Distributional Properties of Nouns and Verbs in Turkish Child-directed Speech*

F. Nihan Ketrez
University of Southern California
ketrez@usc.edu

This study investigates the distribution of nouns and verbs in Turkish child-directed speech. The goal of the study is to determine whether or not there is any regularity in noun and verb distributions that may be useful to children in acquiring these categories, even in a relatively free-word order language. Adopting the model developed by Mintz *et al.* (2002), three analyses were conducted on four children's corpora, and the distribution of nouns and verbs were analyzed in three different window sizes. The results of the analyses showed that, while there was no regularity in the noun distribution, no matter how big the window size was, verbs displayed a regular linear distribution, and even the smallest change in the window size made a significant difference in clustering.

1. Introduction

In this study, I look at noun and verb distributions in Turkish child-directed speech (CDS). The aim of the study is to investigate whether or not there is any regular distributional information available in Turkish CDS that may be helpful to a child in acquiring noun and verb categorization.

We know that children are sensitive to statistical regularities in the languages that they hear from very early ages (Aslin, Safran and Newport, 1998; Saffran, Aslin and Newport, 1999; Marcus, Vijayan, Rao and Vishton, 1999, among others). Previous research has also shown that distributional properties of input utterances give regular cues to children that may be helpful to build up grammatical categories (Cartwright and Brent, 1997; Redington, Chater and Finch, 1998; Mintz, *et al.* 2002). Mintz *et al.*, whose methodology I adopt in this study, propose that preceding and following contexts of words provide the learner with regular information to acquire the grammatical categories of nouns and verbs. Mintz *et al.* (2002), like other distributional studies cited above, is based on only English and works very well for a language that has a relatively fixed word order. Such studies leave open the question of whether their results have cross-linguistic validity. In the present study, I examine Turkish CDS, in order to see whether or not there are distributional regularities that may be used by children for category learning. Turkish is especially interesting for such research because of its typological differences from English and also because it is considered to be a complex "puzzle" for a learner due to its flexible word order patterns (Küntay and Slobin, 1996).

I look at the preceding and following contexts of the words and show that, while there is no evidence that nouns have any distributional regularity in Turkish, verb distribution is surprisingly better than chance and provides evidence that, even in a language with a flexible word order, distributional properties of one of the categories may provide potentially useful cues for children. These results contradict Küntay and Slobin's (1996, 1999) conclusion that neither nouns nor verbs have informative distributional patterns in Turkish child-directed speech.

* The data analyzed in this study were collected for the project "A longitudinal study of the acquisition of Turkish," which was supported by a grant from the Boğaziçi University Research Fund (Project # 98 HB 702). I am grateful to Ayhan Aksu-Koç, the project director, for the data and Toben H. Mintz for providing me with the analysis programs and also for the discussion and detailed comments on earlier versions. I benefited greatly from the comments and questions I received from the audience at the First Student Workshop at USC (April 2001), BOOT-LA at Indiana University (April 2003) and Psychobabble at UCLA (April 2003). This is also a version of the paper that was presented at ICTL-2002, which was held at the Eastern Meditarrenean University at Northern Cyprus (August 2002).

2. An overview: How does a child know that a word is a *verb*?

Categorization of nouns and verbs is one of the most important and basic steps in language learning. Although it is potentially very complex and difficult to master, children do not seem to have problems in the acquisition of category distinctions. Such errors as using a noun as a verb (e.g., *he is keying the door* [3:0], Clark, 1982) are limited to languages that use one form serving two functions, as in the case of English, and are considered "legitimate" in appropriate contexts and "innovative" examples of a "later phase" of language acquisition (Clark, 1982:403). Therefore, they are interpreted as children's overgeneralization of such uses that already exist in the particular language. Such uses are never reported for languages such as Turkish or Arabic. How does a child figure out that a word belongs to one category, and not to the other? Even approaches that assume that noun and verb categories are innate and need not be learned (Chomsky, 1965) have to account for how a child assigns the categorical features to a word that s/he hears. There have been various proposals regarding how children master the acquisition of categories based on studies that investigate semantic, phonological, and distributional properties (or an interaction of all three).

2.1 Semantics helps

The semantics approach draws attention to the strong correlations between the meanings of words and their categories, suggesting that children follow these correlations (Brown, 1957; Macnamara, 1972, 1982; Schlesinger, 1974; Bowerman, 1973; Bates and Macwhiney, 1982). In such an approach, words that refer to actions are verbs, those that describe objects are nouns, and the child groups words such as *book, cup* and *pen* in the same category because they all refer to concrete objects. Only later in development, can these categories be generalized to more abstract and less proto-typical nouns such as *thought* with the aid of some other cues provided by other domains, such as word order, or more specifically, the distributional similarity of the category members across utterances.

Based on this approach, the Semantic Bootstrapping Hypothesis is proposed and developed by Grimshaw (1981), Macnamara (1982) and Pinker (1984, 1987). According to the Semantic Bootstrapping Hypothesis, children are provided with the categories of nouns and verbs by the universal grammar and then they need to "link" some words that they hear to these categories. This linking process is based on the meanings or the semantic context of the words, that is, the names of actions are "linked" to the verbs and the names of objects are "linked" to the nouns. Although semantics may help the child to figure out some distinctions, it is not sufficient by itself. The proposal runs into problems, especially because it cannot account for how exactly a child knows whether or not the word s/he hears refers to the action performed and not to the performer of the action or the object undergoing the action (Gleitman 1990; Gleitman and Gillette, 1995). As Maratsos and Chalkey (1980) also states, there has been no purely semantic approach that can explain the whole course of language acquisition from the very beginning to the actual end.

2.2 Phonology/ Phonetics may be helpful

Phonological form and stress patterns are considered to be helpful because there are observed to be some language specific correlations between word categories and the prosodic/phonological structure of the language such as the stress assignment and the number of syllables (Cassidy and Kelly, 1991; Kelly, 1992; 1996). In English, for example bi-syllabic nouns tend to have primary stress in their first syllable, whereas bi-syllabic verbs have their primary stress in their final syllable (e.g., permit). Nouns have more low vowels whereas verbs have more high vowels, nouns have more syllables, have a longer duration and contain more phonemes (Kelly, 1992; 1996). Kelly (1992) reports similar regularities in languages typologically different from English, such as Russian, as well.

Are children aware of such subtle differences? Research done by Jusczyk and his colleagues has shown that they are, even when they are as young as 9 months old. Jusczyk, Cutler and Redanz (1993) show that children are sensitive to the position of the primary stress and American children, at the age

of 9 months, prefer listening to words that have primary stress in the first syllable, thus showing a sensitivity to the dominant stress pattern in their native language. Jusczyk *et al.* (1993) and Jusczyk, Charles-Luce and Luce (1994) state that Dutch and American infants are sensitive to the sound sequences that exist in their own language and to those that are relatively frequent in their native language. Additional support comes from Morgan, Shi and Allopenna (1996) who also show that content and function words can be classified according to their phonological properties. These studies show that there are correlations between grammatical categories and the phonological structure of words and that children are sensitive to such subtle phonological cues in the language, which may potentially help them in language acquisition.

2.3 Distributional approaches

In addition to the proposals based on the semantics and phonology/phonetics domains, there have also been others that are based on studies that investigate the linear distribution of the words (or smaller units such as inflections) in the input speech. The argument that distributional properties of words signal their grammatical category starts with Bloomfield (1933) and develops with the proposals of Harris (1951) and Fries (1952). According to this view, anything that occurs in the slot in "the ___ is ...", for example, is a noun. A more recent proposal based on this approach is seen in Maratsos and Chalkley (1980) who treat the verbal suffixes *–s* and *–ed* as cues for the syntactic category of the preceding words. Hence anything that precedes the past tense *–ed* is a verb.

Starting with this idea of distributional regularities, computational approaches to category acquisition developed (Cartwright and Brent, 1997; Redington, Chater and Finch, 1998; Mintz *et al.*, 2002 among others). With these studies, it has been shown that there are statistically significant distributional regularities in input speech, which may provide children with helpful paths into the grammatical structure of their language. Mintz *et. al*, for instance, show that nouns and verbs in English have a regular distributional pattern. They look at the distribution of individual categories in one, two and eight word contexts, that is they look at the preceding and following words, increasing the window size in each analysis, with the hierarchical cluster analysis method (HCA) and report that both nouns and verbs cluster well in English, and thus input language provides the learner with helpful categorical information. Although they present clear evidence that there are usable distributional cues provided in adult speech to children, they do not propose a learning model.

In the present study, I conduct a similar distributional analysis on a typologically different language to see whether or not there are regularities with respect to the distribution of words. I apply the methodology of Mintz *et al.* to Turkish corpora so that I can compare the results of Turkish with that of English directly.

In the following section, Section 2 below, I present the structure of Turkish and previous work on Turkish and I discuss the reasons why it is interesting to look at this particular language. In section 3, I describe the corpora and the methodology. In sections 4 and 5, I present the results and discuss the implications of the study.

3. Why is Turkish interesting for a distributional analysis?

Turkish has syntactic and morphological features that make it a very challenging language for an approach that is based on the linear distribution of words. These are flexible word order, pro-drop, rich morphology and a lack of function words. Another relevant property of Turkish is the very early evidence for grammatical categories in children's speech, which is challenging for any learning model.

3.1 Structural properties

3.1.1 Turkish has flexible word order

The canonical word order is SOV in Turkish, as exemplified in (1), but all other six orders are also possible (2) and are used for various communicative functions (Erguvanlı, 1984; Kural, 1992). These variations are not necessarily different in terms of their prosodic features.

(1)	Deniz kitab-ı oku-du	SOV	(2)	a. Deniz okudu kitabı	SVO
	Deniz book-acc read-past[1]			b. Kitabı Deniz okudu	OSV
	Deniz read the book			c. Kitabı okudu Deniz	OVS
				d. Okudu Deniz kitabı	VSO
				e. Okudu kitabı Deniz	VOS

The flexibility of word order could make it difficult to form a regular distributional pattern for particular target words, that is, the linear word order of Turkish might be problematic since the preceding and the following words are very flexible across utterances especially when they are compared to that of a language such as English that has a fixed word order, which allows flexibility only in restricted situations.

In these examples above (1-2), we see the constituents scrambling. As seen in the following examples (3-4), flexibility also applies across clause boundary. The possessed noun in a noun phrase, for instance, as in (3), can move to the end of the sentence, leaving the possessor noun in its original position. In this way, the subject NP splits into two parts in the linear order. The sentence in (4), similarly, shows an instance where the possessor moves instead, leaving the possessed NP. These sentences are examples from the actual corpora I am analyzing.

(3)	STR:	sen-in de var mı	kaşığ-ın ?	(Mine corpus)
		you-gen&2sg too there-is que	spoon-poss&2sg	
		Do you, too, have a spoon?		
(4)	MOT:	a:bi-nin ad-ı	ne sen-in ?	(Mine corpus)
		brother-gen&2sg name-poss&3sg	what you-poss&2sg	
		What is your brother's name?		

Previous research has shown that all these varieties of word order exist in Turkish child directed speech (Slobin and Bever, 1982; Slobin and Talay, 1986; Küntay and Slobin 1996, 1999) and children, from early ages onwards, show a sensitivity to these word order patterns and their functions (Ekmekçi, 1986; Aksu-Koç, 1994; Kornfilt, 1994; Batman-Ratrosyan and Stromswold, 1998).

Küntay and Slobin (1996, 1999) look at the distribution of nouns and verbs in a Turkish mother's speech, more specifically, whether or not verbs occur in a particular position, utterance initial or final, for example. This restricted distributional analysis is done to compare Turkish CDS with that of the languages such as Mandarin Chinese (Tardif, Shatz and Naigles, 1996), where verbs consistently appear at the end of the utterances and are thus, argued to be more salient for learners and English (Goldfield, 1993), in which nouns are proposed to be privileged because they occur in these salient positions. Küntay and Slobin (1996) conclude that they do not see such a regular pattern with respect to the positions in Turkish and therefore, word order cannot be providing a learner with any kind of helpful information as it presents nothing but a complex "puzzle."

3.1.2. Turkish is a pro-drop language

In addition to a flexible word order pattern, Turkish also has a pro-drop mechanism (Kornfilt, 1984; Enç, 1986; Özsoy, 1987), that is, the subject and object nouns are omitted and their interpretations are based on either the çontext of the utterance or, in the case of the subject, on the overt agreement marker on the predicate. In the examples below (5-7), for instance, the preceding context of the verb *oku* 'read' varies. In (5) it is preceded by the subject *Deniz*, in (6) it is preceded by the object *kitabı* 'book-acc' and in the last example, it does not have a preceding context.

(5)	Deniz ∅ oku-du	(6) ∅ kitab-ı oku-du	(7) ∅ ∅ oku-du
	Deniz read-past	book-acc read-past	read-past
	Deniz read (the book).	(Deniz) read the book.	(Deniz) read (the book).

[1] List of abbreviations are presented in Appendix-1.

In this way, through the pro-drop mechanism, the contextual information is reduced if not totally omitted. Although the sentence has the same interpretation in each case, the verb has three different preceding contexts and thus three different distributional patterns.

3.1.3 Function words are rare

Turkish lacks function words such as prepositions, auxiliary verbs, and the definite article, which signal the categorical features of the words in their contexts. It has case and tense affixes attached to the content words instead, which change the morphological structure of the word, without adding anything to the word order. The function words that exist in the language are the indefinite article *bir* 'a' (*e.g. bir elma* "an apple") and the demonstrative articles (*bu* 'this', *su* 'that', o 'it'). *Bir* precedes nouns as an indefinite article, but it also precedes adverbs (*e.g. bir daha* "once more") and verbs (*e.g. bir gelse* "if only he comes") when it means "once" or "if only". The demonstrative articles also appear as pronouns. Postpositions such as *gibi* 'like' and *icin* 'for' are very rare (Only one postposition occurs in the most frequent 200 words in one of the corpora I analyze and did not occur in any other corpora analyzed). Turkish also has a question particle but its position in the structure is not fixed. It can be attached to any phrase (Besler, 1999), thus, it cannot be an indicator of any grammatical category.

Because of all these characteristics, namely flexible word-order, pro-drop and the lack of function words, Turkish is challenging for any model that is based only on word order. However, some language particular characteristics may also turn out to be useful and informative for categorization.

3.1.4 Can complexity turn out to be helpful?

The agglutinative word structure and rich inflectional system may have a positive affect on clustering. Every single alternation of the word produces a new word in this system. Verbs may have tense/aspect, modality, person/number agreement, negation, and voice morphemes (8). Nouns, on the other hand, have plurality, case, and possessive markers (9). In addition, when they are in a predicate position, they have agreement and/or tense/aspect marking as in (10).

(8) aç-tır-a-mı-yor-du-k	(9) kitap-lar-ı-ndan	(10) bu-nlar kitap-tı
open-caus-abil-neg-prog-past-1pl	book-plu-poss&3sg-abl	this-plu book-past
We were not able to make them open.	From his/her books	These were books.

Because every change in the morphological structure of the word (a person marker, for instance) will result in a different word type, the frequency of the verbs and the nouns, and their proportion in the most frequent words is much larger in Turkish. This property increases the amount of information about the categories because the same verb (for example, the verb *oku* 'read') that is marked with different tense markers is very likely to appear in the same context (preceded by the accusative marked noun *kitabı* 'book,' for example).

In addition, with the pro-drop mechanism, there is a reduction in the number of words preceding the verbs. In that way, the subjects are omitted and the verbs that are marked with different person markers (second person and first person, for example) have the same contextual information, therefore the possibility to be grouped together increases.

Thus, although Turkish has properties that may be problematic for a distributional analysis that is based on word order, it has some other properties that may turn out to be helpful for categorization of words, especially for verbs, as described above.

3.2 Developmental properties

Another challenging property of Turkish is that the emergence of inflections, thus the evidence for grammatical categories, is observed to occur at as early as 1;3 in Turkish-speaking children's speech (Ekmekçi, 1979; Aksu-Koç and Slobin, 1985). The productive use of inflections starts around 1;6 (Ketrez, 1999, Ketrez and Aksu-Koç, 1999; 2000) and the acquisition of the whole inflectional system is completed around 2:0-2:6, the age when English-speaking children start producing two word

utterances (Bloom, 1970; Brown, 1973), and thus are considered to be starting to develop categories. Turkish is challenging for any learning model that is based *only* on the statistical regularities of *word* distributions in input because that model should be able to explain the very early emergence of inflectional morphology and the category distinctions in Turkish. The early acquisition of categories is related to particular properties of the language such as the saliency and regularity of affixing (Aksu-Koç and Slobin, 1985). Linear word order, that is the distribution of words across utterances, is not considered helpful.

To summarize, in terms of both its structural features and the developmental properties, Turkish appears to be a challenge for a learning model that is based only on the regularity of the linear order of words and that does not take into consideration any other structural property, such as word morphology. In the present study, I investigate whether or not there are any distributional cues that may be used by a Turkish child, together with other cues that are proposed in the literature. I report that despite the typological challenges, Turkish *does* have some distributional information potentially available for infants, although it is quite limited when compared to the information provided in English child-directed speech.

4. Distributional Analysis Method and the corpora

I apply the methodology developed by Mintz *et al.* to the Turkish corpora. In sections 4.1. and 4.2, below, I describe the subjects and the procedure. In section 4.3., I discuss how the transcriptions are made, together with more information about the structure of the Turkish language.

4.1 Subjects

4.1.1 Speakers

The analysis is based on adult speech directed to four monolingual Turkish children between the ages 1;1 and 2;0. There are in total 12 adult speakers. The data were collected in İstanbul for a longitudinal study of the acquisition of Turkish conducted at Boğaziçi University Psychology Department. The recordings involve natural adult-child conversations recorded on an audiotape at home by the adult herself/himself. No special material or research design is used for data collection. During the sessions children were engaged in daily activities and games introduced by the adults. The adults that took part in the research include middle-class university-educated parents, grandparents, a caretaker and a stranger, who all speak the standard dialect of Turkish. As shown in Table-1, there are totally 15,122 utterances and 76,284 tokens analyzed in the study. Each child's corpus is analyzed individually.

Table 1: Speakers

Child	**Age (child)**	**Speakers (n=12)**	**# of Utterances**	**# of Tokens**
Azra	1;1-2;0	Mother, father, grandmother	2525	11999
Deniz	1;3-2;0	Mother, father, grandmother	7342	38785
Mine	1;6-1;11,23	Mother, father, stranger	1683	8368
Tuna	1;3-2;0	Mother, father, caretaker	3572	17132
		Total:	15122	76284
		Mean:	3780	19071

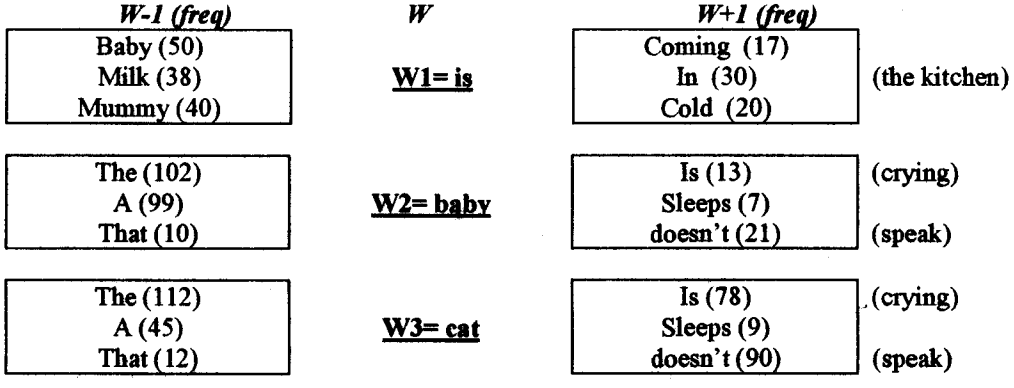
4.1.2 Children

The speech analyzed in the present study is directed to children between 1;1 and 2;0. In the first session analyzed, Azra is in her pre-linguistic stage, Deniz, Tuna, and Mine are in their pre-morphological stages; produce single, uninflected or onomatopoeic words. In the last session analyzed, Azra, Mine and, Deniz are in their morphological stage, producing inflections and multi-word utterances. Deniz, when compared to the others, is much more advanced in terms of her morpho-syntactic development. Tuna, however, can be considered to be in the morphological stage as well, but her development is much slower than that of the others (Ketrez, 1999).

4.2 Procedure

The model I adopt, groups words based on the similarity of their preceding and following lexical contexts. First, the program builds up a list of all the different words that appear in the corpus. Then the words that come immediately before and after each word and their frequency of occurrence in that particular position are recorded. Figure-1 below shows a schema of the distributional contexts of three example words in English[2]. The words in the middle under *W* are the target words whose distribution is analyzed. The words on the left, under *W-1* are the preceding context words and those that are on the right, under *W+1* are the following words. The boxes contain the context words of each target word. The numbers in the parentheses are the frequencies of occurrence.

Figure 1: A schema of the distributional contexts

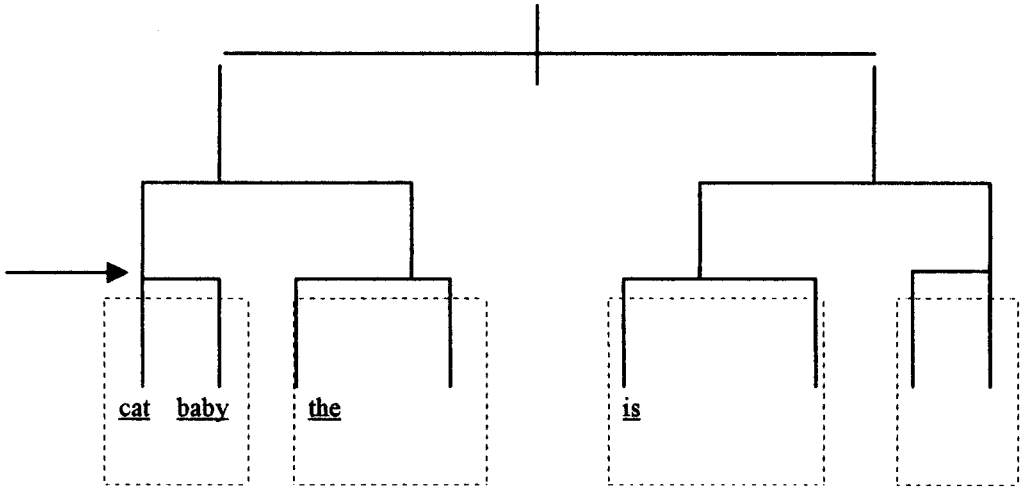


Only the contexts of 200 most frequent words are recorded as targets and only the 200 most frequent words are included in determining the target word contexts. I pick 200 words to be able to compare my results directly with that of the English study, which is also based on the most frequent 200 words. If an infrequent word occurs in a context, for example if the word *sleeps* in *the cat sleeps* is not one of the most frequent words, it is not included as a context word, no entry will be added to that slot. If a target word is the first word of the utterance no entry is made in the *w-1* list of that word. Then the *w-1* list and *w+1* list are taken together and evaluated as the context for a target word.

After the preceding and following contexts are recorded, the context similarity of words is determined. This is done by concatenation of preceding and following contexts of two target words as a vector. Then the context similarity of two words can be determined by computing the angle between the context vectors in a 400-dimentional space. The smaller the angle is, the more similar are the two distributional contexts.. When each target word is paired with all the other target words and their distributional contexts are compared, the result is a similarity rating for each pair of word. Then this similarity information is submitted to a HCA. Figure-2 represents such a hierarchy. The similarity of two words is represented by the height in the tree of the lowest common branch. The lower they are connected, the more similar they are. In this schematic hierarchy, words *baby* and *cat*, for instance, are very similar and they are both different from the word *is*.

[2] Figures are adapted from Mintz *et. al.* (2002). The examples, the words in the figures, are given for illustration purposes. They are not actual analysis results.

Figure 2: A representative result of HCA and selection of levels



In this hierarchy, similar words, that is, the words that have the same distributional contexts cluster together. The target words are then labeled according to their actual grammatical category, as nouns, verbs or adjectives, to determine whether or not the words belonging to the same grammatical category are grouped together.

First, the analysis is based only on the immediate contexts. Only the words preceding and following the targets are included in the distributional context information. In the other analyses, the context window is increased to six words preceding and following the targets in order to form more natural contexts and avoid the possible arbitrariness of the immediate contexts. Increased window sizes are especially important in a language like Turkish, because in this way, we can include the words that are switched in linear order. In examples (11-12) below, the words *kitabı* and *Deniz* are both included in the preceding context of *okudu* in the two-word analysis, whereas in one word analysis, they are treated as two different contexts. In determining the maximum window size, I looked at the maximum utterance length. The longest utterance is 13 words long in the corpus, although the average MLU (words) of all the adults is 2.60. Therefore, the largest context window I have in the analysis is 6-word long.

(11) Deniz kitab-ı oku-du
Deniz book-acc read-past
Deniz read the book.

(12) Kitab-ı Deniz oku-du
book-acc Deniz read-past
Deniz read the book.

Having determined the context similarities, the success of the analysis is evaluated qualitatively and quantitatively. Both qualitative and quantitative methods are used to observe how well the analysis grouped the words belonging to the same category together. In order to see this, a certain level in the hierarchy is determined (shown by the arrow in Figure-2) and all the clusters that are below this determined point are evaluated. In Figure-2, for instance, we have four clusters below this particular level that the arrow points to. They are shown circled by the broken lines.

The cluster attachment height is determined with a measure based on the notion of Unique Entropy developed by Mintz (2000). The level selected should include the number of clusters that will group sufficient words together, but on the other hand, it should not be very high so that the categorization can be meaningful. In Figure-2 for instance the words *cats* and *baby* are considered similar/ belonging to the same category and they are not similar to *the* or *is*. The level selected for analysis is marked with the arrows. Below this point *cat* and *baby* are clustered together and the other words, *is* and *the*, are left outside of the cluster.

The quantitative evaluation measure is called *Purity* (Mintz *et.al.*, 2002). The purity value is calculated individually for nouns and verbs. Purity of a noun category, for instance, represents the degree to which these clusters contain only nouns. In other words, it shows the degree to which this cluster is a *pure* cluster of nouns. In the selection shown in Figure-3, the purity of the cluster that has

cat and *baby* (assuming that there is no other word in this cluster) is quite high. It is a pure cluster of nouns.

Having followed all these steps, we end up with a set of purity values for nouns and verbs. In order to determine whether or not purity values show a regular distributional pattern for nouns and verbs, we determine chance purities, which are based on the pseudo-corpora created from each child's actual corpus randomly. These corpora have the same number of utterances and frequency of words as the original corpus and have a random word order pattern. Ten such corpora are created and the purity values are calculated. Then an average of these values, which represent the chance values, are compared with the actual purity values.

4.3 Transcriptions

4.3.1 Orthography and phonological properties

The data are transcribed orthographically, according to the CHAT conventions of the CHILDES project (MacWhinney and Snow, 1985; 1990), but paying attention to the phonological properties in individual speaker's speech, such as consonant deletions (e.g., [r] deletions in words such as *gidiyo(r)lar* 'they are going'), contractions (e.g. *na:pıyolar* in stead of *ne yapıyorlar* 'what are they doing?') and vowel length (e.g. *Deni::z,*) which are not reflected in orthography. Other than these, Turkish orthography reflects the phonological system of the language, that is, each character uniquely corresponds to a phoneme, and thus, orthographic script is very close to the phonological transcription.

4.3.2 Question particle

I analyze the question particle *-mI,* which is analyzed as a part of the word it follows by Küntay and Slobin (1996), as a distinct word rather than a bound morpheme for the following reasons. Despite the fact that it undergoes vowel harmony, which suggests that it is a bound morpheme, it is considered outside of the *phonological word* (Kabak and Vogel, 2000) in terms of its prosody. It is never stressed when it is the rightmost morpheme of the words, although the rightmost syllable has the primary word stress in Turkish (Lees, 1961; Lewis, 1967; Sezer, 1982). Secondly, its position is quite flexible as opposed to bound morphemes, which have fixed positions within the word structure. It can be attached to any clause (Besler, 2000). In addition, we also know that, in speech segmentation, native speakers of harmonic languages follow the prosodic cues rather than vowel harmony when these cues are in contradiction with one another (Vroomen *et.al.*, 1998). Based on these features, the question particle is analyzed as a separate word in the present analysis.

4.3.3. Ambiguous forms

The ambiguous forms such as the verb *yedi* 'he ate' and the numeral *yedi* 'seven' are excluded in the list of target words, but remained as context words when they appear as a member of one of the categories in the corpora less than 90% of the time. In other words, I did not include the word *yedi* when I was looking at the verb distributions, but I did not totally it in the analysis as it appeared as a context word.

5. Results

Table-2 shows the total number of words included in each analysis as well as the number of nouns and verbs in each corpus. The last column in the table shows the number of words that belong to other lexical categories such as adjectives or conjunctions. As shown in this table, nouns and verbs constitute almost 50% of the total number of words analyzed in the study.

Table 2: The number of words, nouns and verbs (types)

Child	No. of words	Nouns	Verbs	Other
Azra	206	57	47	102
Deniz	199	49	41	109
Mine	189	46	49	94
Tuna	204	47	58	99
Mean:	199.5	49.75	48.75	

The categorization of verbs and nouns are analyzed quantitatively, by looking at the Purity values of each category at different window sizes, and qualitatively by an examination of the individual words in clusters.

5.1 Purity values

The purity values are determined through three different analyses conducted on each of the four corpora individually. The analyses look at the noun and verb contexts in one, two and six word window sizes.

Table 3 shows the purity values. The scores shown in the column on the left are the mean purity values of four children and are based on the most frequent 200 words in each child's corpora. The chance purity for nouns and verbs are obtained from the analysis conducted on the pseudo corpora.

Table 3: Purity values in 1-, 2- and 6-word window analyses

Child	1-word		2-word		6-word	
	Nouns	**Verbs**	**Nouns**	**Verbs**	**Nouns**	**Verbs**
Azra	.47	.48	.45	.49	.45	.54
Deniz	.41	.51	.42	.54	.42	.51
Mine	.46	.47	.46	.51	.47	.49
Tuna	.50	.56	.53	.53	.51	.58
Mean:	**.47**	**.49***	**.47**	**.52***	**.47**	**.54***
Chance Purity:	*.44*	*.44*	*.45*	*.45*	*.44*	*.45*

*$p < .05$

As seen in Table 3, in one word context analysis, the purity value of verbs is .49 and it is significantly better than the chance purity value, which is .44 ($t(3)=3.87$, $p=.030$). This result shows that the verbs in the Turkish corpora display a regular distributional pattern according to their immediate contexts.

The purity value of nouns is recorded as .47. It is very close to the purity value of verbs (.49) and the chance purity for noun (.44) is the same with the one recorded for verbs. Despite these similarities between the purities of nouns and verbs, the noun purity is not significantly different from chance purity ($t(3)=0.90$, $p=.430$). The difference in the significance of nouns and verbs, despite the similar values, suggests that there is individual difference and therefore more variation across the four corpora I am analyzing. Consequently, as opposed to verbs, nouns in the corpora do not cluster together according to their immediate context information.

In summary, noun distribution does not display any regularity in Turkish but verbs cluster significantly well. Considering the language particular properties, it is quite surprising to get significant results at least in one of the categories.

In the second analysis, I increase the context size to two words, that is, I look at two words, rather than only one word, preceding and following the target word. According to this analysis, the purity values recorded for nouns and verbs are .47 and .52, respectively, and the chance purity value that they are compared to is .45 for both categories. When the window size is increased minimally in this way, the verb purity is still significantly better than chance ($t(3)=4.75$, $p=.016$). This minimal change in the context size does not have a significant influence on the noun purity, it is still not significantly better than chance ($t(3)=0.49$, $p=.65$).

Then I compare the purity values of the one-word window analysis with that of the two-word window analysis to see whether or not there is a significant difference between the two window sizes.

The difference between the purity values of the two windows is observed to be significant for verbs ($t(3)=4.25$, $p=.023$), it is not significant for nouns ($t(3)=0.82$, $p=.47$). In other words, a slightly larger window size is better for verb purity, but it does not have a significant effect on the noun purity.

In the third analysis, the window size is increased to six words. The noun and verb purities are recorded as .47 and .54. The chance purity is .44 for nouns and .45 for verbs. The verb purity is still significantly better than chance ($t(3)=4.55$, $p=.019$) in this analysis. Noun purity, however, does not improve. It is recorded as .47 in both two-word and six-word analyses and thus this difference of purity values in two window sizes is not statistically significant ($t(3)=-0.28$, $p=.79$). The noun purity is still not better than chance ($t(3)=1.22$, $p=.30$), either.

In summary, the results based on the quantitative analysis of the noun and verb clustering show that nouns do not cluster together according to the contextual information in their immediate context whereas the clustering pattern of verbs have a regularity that is better than chance. A minimum increase in the window size has a positive effect in verbs, resulting in a significant increase in the verb purity. The verb purity gets even better at six-word window size. We do not see a similar improvement in nouns. The noun purity is never better than chance, and the change in context size does not have any significant influence on the distributional information of nouns and purity scores.

5.2 Examination of clusters

The second phase of the analysis evaluates the individual nouns and verbs in clusters. Table 4 below, exemplifies actual clusters of the corpora of Deniz. In this table, those words that are shown in bold face are verbs and those that are in italics are nouns. Each numbered line shows one cluster and each cluster is separated from the others with a solid line. The gloss and the translation of the words are given in Appendix II. Deniz's sample in Table 4 gives a picture of clustering. In this sample, verbs are grouped together, as in cluster (10). Nouns appear in big clusters, too, such as (1) and (3). The rest of the nouns and verbs display a spread distribution across clusters. In general, we see small groups of verb and noun clusters and only one or two big clusters. There are both nouns and verbs in almost all clusters. There is no cluster, which contains only nouns or only verbs exclusively.

Table 4: Clusters in Deniz's 1-word window analysis

(1) *çocuğun abinin dedenin havuç kitabı kitap teyp zaman* cici da mu ne kırmızı kim kimin benim
(2) *baba kızım* bunu bunun o onlar onları onu onun orası orda ya
(3) **bitti** *teybi anne aydede kalem kedi omi ekmek ev* değil di: diye beni gene tabi yok öyle
(4) **aldım** de hiç mi sana seni
(5) **bak** *adı ay dede deniz denizcim* e evet ha hayır hoppa oh peki başka bu şimdi a a: aa aaa acaba ama
(6) hıh
(7) **na:pıyosun oku söyle yap** *abla mariya* bana bişey hadi lütfen nerdesin niye
(8) **alalım bakalım gel** *canım elma mama* tamam buraya
(9) **bakiym benziyo**
(10) **geldi istiyosun olmuş olsun yapalım yapiym yaptın yapıyo yapıyosun** *renk resmi* var vardı varmış demek diyo güzel kadar
(11) *saniye tane yer dakka balon* daha iyi
(12) *hayatım* çünkü ki
(13) **gitti** allah
(14) **ver koy** *doktor* mı
(15) **diyorum yaptım al** aferin bugün
(16) *kuş* minik
(17) şöyle hah
(18) **na:pıyo na:pıyolar** *çocuk abi bebek* burda işte ah
(19) **istiyo oldu olur uyuyo** *palyaço su top*
(20) *annesi büdü edi* ben bence bi biz sen senin ve
(21) **dur** böyle
(22) ayı nerde
(23) kime nereye

(24) biraz ondan
(25) burası tık
(26) ıh
(27) şu
(28) çok
(29) artık
(30) **buyrun**
(31) **düştü**
(32) efendim
(33) gıdı
(34) **geliyo**
(35) hı
(36) hiç
(37) ham
(38) mısın
(39) mı
(40) misin
(41) musun
(42) nası
(43) pıt
(44) *saat*
(45) sonra

Although clustering does not seem to be very regular because of the spread of the members of the categories across clusters, it is still very different from the example in Table 5, which is one of the randomized corpora derived from Deniz's samples. When we compare the actual corpora clusters in Table 4 and Table 5 with this sample, we see the structural difference of the patterns. In the randomized sample, almost half of the words appear in one big cluster and therefore, result in a low purity score for both categories. Although nouns are all together, the purity scores go down because there are also other categories in this cluster. The rest of the words appear in small, mostly one-word clusters. Thus, there is a patterning in the actual corpora clusters and these patterns in the corpora might be useful to children in categorization.

Table 5: Randomized sample clusters from Deniz corpus.

(1) **na:pıyolar yapıyım** yer
(2) **al alalım aldım bak bakalım bakiym buyrun istiyosun oku oldu olmuş gel dur yapıyo yapalım yaptım** *anne büdü bebek saniye balon dakka deniz denizcim elma havuç kızım kedi kitabı kitap saat* şimdi şu günkü gok a: aa aaa adı ah ama ay böyle bana ben benim bi bişey bu bunu burası buraya burda da daha de değil demek di: e efendim evet güzel hı hıı hıh ha hadi hah ham hayır işte iyi kırmızı ki kim lütfen mı mi misin mu ne nerde nereye niye o onları onu onun orası orda pıt peki sana sen senin sonra tamam var varmış ya yok
(3) *abi* artık
(4) *teybi çocuk* biraz bunun cici
(5) a biz
(6) *ayı dede edi teyp zaman* gitti tık
(7) **yaptın geliyo** *hayatım* başka kadar olur onlar vardı
(8) **düştü** mü
(9) öyle *abla* ve
(10) **yapıyosun** hiç seni
(11) *resmi* gıdı
(12) **na:pıyosun** *mama* mısın nerdesin oh
(13) *top aydede* beni
(14) **benziyo** *renk* diye
(15) *canım dedenin*

(16) *doktor* kimin nası
(17) *kalem* aferin diyo
(18) *çocuğun* ondan
(19) **koy ıh**
(20) na:pıyo ver
(21) *su* tabi
(22) şöyle
(23) *abinin*
(24) acaba
(25) **allah**
(26) *annesi*
(27) *baba*
(28) **bitti**
(29) bugün
(30) **diyorum**
(31) *ev*
(32) **geldi**
(33) gene
(34) hoppa
(35) **istiyo**
(36) *mariya*
(37) minik
(38) musun
(39) **olsun**
(40) *omi*
(41) *palyaço*
(42) **söyle**
(43) *tane*
(44) **uyuyo**
(45) **yap**

In summary, although the clustering patterns of nouns and verbs in the actual corpora do not seem to be regular, the difference between the randomized corpora and the actual corpora obviously display that there is still some patterning in Turkish.

Another property of clustering that is worth mentioning is the role of word morphology in clustering. When we have a closer look at the actual words in the clusters, we see that some words in some clusters are the derivations from the same stems. In Table-4, in (1) for instance, the words *kitabı* book-acc and *kitap* 'book' are different morpho-syntactic forms derived from the same stem *kitap* 'book.' In (10), in the same table, the verbs *yapalım* 'do-opt-3p,' *yapıym* 'do-opt-1s' *yaptın 'you did (it)' yapıyo* 'he did (it)') *yapıyosun* 'you are doing (it)' appear together. Henceforth, I will call the words/stems that appear in only one form, *singletons*. The word *abla* 'sister' in Deniz's sample, for example, is a singleton because we do not have any other word that is derived from the same stem *abla* 'sister'. The stems that appear inflected by various suffixes, such as the words *kitap* and *kitabı* that share the same stem *kitap* will be called *non-singletons* in the analysis below. The groups of words such as *kitab* and *kitabı*, that is the set of words sharing a stem will be called *non-singleton sets*.

Both noun and verb *non-singletons* co-occur with the verbs and nouns with which they share the same stem in some clusters. However, this is more often seen in verbs, rather than nouns. The difference observed is due to the fact that nouns have less morphological variation and thus are more likely to appear as singletons when compared to verbs in Turkish. In Deniz corpus, for example, no nouns with more than two different forms are recorded in 200 words. Verbs, however, have richer morphology, and consequently have more morphological variants. They are more likely to appear as non-singletons. The different verb forms appear in as many as eight forms (in the case of *yap*' do/make'). As we see in Table 4, some of the members of non-singleton sets are grouped together.

This picture raises the question of whether or not the regular clustering of verbs result from this morpho-synactic property? Does the difference between the purity of the verbs and nouns result from

the fact that verbs have a richer morphology, more non-singleton sets, and therefore have a higher potential to cluster together?

To answer these questions I look at the singletons versus non-singletons in the corpora. Table 6 shows the percentages of the singletons and non-singletons.

Table 6: Singletons versus non-singletons (percentages)

Child	**Verbs**		**Nouns**	
	Non-singleton	**Singleton**	**Non-singleton**	**Singleton**
Azra	76 (n=36)	24 (n=11)	46 (n=26)	54 (n=31)
Deniz	71 (n=29)	29 (n=12)	29 (n=14)	71 (n=35)
Mine	45 (n=22)	55 (n=27)	48 (n=22)	52 (n=24)
Tuna	50 (n=29)	50 (n=29)	21 (n=10)	78 (n=37)
Mean:	60.5	39.5	36	63.7

An example of this distribution is seen in Table 7, which displays the noun and verb clustering in Deniz's one word-window analysis. In the first column, the nouns *abi* 'brother' and *abi-nin* 'brother-gen&3sg' and *anne* 'mother' and *anne-si* 'mother-poss&3sg' have the same stem, and are therefore non-singletons. Similarly the words listed on the third column, such as *ist-iyo* 'want-prog', *isti-yo-sun* 'want-prog-2sg', are the verb non-singletons.

In Table 6, we see that an average of 36% of the nouns are non-singletons whereas 63.7% of the nouns are singletons. This distribution is the opposite in verbs. 60.5% of verbs are non-singletons and 39.5% of the verbs are singletons. This is the general picture of all the corpora. As seen in the table, there are individual differences among children. Although in Azra, Deniz and Tuna, the morpho-syntactic variation is a property of verbs rather than nouns, in Mine corpus nouns and verbs almost have an equal degree of variation.

Does this morpho-syntactic characteristic affect the clustering of lexical items and consequently the purity values of the words? Do we have the members of the non-singleton sets grouped together and is this the reason why we have better purity values in verbs?

Table 7: An example. Singletons and non-singletons in Deniz's 1-word window clusters

Nouns	**Non-singletons (sets) (n=14):** abi, abinin, anne, annesi, çocuk, çocuğun, dede, dedenin, deniz, denizcim, kitabı, kitap, teyp, teybi. **Singletons (n=35):** abla, adı, ay, aydede, ayı, baba, balon, bebek, büdü, dakka, doctor, edi, ekmek, elma, ev, havuç, hayatım, kalem, kedi, kuş, kızım, mama, mariya, omi, palyaço, su, tane, top, zaman, renk, resmi, saat, saniye, canım, yer
Verbs	**Non-singletons (sets) (n=29):** istiyo, istiyosun, na:pıyo, na:pıyolar, na:pıyosun, oldu, olmuş, olsun, olur, yap, yapalım, yapiym, yaptım, yaptın, yapıyo, yapıyosun, yapıyım, demek, de, al, alalım, aldım, bakalım, bakiym, diyo, diyorum, gel, geldi, geliyo **Singletons (n=12):** bak, benziyo, bitti, buyrun, dur, düstü, gitti, koy, oku, uyuyo, ver, söyle

To answer these questions, I look at the individual verbs and nouns clustering in each child's corpus and see whether or not non-singleton sets cluster most of the time. Table 8 shows the results of the one-word window clusters. In summary, the majority of the verbs that cluster with other verbs are not necessarily the members of the same non-singleton sets, that is, they are verbs derived from different stems.

Table 8: Singletons and non-singletons in their clusters (percentages)

	Verbs		**Nouns**	
	non-singleton (sets)	**singleton**	**non-singleton (sets)**	**singleton**
Azra	18 (n=7)	82 (n=32)	14 (n=6)	86 (n=39)
Deniz	33 (n=12)	67 (n=24)	10 (n=4)	90 (n=35)
Mine	6 (n=2)	94 (n=32)	16 (n=6)	84 (n=32)
Tuna	18 (n=9)	83 (n=41)	5 (n=2)	95 (n=40)
Mean:	19	81	11	89

In this analysis, only the verbs and nouns that cluster with the words belonging to their own lexical category are included. A verb, for instance, is excluded if it is the only verb in its cluster. When two members of a non-singleton set (*e.g., yaptı, yaptım*) are in the same cluster they are added to the non-singleton list (one score per verb). If there are other verbs in that cluster (*e.g., koy, ver*), they are added to the singleton category, too, because they also group with other verbs. In this way, we look at both the clustering of both singletons and the non-singletons.

As seen in Table 9, the non-singletons constitute only 19% of the clustering verbs. The majority (81%) of verbs cluster with singletons. This picture suggests that although the word morphology has an affect on the purity values, it cannot be the only reason why verb purities are high when compared to the nouns. When we look at the actual clusters we also see that the members of the non-singleton sets do not necessarily cluster with each other. In Deniz's samples, for example, although the majority of the *yap* 'do/make' verbs are grouped together in cluster (9), *yap* 'do/make', the imperative form appears in cluster (7) and *yaptım* 'I did' appears in (14). Similarly, the two verbs *olmuş* 'it happened' and *olsun* 'let it be/happen' are grouped together, the two other forms of the same verb, *oldu* 'it happened' and *olur* 'it happens' appear in another cluster. The word *bak* 'look' never clusters with the other variants of the same verb, for instance. Actually, it does not cluster with other verbs, rather it appears with the words, such as *yes, no* in all the corpora.

In short, it is not necessarily the case that the members of the non-singleton sets cluster together. The reason why we have somewhat high purity values is not the consequence of verb morphology alone. However, this morpho-syntactic property *does* seem to have an influence on clustering, although it is not the main reason for higher verb purities.

In addition to the rich verb morphology, another language particular contribution to the clustering is the pro-drop and flexible word order. It is interesting to notice that the verbs marked with different person markers are grouped together. In Table-4, for example the words *yaptın* 'you did it', *yapalım* 'let us do it', *yapayım* 'let me do it' are clustered together. This might be a result of the pro-drop mechanism in the language as well as the flexibility of nouns. Because the nouns/subjects are not stable or not present in the surface at all, the contexts of the verbs remain the same although they have different subjects. In this respect, the absence of the subjects is helpful in the clustering of the same type of verbs.

In short, all these language particular factors, the rich inflectional system, word order flexibility and pro-drop affect the clustering of verbs, although they may be considered to be problems for such an analysis. As discussed above, these characteristics affect verbs more than nouns. Morphology is much richer in verbs, so verbs have more morpho-syntactic varieties of the same form. Reduction or the displacement of the arguments also help the verbs appear in similar contexts no matter what their subject are. Therefore, it is not surprising that we see a difference in the degree to which clustering is much more regular in verbs.

6. Discussion

In this study, I looked at the distribution of nouns and verbs in adult speech directed to four Turkish children in order to see whether or not there was any regularity in the distribution of these categories. I adopted the method developed by Mintz *et al.* (2002) so that I could compare my results with that of the English study directly. Turkish was especially interesting and challenging for such an analysis based on the word distributions because of its flexible word order pattern and other language particular and developmental properties discussed in the relevant sections above.

First, I looked at the words in their one word, two word and six word window contexts. The results of the analyses suggested that there was a difference in the distributional regularity of nouns and verbs in Turkish. Nouns did not cluster well no matter how large the window size was. Verbs, however, clustered well, and the clustering degree improved significantly even with a minimal increase in the context size (from one word window to two word window).

The results contrast with the results reported for English in Mintz *et al.* (2002). In English, both nouns and verbs cluster significantly well and a minimum increase in the window size does not affect the clustering pattern. The clustering is much better in English not only because the purity values are much higher (.76 for nouns and .66 for verbs) but also because the chance purity obtained form the pseudo-corpora is low (.34 for nouns and .30 for verbs). Thus a comparison of the two sets of purities (that of the actual corpora and the pseudo corpora) result in a much more significant difference and shows a much more regular distributional pattern. In Turkish the purity values of the actual corpora are lower (.48 for nouns and .49 for verbs) and the purity of the pseudo corpora is higher (.44) when compared to English. Consequently, Turkish does not have a distributional pattern as regular as the one observed in the English analyses.

In Turkish, even a minimum change in the window size had an influence on the verb purity. Such a change does not affect the purity values in English. It is not surprising to see that English distributional information is not affected while we see a significant improvement in Turkish. As discussed above, because of scrambling, words in the contexts of the targets change their positions in Turkish. When we increase the window size we include the words that switch their order, and the linear order of the words become irrelevant. Such a change does not have an effect in English because word order is not as flexible as it is in Turkish and also because, in the one-word window analysis purity values are already very high and it is hard to get even higher results.

Secondly, I examined the individual words and their clustering patterns qualitatively. These analyses suggested that rich morphology had affected the clustering pattern, that is, some of the non-singleton words, the words that were the morpho-syntactic variants of some other words in the samples analyzed, clustered together. However, this property did not seem to be the only factor in the clustering of verbs as I discussed in the previous section.

In the examination of the clusters, I also saw that nouns and verbs in Turkish corpora had a spread patterning, that is, they appeared in small clusters together with words belonging to other categories. Whereas in English (as reported by Mintz *et al.* (2002), all the nouns are grouped together in one big noun cluster and all the verbs appear in another cluster that contains exclusively the verbs. There is no instance where a noun and verb are clustered together in English, whereas in Turkish I saw that those clusters that contained a member of the noun category might also have verbs. Despite that difference in clustering, I still have some regular results for verbs in Turkish.

In summary, clustering, and thus the categorization of nouns and verbs, was not as regular as it was in English. Due to language particular factors, this result was not surprising. What was quite unexpected was to see that verbs clustered significantly well. Thus, the overall results obtained from the verbs showed that, even in a language like Turkish, there were regular cues that might help a child in category recognition.

My results contradict Küntay and Slobin's (1996) conclusion that distributional properties of verbs and nouns in Turkish child directed speech cannot be informative. However, there is a methodological difference in the way I look at the Turkish data. What they are concerned with is the question whether or not verbs occur in particular positions, utterance initial or final, for example. They do not look at the surrounding contexts of nouns and verbs. In the present study, I show that a less restricted analysis that looks at all the words that the verbs co-occur with in any position in the utterance leads us to a conclusion that verbs have a regular distributional pattern. The results I obtained from nouns, however, shows that, there is no regular distribution displayed by the nouns just as Küntay and Slobin (1996, 1999) state.

In this study, I looked at the distribution of the words without taking into account any other information that may be helpful in categorization because the aim of the study was to see what word order alone can provide. I did not consider the phonetic similarity of the words deriving from the same roots for instance. In this analysis, the words *yaptım* and *yaptın* were treated as different words because they differed in one sound. Each of these words was evaluated as equally different from a word such as *kitap* as from one another. Children are however sensitive to the similarity between the words *yaptım* and *yaptın* and their difference from *kitap*. In addition, they also receive other phonological and

semantic cues that may be helpful to distinguish categories. For example, although both nouns and verbs are inflected by tense, the tense marker in verbs is stressed, that is it has the primary stress in the word (e.g. *at-tı* 'he threw' versus *at-tı* 'he was a horse'), whereas in nouns it never has the primary stress[3]. When looking at the distributional properties without taking into consideration any other information available to a child, I ignore the capacities of the child and other kinds of helpful information around him/her. However, the present study shows that following the distributional patterns of words in the language, a child may arrive at some conclusions about the grammatical structure and category patterning of the language s/he is hearing. Due to its limitations, distributional information can be used only together with some other cues that are more regular and robust such as word morphology. Without assuming the existence of any cues other than word order, we cannot account for the early acquisition of grammatical categories in Turkish. Inflections, that is, the evidence for categories, start to appear around 1:6 the latest in Turkish children's speech. In English, however, the two word utterances appear around 2:0-2:6, which is considered to be the time when child speech provides evidence for category distinctions. In other words, the overall results of the study suggest that a learning model that limits the information available to the learner with the distribution of whole words across utterances cannot hold. Further research will investigate the ways the distributional information can be used together with other kinds of information, such as the morphological cues.

Using the same methodology I used in this study, another set of distribution analyses can be conducted on the morpheme orders across utterances. Earlier research shows that the functional categories (inflections) can be separated from the lexical categories (stems) (Shi, Morgan and Allopenna, 1998) following distributional, phonological and acoustic cues in Turkish child directed speech. We also know that children's early utterances provide evidence that children are actually aware of the stem-inflection distinctions (Ketrez, 1997; Borer and Rohrbacher, 1998). Therefore we *do* have evidence that children are sensitive to the morpheme boundaries and further distributional research based on the order of inflectional morphology will show us a complete picture of what is available for the child in the input utterances.

A final point I should mention is the size of the corpora I analyzed. The total number of utterances I analyzed was about 15,000. The biggest corpus (Deniz) had about 7,000, and the smallest one (Mine) had 1,683 utterances. This is a very small database when compared to that used for the English study (Mintz *et al.*, 2002), which include an average of 14,00 utterances for each child. In order to be able to make a complete comparison with English, the analysis of Turkish utterances should come from a bigger database. Future research will show to what extend the results of the present study can be generalized to Turkish child directed speech.

7. Conclusion

The present research, which looked at the distribution of words as atomic units in a highly agglutinative and flexible-word-order language, has found that despite the challenges of the language, there might still be some regularities which may be used by children in learning that language. It was surprising to find that there is any regularity at all in this restricted picture. A more realistic analysis, based on the distribution of the morphemes across utterances, will provide us with the complete account of the useful information available to a child. Future research, based on a larger database, will investigate the regularity and the nature of this information.

[3] They are not analyzed as the same tense marker in literature. Kornfilt (1996), for instance, suggests that there is an empty copula following the nouns (and other nominals) and tense is attached to that copula. Because there is a word boundary between the noun and the empty copula, these morphemes cannot take the primary stress. What is important in our analysis here is that both tense markers has the same form (*-DI*) and in an analysis such as the one I am conducting, they are treated as the same, although they do not actually sound the same in actual speech.

Appendix I: List of abbreviations

Abl	ablative	Fut	future	Opt	optative	Sg	singular
Acc	accusative	Gen	genitive	Past	past	Pl	plural
Aor	aorist	Inch	inchoative	Perf	perfective	&	contraction, fusion
Dat	dative	Inf	infinitive	Poss	possessive	-	morpheme boundary
Dim	diminutive	Loc	locative	Prog	progressive	+	compound boundary

Appendix II: Glossary

Below is the list of the words that appear in the tables in this paper. The list includes the 200 most frequent words in the Deniz and Mine corpora only, it is not a complete list of the words analyzed. See Appendix I for the abbreviations.

Nouns

words	gloss	translation	category
abi	elder brother	elder brother	noun
abinin	elder brother-gen&1sg	of elder brother	noun
abla	elder sister	elder sister	noun
adı	name-poss&3sg	his/her/its name	noun
anne	mother	mother	noun
annesi	mother-poss&3sg	his/her/its mother	noun
aydede	Aydede	Aydede	noun (proper)
ayı	(1) moon-acc	moon	noun
	(2) moon-poss&3sg	his/her/its moon	noun
	(3) bear	bear	noun
baba	dad	dad	noun
balon	baloon	baloon	noun
bebek	baby	baby/doll	noun
büdü	büdü	Büdü	noun (proper)
canım	heart-poss&1sg	my heart/honey	noun
çocuğun	(1) child-gen3sg	the child's	noun
	(2) child-poss&2sg	your child	noun
çocuk	child	child	noun
dakka	minute	minute	noun
dede	grandfather	grandfather	noun
dedenin	grandfather-poss&3sg	grandfather's	noun
deniz	(1) Deniz	Deniz	noun (proper)
	(2) sea	deniz	noun
denizcim	Deniz-dim-poss&1sg	my dear/little Deniz	noun
doktor	doctor	doctor	noun
edi	Edi	Edi	noun (proper)
ekmek	bread	bread	noun
elma	apple	apple	noun
ev	house	house	noun
havuç	carrot	carrot	noun
hayatım	life-poss&1sg	my life/honey	noun
kalem	pen	pencil/pen	noun
kedi	cat	cat	noun
kızım	daughter-poss&1sg	my daughter	noun
kitabı	(1) book-poss&3sg	his/her/its book	noun
	(2) book-acc	book	noun
kitap	book	book	noun
kuş	bird	bird	noun
mama	food (baby talk)	food	noun
mariya	Maria	Maria	noun (proper)
omi	Omi	Omi	noun (proper)
palyaço	clown	clown	noun

renk	colour	colour	noun
resmi	(1) picture-poss&3sg	his/her/its picture	noun
	(2) picture-acc	picture	noun
saat	clock/hour	clock/hour	noun
saniye	second	second	noun
su	water	water	noun
tane	number/piece/grain	number/piece/grain	noun
teybi	cassette player-acc	cassette player	noun
teyp	cassette player	cassette player	noun
top	ball	ball	noun
zaman	time	time	noun

Verbs

al	take	take	verb
alalım	take-opt-1pl	let's take (it)	verb
aldım	tak-past-1sg	I took (it)	verb
bak	look	look	verb
bakalım	look-opt&1pl	let us look	verb
bakiym	look-opt&1sg	let me look	verb
bakıyo	look-prog	(he/she/it) is looking	verb
benziyo	resemble-prog	(it) is resembling	verb
bitti	finish-past	(it) finished	verb
buyrun	come&in-opt-2sg	(you) come in!	verb
diyo	say-prog	(he/she/it) is saying	verb
diyorum	say-prog-1sg	(I) am saying	verb
dur	stop	stop	verb
düştü	fall-past	(he/she/it) fell	verb
gel	come	come	verb
gelcek	come-fut	he/she/it will come	verb
geldi	come-past	(he/she/it) came	verb
geliyo	come-prog	(he/she/it) is coming	verb
gitti	go-past	(he/she/it) went	verb
istiyo	want-prog	(he/she/it) wants	verb
istiyosun	want-prog-1sg	(you) want	verb
koy	put	put	verb
na:pıyo	what&do/make-prog	what is s/he/it doing?	verb
na:pıyolar	what&do/make-prog-3pl	what are they doing?	verb
na:pıyosun	what&do/make-prog-2sg	what are you doing?	verb
oku	read	read	verb
oldu	be/happen/fit-past	it happened/fit	verb
olmuş	be/happen/fit-past	it happened/fit	verb
olsun	be/happen/fit-opt	let it be/happen	verb
olur	be/happen/fit-aor	it is/happens/fits	verb
söyle	tell	tell	verb
uyuyo	sleep-prog	he/she/it is sleeping	verb
ver	give	give	verb
yap	do/make	do/make	verb
yapalım	do/make-opt-1pl	let us do (it)	verb
yapıyım	do/make-opt-1sg	let me do (it)	verb
yapıyo	do/make-prog	he/she/it is doing (it)	verb
yapıyosun	do/make-prog-2sg	you are doing (it)	verb
yaptım	do/make-past-1sg	I did (it)	verb
yaptın	do/make-past-2sg	you did (it)	verb

Ambiguous forms

yer	(1) eat-aor	s/he/it eats	verb
	(2) place	place	noun
ay	(1) oh/oops	oh/oops	interjection
	(2) moon	moon	noun

de	(1) say	say	verb
	(2) too	too	conjunction
demek	(1) so/thus in that case	so/thus/in that case	adverb/conjuction
	(2) say-inf	to say	verb

Others

a		oh	interjection
a:		o:h	interjection
aa		o:h	interjection
aaa		o:h	interjection
acaba	I&wonder	I wonder	adverb
aferin	well&done (baby talk)	well done!	other
ah		o:h	interjection
allah	god	God!	interjection
ama	but	but	conjunction
artık	anymore	anymore	adverb
bana	I-dat	to me	pronoun
başka	other	other	adjective
ben	I	I	pronoun
bence	I-according+to	if you ask me	adverb
beni	I-acc	me	pronoun
benim	I-gen&1sg	my	pronoun
bi	one/a	one/a	article/numeral
bişey	something	something	pronoun
biz	we	we	pronoun
böyle	like+this	like+this	adverb/pronoun
bu	this	this	pronoun (dem)
bugün	this+day	today	adverb
bunu	this-acc	this	pronoun (dem)
bunun	this-gen&3sg	of this	pronoun (dem)
burası	here-poss&3sg	here	pronoun
burda	here-loc	in here	pronoun
cici	cute (baby talk)	cute/nice/beautiful	adjective
çok	very/a lot of	very/a lot of	adverb/adjective
çünkü	because	because	conjunction
da	too	too	conjunction
de	too	too	conjunction
daha	more	more/yet	adverb/adjective
değil	no/not	no/not	negation
di:	no/not	no/not	negation
diye	because/so that	because/so that	complementizor
e	well/all right/then	well/all right/then	adverb
efendim	master/sir-poss&1sg	yes	other
evet	yes	yes	other
gene	again	again	adverb
gıdı		tickling sound	interjection
güzel	nice	nice/nicely	adjective/adverb
ha	either/or	either/or	conjuction
hadi	come+on	come on!	interjection
ham		eating sound	interjection
hangisi	which-poss&3sg	which one?	pronoun (wh-word)
hayır	no	no	other
hepsi	all	all	adjective
hı		excuse me?	interjection
hah		ok/now it is done	interjection
hıh		ok/now it is done	interjection
hıı		I see	interjection
hiç	any/ever	any/ever	adverb/adjective
hoppa		oops	interjection
işte	here it is, there	here it is, there	adverb

iyi	good	good/well	adjective/adverb
kadar	until	until	postposition
kırmızı	red	red	adjective
kim	who	who?	pronoun (wh-word)
kime	who-dat	to whom?	pronoun (wh-word)
kimin	who-gen&3sg	whose?	pronoun (wh-word)
lütfen	please	please	other
mı	que		clitic
mısın	que-2sg		clitic
mi	que		clitic
minik	small	small	adjective
mu	que		clitic
musun	que-2sg		clitic
nası	how	how?	adverb (wh-word)
ne	what	what?	pronoun (wh-word)
nerde	where	where?	pronoun (wh-word)
nerdesin	where-2sg	where are you?	pronoun (wh-word)
nereye	where-dat	to where?	pronoun (wh-word)
niye	why	why	pronoun (wh-word)
o	he/she/it	he/she/it	pronoun
oh			interjection
ondan	he/she/it-abl	from him/her/it	pronoun
onlar	he/she/it-plu	they	pronoun
onları	he-she-it-plu-acc	them	pronoun
onu	he/she/it-acc	him/her/it	pronoun
onun	he/she/it-gen&3sg	his/her/its	pronoun
orası	there	there	pronoun
orda	there-loc	in there	pronoun
öyle	like this	like this	adverb
peki	ok	ok	other
pıt		hitting sound	interjection
sana	you-dat	to you	pronoun
sen	you	you	pronoun
seni	you-acc	you	pronoun
senin	you-gen&1sg	your	pronoun
sonra	later	later	adverb
şimdi	now	now	adverb
şöyle	like that	like that	adverb
şu	that	that	pronoun (dem)
tabi	natural	naturally/of course	other
tamam	ok	ok	other
tık		knocking sound	interjection
var	exist	there is	existential
vardı	exist-past	there was	existential
varmış	exist-perf	there was	existential
ve	and	and	conjunction
ya	either/or	either/or	conjunction
yok	exist&not	there is not	existential

References

Aksu-Koç, A. (1994). Development of linguistic forms: Turkish. In R. Berman and D. I. Slobin (Eds.), *Relating Events in Narrative: A Crosslinguistic Developmental Study*, 329-385. New Jersey: Lawrence Erlbaum Associates, Publishers.

Aksu-Koç, A. and Slobin, D. I. (1985). The acquisition of Turkish. In D. I. Slobin (Ed.), *The Crosslinguistic Study of Language Acquisition*. Volume 1: The Data, 839-878. New Jersey: Lawrence Erlbaum Associates, Publishers.

Aslin, R. N., Safran J. R. and Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321-234.

Bates, E. and MacWhinney, B. (1982). Functionalist Approaches to grammar. In E. Wanner and L. R. Gleitman (Eds.), *Language Acquisition: The State of the Art*. Cambridge: Cambridge University Press.

Batman-Ratyosyan, N. and Stromswold, K. (1998). *What Turkish acquisition tells us about underlying word order and scrambling*. Ms., Rutgers University.

Besler, D. (1999). *Question particle and movement in Turkish*. Unpublished MA thesis, Bogazici University, Istanbul.

Bloom, L. (1970). *Language Development: Form and Function in Emerging Grammars*. Cambridge, MA: MIT Press.

Bloomfield, L. (1933). *Language*, New York: Holt, Reinehart and Winston.

Brown, R. (1957). Linguistic determinism and the part of speech. *The Journal of Abnormal and Social Psychology, 55*(1), 1-5.

Brown, R. (1973). *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Borer, H. and Rohrbacher, B. (1998). Minding the absent: Argument for the full competence hypothesis, to appear in U. Laksmanan (Ed.), *Acquisition within the minimalist framework* (tentative title).

Bowerman, M. (1973). Structural relationship in children's utterances: Syntactic or semantic. In T. Moore (Ed.), *Cognitive Development and the Acquistion of Language*. New York: Academic Press.

Cartwright, M. R. and Brent, M. (1997). Syntactic Categorization in early language acquisition: Formalizing the role of distributional analysis. *Cognition, 63*, 121-170.

Cassidy, K. W. and Kelley, M. (1991). Phonological information for grammatical category assignments. *Journal of Memory and Language, 30*(3), 348-369.

Chomsky, N. (1965). *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Clark, E. V. (1982). The young word maker: A case study of innovation in the child's lexicon. In E. Wanner and L. Gleitman (Eds.), *Language Acquisition: The State of the Art*, New York: Cambridge University Press, 390-425.

Ekmekçi, Ö. (1979). *Acquisition of Turkish: A longutidunal study on the early language development of a Turkish child*. Unpublished Ph.D. Dissertation, University of Texas, Austin.

Ekmekçi, Ö. (1986). The Significance of word order in the acquisiton of Turkish. In D. I. Slobin and K. Zimmer (Eds.), *Studies in Turkish Linguistics*, 265-272. Amsterdam: John Benjamins Puslishing Company.

Enç, M. (1986). Topic switching and pronominal subjects in Turkish. In D. I. Slobin and K. Zimmer (Eds.), *Studies in Turkish Linguistics*, 195-231. Amsterdam: John Benjamins Puslishing Company.

Erguvanlı. E. (1984). *Function of Word Order in Turkish*. Berkeley: University of California Press.

Fisher, C. and Tokura, H. (1996). Acoustic cues to grammatical structure in infant- directed speech: Cross-linguistic evidence. *Child Development, 67*, 3192-3218.

Fries, C. C. (1952). *The structure of English*. New York: Harcourt, Brace and Company.

Gleitman, L. R. (1990). The structural sources of verb meaning. *Language Acquisition, 1*, 3-55.

Gleitman L. R. and Gillette, J. (1995). The role of syntax in verb learning. In P. Fletcher and B. MacWhinney (Eds.), *The Handbook of Child Language*, 413-427. Cambridge, MA: Blackwell.

Gleitman, L. R. and Wanner, E. (1982). Language acquisition: The state of the art. In L. R. Gleitman and E. Wanner (Eds.), *Language Acquisition: The state of the art, Cambridge*, UK: Cambridge University Press. 1-48.

Goldfiled, B. A. (1993). Noun bias in maternal speech to one-year-olds. *Journal of Child Language, 20*, 85-99.

Grimshaw, J. (1981). Form, function and the language acquisition device. In C. L. Baker and J. J. MacCarty (Eds.), *The Logical Problem of Language Acquisition*, 165-182. Cambridge, MA: MIT Press.

Harris, Z. S. (1951). *Methods in Structural Linguistics*. Chicago: University of Chicago Press.

Jusczyk, P. W., Hirsh-Pasek, K., Nelson, D. G. K. , Kennedy, L. J., Woodward, A. and Piwoz, J. (1992). Perception of acoustic correlates of major phrasel units by young infants. *Cognitive Psychology, 24*, 252-293.

Jusczyk, P. W., Friederici, A. D. , Wessels, J. M. , Svenkerud, V. Y. and Jusczyk, A. M. (1993). Infants sensitivity to sound patterns of native language words. *Journal of Memory of Language, 32*, 402-420.

Jusczyk, P. W., Cutler, A. and Redanz, N. J. (1993). Infants preference for predominant stress patterns of English Words. *Child Development, 64*, 675-687.

Jusczyk, P. W., Luce, P. A. and Charles-Luce, J. (1994). Infants' sensitivity to phonotactic patterns in the native language. *Journal of Memory and Language, 33*, 630-645.

Kabak, B. and Vogel, I. (2000). *The phonological word and stress assignment in Turkish.* Paper presented at Marburg 2000, 22nd Annual Conference of the Linguistics Association of Germany, March 2000.

Kelly, M. (1992). Using sound to solve syntactic problem: The role of phonology in grammatical category assignment. *Psychological Review, 99*(2), 349-364.

Kelly, M. (1996). The role of phonology in grammatical category assignments. In J. L. Morgan and K. Demuth (Eds.), *Signal to Syntax: Bootsrapping from Speech to Grammar in Early Acquisition,* 249-262. Hillsdale, NJ: Lawrence Earlbaum.

Ketrez, F. N. (1997). Bir çocuğun dil ediniminde ses aktarımı [metathesis in the language acquisition period of a child]. In Ş. Ruhi and D. Zeyrek (Eds.), *XI. Dilbilim Kurultayı: Bildiriler,* Ankara: Ortadoğu Teknik Üniversitesi Yayınları. 83-93.

Ketrez, F. N. (1999). *Early verbs and the acquisition of Turkish argument structure.* Unpublished MA Thesis, Bogazici University, Istanbul Turkey.

Ketrez, F. N. and Aksu-Koç, A. (1999). *Once there was (not) a verb.* Paper presented at the 8th Congress on the study of the child language, University of the Basque Country, Spain.

Ketrez, F. N. and Aksu-Koç, A. (2000). *Acquisition of nouns and verbs in Turkish.* Paper presented at the 10th International Conference on Turkish Linguistics, Bogazici University, Istanbul.

Kornfilt, J. (1984). *Case marking, agreement and empty categories in Turkish.* Ph.D. Dissertation, Harvard University, Cambridge, MA.

Kornfilt, J. (1994). Some remarks on the interaction of case and word order in Turkish: Implications for acquisition. In B. Lust, M. Suner and J. Whitman (Eds.), *Syntactic Theory and First Language Acquisition: Cross-linguistic Perspectives.* Vol. 1: Heads, Projections and Learnability, 171-199. Hillsdale, NJ: Lawrence Earlbaum.

Kornfilt, J. (1996). On some copular clitics in Turkish. In A. Alexiadou *et al.* (Eds.), *ZAS Papers in Linguistics, 6,* 96-114. Berlin: Zentrum für Allgemeine Sprachwissenschaft.

Kural, M. (1992). *Properties of scrambling in Turkish.* Ms., UCLA.

Küntay, A. and Slobin, D. I. (1996). Listening to a Turkish mother: Some puzzles for acquisition. *Social Interaction, Social Context and Language: Essays in Honor of Susan Ervin-Tripp,* 265-286.Hillsdale: Lawrence Erlbaum Associates Publishers.

Küntay, A. and Slobin, D. I. (1999). *Discourse behavior of lexical categories in Turkish child directed speech: Nouns vs. verbs.* Paper presented at the VIIIth IASCL Congress, San Sebastian, Basque Country, Spain.

Lees, R. (1961). *The Phonology of Modern Standard Turkish.* Uralic and Altaic Series 6. Bloomington: Indiana University Press.

Lewis, G. (1967). *Turkish Grammar.* Oxford: Oxford University Press.

Macnamara, J. (1972). Cognitive basis of language learning in infants. *Psychological Review, 79,* 1-14.

Macnamara J. (1982). *Names for things: A study of child language.* Cambridge, MA: MIT Press.

MacWhinney, B. (1995). *The CHILDES Project: Tools for Analyzing Talk.* New Jersey: Lawrence Erlbaum Associates, Publishers.

MacWhinney, B. and Snow, C. (1985). The child language data exchange system: An update. *Journal of Child Language, 12* , 271- 296.

MacWhinney, B. and Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language, 17* , 457- 472.

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

Marcus, G. F., Vijayan, S., Rao, S. B., and Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science, 283,* 77-80.

Mintz, T. H. (2000). Unique entropy as a model of linguistic classification. *Proceedings of the Twenty Second Annual Cognitive Science Society.* Mahwah, NJ: Laurence Erlbaum Associates, Publishers.

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

Morgan, J. and Newport, E. (1981). The Role of Constituent Structure in the induction of an artificial language. *Journal of Verbal Bahavior, 20,* 67-85.

Morgan, J., Shi, R. and Allopenna, P. (1996). Perceptual bases of rudimentary grammatical categories: Toward a broader conceptualization of bootsrapping. In J. L. Morgan and K. Demuth (Eds.), *Signal to Syntax: Bootsrapping from Speech to Grammar in Early Acquisition,* 263-283. Hillsdale: Laurence Erlbaum Associates, Publishers.

Ozsoy, A. S. (1987). Null subject paramater and Turkish. In H. E. Boeschoten and L. T. Verhoven (Eds.), *Studies on Modern Turkish, Proceedings of the Third Conference on Turkish Linguistics.* Tilburg University Press.

Pinker, S. (1984). *Language Learnability and Language Development.* Cambridge, MA: Harvard University Press.

Pinker, S. (1987). The Bootsrapping problem in Language acquisition, In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition*, 399-441.Hillsdale: Laurence Erlbaum Associates, Publishers.

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

Saffran, J. R. , Aslin, R. N. and Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science, 274*, 1926-1928.

Schlesinger, I. M. (1974). Relational concepts underlying language acquistion. In R. Schiefelbusch and L. Lloyd (Eds.), *Language Perspectives: Acquisition, Retardation and Intervension*. Baltimore, Maryland: University Park Press.

Sezer, E. (1983). On non-final stress in Turkish. *Journal of Turkish Studies, 5*, 61-69.

Shi, R. Morgan, J. and Allopenna, P. (1998). Phonological and acoustic bases for grammatical category assignment: A cross-linguistic perspective. *Journal of Child Language, 25*, 160-201.

Slobin, D. I and Bever, T. G. (1982). Children use canonical sentence schemas: A cross- linguistic study of Word Order and Inflections. *Cognition, 12*(3), 229-265.

Tardif, T., Shatz, M. and Naigles, L. (1996). *The influence of care giver speech on children's use of nouns and verbs: A comparison of English, Itailan and Mandarin*. Paper presented at the VIIth International Conference for the Study of Child Langauge, Istanbul.

Vroomen, J. , Tuomainen, J. and de Gelder, B. (1998). The Role of Word Stress and Vowel Harmony in Speech Segmentation. *Journal of Memory and Language, 38*, 133-149.