FROM SYLLABLE TO MEANING: EFFECTS OF KNOWLEDGE OF SYLLABLE IN LEARNING THE MEANING BEARING UNITS OF LANGUAGE

A THESIS SUBMITTED TO
THE INFORMATICSINSTITUTE
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÇAĞRI CÖLTEKİN

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

DECEMBER 2006

Approval of the Graduate School of Informatics.

Assoc. Prof. Dr. Nazife Baykal
Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Prof. Dr. Deniz Zeyrek
Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Dr. Cem Bozşahin
Supervisor

Examining Committee Members

Prof. Dr. Deniz Zeyrek	(METU, FLE & COGS)	_____
Assoc. Prof. Dr. Cem Bozşahin	(METU, CENG & COGS)	_____
Dr. Ayşenur Birtürk	(METU, CENG)	_____
Assist. Prof. Dr. Annette Hohenberger	(METU, COGS)	_____
Assist. Prof. Dr. Bilge Say	(METU, COGS)	_____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: Çağrı Çöltekin

Signature :

ABSTRACT

FROM SYLLABLE TO MEANING: EFFECTS OF KNOWLEDGE OF SYLLABLE IN LEARNING THE MEANING BEARING UNITS OF LANGUAGE

Çağrı Çöltekin

M. S., Cognitive Science

Supervisor: Assoc. Prof. Dr. Cem Bozşahin

December 2006, 48 pages

This thesis aims to investigate the role of the syllable, a non-meaning bearing unit, in learning high level meaning bearing units—the lexical items of language. A computational model has been developed to learn the meaning bearing units of the language, assuming knowledge of syllables. The input to the system comprises of words marked at syllable boundaries together with their meanings. Using a statistical learning algorithm, the model discovers the meaning bearing elements with their respective syntactic categories. The model's success has been tested against a second model that has been trained with the same corpus segmented at morpheme boundaries. The lexicons learned by both models have been found to be similar, with an exact overlap of 71%.

Keywords: Language Acquisition, Syllable, Morpheme

ÖZ

HECEDEN ANLAMA: ANLAM TAŞIYAN DİLBİLİMSEL BİRİMLERİN
ÖĞRENİMİNDE HECE BİLGİSİNİN ETKİLERİ

Çağrı Çöltekin
Yüksek Lisanas, Bilişsel Bilimler Bölümü
Tez Yöneticisi: Doç. Dr. Cem Bozşahin

Aralık 2006, 48 sayfa

Bu tez, anlam taşımayan bir birim olan hecenin, anlam içeren dilbilimsel birimlerin öğrenimindeki etkisini araştırmaktadır. Bu çalışma için, hece bilgisinden yola çıkarak anlam taşıyan dilbilimsel birimlerin öğrenimini hedefleyen bir model tasarlanıp, bu model bilgisayar ortamında gerçeklenmiştir. Model, hecelere bölünmüş sözcükler ve anlamlarından oluşan bir girdiden, sözcükten daha kücük, anlam içeren birimler ve bu birimlerin sözdizimsel kategorilerini öğrenmektedir. Bu modelin başarısını tartmak için, aynı öğrenme yöntemiyle çalışan, ancak girdi olarak heceler yerine biçimbirimlere bölünmüş sözcükleri alan ikinci bir model kullanıldı. Her iki modelin öğrenme sonrası ürettikleri sözlüklerdeki birimlerin %71 oranında örtüştüğü gözlendi.

Anahtar Kelimeler: Dil Edinimi, Hece, Biçimbirim

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

A valid utterance in any natural language consists of a set of linguistic units that are associated with some meaning. Native speakers have the knowledge of these units, along with their meaning and permissible combinations, to produce and comprehend interpretable utterances. Acquiring a language requires the ability to recognise and recall these units—most likely the items in the lexicon— from a seemingly continuous and ambiguous language input.

This study is concerned with the contribution of knowledge of a basic linguistic unit, the syllable, to the process of acquiring the meaning bearing elements of the language.

The problem that the learner needs to solve is associating the sound signal with meaning.

One of the first difficulties children face is extracting the lexical units of the language from speech. The speech signal does not contain markers analogous to spaces or punctuation marks as in written language. Research so far shows that there is no single strategy used exclusively for extracting lexical units from continuous speech signal. Children seem to be sensitive to a number of cues[1] in their language, which have been shown to help identify the lexical units.

Segmenting continuous speech into the lexical units is not the only requirement for acquiring language. The child also has to relate the unit with its semantic content and syntactic role. A lexicon without syntactic and semantic aspects of the units would not be very useful for either understanding or producing meaningful and well formed sentences.

In this study, we investigate the use of syllables in identifying morphemes from linguistic data, i.e. learning a morphemic lexicon. Syllable is one of the basic units of language.

[1] Among others, prosody, distributional regularities of phonotactic units and lexical knowledge are known to be useful. See Section 2.1 for a brief discussion of these cues.

Children are known to develop the ability to identify syllables at a very early stage of language acquisition (Dehaene-Lambertz & Houston, 1998; Mehler et al., 1996). Morpheme is defined as the smallest meaning bearing unit of language.[2] There is little doubt that children learn morphemes to understand and build more complex utterances. Even though most research on the topic focuses on words as lexical items, there are several reasons for assuming a *morphemic lexicon*, i.e. a lexicon having morphemes as base units.[3]

Learning to identify morphemes has some additional difficulties compared to learning to identify words. For instance, words frequently appear in isolation. It is also known that child directed speech contains a large number of single-word utterances (Brent & Siskind, 2001). However, most morphemes are always bound to other morphemes.[4]

In this study, a computational model that learns a *word grammar*, using a statistical learning method, has been developed. The model gets words segmented at syllable boundaries and the logical form of the words as input, and builds a morphemic lexicon. The lexicon maps the phonetic forms of the morphemes with appropriate atomic logical forms, and also assigns a syntactic category to each lexical item. For example, for the input (1),[5] it produces the lexical items in (2).[6]

(1) `ev-de-ki` : $relative(locative(house))$

(2) `ev` := N : $house$
`de` := $N_{loc}\backslash N$: $\lambda x.locative(x)$
`ki` := $N_{rel}\backslash N_{loc}$: $\lambda x.relative(x)$

The same method of learning is used in two different models. One model gets words marked at syllable boundaries (here after the *syllable-based model*), and the other one gets

[2]This is the most common textbook or dictionary definition of *morpheme*, e.g. Fromkin & Rodman (1993, page 44).

[3]See Chapter 2 for a brief discussion, see Marslen-Wilson (1999) for psycho-linguistic reasons and Bozşahin (2002) for computational reasons for assuming a morphemic lexicon, even for morphologically simpler languages such as English.

[4]We use the term *word* as the minimal free form, while *morpheme* is the minimal bound form.

[5]In this example, there is a one-to-one match between syllables and morphemes. However, it should be noted that this is not usually the case, and we do not assume that syllables bear meaning.

[6]We use a subset of Combinatory Categorial Grammar (CCG) formalism, and CCG notation in this work. See Section 2.6 for a short overview of CCG and its use in this study.

input words marked at morpheme boundaries (here after the *morpheme-based model*). Both models are expected to learn lexical items consisting of triplets of *phonetic form* (PF), *syntactic category* (CAT) and *logical form* (LF) as in the items shown in (2). The results of the morpheme-based model are used as an upper bound for what the system can learn. If the knowledge of the syllable is useful in developing the notion of morpheme, we expect the results of the syllable-based model to be comparable with the results of the morpheme-based model.

The input to the models are segmented words, and the logical representation of the whole word. The models are expected to learn which parts of the segmented input are associated with which parts of the possibly complex logical form.

1.1 Limitations

The system we have developed is generic enough to learn other syntactic categories too. However, for practical reasons, the tests in this study have been restricted to a fragment of Turkish nominal morphology.

It should also be noted that we do not claim the cognitive reality of the learning method in this study. The main interest of the study is to test the learnability of the minimal meaning bearing units from the syllable, using a statistical learning system. On the other hand, there is considerable evidence that statistical learning plays an important role in cognitive development, including language acquisition (Kuhl, 2004; Saffran et al., 2001; Altmann, 2002). Children's use of statistical properties of the language they are acquiring is shown by many studies, including Saffran (2003) and Christiansen et al. (2006).

CHAPTER 2

BACKGROUND

In this study, we investigate the effects of syllables on learning morphemes, i.e. the minimal meaningful units of language. This chapter gives a brief review of the current literature on related areas of research, and a short introduction to the tools and formalisms used.

The first subject that is related to the problem at hand is segmenting the speech signal into units that can be mapped to lexical items. The syllable's role in this task is of particular interest to this work. Lexicon, its structure and the acquisition of it by both children and cognitively inspired computational models are also closely related to this thesis. Besides reviewing the research in these areas, this chapter gives a very short introduction to part of the Turkish morphology that is relevant to the data used, and the Combinatory Categorial Grammar (CCG), the theory of grammar that our computational model is based on.

2.1 Speech Segmentation

One of the first challenges of arriving at the meaning bearing units from the speech input is segmenting the seemingly continuous speech into lexical units.

Unlike written language, where words are separated with spaces,[1] spoken language does not have discernible marks between lexical items.

Speech segmentation is a relatively well studied subject, and it is known that we do make use of a number of cues to find the boundaries of lexical units:

[1]Some languages, e.g. Japanese, do not have markers between words. So, segmenting the written language input is also a problem for processing the written form of some languages.

- Lexical knowledge is one of the cues which helps identifying units in speech input. Having a fairly complete lexicon helps matching the input with the lexicon and reducing the possible lexical hypothesis. Cohort (Marslen-Wilson, 1973) and TRACE (McClelland & Elman, 1986) are examples of early models applying this principle to word recognition. Although useful, lexical information is not enough by itself to do all the segmentation task, due to the number of possible parallel hypotheses (Shillcock, 1990). Furthermore, since children acquiring language have a rather limited lexicon, lexical knowledge is not very useful during early stages of language acquisition.

- Prosodic cues—stress, pauses, segmental lengthening, metrical patterns, and intonation contour— are also known to be helpful in identifying lexical units from continuous speech input. The use of prosodic cues to segment speech by infants and adults is confirmed by a number of empirical studies, including (Jusczyk et al., 1993; Morgan, 1996; Jusczyk et al., 1999b; Mattys & Samuel, 2000; Jusczyk et al., 1999b), and used in models of speech segmentation (Cooper & Paccia-Cooper, 1980; Glietman et al., 1988).

- Phonotactic constraints, and the permitted sound patterns of the language are also shown to be used by both adults and children in speech segmentation (Saffran et al., 1996; Mattys & Jusczyk, 2001). A number of computational studies (Brent & Cartwright, 1996; Allen & Christiansen, 1996; Christiansen & Allen, 1997) also report better performance in the segmentation process and success in learning to segment when phonotactic constraints are included.

- Distributional regularities—a special case of phonotactic constraints— are also known to contribute to the lexical segmentation task (Mattys & Jusczyk, 2001). It is based on the fact that in every language some sound patterns occur very frequently, some rarely, some never. This fact is also used by some computational models (e.g. Brent et al. (1991); Venkataraman (1999); Allen & Christiansen (1996)) to get better performance.

- Allophonic differences is yet another cue that is found to be helpful in segmentation task (Jusczyk et al., 1999a). An example of this phenomenon is the stop consonants

(e.g. /t/ phoneme) in English, which tend to be aspirated when they are word-initial (Church, 1987).

All of the above listed phenomena are useful for segmenting speech into lexical units, and all recent models of lexical segmentation of speech use multiple cues.

All of the studies cited above consider the *word* as the lexical unit. Although word and morpheme segmentation tasks are related, morpheme segmentation and learning morphology, have additional difficulties. A major difficulty is that, bound morphemes are always attached to other morphemes, whereas words can be seen in isolation. Similarly, some of the cues above do not help identifying morpheme boundaries. For example, vowel harmony may be helpful with identifying word boundaries in Turkish,[2] while it would not be useful for identifying morpheme boundaries.

2.2 The Syllable as the Basic Unit of Recognition

In this thesis we investigate the role of a basic segmental unit in discovering the meaning bearing elements of language. While traditionally *phoneme* has been regarded as the basic unit of recognition, some of the recent research in diverse fields favour the *syllable* as the basic unit of recognition (for a comprehensive summary, cf. Wu, 1998, chapter 2).

The literature on the subject so far has not provided a clear stance on the issue of what the basic unit of speech perception is. However, it is likely that the recognition process does not rely on a single source. The process makes use of multiple units and information sources in multiple levels for mapping continuous sound patterns to meaning bearing units of the language. Wu (1998, pp. 20-28) provides a review of the subject.

The support for syllable as a unit of speech perception, *syllable-effect*, mostly comes from research on French (Mehler et al., 1981), and other Romance languages (Sebastian-Galles et al., 1992; Bradley et al., 1993). However, the same effect does not seem to be applicable to English (Cutler et al., 1986) and Dutch (Vroomen & de Gelder, 1994). In the former group of languages, the durations of syllables appears to be roughly constant, whereas the latter

[2]While there is no study on the role of vowel harmony for segmentation of Turkish, it is known to be used by speakers of other languages, e.g. Finnish (Suomi et al., 1997), for identifying word boundaries.

group have inter-stress interval roughly constant. Traditionally, the former group is classified as *syllable-timed* and the latter group is classified as *stress-timed* (Pike, 1945). Further cross-linguistic studies show that, the use of syllable or stress as the primary cue is likely to be language dependent (Cutler et al., 1986; Cutler, 1997).

Although stress-timed languages seem to rely mostly on the stress pattern of the language, findings of Aslin et al. (1998) indicate that 8-month-old children acquiring English do use the statistical arrangement of syllables. Even if not as much as the syllable-timed languages, syllable has been found to be useful and used by the speakers of stress-timed languages, as well (Cutler, 1997). Statistical regularities between syllables, or other phonetic units may even be the source of information for children to learn the stress pattern of their language (Thiessen & Saffran, 2003).

Recent studies of automated speech recognition (ASR) have also been making use of syllable as a unit of perception (Wu, 1998). In addition to ASR research, research on the acoustic properties of human speech, e.g. speech intelligibility studies, also strengthens the syllable's role as a unit of speech recognition (Greenberg, 1996; Greenberg et al., 2003).

Despite the disagreement among researchers, there is considerable evidence supporting that the syllable is a major representational form. Especially with syllable-timed languages, the syllable's important role as a major segmental unit of speech understanding is well established.

There are some speculations (Reimers, 2005; Ramus, 2001; Mehler et al., 2000) that Turkish is a syllable-timed language, yet there has been no conclusive study determining the place of Turkish regarding this classification. Nonetheless, the syllable structure of Turkish is more similar to syllable-timed languages: syllable boundaries in Turkish are reliably marked and the syllables are not too complex.[3]

This thesis focuses on the syllable as a source information for learning lexical items. This, however, does not undermine the possible uses of stress or other prosodic cues. From the literature on the subject, it seems very likely that both play an important role for speech recognition as well as language acquisition. The segmentation strategy of children learning

[3]Theoretically, the most complex form of a Turkish syllable is probably CCVCC (e.g. borrowed word *trend*). In practice, however, most syllables are even simpler. The syllables of the forms CV, CVC, V, and VC constitute 99.5% of the syllables. The other 4 forms present in the data, in order of frequency, are CVCC, CCVC, VCC, and CCV.

Turkish is likely to be more similar to the strategy used by learners of syllable-timed languages. However, this does not rule out the role of stress. The regular primary word stress in Turkish is on the last syllable. The irregular stress is also predictable to some extent (Kornfilt, 1997, p. 514). A number of recent studies (e.g. Kabak & Vogel, 2001; Inkelas & Orgun, 2003)[4] capture both irregular and regular stress of the language by a unified analysis.

The predictability of word stress is certainly useful for word segmentation. Its use in morpheme segmentation, on the other hand, is not clear. The models developed in this thesis do not make use of stress, however, we see it as a possible extension to this study.

2.3 Learning Morphology

Children start using morphology as early as 12 to 20 months, and approximately by the age of 2 they start using it systematically (Clark, 1998). The acquisition of accurate knowledge of morphology comes even earlier for children learning Turkish (Aksu-Koç & Slobin, 1985).

Most of the studies in the literature focus on children's production of morphology (e.g. Berko, 1958; Clark, 1998). However there is no clear account of *how* children acquire morphology of their language. The studies with models of acquisition of morphology generally consist of computational models.

A number of studies on learning morphology use a method similar to Harris (1955). The related morphemes and morpheme boundaries are detected by making use of the fact that common word beginnings and endings are most likely morphemes. The words branching from the same root are considered to be morphological variants, and branching points are considered to be the morpheme boundaries. This approach, although primitive, seems to find a considerable amount of morphological variants. However, this strategy can lead to some erroneous classifications, such as classifying 'all' and 'ally' as morphological variants, or overlooking the not very frequent suffixes like 'dirt-y' (Schone & Jurafsky, 2000).

Schone & Jurafsky (2001) extend the method mentioned above with semantic information. They show that, incorporating semantic relatedness information they get using the Latent Semantic Analysis improves the success of the system considerably. They extend their work

[4]Even though the cited studies disagree with with each other's approach, the disagreement is on the methodology. The differing analyses agree on relatively simple treatment of the irregular stress patterns in Turkish.

further (Schone & Jurafsky, 2001) by making use of frequently occurring segments at the word beginnings and endings, the local syntactic context, and by extending the semantic information with transitive closure.[5] .

Neuvel & Fulop (2002), however, take a different approach, where they discover similar sound—or grapheme— segments between words in a lexicon containing the orthographic form of the words, associated with syntactic category information. They exhaustively search their lexicon to find morphological relationships. Their *whole word morphology* is likely to employ a lexicon with a full listing of inflected words, and a rule set that identifies morphological relationships between the words.

All of the above mentioned studies of learning morphology are computational studies. Some of them, as a by-product, create a morphemic lexicon. However, their aim is learning morphology rather than identifying the lexical items of the language.

Aronoff et al. (2006) claim that frequently occurring sound sequences are morphemes, and this alone can be used to bootstrap morphology acquisition. Their work on Spanish CHILDES data gives positive results(MacWhinney & Snow, 1990).

2.4 Turkish Morphology

This section is intended to provide brief information on the subset of the Turkish morphology used in this study. For a recent comprehensive reference, see Göksel & Kerslake (2005) or Kornfilt (1997).

Turkish has a rather complex morphological system. A single word in Turkish may correspond to a whole sentence in English. Word formation is achieved by suffixation, and sometimes the number of suffixes used may get quite high. Despite the seemingly complex morphology, children acquiring Turkish seem to be learning morphology at a very young age, and free of errors. Children as young as 15 months old use inflections productively (Aksu-Koç & Slobin, 1985), and by the age of 2 they acquire the complete noun inflection system (Topbaş et al., 1997). The early acquisition of morphology is attributed to the regularity of

[5] For three different words w_1, w_2 and w_3; if w_1 and w_2 are semantically related to w_3, w_2 and w_3 are also semantically related.

the system, and the necessity of acquiring morphology to be able to comprehend the language (Aksu-Koç & Slobin, 1985).

The model in this study is designed to learn only a part of the Turkish nominal morphotactics. Although the learning method used is not restricted, we chose to restrict our tests with this system to nominal morphotactics for practical reasons.[6]

Our system is trained to learn an inventory of the Turkish nominal inflections listed in Table 2.1. For the sake of simplicity, we have left out the inflection `-ki`. The word-grammar to be learned allows morpheme sequences of the nominal root, the plural suffix, the possessive suffix, and the case suffix.[7] All suffixes are optional, however, when they are present they have to follow this order.[8]

Table 2.1. Inventory of the Turkish nominal inflections learned by the computational model. The notation follows the convention used in the recent literature in Turkish Linguistics. Capital letters denote one of the possible phonological alternations, 'A': 'a' or 'e', 'D': 'd' or 't', 'I': 'ı', 'i', 'u', or 'ü'.

-lAr	Plural
-(I)m, -(I)mIz, -(I)n, -(I)nIz, (s)I, -lArI	Possessive Markers
-(y)I, -nI	Objective Case
-(n)In	Genitive Case
-(y)A, -nA	Dative Case
-DA -nDA	Locative Case
-DAn -nDAn	Ablative Case
-(y)lA	Instrumental/commutative Case

2.5 The Lexicon

The lexicon is an important part of our knowledge of language. Even though the structure of it is not clearly understood yet, we expect it to include—or contain some association of—

[6]The processing of the data, segmenting and tagging the words, and comparing the results are labor intensive.

[7]These may be followed by the relative `-ki` which may cause arbitrarily long words in theory. However, we have excluded `-ki` from the training and test sets used in this study, because of potentially high computational requirements that it may cause. However, it's occurrences in the corpus were relatively rare.

[8]See Oflazer et al. (1994) for a detailed description of the morphotactic process for both nouns and verbs.

the phonological form, the meaning (logical form), and syntactic category information of each lexical unit.

Although there is considerable debate on whether the lexicon contains only words or morphemes, there are strong psycholinguistic arguments (Hankamer, 1986; Marslen-Wilson, 1999) which suggest that the lexicon should contain morphemes —at least to some degree. For languages like Turkish, this argument is further supported by the huge amount of possible enumerations of all words in the language, recursive morphological rules mentioned above, and the productivity of constructing novel word forms. Especially for morphologically rich languages, like Turkish, it seems impossible to store all the inflected forms of the words in the lexicon.

Besides efficient storage, there are also other reasons for postulating a morphemic lexicon, such as productivity. Even for languages which are not morphologically complex, e.g. English, the way speakers come up with new words suggests that they are aware of how to compose these words from smaller units. For example, the English past tense over-generalisation errors by children led researchers to come up with many different theories of language development (Rumelhart & McClelland, 1986; Pinker, 1991). While which theory of language acquisition is supported by this phenomenon is still arguable, the over-generalisation errors certainly support the idea of the morpheme as a unit of the lexicon.

Inflectional morphemes that operate on larger linguistic units than words also support a morphemic lexicon. This suggests that the morphemes' scope is not always limited to words, and has to be dealt with during syntactic and semantic processing of larger-than-word units. The example (3) below, taken from Bozşahin & Göçmen (1995), illustrates this phenomenon.

(3)
iyi	okumuş	çocuk
well	read-REL	child

'well educated child'

The correct semantic bracketing of the phrase is *[[[iyi oku]muş] çocuk]*. The correct interpretation cannot be obtained without providing the relative suffix a wide scope, which supports the morphemic lexicon. See Bozşahin & Göçmen (1995) for further examples, and a computational treatment of the subject.

The above examples point strongly towards a high amount of interaction between morphology and syntax, favouring a morphemic lexicon. On the other hand, there is not a widely held consensus among the formal theories of grammar on the place of morphology. Derivational morphology is generally considered to be internal to lexicon. Inflectional morphology, on the other hand, is considered as part of the syntax by some formalisms of the grammar, including LFG (Kaplan & Bresnan, 1982), and more recently the morphosyntactic framework by Bozşahin (2002) which is based on Combinatory Categorial Grammar (Steedman, 2000).

2.6 Combinatory Categorial Grammar

This section gives an informal introduction to part of Categorial Grammar (CCG) (Steedman, 2000) that is used in this study.

CCG is a theory of grammar that favours lexicalism. All language specific syntactic and semantic information is stored in the lexicon. The set of rules operating on the syntactic and semantic information provided by the lexicon is universal.

The model developed in this study learns a CCG lexicon. Every item in a CCG lexicon consists of a phonological form (PF), a syntactic category (CAT) and a logical form (LF) of the lexical item. The example entries in (4) below are typical entries that the system used in this study tries to learn from the input. The first information, before '`:=`', is the phonological form; the second information is the syntactic category; and the third, after '`:`', is the logical form.

(4) `adam := `N `:` man
`lar := `$N_{plu}\backslash N$ `:` $\lambda x.plural(x)$

The first entry in (4) is a basic type, N, which refers to an object, *man*. The second entry is a more complex type. The syntactic category of the second item indicates that when this entry is preceded by a syntactic category of type N, it produces a syntactic category of N_{plu}. The semantics of the second item says that this is a function that makes its argument plural. The derivation of `Adamlar` using these two lexical entries in CCG is shown in (5).

(5)
$$\dfrac{\dfrac{adam : man}{N : man} \quad \dfrac{lar : \lambda x.plural(x)}{(N_{plu}\backslash N) : \lambda x.plural(x)}<}{N_{plu} : plural(man)}$$

For syntactic types X, Y, and Z; semantic functions (predicates) f and g, and argument a, the implementation in this study makes use of the following CCG operations:

- Forward Application $>$: $X/Y : f\ Y : a \Rightarrow X : fa$
- Backward Application $<$: $Y : a\ X\backslash Y : f \Rightarrow X : fa$
- Forward Composition $>$ B: $X/Y : f\ Y/Z : g \Rightarrow X/Z : \lambda x.f(gx)$
- Forward Composition $>$ B: $Y\backslash Z : g\ X\backslash Y : g \Rightarrow X\backslash Z : \lambda x.f(gx)$

Other CCG operations which are not necessary for Turkish inflectional morphology, e.g. type raising, are not implemented in this thesis (cf. Steedman, 2000).

We use the basic categories N, N_{plu}, N_{pos}, N_{obj}, N_{gen}, N_{dat}, N_{loc}, N_{abl}, N_{ins}. The first category is the base noun, the second is plural, and the third is possessive. The rest of the categories are the cases[9] objective, genitive, dative, locative, ablative, and instrumental, respectively. Instead of assigning basic categories for each inflection, a better approach would be to use modalities as in Bozşahin (2002). However, the approach used is chosen for its simplicity.

The grammar to be learned in this study is extremely simple. Due to the simple nature of the grammar the models learn, we do not make use of the full power of CCG. The only CCG rules we use in this study are the forward and backward function application rules.

Despite the limitations we have introduced, the use of a powerful formalism makes it easier to extend the system for addressing more general cases of learning a language.

[9] `-(y)lA` also has a free form, `ile`, and is not generally regarded as a case by most descriptive grammars of Turkish. Since the bound form behaves like other inflections considered as cases, we'll refer to it as 'instrumental' case.

2.7 Models of Learning Lexicon Using Phonetic and Semantic Information

This study uses a computational model to learn a lexicon from the given language input which contains phonetic and logical forms of words. The model, while finding the associations between phonetic and logical forms in the input, also learns syntactic properties of the given input language. We follow a method similar to the models used by Zettlemoyer & Collins (2005), and Jack et al. (2006).

The computational model of Zettlemoyer & Collins (2005) learns a Probabilistic Combinatory Categorial Grammar (PCCG). Since learning PCCG requires learning the lexicon of the input language, their system learns a lexicon for the linguistic domain they are working on. The system learns logical forms and syntactic categories for individual lexical items. Syntax is a hidden variable in the system. Their domain is specific,[10] and their motivation is solving an engineering problem. However, the learning method is cognitively relevant. Children acquiring languages also have a similar setting: they hear the phonological form of the utterance directed at them, generally in a setting where the objects or the events in the utterance are accessible in their immediate surroundings.

Jack et al. (2006) also use a similar computational model to learn lexical items from sentences describing simple event–object relationships. They also provide the learning system with utterances and encoding of the meaning of each utterance. However, they mark the utterances at syllable boundaries. Their system develops a simpler lexicon, and a separate syntactic analysis unit.

[10] Queries to a database of Unites States geography and a database of job listings.

CHAPTER 3

LEARNING MORPHEMES FROM SYLLABLES: A COMPUTATIONAL MODEL

This chapter describes the computational model for learning morphemes from syllabified words. The model's target is to learn meaning bearing items smaller than words, given the word and its semantics. The input word is marked on syllable boundaries. The system is expected to develop a lexicon where each lexical item—extracted from the input words— is associated with its logical form and syntactic category. We call this model, the *syllable-based model*.

We also test the system using input words marked at morpheme boundaries. This model, called the *morpheme-based model* in this text, is used for testing the performance of the syllable-based model. The results from both models are compared to assess the effectiveness of the syllable-based model.

3.1 The Syllable-Based Model

The input to the model are pairs of phonetic forms marked at syllable boundaries and the logical form of the word. Two examples of such pairs are given in (6).

(6) `a-dam-la-ra :` $dative(plural(man))$
`ev-de-ki-le-re :` $dative(plural(relative(locative(house))))$

From (6), ideally,[1] both the syllable-based model and the morpheme-based model are expected to learn the following lexicon:

(7) `adam` := N : man
`ev` := N : $house$
`lar` := $N_{plu}\backslash N$: $\lambda x.plural(x)$
`ler` := $N_{plu}\backslash N$: $\lambda x.plural(x)$
`a` := $N_{dat}\backslash N_{plu}$: $\lambda x.dative(x)$
`e` := $N_{dat}\backslash N_{plu}$: $\lambda x.dative(x)$
`de` := $N_{loc}\backslash N$: $\lambda x.locative(x)$
`ki` := $N_{rel}\backslash N_{loc}$: $\lambda x.relative(x)$

The lexicon contains phonological, syntactic and semantic information for each lexical item. We also keep a weight for each entry. Learning is based on adjusting this weight during the training of the system (See Section 3.3 for the use of weight and for the details of learning algorithm).

The difference in the lexicons of syllable- and morpheme-based models are the type assignments to units where a morpheme is split across syllable boundaries. For example, in the input `a-dam-la-ra`, the morpheme `lar` is split between two syllables, and the syllable `ra` contains part of the morpheme `lar` and the morpheme `a`. For such cases, we expect to get reasonable correlations, e.g. for the example `a-dam-la-ra`, the set of lexical entries in (8) is considered success[2]. Given enough exposure, semantics of `-la` is expected to be covered by that of `-lar` in the model. In general we expect semantics of the syllable-based model to be subsumed by that of the morpheme-based model.

(8) `adam` := N : man
`la` := $N_{plu}\backslash N$: $\lambda x.plu(x)$
`ra` := $N_{dat}\backslash N_{plu}$: $\lambda x.dat(x)$

[1]The lexicon of the syllable-based model diverges from this ideal lexicon. See Section 3.3 for explanation.

[2]Compared to learning lexical items like `la` := $N_{dat}\backslash N$: $\lambda x.dative(x)$ or `ra` := $N_{plu}\backslash N$: $\lambda x.plural(x)$.

The assumption for the model is that, due to other input from their surroundings, children have the logical forms of the utterances also available to them. The task they face is relating the semantics of the real-world circumstance they are in with the phonological form of the utterance they are hearing. They have the phonological input—including the knowledge of syllable— and the logical form at their disposal. However, they do not know exactly which speech segment corresponds to which logical form.

3.2 The Morpheme-based Model

The morpheme-based model differs from the syllable-based model with respect to the input given to the system. The morpheme-based model is given words marked at morpheme boundaries. Since the system gets its input segmented at the boundaries of the units it is trying to learn, the task is easier compared to the syllable-based model. However, the model still needs to associate the right morpheme in the input with the right logical form.

The input words in example (6) for the syllable-based model would be marked as shown in (9) for the morpheme-based model.

(9) `adam-lar-a` : $dative(plural(man))$
`ev-de-ki-ler-e` : $dative(plural(relative(locative(house))))$

3.3 Learning

This section describes the algorithm used in syllable- and morpheme-based models described in the previous sections. Both models use the same learning algorithm described in this section.

The learning algorithm is similar to the algorithms used by Zettlemoyer & Collins (2005) and Jack et al. (2006), with the following differences:

- Even though the linguistic domains that both studies use are very limited,[3] both systems are designed to learn the syntax from sentences. We restrict ourselves here to the morphotactics, i.e. the syntactic structure of words.

- Zettlemoyer & Collins (2005) use a pre-determined table of possible lexical items (including syntactic categories) based on input logical form. Jack et al. (2006) also use a list of heuristics in their syntactic analysis unit. Both studies eliminate a large number of possible lexical hypotheses using 'hard-coded' heuristics. The system used in this study uses all possible lexical items that can be derived from the input, constrained only by two basic principles:

  1. Basic universal directionality constraints,
  2. Principle of Categorial Type Transparency (Steedman, 2000, p 36).[4]

  Thus, we use CCG as a theory that shapes prior and likelihood probabilities in the learning process.

The learning starts with an initial lexicon L_0, which may be empty. For each input word, the system generates all possible hypotheses, selects the hypotheses to be added to the lexicon, and updates the lexicon using the new evidence provided by the input word.

3.3.1 Hypothesis Generation

For every input word, all possible lexical hypotheses are generated. For instance, for the previous example input of '`adam-lar-a` : $dative(plural(man))$', the morpheme-based model generates the following list of lexical hypotheses:

(10) `adam` := $N : man$
`adam` := $N_{plu}/N : \lambda x.plural(x)$
`adam` := $N_{plu}/N_{dat} : \lambda x.plural(x)$

[3] Training set used by Zettlemoyer & Collins (2005) consists of queries to databases in English. The application is intended for an NLP interface for database queries. Jack et al. (2006) also use a limited language describing a small set of events on a small number of objects.

[4] A more detailed description is provided on page 22.

$\texttt{adam} := N_{dat}/N : \lambda x.dative(x)$

$\texttt{adam} := N_{dat}/N_{plu} : \lambda x.dative(x)$

$\texttt{lar} := N_{plu}\backslash N : \lambda x.plural(x)$

$\texttt{lar} := N : men$

$\texttt{lar} := N_{plu}\backslash N_{dat} : \lambda x.plural(x)$

$\texttt{lar} := N_{plu}/N_{dat} : \lambda x.plural(x)$

$\texttt{lar} := N_{plu}/N : \lambda x.plural(x)$

$\texttt{lar} := N_{dat}\backslash N : \lambda x.dative(x)$

$\texttt{lar} := N_{dat}\backslash N_{plu} : \lambda x.dative(x)$

$\texttt{lar} := N_{dat}/N_{plu} : \lambda x.dative(x)$

$\texttt{lar} := N_{dat}/N : \lambda x.dative(x)$

$\texttt{a} := N : men$

$\texttt{a} := N_{dat}\backslash N_{plu} : \lambda x.dative(x)$

$\texttt{a} := N_{dat}\backslash N : \lambda x.dative(x)$

$\texttt{a} := N_{plu}\backslash N_{dat} : \lambda x.plural(x)$

$\texttt{a} := N_{plu}\backslash N : \lambda x.plural(x)$

At this stage, the algorithm creates a large number of wrong hypotheses, such as $\texttt{lar} := N_{dat}/N_{plu} : \lambda x.dative(x)$. Only the hypotheses that are impossible due to universal constraints like $\texttt{a} := N_{dat}/N_{plu} : \lambda x.dative(x)$ are not created.

The syllable-based model needs to take into account the fact that the number of elements in the logical form may not match with the number of elements in the segmented word input. In such cases, we form consecutive clusters of the input units matching the number of elements in the input logical form. For the above example, $\texttt{a-dam-la-ra} : dative(plural(adam))$, all hypotheses for (11)[5] are created (12).

(11) $\texttt{a.dam-la-ra} : dative(plural(man))$

$\texttt{a-dam.la-ra} : dative(plural(man))$

$\texttt{a-dam-la.ra} : dative(plural(man))$

[5]In this example, '–' indicates a possible morpheme boundary, and '.' indicates a syllable boundary.

(12) $\texttt{a.dam} := N : man$

$\texttt{a.dam} := N_{plu}/N : \lambda x.plural(x)$

$\texttt{a.dam} := N_{plu}/N_{dat} : \lambda x.plural(x)$

$\texttt{a.dam} := N_{dat}/N : \lambda x.dative(x)$

$\texttt{a.dam} := N_{dat}/N_{plu} : \lambda x.dative(x)$

$\texttt{a} := N : man$

$\texttt{a} := N_{plu}/N : \lambda x.plural(x)$

$\texttt{a} := N_{plu}/N_{dat} : \lambda x.plural(x)$

$\texttt{a} := N_{dat}/N : \lambda x.dative(x)$

$\texttt{a} := N_{dat}/N_{plu} : \lambda x.dative(x)$

$\texttt{dam} := N_{plu}\backslash N : \lambda x.plural(x)$

$\texttt{dam} := N : man$

$\texttt{dam} := N_{plu}\backslash N_{dat} : \lambda x.plural(x)$

$\texttt{dam} := N_{plu}/N_{dat} : \lambda x.plural(x)$

$\texttt{dam} := N_{plu}/N : \lambda x.plural(x)$

$\texttt{dam} := N_{dat}\backslash N : \lambda x.dative(x)$

$\texttt{dam} := N_{dat}\backslash N_{plu} : \lambda x.dative(x)$

$\texttt{dam} := N_{dat}/N_{plu} : \lambda x.dative(x)$

$\texttt{dam} := N_{dat}/N : \lambda x.dative(x)$

$\texttt{dam.la} := N_{plu}\backslash N : \lambda x.plural(x)$

$\texttt{dam.la} := N : man$

$\texttt{dam.la} := N_{plu}\backslash N_{dat} : \lambda x.plural(x)$

$\texttt{dam.la} := N_{plu}/N_{dat} : \lambda x.plural(x)$

$\texttt{dam.la} := N_{plu}/N : \lambda x.plural(x)$

$\texttt{dam.la} := N_{dat}\backslash N : \lambda x.dative(x)$

$\texttt{dam.la} := N_{dat}\backslash N_{plu} : \lambda x.dative(x)$

$\texttt{dam.la} := N_{dat}/N_{plu} : \lambda x.dative(x)$

$\texttt{dam.la} := N_{dat}/N : \lambda x.dative(x)$

$\texttt{la} := N_{plu}\backslash N : \lambda x.plural(x)$

$\texttt{la} := N : man$

$\texttt{la} := N_{plu}\backslash N_{dat} : \lambda x.plural(x)$

$\texttt{la} := N_{plu}/N_{dat} : \lambda x.plural(x)$

$\texttt{la} := N_{plu}/N : \lambda x.plural(x)$

$\texttt{la} := N_{dat}\backslash N : \lambda x.dative(x)$

$\texttt{la} := N_{dat}\backslash N_{plu} : \lambda x.dative(x)$

$\texttt{la} := N_{dat}/N_{plu} : \lambda x.dative(x)$

$\texttt{la} := N_{dat}/N : \lambda x.dative(x)$

$\texttt{ra} := N : man$

$\texttt{ra} := N_{dat}\backslash N_{plu} : \lambda x.dative(x)$

$\texttt{ra} := N_{dat}\backslash N : \lambda x.dative(x)$

$\texttt{ra} := N_{plu}\backslash N_{dat} : \lambda x.plural(x)$

$\texttt{ra} := N_{plu}\backslash N : \lambda x.plural(x)$

$\texttt{ra} := N : man$

$\texttt{la.ra} := N : man$

$\texttt{la.ra} := N_{dat}\backslash N_{plu} : \lambda x.dative(x)$

$\texttt{la.ra} := N_{dat}\backslash N : \lambda x.dative(x)$

$\texttt{la.ra} := N_{plu}\backslash N_{dat} : \lambda x.plural(x)$

$\texttt{la.ra} := N_{plu}\backslash N : \lambda x.plural(x)$

$\texttt{la.ra} := N : man$

It is assumed for a given input that the number of components in the phonetic form are always greater than or equal to the number of components in the logical form. This assumption is fully supported by the corpus used: none of the words had fewer number of syllables than the number of components in the corresponding LF. This assumption is also supported by some rare irregularities in Turkish morphophonetics. For instance, morphemes consisting of single phonemes (like first person possessive `-(I)m`) gets an exceptional buffer consonant when it is attached to `su` (= *water*). So, the inflected form is `su-yum` instead of `*su-m` as the regular morphophonetics of the language suggests. This prevents the resulting word to have more LF components than the number of syllables it has. The similar phenomena are also observable with other short root forms like `o`, `şu`. Due to buffer vowels, additional inflections also preserve this condition.

All of these lexical hypotheses are processed with the rules explained below. It should be noted that two constraints are applied to the enumeration of possible lexical hypotheses. First, we leave out the hypotheses that are not possible due to the universal directionality principle

of grammar. For example in (12) above '`a.dam` := $N_{plu}/N : \lambda x.plural(x)$' is not included, since '`a.dam`' being the leftmost element, does not have anything on its left. The other constraint on the generation of lexical hypotheses is based on the Principle of Categorial Type Transparency (PCTT). This constraint enforces a basic syntactic category, such as N, for an 'argument' logical form, such as `man`. Whereas a functor logical form, such as $\lambda.xplural(x)$, is mapped to a complex syntactic category, like $N_{plu}\backslash N$.[6] For example, hypotheses like '`adam` := $N : \lambda x.plural(x)$' or '`adam` := $N_{plu}\backslash N : man$' are not allowed.[7]

The learning algorithm takes a sequence of segmented words, and produces a lexicon that contains a 4-tuple for each lexical item. Every item is composed of a *phonetic form*, a *syntactic category*, a *logical form*, and a *weight*. Example (13) presents two of such lexical items.

(13) `ev` := $N : house$; 0.98435
`ler` := $N_{plu}\backslash N : \lambda x.plural(x)$; 0.82113

The *phonetic form* (PF) is in fact the *orthographic* form taken from the CHILDES database. Due to the relative orthographic transparency of Turkish, using orthographic transcriptions is a common practice in studies analysing Turkish language data (e.g. Ekmekçi, 1982; Küntay & Slobin, 1994).

The *syntactic category* (CAT) can either be one of the basic categories or a complex category. We have used multiple basic categories, for all possible morphosyntactic categories in the data. A basic category is one of the following:

- A nominal root (or nominative case), N.
- A plural noun, N_{plu}.
- Possessive forms, N_{p1p}. N_{p1s}, N_{p2p}, N_{p2s}, and N_{p3p}.
- Cases or case like forms, N_{loc}, N_{abl}, N_{dat}, N_{acc}, N_{gen}, N_{ins}.

[6] This is true in our simplified models that are used in this thesis. Otherwise, the limitations that the PCTT enforce would be different.

[7] Note that this constraint only checks the relationship between LF and CAT. For example, '`adam` := $N_{plu}\backslash N : \lambda x.plural(x)$' is perfectly fine with regard to this constraint.

The use of multiple basic forms is motivated by clarity of expression. Complex categories are any combinations of X/Y or $X\backslash Y$ for any two basic categories X and Y.

The *logical form* (LF) either refers to an object, or a function. The models always map functors to complex CATs, and arguments to basic CATs.

The *weight* is the probability of the 3-tuple $< \mathrm{PF}, \mathrm{CAT}, \mathrm{LF} >$ being a lexical item in the language.

3.3.2 Bayesian Learning

Learning is achieved by updating the weights based on new input. The model follows *Bayesian inference* while updating the weights. The weights in the lexicon are the probability, or the system's belief, that the lexical item in question is correct. Each weight update consists of determining the new weight, the probability of the lexical hypothesis h given the new evidence E. The new evidence, E is the input word presented to the system. So, the weight of the lexical item after seeing the input is,

$$P(h \mid E) = \frac{P(E \mid h)P(h)}{P(E)}$$

$P(h)$, the *prior* probability, is the probability (weight) of the lexical hypothesis before seeing the input E. This means, the higher the previous weight value is, the higher the new weight will be. $P(E \mid h)$, the likelihood, is calculated as the number of parses of the input word that the h is used divided by the total number of parses. This determines the contribution of the new input to the posterior probability. The higher the number of parses in the input that the hypothesis supports, the higher the likelihood value will be. So, if the hypothesis is used by all possible parses of the input, the value is 1. The value gets smaller with the parses of the input word that do not include the hypothesis. Calculating $P(E)$ or finding a distribution for it is rather difficult. However, since it is constant for all the hypotheses being considered, it can be ignored. Posterior probability of our hypothesis is directly proportional to the prior, $P(h)$, and the likelihood, $P(E \mid h)$.

Bayesian inference stipulates how learners should update their beliefs with the new evidence. It can be applied to any model learning from data. Bayesian inference has been successfully applied to modelling diverse areas of cognition, including word recognition (Norris,

2006), word learning (Tenenbaum & Xu, 2000), vision (Yuille & Kersten, 2006) and sensorimotor control (Körding & M.Wolpert, 2004). Griffiths et al. (pear) give a review of Bayesian models applied to cognitive science (see MacKay, 2003, for reference and other applications).

Algorithm 3.1 Learning algorithm used for training the models.

1. The algorithm takes initial lexicon L_0 and a sequence of input items that consist of phonetic and logical encodings of a word.
2. After the n^{th} input, the updated lexicon L_n is determined by the following procedure:
   (a) All possible lexical hypotheses from the given input are generated with the rules described above.
   (b) Generated hypotheses are placed in a temporary lexicon, L_T. The weights of the items are obtained from the current lexicon L_{n-1}. If the lexical hypothesis is not in L_{n-1}, an initial weight w_0 is assigned for the weight of the item in L_T. In the experiments reported in Section 3.5, an initial weight of 0.1 is used.
   (c) All possible parses of the word using L_t are produced. Each parse is assigned a weight proportional to the weights of all the lexical items used for derivation.
   (d) All the hypotheses that entertain the parse with the highest weight are inserted into L_{n+1}, with the new weight determined by the calculation described above.

3.4 Data

The primary corpus used in this study is from the CHILDES database (MacWhinney & Snow, 1990). This section presents a brief description of the data used. Further details can be found in Aksu-Koç & Slobin (1985) and Slobin (1990).

The data contain 51 recording sessions with 33 children. The ages of the children vary between 2;0 to 4;8. The average age of children in all the recording sessions is 3;4. Sessions are recorded by two investigators: 36 by one, and 15 by the other.

The number of some units of interest in the corpus and the distributions of them per recording session are presented in Table 3.1. For this study, we were only interested in nouns in the child directed speech (CDS). CDS in this corpus consists mainly of utterances by the experimenter (88%). 1% of the utterances reported in Table 3.1 belonged to non-native or non-adult speakers, and were not included in the the CDS set used. The average number of words per utterance in CDS is 3.13, and does not show a correlation with the age of the children.

Table 3.1. Counts of various units of interest in the data. Child directed speech is abbreviated as 'CDS'. **min**, **average**, **max** values are minimum, average and maximum per recording session. The last column lists the total number of the unit described by the first column in the whole corpus. Words other than nouns and verbs are not classified, and are referred to as 'Other CATs' in the table. These include the categories like adjectives, as well as the words/utterances that cannot be categorised (e.g. baby-talk).

	min	**average**	**max**	**total**
Utterances	95	469	1190	23932
Utterances in CDS	37	200	500	10206
Utterances by children	58	269	690	13726
Words in CDS	91	655	1806	33450
Nouns in CDS	33	243	627	12389
Verbs in CDS	30	179	507	9142
Other CATs in CDS	28	233	672	11919
Morphemes in nouns in CDS	55	402	985	20734
Non-inflected nouns CDS	11	108	338	5542
Inflected nouns in CDS	20	132	326	6762
Syllables in nouns in CDS	200	1494	4207	76238

All the nouns in the CDS have been segmented at morpheme boundaries and tagged with a logical form representing the semantic content of the word.[8] Each segmented and tagged word looks like one of the items in example (8) on page 17. The derivational process has been ignored. All derivational morphemes are considered as part of the nominal root. For the rest of this section, the term *morpheme* refers to either a nominal root—possibly with derivations—, or an inflectional morpheme. Due to the nature of the CHILDES transcriptions, automated segmenting and tagging was not practical. The segmentation and tagging is mostly done by hand. Marking the syllable boundaries is done automatically.

Of the 12389 nouns reported in Table 3.1, 30 of them contained the morpheme `-ki`, and were excluded from the data set. The set of words used in the study contains 885 unique root forms. 272 of the root forms only appeared in inflected forms in this corpus, and 323 of them only appeared in root form. The most frequent free morphemes are pronouns (like `sen`, `o`, `ben`). The total number of phonetic alternations for all nominal inflection types in the corpus, excluding `-ki`, is 76 (see Table A.1 in Appendix A for a detailed listing of all inflections).

[8] All the words that fill a noun position in syntactic structure are classified as nouns.

As for the nouns, the average number of morphemes per word in the corpus is 1.68.[9] Even though there is no correlation between the age of the child and the number of words per child directed utterance,[10] the average number of morphemes per noun in child directed speech slightly increases with the age of the children.

The average number of syllables per word—including all syntactic categories— is 2.28, the average number of syllables per utterance is 4.19. The number of syllable per word, and hence the number of syllables per utterance, in child directed speech also follows a slight increase with the age of the children.[11]

The training set consists only of nouns extracted from the child directed speech. Randomly selected 1389 (approximately 10%) nouns are set aside for testing purposes, and were not used in training. The training set contained the rest of the 11000 nouns present in the corpus.

Additionally, 500 distinct nouns from the utterances of children were picked at random. These words were segmented and tagged like the training data, and were used in *production tests*. These tests and use of the test sets is described in Section 3.5.3.

3.5 Tests and Results

Both morpheme- and the syllable-based models have been trained using the input described in Section 3.3. Both models' outputs are lexicons. The lexicon learned by the syllable-based model is referred to as L_s. We call the lexicon learned by the morpheme-based model L_m. We have constructed another lexicon, which is the 'golden standard' that can be extracted from our training set. We use this lexicon L_r as reference while assessing the performance of our models.

The lexicons learned by the models are different due to differences in the input format. However, the expectation is that the syllable-based model approximates the morpheme-based

[9] This is lower than 1.96 morpheme per word reported by Küntay & Slobin (1994) for the child directed speech in their corpus. The difference may be due to their account of derivational morphology, or the differences between the settings of the recording sessions.

[10] This fact should only be taken for the data at hand. Due to similar 'experimental' setting of the recordings, and due to the high ages of the children, the data may not represent child directed speech in general.

[11] Even though we are looking at the child directed speech rather than utterances produced by children, these findings are in-line with findings of Ekmekçi (1982), that length of utterances in syllables and morphemes (rather than words) is a better indication of Turkish children's development of mean length of utterance.

model. It should be noted that the knowledge of which morpheme means what cannot be assumed in the morpheme-based model either. The child has to learn this association; what is assumed to be innate is the ability to make the form-meaning association in a combinatory fashion.

3.5.1 The Morpheme-Based Model

The morpheme-based model has two uses in this study. First, it is used to assess the success of the learning algorithm. Second, the results obtained from the syllable-based model were compared with the morpheme-based model's results. To test the learning algorithm, we compared the lexicons L_r, the reference lexicon and the lexicon learned by the morpheme-based model.

The reference lexicon, L_r, is constructed using the training data. It contained 857 nominal root forms and 150 inflections present in the training set. The inflections are listed in Appendix A.

The morpheme-based model, after being trained with the training data once, fails to find only 16 nominal roots, and 7 inflectional forms, listed in (14). In addition, the model also learns 4 inflectional forms, listed in (15), that are not in the reference lexicon.

(14) in := $N_{gen}\backslash N_{p2p} : \lambda x.genitive(x)$
nı := $N_{acc}\backslash N_{p3p} : \lambda x.accusative(x)$
la := $N_{ins}\backslash N_{p1s} : \lambda x.instrumental(x)$
muz := $N_{p1p}\backslash N : \lambda x.pos1p(x)$
mız := $N_{p1p}\backslash N : \lambda x.pos1p(x)$
u := $N_{acc}\backslash N_{p1p} : \lambda x.accusative(x)$
ın := $N_{gen}\backslash N_{p1p} : \lambda x.genitive(x)$

(15) in := $N_{p2s}\backslash N_{gen} : \lambda x.pos2s(x)$
nı := $N_{p3s}\backslash N_{acc} : \lambda x.pos3s(x)$
ni := $N_{p3s}\backslash N_{acc} : \lambda x.pos3s(x)$
nu := $N_{p3s}\backslash N_{acc} : \lambda x.pos3s(x)$

After a second run through the training set, L_m contains all the root forms. In addition to this, it also learns the first two lexical items in (14). Subsequent training runs do not add new items to the lexicon. However, the weights are adjusted in such a way that the weights of the alternatives to initial "mistakes" get higher, while most of the mistakenly learned items remain with low weights.

Due to different phonological realisation of the lexical items, the morpheme-based model assigns more than one phonological form to a single semantic and syntactic function. For the root forms, this happens especially due to phonological alternations to root final consonants (e.g. `kitap` and `kitab` both associated with the same logical item *book*). Since the system is not aware of any phonological process, this is considered normal. In addition to phonological alternations, there are a number of different phonological realisations either due to baby-talk (e.g. `kedi` and `pisi`, both referring to *cat*), or differences in dialect, accent (e.g. `nene` and `nine`, both referring to *grandmother*). Out of 806 different 'senses' in L_m, 79 of the root forms have 2, and 1 of them has 3 phonological realisations. Alternate phonological forms are more common for inflections as they also change form due to vowel harmony and buffer consonants. The details of these alternations for each semantic/syntactic form for the morpheme-based model are summarised in Appendix A.

Except the lexical items that are not frequent enough to be learned by the morpheme-based model, most of the 'mistakes' made by the model are due to ambiguous inflections.

3.5.2 The Syllable-Based Model

The same training method is applied to the syllable-based model. The main difference between the two models is the input provided to them. The morpheme-based model gets the phonetic form already marked at morpheme boundaries. However, for the syllable-based model, the input words are marked at syllable boundaries. So, instead of the morphemes, the syllable-based model tries to assign semantic/syntactic roles to individual syllables or syllable groups.

It takes 3 training runs through the corpus for the syllable-based model to stop learning new lexical items. After the first run through the corpus, the syllable-based model learns 753 root forms. It adds 16 more on the second run, after which it stops adding any further root

forms. As for the inflections, it learns 82 of them on the first run, and adds 17 and 3 more inflections on the second and third run, respectively. The syllable based model stabilizes at 871 lexical items, of which 769 are root nouns and 102 are inflections.

For the nominal roots, the syllable model was not able to find any phonetic forms for 103 of the logical forms present in L_m. However, for the rest (86%) of the forms present in L_m, it was able to assign at least one association between logical form and phonetic form.

In addition to phonological alternations (e.g. both `-ler` and `-lar` for the plural suffix) which are also present in L_m, the syllable-based model assigns multiple phonological forms to the same 'sense' (e.g. `-le` and `-la` in addition to `-ler` and `-lar`), since the lexical items do not always end at syllable boundaries. Despite this shortcoming, it can still find 719 (71%) of the 1006 lexical items learned by the morpheme-based model. For the root nouns, the match between L_m and L_s is even higher, with 77%. The syllable-based model, on the other hand, learns slightly more duplicated phonetic forms. Besides the 579 lexical items with a single phonetic form, number of items with 2, 3 and 4 phonetic forms are 96, 8 and 4, respectively. The number of items with more than 4 phonetic forms is 6. Examples of duplicated phonetic forms that exist in L_s, but not in L_m, are `tre` and `tren` for *train*, `ça` and `çay` for *tea*.

The syllable-based model performs worse on inflections. The exact overlap between the models is 58 lexical items (38% of the inflections in L_s, 58% of inflections L_m). 37% of the inflections that syllable based model failed to learn were morphemes with a single phoneme, e.g. '`i` := $N_{acc}\backslash N : \lambda x.accusative(x)$'. For the accusative marker `-i`, the syllable-based model learns multiple lexical items, such as '**n**`i` := $N_{acc}\backslash N : \lambda x.accusative(x)$', '**s**`i` := $N_{acc}\backslash N : \lambda x.accusative(x)$'. A detailed list of inflections learned by the syllable-based model is given in Appendix A. Using a 'loose' match criterion that matches the phonetic forms with a single phoneme deletion or insertion increases the overlap between L_m and L_s drastically. 75% of the inflections in L_s (53% of L_m) match with the loose match criteria.

Comparing all the results together, including both bound and free morphemes, gives an exact match[12] of 81% of the lexical items in L_s which corresponds to 71% in L_m. Applying the loose match to the complete lexicons (including both root and inflectional forms) increases

[12]Note that *exact match* refers to exact match between individual lexical items, not the whole complete lexicons

the match up to 96% of the items in L_s (84% of L_m). A summary of the number of items in each lexicon and number of matching entries are given in Table 3.2.

The similarity between the two systems is even more visible for the items with higher weights. A comparative list of highest scoring 50 items learned by both models is given in Appendix B.1.

Table 3.2. Overall summary of the comparison of the lexicons. Exact match the count of matching items with all fields (PF. LF, CAT) in the lexicons being compared. Loose match ignores a single phoneme difference in PF. LF/CAT match ignores the PF completely.

	Root Forms	Inflections	All Morphemes
Number of items in L_r	857	150	1007
Number of items in L_m	857	149	1006
Number of items in L_s	769	102	871
Exact match between L_r & L_m	857	145	1002
Exact match between L_m & L_s	661	58	719
Loose match between L_m & L_s	746	100	846
Matching LF/CAT in L_r & L_m	857	145	1002
Matching LF/CAT in L_m & L_s	696	82	778

3.5.3 Recognition and Generation

In addition to comparing the lexicons learned by the models, the performance of the system is tested by checking how good the learned lexicon recognises or generates the forms present in the data.

For the tests of recognition, a test set of 100 words from child directed speech was presented to the system. Certain automatically obtainable results are further compared with 4 more sets of same size selected randomly from the same data. For generation tests, another 100 word test set picked randomly from utterances by the children was presented to the system. Like the test set used for recognition, 4 more sets obtained in a similar fashion were used to check the validity of the results presented below.

We designed two routines for the test: *generate* and *recognise*. *Generate* takes a lexicon and a logical form and generates the possible surface form. *Recognise* takes a lexicon and the phonetic form of a word, and displays all possible parses allowed by the lexicon.

The morpheme-based model was presented with words marked at morpheme boundaries. The model failed to recognise 4 words in the test set. For the words which had multiple parses, the system almost always scored the correct form with the highest score. There was only one case, where the highest scored parse was wrong syntactically. The lexicon favoured parse (16), instead of (17). Parse (16) is syntactically wrong—if both the possessive and the case marker are present, the possessive marker has to precede the case marker. This is one of the ambiguous cases for the learner, the third person possessive suffix and the accusative case marker having the same phonetic realisations. For such cases, despite the failure of getting the right syntactic from, the logical form still 'makes sense'.

(16)

Kanat	-lar	-ı	-nı
$N : wing$	$(N_{plu}\backslash N)$:$\lambda x.plural(x)$	$(N_{acc}\backslash N_{plu})$:$\lambda x.accusative(x)$	$(N_{p3s}\backslash N_{acc})$:$\lambda x.pos3s(x)$
N_{plu}:$plural(wing)$ <			
N_{acc}:$accusative(plural(wing))$ <			
N_{p3s}:$pos3s(accusative(plural(wing)))$ <			

(17)

Kanat	-lar	-ı	-nı
$N : wing$	$(N_{plu}\backslash N)$:$\lambda x.plural(x)$	$(N_{p3s}\backslash N_{plu})$:$\lambda x.pos3s(x)$	$(N_{acc}\backslash N_{p3s})$:$\lambda x.accusative(x)$
N_{plu}:$plural(wing)$ <			
N_{p3s}:$pos3s(plural(wing))$ <			
N_{acc}:$accusative(pos3s(plural(wing)))$ <			

The syllable-based model, when presented with syllabified words as input, failed to recognise 31 of the items. All the failures were due to the lexical items that were not in the lexicon. Examples include `ka-pa-ğı-nı` and `de-ni-ze`. The former was not recognised because it was a rare word in the corpus, it did not make it to the lexicon learned by the syllable-based model. For the latter, the morphemes `deniz`, and `-e` were both in the lexicon, however, the tokenisation of the input did not allow the syllable-based model to recognise the word.

The syllable-based model, in total recognized 69 of the input words, where the morpheme-based model was able recognize 94 of them. The average number of parses for the recognized items were similar for both models; 1.10 for the morpheme-based model and 1.09 for the syllable-based model. These numbers are similar among all 100 word sets used for cross

validation. On average, the morpheme based model recognized 97.6%, and the syllable based model recognized 63.4% of the given input. When averaged over all the sets, the number of parses for each recognized input was 1.13 for both models.

For the generation tests, the system is presented with a logical form, e.g. '*plural(man)*', and is expected to produce the phonetic form using the lexicon it learned. A score— proportional to the weights of the individual lexical units used in generation— is calculated for all the generated phonetic forms. For these tests the 100-word test set extracted from the utterances of the children is used.

Due to multiple phonetic alternations, and no phonetic knowledge built into the models, the generation test always over-generates. For example, the logical form '*plural(man)*' generates both '`adam-lar`' and '`adam-ler`'. The over-generation rate for the morpheme-based model was 5.79. The outputs obtained using L_m and L_r were exactly the same. The syllable-based model on the other hand, generates slightly less than the morpheme-based model. The average number of SFs generated by the syllable-based model per LF was 5.55.

The models performed similarly on the generation task. The overlap of highest scoring surface forms generated by both models was 62. Most of the errors made by both models were due to lack of the knowledge of morphophonological alternations. The models would have generated more similar outputs if they were made aware of phonetic alternations.

To cross check the results with more data, we used 4 more sets of unique 100 words which are randomly selected from child speech. They produced similar results for average number of items generated for both models. The average number of items generated overall was 5.62 for the syllable-based model, and 6.02 for the morpheme-based model.

3.5.4 Variations in Input Order

All of the numbers reported in this section were obtained by tests where the order of input is the order of the CHILDES data files. Although we did not find any strong correlation between linguistic units per utterance, one still wonders if the system can detect any regularities that are not apparent. To test this, the system is trained with input sorted by the age of children (both in ascending and descending order). However, the results from different orders of input did not show any significant difference. There were only a slight differences—without any

special correlation with the order of input— on learning rate, and the weights of some lexical items in the final lexicons.

CHAPTER 4

CONCLUSIONS

This thesis is an attempt to investigate the role of the ability of *syllable segmentation* on discovering morphemes—the minimal meaning bearing units of language. A computational model has been developed to learn a morphemic CCG lexicon using a Bayesian learning algorithm. The input consists of the segmented phonetic form and the semantic content of a given word. The resulting lexicon contains entries of the form '`phonetic form` := $syntactic\ category : logical\ form$" for each meaning bearing element discovered in the input. The model assumes pre-existing ability of extracting the syllables from the speech signal, and the availability of semantic content of the speech signal being processed. The end-product is a set of lexical entries which are the associations between the segments of the phonetic form of the input with the atomic elements of the semantic content. Since we have used CCG, a linguistic formalism that assumes transparent relationship between semantics and syntax, the syntactic category of the item is thus obtained with minimal effort. Hence, the resulting lexicon is a CCG lexicon ready to recognise and generate the linguistic domain—for the purpose of this study, Turkish nominal morphotactics. Additionally, the statistical nature of the learning algorithm helps to grade ambiguous parses of the same word.

To test the target model, which is referred to as the *syllable-based model*, another model has been created that uses the same learning algorithm, yet takes an input that is marked at morpheme boundaries. The main use of this second model, the *morpheme-based model*, has been to compare it with our main target model, the syllable-based model.

The data used for training both models is Turkish data from CHILDES. Only the nouns present in child directed speech in the corpus are used to train the models. The restriction to nouns has been due to practical reasons, mainly to avoid the laborious process of marking the

input, and comparing the output lexicons. Otherwise, the models are capable of processing other categories as well. Both models were trained and tested using the same data. The results obtained from both models are compared and reported in Section 3.5.

A test of the morpheme-based model against a reference lexicon shows that, both syntactic and semantic associations are almost perfect. The morpheme-based model learns all of the root forms perfectly. It fails to find only 5 inflections and erroneously generates 4. The learning algorithm, for the task at hand works reasonably good. In addition to its success, the morpheme-based model converges very fast.

The syllable-based model was not designed to outperform the morpheme-based model. It was expected from the start that the syllable-based model would perform worse than the morpheme-based model, whose input contains far richer information than the syllable-based model. On the other hand, it was expected of the syllable-based model to achieve comparable results with the morpheme-based model. There are differences between the lexicons learned by the syllable-based model and the morpheme-based model, since morphemes are presented to the syllable-based model in all kinds of phonological environments, e.g. -le, `-la`, `-re`, `-ra` and `-ler` `-lar` in abundance. However, the results indicate that the lexicons learned by both models overlap considerably. The exact overlap between the lexicons is a significant 71%. The syllable-based model's failure to learn the inflectional morphemes was mostly due to the shorter and more ambiguous forms of the inflections. The similarity between lexicons learned by the models is rather low (38%) at first sight. Considering the close phonetic forms with the same logical form and syntactic category (e.g. taking into account that `-le` and `-ler` are similar) the overlap of the inflectional forms increases to 58%.

The results from recognition and production tests also show that the syllable-based model is reasonably successful in these tasks compared to morpheme-based model.

The overall results suggest that knowledge of the syllable is useful for learning the meaning bearing units. This finding supports the idea that the segmental units that children are aware of during the early stages of language acquisition can be useful for learning more abstract and higher level linguistic units.

4.1 Future work

The preference of the syllable as the basic unit of speech sounds took us into the debate of syllable vs. stress as the primary cue for early speech recognition. This thesis does not consider syllable and stress as mutually exclusive sources of information. On the contrary, both have a different function, and can even act in a complementary manner. One of the possible extensions of this work would be to studying the differences that the addition of stress into the input might introduce.

The learning method used is rather primitive, and is not likely to scale for more complex parts of language acquisition. Improving the learning system and using it for learning a lexicalised grammar, not just a word grammar, is another possible extension for this work.

REFERENCES

Aksu-Koç, A. & Slobin, D. (1985). The acquisition of Turkish. In Slobin, D., editor, *The Crosslinguistic Study of Language Acquisition, Vol 1: The data.*, pp. 839–878. Erlbaum.

Allen, J. & Christiansen, M. H. (1996). Integrating multiple cues in word segmentation: A connectionist model using hints. In *Proceedings of the 18th Annual Cognitive Science Society Conference*, pp. 370–375.

Altmann, G. T. M. (2002). Statistical learning in infants. In *Proceedings of the National Academy of Sciences*, 99, p. 15250 –15251.

Aronoff, J. M., Giralt, N., & Mintz, T. H. (2006). Stochastic approaches to morphology acquisition. In *Selected Proceedings of the 7th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages*, pp. 110–121.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month old infants. *Psychological Science*, 9, 321–324.

Berko, J. (1958). The child's learning of English morphology. *Word*, 14, 150–177.

Bozşahin, C. (2002). The combinatory morphemic lexicon. *Computational Linguistics*, 28(2), 145–186.

Bozşahin, C. & Göçmen, E. (1995). Categorial framework for composition in multiple linguistic domains. In *Proceedings of the 4th Int Conf on Cognitive Science of NLP*, Dublin.

Bradley, D. C., Sanchez-Casas, R. M., & Garciá-Albea, J. E. (1993). The status of the syllable in the perception of Spanish and English. *Language and Cognitive Processes*, 8, 197–234.

Brent, M. R. & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition*, 61, 93–125.

Brent, M. R., Gafos, A., & Cartwright, T. A. (1991). Phonotactics and the lexicon: Beyond bootstrapping. In Clarck, E., editor, *Proceedings of the 1994 Standford Child Language Research Forum*. Cambridge, UK: Cambridge University Press.

Brent, M. R. & Siskind, J. M. (2001). The role of exposure to isolated words in early vocabulary development. *Cognition*, 81, 33–44.

Christiansen, M. H. & Allen, J. (1997). Coping with variation in speech segmentation. In *Proceedings of GALA'97 Conference on Language Acquisition: Knowledge Representation and Processing*, pp. 327–332. University of Edinburgh.

Christiansen, M. H., Reali, F., & Chater, N. (2006). The Baldwin effect works for functional, but not arbitrary, features of language. In *Proceedings of the 6th evolution of language conference*.

Church, K. W. (1987). Phonological parsing and lexical retrieval. *Cognition*, 25, 53–69.

Clark, E. V. (1998). Morphology in language acquisition. In Spencer, A. & Zwicky, A. M., editors, *The Handbook of Morphology*. Oxford: Blackwell.

Cooper, W. E. & Paccia-Cooper, J. M. (1980). *Syntax and Speech*. Cambridge MA: Harvard University Press.

Cutler, A. (1997). The syllable's role in the segmentation of stress languages. *Language and Cognitive Processes*, 12, 839–845.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language*, 25, 385–400.

Dehaene-Lambertz, G. & Houston, D. (1998). Faster orientation latency toward native language in two-month old infants. *Language and Speech*, 41(1), 21–43.

Ekmekçi, F. O. (1982). Language development of a Turkish child: A speech analysis in terms of length and complexity. *METU Journal of Human Sciences*, 1(2), 102–112.

Fromkin, V. & Rodman, R. (1993). *An Introduction to Language*. Harcourt Brace Collage Publishers, fifth edition.

Glietman, L. R., Glietman, H., Landau, B., & Wanner, E. (1988). Where learning begins: Initial representation for language learning. In Newmayer, F. J., editor, *Linguistics: The Cambridge Survey*, 3, pp. 150–193. Cambridge, UK: Cambridge University Press.

Greenberg, S. (1996). Understanding speech understanding - towards a unified theory of speech perception. In *Proceedings of the ESCA Tutorial and Advanced Research Workshop on the Auditory Basis of Speech Perception*, pp. 1–8.

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—a syllable-centric perspective. *Journal of Phonetics*, 31, 465–485.

Griffiths, T. L., Kemp, C., & Tenenbaum, J. B. (to appear). Bayesian models of cognition. In Sun, R., editor, *Cambridge Handbook of Computational Cognitive Modeling*.

Göksel, A. & Kerslake, C. (2005). *Turkish: A Comprehensive Grammar*. London: Routledge.

Hankamer, J. (1986). Morphological parsing and the lexicon. In Marslen-Willson, W., editor, *Lexical Representation and Process*, pp. 392–408. Cambridge, MA: MIT Press.

Harris, Z. S. (1955). From phoneme to morpheme. *Language*, 31, 190–222.

Inkelas, S. & Orgun, C. O. (2003). Turkish stress: a review. *Phonology*, 20, 139–161.

Jack, K., Reed, C., & Waller, A. (2006). From syllables to syntax: Investigating staged linguistic development through computational modelling. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*.

Jusczyk, P. W., Cutler, A., & Redanz, N. (1993). Preference for the predominant stress pattern of English words. *Child Development*, 64, 675–687.

Jusczyk, P. W., Hohne, E. A., & Baumann, A. (1999a). Infants' sensitivity to allophonic cues for word segmentation. *Perception & Psychophysics*, pp. 1465–1476.

Jusczyk, P. W., Hohne, E. A., & Newsome, M. (1999b). The beginnings of word segmentation by English learning infants. *Cognitive Psychology*, 39, 159–207.

Kabak, B. & Vogel, I. (2001). The phonological word and stress assignment in Turkish. *Phonology*, 18, 315–360.

Kaplan, R. M. & Bresnan, J. (1982). Lexical-functional grammar: A formal system for grammatical representation. In Bresnan, J., editor, *The Mental Representation of Grammatical Relations*, pp. 173–281. MIT Press, Cambridge, MA.

Kornfilt, J. (1997). *Turkish*. London and New York: Routledge.

Kuhl, P. K. (2004). Early language acquisition: Cracking the speech code. *Nature Reviews Neuroscience*, 5(11), 831–843.

Körding, K. P. & M.Wolpert, D. (2004). Bayesian integration in sensorimotor learning. *Nature*, 427, 244–247.

Küntay, A. & Slobin, D. I. (1994). Nouns and verbs in Turkish child directed speech. In MacLaughlin, D. & McEwen, S., editors, *Proceedings of the Boston University Conference on Language Development 19*, pp. 323–334.

MacKay, D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, Cambridge, U.K.

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

Marslen-Wilson, W. (1999). Abstractness and combination: Morphemic lexicon. In Garrod, S. & Pickering, M., editors, *Language Processing*, pp. 101–119. Psychology Press.

Marslen-Wilson, W. D. (1973). Linguistic structure and speech shadowing at very short latencies. *Nature*, 244, 522–523.

Mattys, S. L. & Jusczyk, P. W. (2001). Phonotactic cues for segmentation of fluent speech by infants. *Cognition*, 78, 91–121.

Mattys, S. L. & Samuel, A. G. (2000). Implications of stress pattern differences in spoken word recognition. *Journal of Memory and Language*, 36, 87–116.

McClelland, J. L. & Elman, J. L. (1986). Trace model of speech perception. *Cognitive Psychology*, 18, 1–86.

Mehler, J., Christophe, A., & Ramus, F. (2000). How infants acquire language: some preliminary observations. In Marantz, A., Miyashita, Y., & O'Neil, W., editors, *Image, Language, Brain: Papers from the first Mind-Brain Articulation Project symposium*, pp. 51–75. Cambridge, MA: MIT Press.

Mehler, J., Dommergues, J. Y., Frauenfelder, U., & Segui, J. (1981). The syllable's role in speech segmentation. *Journal of Verbal Learning and Verbal Behavior*, 20, 298–305.

Mehler, J., Dupoux, T., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. In Morgan, J. L. & Demuth, K., editors, *Signal to syntax*. Lawrence Erlbaum.

Morgan, J. L. (1996). A rhythmic bias in preverbal speech segmentation. *Journal of Memory and Language*, 35, 666–688.

Neuvel, S. & Fulop, S. (2002). Unsupervised learning of morphology without morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning*. ACL.

Norris, D. (2006). The Bayesian reader: Explaining word recognition as an optimal bayesian decision process. *Psychological Review*, 113, 327–357.

Oflazer, K., Göçmen, E., & Bozşahin, C. (1994). An outline of Turkish morphology. Technical report, METU, Bilkent.

Pike, K. L. (1945). The intonation of American English. In Bolinger, D., editor, *Intonation*, pp. 53–83. Harmondsworth: Penguin.

Pinker, S. (1991). Rules of language. *Science*, 253, 530–535.

Ramus, F. (2001). Languages' rhythm and language acquisition. http://www.physik.uni-bielefeld.de/complexity/ramus.pdf.

Reimers, P. M. (2005). The basic syllable in first language acquisition. In *Proceedings of Essex Graduate Student Papers in Language and Linguistics*.

Rumelhart, D. E. & McClelland, J. L. (1986). On learning the past tenses of English verbs. In McClelland, J. L. & Rumelhart, D. E., editors, *Parallel distributed processing: vol 2: psychological and biological models*, 2, pp. 216–271. MIT Press, Cambridge, MA, USA.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, 12, 110–114.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical cues in language acquisition: Word segmentation by infants. In *Proceedings of 18th Annual Cognitive Science Society Conference*, pp. 376–380. Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Saffran, J. R., Senghas, A., & Trueswell, J. C. (2001). The acquisition of language by children. In *Proceedings of the National Academy of Sciences*, 98, pp. 12874–12875.

Schone, P. & Jurafsky, D. (2000). Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL-2000)*, pp. 67–72.

Schone, P. & Jurafsky, D. (2001). Knowlege-free induction of inflectional morphologies. In *Proceedings of the North American chapter of the Association for Computational Linguistics (NAACL-2001)*.

Sebastian-Galles, N., Dupoux, E., Segui, J., & Mehler, J. (1992). Contrasting syllabic effects in Catalan and Spanish. *Journal of Memory and Language*, 31, 18–32.

Shillcock, R. (1990). Functional parallelism in spoken word recognition. In Altman, G. T. M., editor, *Cognitive Models of Speech Processing*, pp. 24–49. Cambridge, MA: MIT Press.

Slobin, D. (1990). Universal and particular in the acquisition of language. In Wanner, E. & Gleitman, L., editors, *Language acquisition: The state of the art*, pp. 128–172. Cambridge University Press.

Steedman, M. (2000). *The syntactic process*. MIT Press, Cambridge, MA, USA.

Suomi, K., McQueen, J. M., & Cutler, A. (1997). Vowel harmony and speech segmentation in Finnish. *Journal of Memory and Languae*, 36, 422–444.

Tenenbaum, J. B. & Xu, F. (2000). Word learning as Bayesian inference. In Glietmen, L. R. & Joshi, A. K., editors, *Proceedings of 22nd Annual Conference of Cognitive Science Society*, pp. 517–522. Erlbaum.

Thiessen, E. D. & Saffran, J. R. (2003). Wnen cues collide: Use of stress and statistical cues to word boundaries by 7- to 9-mohth-old infants. *Developmental Psychology*, 39(4), 706–716.

Topbaş, S., Maviş, I., & Başal, M. (1997). Acquisition of bound morphemes: Nominal case morphology in Turkish. In *Proceedings of the 8th International Conference on Turkish Linguistics*, pp. 127–137. Ankara University.

Venkataraman, A. (1999). A statistical model for word discovery in child directed speech. Technical report, Computer Science, IIST, Massey University.

Vroomen, J. & de Gelder, B. (1994). Speech segmentation in dutch: No role for the syllable. In *Proceedings of the Third International Conference on Spoken Language Processing*, 3, p. 1135–1138, Yokohama.

Wu, S.-L. (1998). Incorporating information from syllable-length time scales into automatic speech recognition. Technical Report TR-98-014, International Computer Science Instutute, Berekeley, California.

Yuille, A. & Kersten, D. (2006). Vision as Bayesian inference: analysis by synthesis? *Psychological Review*, 113, 301–308.

Zettlemoyer, L. S. & Collins, M. (2005). Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Twenty First Conference on Uncertainty in Artificial Intelligence (UAI-05)*.

APPENDICES

APPENDIX A: THE INFLECTIONAL FORMS

These are the listings of inflectional forms learned from the same input.

- *Table A.1* lists the inflections obtained manually.
- *Table A.2* lists the inflections learned by the morpheme-based model.
- *Table A.3* lists the inflections learned by the syllable-based model.

Table A.1. Inflections in the hand-made lexicon.

Logical Form	Syntactic Category	Phonetic Form(s)
$\lambda x.plural(x)$	$N_{plu} \backslash N$	lar, ler, nlar
$\lambda x.pos1p(x)$	$N_{p1p} \backslash N$	imiz, ımız, mız, muz
	$N_{p1p} \backslash N_{plu}$	imiz
$\lambda x.pos1s(x)$	$N_{p1s} \backslash N$	m, ım, im, um, üm
	$N_{p1s} \backslash N_{plu}$	ım, im
$\lambda x.pos2p(x)$	$N_{p2p} \backslash N$	ınız, iniz, nız, niz, unuz
	$N_{p2p} \backslash N_{plu}$	iniz
$\lambda x.pos2s(x)$	$N_{p2s} \backslash N$	ın, in, n, un, ün, yun
	$N_{p2s} \backslash N_{plu}$	ın, in
$\lambda x.pos3p(x)$	$N_{p3p} \backslash N$	ları, leri
$\lambda x.pos3s(x)$	$N_{p3s} \backslash N$	ı, i, sı, si, su, sü, u, ü
	$N_{p3s} \backslash N_{plu}$	ı, i
$\lambda x.locative(x)$	$N_{loc} \backslash N$	de, da, nda, ta, te
	$N_{loc} \backslash N_{plu}$	de, da
	$N_{loc} \backslash N_{p1s}$	de, da
	$N_{loc} \backslash N_{p2s}$	de, da
	$N_{loc} \backslash N_{p3s}$	nde, nda
	$N_{loc} \backslash N_{p2p}$	de
$\lambda x.ablative(x)$	$N_{abl} \backslash N$	den, dan, ndan, tan, ten
	$N_{abl} \backslash N_{plu}$	den, dan
	$N_{abl} \backslash N_{p1s}$	dan
	$N_{abl} \backslash N_{p2s}$	den, dan
	$N_{abl} \backslash N_{p3s}$	nden, ndan
	$N_{abl} \backslash N_{p3p}$	nden
$\lambda x.dative(x)$	$N_{dat} \backslash N$	e, a, ye, ya, na
	$N_{dat} \backslash N_{plu}$	e, a
	$N_{dat} \backslash N_{p1s}$	e, a
	$N_{dat} \backslash N_{p2s}$	e, a
	$N_{dat} \backslash N_{p3s}$	ne, na
	$N_{dat} \backslash N_{p1p}$	e
	$N_{dat} \backslash N_{p2p}$	e, a
	$N_{dat} \backslash N_{p3p}$	ne
$\lambda x.accusative(x)$	$N_{acc} \backslash N$	ı, i, ni, nu, u, ü, yı, yu, yü
	$N_{acc} \backslash N_{plu}$	ı, i
	$N_{acc} \backslash N_{p1s}$	ı, i
	$N_{acc} \backslash N_{p2s}$	ı, i, u, ü
	$N_{acc} \backslash N_{p3s}$	nı, ni, nu
	$N_{acc} \backslash N_{p1p}$	i, u
	$N_{acc} \backslash N_{p2p}$	i
	$N_{acc} \backslash N_{p3p}$	nı, ni
$\lambda x.genitive(x)$	$N_{gen} \backslash N$	ın, in, nın, nin, nun, nün, un, ün
	$N_{gen} \backslash N_{plu}$	ın, in
	$N_{gen} \backslash N_{p1s}$	in
	$N_{gen} \backslash N_{p2s}$	ın, in
	$N_{gen} \backslash N_{p3s}$	nın, nin
	$N_{gen} \backslash N_{p1p}$	ın
	$N_{gen} \backslash N_{p2p}$	in
$\lambda x.instrumental(x)$	$N_{ins} \backslash N$	inle, inlen, la, lan, le, len, nla, nlan, nunla, yla, ylan, ylen, yle
	$N_{ins} \backslash N_{plu}$	la, le, lan
	$N_{ins} \backslash N_{p1s}$	la
	$N_{ins} \backslash N_{p2s}$	la, lan, le, len
	$N_{ins} \backslash N_{p3s}$	nlan, yle

Table A.2. Inflections derived by the morpheme-based model after two runs through the training set. The **bold** items are the items that are in L_m, but not in L_r, and items printed with lighter text colour are the items that are not in L_m but in L_r.

Logical Form	Syntactic Category	Phonetic Form(s)
$\lambda x.plural(x)$	$N_{plu}\backslash N$	lar, ler, nlar
$\lambda x.pos1p(x)$	$N_{p1p}\backslash N$	imiz, ımız, mız , muz
	$N_{p1p}\backslash N_{plu}$	imiz
$\lambda x.pos1s(x)$	$N_{p1s}\backslash N$	m, ım, im, um, üm
	$N_{p1s}\backslash N_{plu}$	ım, im
$\lambda x.pos2p(x)$	$N_{p2p}\backslash N$	ınız, iniz, nız, niz, unuz
	$N_{p2p}\backslash N_{plu}$	iniz
$\lambda x.pos2s(x)$	$N_{p2s}\backslash N$	ın, in, n, un, ün, yun
	$N_{p2s}\backslash N_{plu}$	ın, in
	$N_{p2s}\backslash N_{gen}$	**in**
$\lambda x.pos3p(x)$	$N_{p3p}\backslash N$	ları, leri
$\lambda x.pos3s(x)$	$N_{p3s}\backslash N$	ı, i, sı, si, su, sü, u, ü
	$N_{p3s}\backslash N_{plu}$	ı, i
	$N_{p3s}\backslash N_{acc}$	**nı**, **ni**, **nu**
$\lambda x.locative(x)$	$N_{loc}\backslash N$	de, da, nda, ta, te
	$N_{loc}\backslash N_{plu}$	de, da
	$N_{loc}\backslash N_{p1s}$	de, da
	$N_{loc}\backslash N_{p2s}$	de, da
	$N_{loc}\backslash N_{p3s}$	nde, nda
	$N_{loc}\backslash N_{p2p}$	de
$\lambda x.ablative(x)$	$N_{abl}\backslash N$	den, dan, ndan, tan, ten
	$N_{abl}\backslash N_{plu}$	den, dan
	$N_{abl}\backslash N_{p1s}$	dan
	$N_{abl}\backslash N_{p2s}$	den, dan
	$N_{abl}\backslash N_{p3s}$	nden, ndan
	$N_{abl}\backslash N_{p3p}$	nden
$\lambda x.dative(x)$	$N_{dat}\backslash N$	e, a, ye, ya, na
	$N_{dat}\backslash N_{plu}$	e, a
	$N_{dat}\backslash N_{p1s}$	e, a
	$N_{dat}\backslash N_{p2s}$	e, a
	$N_{dat}\backslash N_{p3s}$	ne, na
	$N_{dat}\backslash N_{p1p}$	e
	$N_{dat}\backslash N_{p2p}$	e, a
	$N_{dat}\backslash N_{p3p}$	ne
$\lambda x.accusative(x)$	$N_{acc}\backslash N$	ı, i, ni, nu, u, ü, yı, yu, yü, **si**
	$N_{acc}\backslash N_{plu}$	ı, i
	$N_{acc}\backslash N_{p1s}$	ı, i, ü
	$N_{acc}\backslash N_{p2s}$	ı, i, u, ü
	$N_{acc}\backslash N_{p3s}$	nı, ni, nu
	$N_{acc}\backslash N_{p1p}$	i, u
	$N_{acc}\backslash N_{p2p}$	i
	$N_{acc}\backslash N_{p3p}$	nı, ni
$\lambda x.genitive(x)$	$N_{gen}\backslash N$	ın, in, nın, nin, nun, nün, un, ün
	$N_{gen}\backslash N_{plu}$	ın, in
	$N_{gen}\backslash N_{p1s}$	in
	$N_{gen}\backslash N_{p2s}$	ın, in
	$N_{gen}\backslash N_{p3s}$	nın, nin
	$N_{gen}\backslash N_{p1p}$	ın
	$N_{gen}\backslash N_{p2p}$	in
$\lambda x.instrumental(x)$	$N_{ins}\backslash N$	inle, inlen, la, lan, le, len, nla, nlan, nunla, yla, ylan, ylen, yle
	$N_{ins}\backslash N_{plu}$	la, le, lan
	$N_{ins}\backslash N_{p1s}$	la
	$N_{ins}\backslash N_{p2s}$	la, lan, le, len
	$N_{ins}\backslash N_{p3s}$	nlan, yle

Table A.3. Inflections derived by the syllable-based model after three runs through the training set. The **bold** items are the items that are in L_s, but not in L_m, and items printed with lighter text colour are the items that are not in L_s but in L_m.

Logical Form	Syntactic Category	Phonetic Form(s)
$\lambda x.plural(x)$	$N_{plu}\backslash N$	lar, ler, nlar
	$N_{plu}\backslash N_{ins}$	**rin**, **riy**, **rım**, **rın**
$\lambda x.pos1p(x)$	$N_{p1p}\backslash N$	**y**imiz, **y**ımız
	$N_{p1p}\backslash N_{plu}$	imiz
$\lambda x.pos1s(x)$	$N_{p1s}\backslash N$	**ne**m, ım , **n**im, **y**um, üm
	$N_{p1s}\backslash N_{plu}$	ım, im
$\lambda x.pos2p(x)$	$N_{p2p}\backslash N$	nız, niz, **n**unuz, ınız, iniz
	$N_{p2p}\backslash N_{plu}$	iniz
	$N_{p2p}\backslash N_{acc}$	**zi**
$\lambda x.pos2s(x)$	$N_{p2s}\backslash N$	yun, **n**in, **n**un, **s**in, **tağın, nen, n , ın , ün**
	$N_{p2s}\backslash N_{acc}$	ni, nu, nı, **rini**
	$N_{p2s}\backslash N_{dat}$	**ne**, **ğına**
	$N_{p2s}\backslash N_{loc}$	**bında**, **ğında**
	$N_{p2s}\backslash N_{ins}$	**rin**
$\lambda x.pos3p(x)$	$N_{p3p}\backslash N$	ları, leri
	$N_{p3p}\backslash N_{dat}$	**lerine**
$\lambda x.pos3s(x)$	$N_{p3s}\backslash N$	sı, si, su, sü, **n**u, **y**ı, **tağı**, **yağı**, ü , i
	$N_{p3s}\backslash N_{dat}$	**rına**
	$N_{p3s}\backslash N_{loc}$	**sında**
$\lambda x.locative(x)$	$N_{loc}\backslash N$	de, da, ta, te, nda
	$N_{loc}\backslash N_{plu}$	de, da
	$N_{loc}\backslash N_{p1s}$	de , da
	$N_{loc}\backslash N_{p2s}$	de, da
	$N_{loc}\backslash N_{p3s}$	nde , nda
	$N_{loc}\backslash N_{p2p}$	de
$\lambda x.ablative(x)$	$N_{abl}\backslash N$	den, dan, tan, ten, ndan
	$N_{abl}\backslash N_{plu}$	den, dan
	$N_{abl}\backslash N_{p1s}$	dan
	$N_{abl}\backslash N_{p2s}$	den, dan
	$N_{abl}\backslash N_{p3s}$	nden , ndan
	$N_{abl}\backslash N_{p3p}$	nden
$\lambda x.dative(x)$	$N_{dat}\backslash N$	ye, ya, na, **ne**, **nara**, **yağa**, **ni**, a
	$N_{dat}\backslash N_{plu}$	e , a
	$N_{dat}\backslash N_{p1s}$	e , a
	$N_{dat}\backslash N_{p2s}$	e , a
	$N_{dat}\backslash N_{p3s}$	ne, na
	$N_{dat}\backslash N_{p1p}$	e , a
	$N_{dat}\backslash N_{p2p}$	e , a
	$N_{dat}\backslash N_{p3p}$	ne
$\lambda x.accusative(x)$	$N_{acc}\backslash N$	ni, nu, yı, yu, yü,**yi**, **n**u, **si**, **liği**, **tabı**, **tağı**, **yağı**, **ne**, ü, ı
	$N_{acc}\backslash N_{plu}$	ı , i
	$N_{acc}\backslash N_{p1s}$	ı , i
	$N_{acc}\backslash N_{p2s}$	ı , i , u , ü
	$N_{acc}\backslash N_{p3s}$	nı, ni, nu
	$N_{acc}\backslash N_{p1p}$	i , u
	$N_{acc}\backslash N_{p2p}$	i
	$N_{acc}\backslash N_{p3p}$	nı, ni
$\lambda x.genitive(x)$	$N_{gen}\backslash N$	nın, nin, nun, nün, **yun**, ın , in , un , ün
	$N_{gen}\backslash N_{plu}$	ın , in
	$N_{gen}\backslash N_{p1s}$	in
	$N_{gen}\backslash N_{p2s}$	ın , in
	$N_{gen}\backslash N_{p3s}$	nın, nin
	$N_{gen}\backslash N_{p2p}$	in
$\lambda x.instrumental(x)$	$N_{ins}\backslash N$	la, le, lan, len, nunla, **n**inle, **neyle**, inlen , nla , nlan , yla , ylan , ylen
	$N_{ins}\backslash N_{plu}$	la, le, lan
	$N_{ins}\backslash N_{p2s}$	la, len, le , lan
	$N_{ins}\backslash N_{p3s}$	nlan , yle

APPENDIX B: THE HIGHEST SCORING LEXICAL ITEMS

Table B.1. Top 50 items learned by the morpheme- and the syllable-based models.

Morpheme-based Model	Syllable-based Model
zaman := $N : time$;1.0000	zaman := $N : time$;1.0000
sen := $N : you$;1.0000	sen := $N : you$;1.0000
o := $N : he/she/it$;1.0000	o := $N : he/she/it$;1.0000
lar := $N_{plu} \backslash N : \lambda x.plural(x)$;1.0000	na := $N_{dat} \backslash N : \lambda x.dative(x)$;1.0000
ben := $N : I$;1.0000	ben := $N : I$;1.0000
başka := $N : other$;1.0000	başka := $N : other$;1.0000
a := $N_{dat} \backslash N : \lambda x.dative(x)$;1.0000	lar := $N_{plu} \backslash N : \lambda x.plural(x)$;0.9999
n := $N_{p2s} \backslash N : \lambda x.pos2s(x)$;0.9999	se := $N : you$;0.9997
in := $N_{gen} \backslash N : \lambda x.genitive(x)$;0.9999	nin := $N_{gen} \backslash N : \lambda x.genitive(x)$;0.9992
ler := $N_{plu} \backslash N : \lambda x.plural(x)$;0.9998	nu := $N_{acc} \backslash N : \lambda x.accusative(x)$;0.9963
anne := $N : mother$;0.9996	sa := $N : you$;0.9929
baba := $N : father$;0.9982	ba := $N : I$;0.9928
kedi := $N : cat$;0.9980	kedi := $N : cat$;0.9867
da := $N_{loc} \backslash N : \lambda x.locative(x)$;0.9967	ler := $N_{plu} \backslash N : \lambda x.plural(x)$;0.9856
i := $N_{acc} \backslash N : \lambda x.accusative(x)$;0.9966	köpek := $N : dog$;0.9845
ev := $N : house$;0.9966	da := $N_{loc} \backslash N : \lambda x.locative(x)$;0.9772
bu := $N : this$;0.9958	birşey := $N : something$;0.9758
köpek := $N : dog$;0.9950	ütü := $N : iron$;0.9740
nu := $N_{acc} \backslash N : \lambda x.accusative(x)$;0.9944	ev := $N : house$;0.9694
ban := $N : I$;0.9924	resim := $N : picture$;0.9693
san := $N : you$;0.9921	baba := $N : father$;0.9681
e := $N_{dat} \backslash N : \lambda x.dative(x)$;0.9904	an := $N : mother$;0.9652
nlar := $N_{plu} \backslash N : \lambda x.plural(x)$;0.9897	nen := $N_{p2s} \backslash N : \lambda x.pos2s(x)$;0.9574
çocuk := $N : child$;0.9888	su := $N : water$;0.9569
ım := $N_{p1s} \backslash N : \lambda x.pos1s(x)$;0.9856	de := $N_{loc} \backslash N : \lambda x.locative(x)$;0.9545
ütü := $N : iron$;0.9799	anne := $N : mother$;0.9527
birşey := $N : something$;0.9791	balık := $N : fish$;0.9519
balık := $N : fish$;0.9791	kuş := $N : bird$;0.9472
kuş := $N : bird$;0.9769	çocuk := $N : child$;0.9443
kız := $N : girl$;0.9749	yemek := $N : food$;0.9433
de := $N_{loc} \backslash N : \lambda x.locative(x)$;0.9736	tane := $N : piece$;0.9394
resim := $N : picture$;0.9712	kız := $N : girl$;0.9367
yer := $N : place$;0.9679	bu := $N : this$;0.9319
ı := $N_{acc} \backslash N_{\lambda} x.plural(x) : \lambda x.accusative(x)$;0.9605	on := $N : he/she/it$;0.9314
ya := $N_{dat} \backslash N : \lambda x.dative(x)$;0.9583	kurt := $N : wolf/worm$;0.9269
su := $N : water$;0.9580	bun := $N : this$;0.9106
can := $N : life$;0.9540	le := $N_{ins} \backslash N : \lambda x.instrumental(x)$;0.8969
im := $N_{p1s} \backslash N : \lambda x.pos1s(x)$;0.9524	be := $N : I$;0.8738
yemek := $N : food$;0.9496	bur := $N : here$;0.8726
araba := $N : car$;0.9461	biz := $N : we$;0.8667
tane := $N : piece$;0.9423	bugün := $N : today$;0.8606
top := $N : ball$;0.9406	nun := $N_{gen} \backslash N : \lambda x.genitive(x)$;0.8514
kurt := $N : wolf/worm$;0.9392	şu := $N : that$;0.8487
biz := $N : we$;0.9288	araba := $N : car$;0.8426
oyun := $N : game$;0.9224	yer := $N : place$;0.8409
ı := $N_{acc} \backslash N : \lambda x.accusative(x)$;0.9151	nim := $N_{p1s} \backslash N : \lambda x.pos1s(x)$;0.8338
u := $N_{acc} \backslash N : \lambda x.accusative(x)$;0.9111	len := $N_{ins} \backslash N : \lambda x.instrumental(x)$;0.8218
le := $N_{ins} \backslash N : \lambda x.instrumental(x)$;0.8893	ni := $N_{acc} \backslash N : \lambda x.accusative(x)$;0.8196
şu := $N : that$;0.8851	bebek := $N : baby$;0.8165
bur := $N : here$;0.8726	oyun := $N : game$;0.8004