFV	209	51	V	100%
ACC__PAST	203	71	V	100%
QUE__FUT	165	34	V	100%
LOC__var	152	52	Wh	55%
ACC__IPFV	151	61	V	99%

3. Discussion and Conclusion

Frequent frames at both levels performed well in terms of accuracy in German. Language specific properties such as the flexible word order, multiple forms of the definite determiners and homophony to other function words are not catastrophic for the frequent frames analysis. However, some of these properties may have resulted in the smaller number of tokens categorized in the word-level analysis compared to the morpheme-level analysis in German.

Although $\_\_F_1F_2$ at morpheme-level obtained moderate accuracy which presumably is caused by the fact that suffixes in German are very informative about the word category, the two non-frame trigram environments ($\_\_F_1F_2$ and $F_1F_2\_\_$) did not achieve comparable performance as the frequent frames, which suggests that it is not only important to have two context units (comparing to bigrams) but the relative positions of the context units and the target unit are also important in accurate categorization. Similar results were reported by Chemla at al. (2009) for French (although in French, neither non-frame trigram achieved accuracies as high as the $\_\_F_1F_2$ context in German). The underlying source for the categorization superiority of the framing relationship was further explored in Wang & Mintz (2009). They compared the syntactic structures associated with frequent frames, bigrams, $\_\_F_1F_2$ and $F_1F_2\_\_$ in English child-directed speech. The results showed that the syntactic structures associated with

[2] ACC=Accusative case, DAT=Dative case, FUT=Future tense, GEN=Genitive case, IPFV=Imperfective (progressive marker), LOC=Locative case, PAST=Past tense, POSS&3S=Possessive and 3rd person singular fused, QUE=Yes-no question particle, ne (what), var (there is)

frequent frames are more restricted and consistent than those associated with the other distributional environments. Inherently, the categories of the target word in frequent frames are tightly constrained and exhibit less variations, which lead to robust categorization. The moderate accuracy of the __F_1F_2 morpheme context in German is consistent with the typological facts about the richer inflectional system of German compared to English.

A word-level frequent frame analysis of German by Stumper & Lieven (2009) produced lower accuracy (0.776) than our current analysis. The difference in accuracy is difficult to pinpoint and could be due to many reasons, such as the size of the corpora, variations among adult speakers or the category labels used. It would be interesting to undertake a morpheme-level analysis of their corpus.

In Turkish, free word order and agglutinative morphology apparently resulted in poor accuracy for word-level frequent frames. However, morpheme-level frequent frames achieved high accuracy. Importantly, it was not simply suffixes bound to the stem that resulted in accurate categories, since __F_1 contexts were not as accurate as morpheme frames. Rather, suffixes on the previous word appear to provide additional distributional information that result in more accurate categories.

Taken together, these analyses suggest that frequent frames provide informative lexical category information in a variety of typologically distinct languages, as long as they are analyzed at the appropriate level of granularity. In Turkish, and to some extent German, the morpheme appears to be the correct level of granularity. For young children acquiring languages with rich morphology like German and Turkish, previous studies have demonstrated that they have access to morpheme-level information very early (Aksu-Koç & Ketrez, 2003; Höhle, Schmitz, Santelmann, & Weissenborn, 2006; Ketrez & Aksu-Koç, 2009). Dutch infants have even been shown to be able to use frequent morpheme frames to categorize nonsense words (Erkelens, 2009). Therefore, it is plausible that children learning morphologically rich languages could use morpheme frequent frames to start building lexical categories.

This possibility raises the question of how young children could determine the right level of granularity for their language. It is conceivable that the appropriate level can be computed from properties of the input itself. For example, Swingley (2005) presented a computation mechanism that can segment word-level sequences from English and Dutch corpora based on the co-occurrence properties of adjacent syllables. One can image a similar mechanism that segments morpheme-like units from languages with rich morphology.

Generalized in this way, frequent frames could be a cross-linguistically viable pattern for generating an initial set of lexical categories and bootstrapping into the language's lexical categories. Once preliminary frame-based categories are formed, other sources of correlated information could be discovered, such as phonological properties. Indeed, other distributional cues that are more language specific could be discovered and leveraged for further categorization, such as the __F_1 bigram pattern in Turkish, and __F_1F_2 in German.

In conclusion, our analyses of German and Turkish child-directed speech have shown that frequent morpheme frames are highly accurate categorization contexts in the two languages. Languages with rich morphology and free word order are not a problem for frequent frames, if analyzed at the right level. Frequent frames could well be a potential universal cue for initial bootstrapping.

References

Aksu-Koç, Ayhan, & Ketrez, Nihan. (2003). Early verbal morphology in Turkish: Emergence of inflections. In W. U. Dressler, D. Bittner & M. Kilani-Schoch (Eds.), *Development of Verb Inflection in First Language Acquisition: A Cross Linguisitc Perspective* (pp. 27-52). Berlin: Mouton de Gruyter.

Bloomfield, Leonard. (1933). *Language*. New York: Henry Holt.

Braine, Martin D. S. (1976). Children's first word combinations. With Commentary by Melissa Bowerman. *Monographs of the Society for Research in Child Development, 41*.

Cai, Xin. (2006). *The acquisition of syntactic categories in Chinese on the basis of distributional information.* MA, Hunan University.

Cartwright, Timothy A., & Brent, Michael R. (1997). Syntactic categorization in early language acquisition: formalizing the role of distributional analysis. *Cognition, 63*(2), 121-170.

Cassidy, Kimberly Wright, & Kelly, Michael H. (1991). Phonological Information for Grammatical Category Assignments. *Journal of Memory and Language, Fl*(JMemL)).

Chemla, Emmanuel, Mintz, Toben H., Bernal, Savita, & Christophe, Anne. (2009). Categorizing words using frequent frames: what cross-linguistic analyses reveal about distributional acquisition strategies. *Developmental Science, 12*, 396-406. doi: 10.1111/j.1467-7687.2009.00825.x

Erkelens, Marian. (2009). *Learning to categorize verbs and nouns: Studies on Dutch*. Utrecht: LOT.

Gomez, Rebecca, & Maye, Jessica. (2005). The Developmental Trajectory of Nonadjacent Dependency Learning (Vol. 7, pp. 183 - 206): Psychology Press.

Grimshaw, Jane. (1981). Form, function, and the language acquisition device. In C. L. Baker & J. J. McCarthy (Eds.), *The logical problem of language acquisition*. Cambridge, Mass.: The MIT Press.

Harris, Z. S. (1951). *Structural linguistics*. Chicago: University of Chicago Press.

Höhle, Barbara, Schmitz, Michaela, Santelmann, Lynn M., & Weissenborn, Jurgen. (2006). The Recognition of Discontinuous Verbal Dependencies by German 19-Month-Olds: Evidence for Lexical and Structural Influences on Children's Early Processing Capacities. *Language Learning and Development, 2*(4), 277-300.

Kelly, Michael H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review, 99*(2), 349-364.

Ketrez, Nihan, & Aksu-Koç, Ayhan. (2009). Early Nominal Morphology: Emergence of Case and Number. In M. Voeikova & U. Stephany (Eds.), *The Development of Number and Case in the First Language Acquisition: A Cross-Linguistic Perspective* (pp. 15-48). Berlin: Mouton de Gruyter.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk*. Mahwah, NJ: Lawrence Erlbaum Associates.

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

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

Mintz, T. H., Newport, E. L. , & Bever, T. (2002). The distributional structure of grammatical categories in speech to young children. *Cognitive Science, 26*, 393-424.

Mintz, Toben H. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition, 30*, 678-686.

Mintz, Toben H. (2006). Finding the verbs: distributional cues to categories available to young learners. In R. M. Golinkoff K. Hirsh-Pasek (Ed.), *Action Meets Word: How Children Learn Verbs* (pp. 31-63). New York: Oxford University Press.

Monaghan, Padraic, Christiansen, Morten H., & Chater, Nick. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology, 55*(4), 259-305.

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

Redington, Martin, Chater, Nick, & Finch, Steven. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science, 22*(4), 425-469.

Schlesinger, Izchak M. (1971). Learning grammar: From pivot to realization rule. In R. Huxley & E. Ingram (Eds.), *Language acquisition: models and methodology*. New York, NY: Academic Press.

Shi, Rushen, Morgan, James L., & Allopenna, Paul. (1998). Phonological and acoustic bases for earliest grammatical category assignment: a cross-linguistic perspective. *Journal of Child Language, 25*(01).

Stumper, Barbara, & Lieven, Elena. (2009). *'Frequent frames' in German child-directed speech: A limited cue to grammatical categories*. Poster presented at the Architectures and Mechanisms for Language Processing, Barcelona.

Swingley, Daniel. (2005). Statistical clustering and the contents of the infant vocabulary. *Cognitive Psychology, 50*(1), 86-132. doi: DOI: 10.1016/j.cogpsych.2004.06.001

Ural, A. Engin, Yuret, Deniz, Ketrez, F. Nihan, Koçbaş, Dilara, & Küntay, Aylin C. (2009). Morphological cues vs. number of nominals in learning verb types in Turkish: The syntactic bootstrapping mechanism revisited. *Language and Cognitive Processes, 24*(10), 1393-1405.

Wang, Hao, & Mintz, Toben. (2009). *From Linear Sequences to Abstract Structures: Distributional Information in Infant-direct Speech.* Proceedings Supplement of the 34th Boston University Conference on Language Development, Boston, MA.

Weisleder, Adriana, & Waxman, Sandra R. (2010). What's in the input? Frequent frames in child-directed speech offer distributional cues to grammatical categories in Spanish and English. [10.1017/S0305000909990067]. *Journal of Child Language, 37*(05), 1089-1108.

Xiao, Ling, Cai, Xin, & Lee, Thomas H. (2006). *The development of the verb category and verb argument structure in children before two years of age.* Paper presented at the Seventh Tokyo Conference on Psycholinguistics, Tokyo.