# POWER OF FREQUENCIES: N-GRAMS AND SEMI-SUPERVISED MORPHOLOGICAL SEGMENTATION IN TURKISH

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS
OF
THE MIDDLE EAST TECHNICAL UNIVERSITY

BY

ÖZKAN KILIÇ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF
DOCTOR OF PHILOSOPHY
IN
THE DEPARTMENT OF COGNITIVE SCIENCE

JUNE 2013

POWER OF FREQUENCIES: N-GRAMS AND SEMI-SUPERVISED
MORPHOLOGICAL SEGMENTATION IN TURKISH

Submitted by **ÖZKAN KILIÇ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in the Department of Cognitive Science**, **Middle East Technical University** by,

Prof. Dr. Nazife Baykal ________________________
Director, Informatics Institute

Prof. Dr. Cem Bozşahin ________________________
Head of Department, Cognitive Science

Prof. Dr. Cem Bozşahin ________________________
Supervisor, Cognitive Science, METU

**Examining Committee Members:**

Prof. Dr. Deniz Zeyrek ________________________
COGS, METU

Prof. Dr. Cem Bozşahin ________________________
COGS, METU

Asst. Prof. Dr. Cengiz Acartürk ________________________
COGS, METU

Prof. Dr. Ferda Nur Alpaslan ________________________
CENG, METU

Asst. Prof. Dr. F. Nihan Ketrez ________________________
English Teacher Education, Istanbul Bilgi University

**Date:** **20 June 2013**

**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 wok.**

**Name, Last name : ÖZKAN KILIÇ**

**Signature : ____________________**

# ABSTRACT

POWER OF FREQUENCIES: N-GRAMS AND SEMI-SUPERVISED
MORPHOLOGICAL SEGMENTATION IN TURKISH

Kılıç, Özkan
Ph.D., Department of Cognitive Science
Supervisor: Prof. Dr. Cem Bozşahin

June 2013, 147 pages

Turkish is an agglutinating language with a non-rigid word order. When communicating, the word internal structure in Turkish is required to be segmented because Turkish morphosyntax is tortuous and it plays a central role in semantic analysis. Distinguishing a sub-word unit actually means performing a morph segmentation task, which is accomplished by children at an astonishing success rate. In this study, morph segmentation of Turkish words was demonstrated with a semi-supervised Hidden Markov Model, which emphasized the power of frequencies and sequences as direct (or indirect negative) evidence for language acquisition. The method achieved .88, .92 and .90 (precision, recall and *f*-score) measures after being trained by the METU Corpus and the METU-Sabancı Turkish Treebank. Additionally, statistical approaches were offered for compound word recognition and segmentation. In order to corroborate the use of frequencies in the cognitive studies, the experimental studies and the corresponding statistical models in Turkish emphatic reduplication and the acceptability of nonce words were also proposed in this study. This study shows that since the probability mass in child-directed speech is skewed toward possible word forms and unlikely morph sequences, this mass can be used by various models to mimic human-level linguistic capabilities. Furthermore, human beings have a statistical learning ability and it is not specific to the faculty of language as claimed by nativists but to general cognition. This allows the plausible and valid use of computational and statistical models to analyze language. Such predictive models can allow a deeper understanding of language.

Keywords: Indirect Negative Evidence; Morph Segmentation; Semi-supervised Learning

# ÖZ

## TEKRARLARIN GÜCÜ: TÜRKÇE'DE N-GRAMLAR VE YARI-DENETİMLİ BİÇİMBİLİMSEL BÖLME

Kılıç, Özkan
Doktora, Bilişsel Bilimler Bölümü
Tez Yöneticisi: Prof. Dr. Cem Bozşahin

Haziran 2013, 147 Sayfa

Türkçe serbest sözcük dizimine sahip bitişimli bir dildir. İletişim sırasında, Türkçe'deki kelimelerin yapısal bölümlerine ayrılması gereklidir; çünkü Türkçe'nin biçimbilimsel sözdizimi karışıktır ve bu durum anlamsal çözümlemede merkezi bir rol oynar. Sözcük-altı parçacıkların ayrıştırılması aslında çocuklar tarafından şaşırtıcı bir başarıyla gerçekleştirilen bir biçimbirim bölme işlemidir. Bu çalışmada, Türkçe kelimelerin biçimbirim ayrıştırılması bir yarı-denetimli Gizli Markov Model'i ile gösterilmiştir. Model, tekrarların ve dizilimlerin gücünü dil ediniminde doğrudan (veya dolaylı olumsuz) kanıt olarak vurgulamaktadır. Yöntem, ODTÜ Türkçe Derlemi ve ODTÜ-Sabancı Türkçe Ağaç Yapılı Derlemi tarafından eğitildikten sonra .88, .92 ve .90 (duyarlık, doğruluk, *f*-değeri) ölçümlerine ulaşmıştır. Ayrıca, bileşik sözcük tanımlama ve bölme için istatistiksel yaklaşımlar önerilmiştir. Bilişsel bilimlerde sıklıkların kullanımını desteklemek amacıyla, Türkçe sıfat pekiştirme ve sahte kelimelerin kabul edilebilirliği ile ilgili deneysel çalışmalar ve ilgili istatistiksel modeller bu çalışmada önerilmiştir. Bu çalışma şunu göstermektedir; çocukları yönlendiren konuşmalarda olası kelime formları ve muhtemel olmayan biçimbirim sıralarına yönelik çarpık bir olasılık yığını olduğu için, bu yığın çeşitli istatistiksel modeller tarafından insan düzeyinde dilbilimsel yetenekleri taklit etmede kullanılabilir. Ayrıca, insanlar istatistiksel bir öğrenme yeteneğine sahiptir ve bu yetenek doğalcıların iddia ettiği gibi dil yetisine has değildir fakat genel bilişsel yeteneklere dahildir. Bu durum dili analiz edecek hesaplamalı ve istatistiksel modellerin anlamlı ve geçerli kullanımlarına olanak sağlamaktadır. Böyle tahminsel modeller dilin derinlemesine anlaşılmasına izin vermektedir.

Anahtar Kelimeler: Biçimbirim Bölme; Dolaylı Olumsuz Delil; Yarı-denetimli Öğrenme

## ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my supervisor Prof. Dr. Cem Bozşahin for his guidance, suggestions, constructive criticism, invaluable encouragement, belief in my ability, and friendly attitude throughout the development of this thesis. He is an inspiration for what academic life should be about in Cognitive Science.

I would like to thank Prof. Dr. Deniz Zeyrek and Prof. Dr. Ferda Nur Alpaslan for their evaluations and concerns during the development of this thesis. I would also like to thank Assoc. Prof. Dr. Aslı Göksel and Asst. Prof. Dr. Nihan Ketrez for their evaluations and professional contributions. My special and friendly thanks also go to Prof. Dr. Mustafa Kuzuoğlu, Asst. Prof. Dr. Murat Şakir Ceyhan, Asst. Prof. Dr. Serkan Kükrer, Dr. Çağrı Çöltekin, Dr. Bilge Say, Duygu & Erkan Vural and Chris Taylor. I also thank TÜBİTAK for their financial support.

I would also like to express my sincere thanks to my best friends, Gökben Kaynar and Yahya Hürol Sipahioğlu, and my dear family. Finally, Ilgar Seyidov deserves very special thanks for his eternal encouragement and altruistic support.

# TABLE OF CONTENTS

ABSTRACT ........................................................ iv

ÖZ ........................................................ v

ACKNOWLEDGMENTS ........................................................ vi

TABLE OF CONTENTS ........................................................ vii

LIST OF TABLES ........................................................ ix

LIST OF FIGURES ........................................................ xi

LIST OF ABBREVIATIONS ........................................................ xii

CHAPTER

1. INTRODUCTION ........................................................ 1

   1.1 From Communication to Morphology ........................................................ 2

   1.2 Aim ........................................................ 5

   1.3 Scope ........................................................ 5

   1.4 Research Questions ........................................................ 5

   1.5 Nature and Nurture ........................................................ 6

   1.6 Material and Method ........................................................ 8

2. LINGUISTIC MORPHOLOGY AND ITS ASPECTS ........................................................ 9

   2.1 Operations in Linguistic Morphology ........................................................ 11

      2.1.1 Concatenation ........................................................ 12

      2.1.2 Morphophonemic Operations ........................................................ 13

   2.2 Ordering of Affixes ........................................................ 15

   2.3 Morphology and Its Interfaces ........................................................ 18

      2.3.1 Morphology and Phonology ........................................................ 19

      2.3.2 Morphology and Syntax ........................................................ 26

      2.3.3 Morphology and Semantics ........................................................ 32

   2.4 Acquisition of Morphology ........................................................ 42

3. COMPUTATIONAL MORPHOLOGY and ITS ASPECTS ........................................................ 48

   3.1 Rule-based Morphological Analyses ........................................................ 50

      3.1.1 Finite State Automata and Transducers ........................................................ 50

      3.1.2 Implementations and Improvements for Rule-Based Models ........................................................ 53

3.2 Statistical Morphological Analyses ................................................................ 55
3.2.1 Supervised Approaches to Morphology ..................................................... 55
3.2.2 Unsupervised and Semi-supervised Approaches to Morphology.................. 55
3.4 Need for Statistical Models of Morphology..................................................... 61
4. SEMI-SUPERVISED MORPH SEGMENTATION in TURKISH............................. 63
4.1 Method and Findings for Turkish Emphatic Reduplication............................... 63
4.2 Method and Findings for Acceptability of Nonce Words in Turkish ................. 69
4.3 The Morph Segmentation Method and Findings............................................... 71
4.3.1 Power of Raw Frequencies ......................................................................... 71
4.3.2 Compound Word Recognition and Segmentation........................................ 79
4.3.3 The HMM and Morph Segmentation ......................................................... 83
4.3.4 Assumptions and Improvements for the HMM........................................... 87
4.3.5 Smoothing and Backoff for the HMM......................................................... 89
4.2 Results ............................................................................................................ 91
4.2.1 Results of the Method before and after 5,010 Words................................... 92
4.2.2 Results of the Method after Assumptions, Improvements and 43,574 Words ............................................................................................................................ 92
5. COGNITIVE ASPECTS OF SEMI-SUPERVISED METHODS ............................... 93
5.1 Cognitive Reflections on Statistical Learning.................................................. 94
5.2 Degree of Supervision in Morphology Learning.............................................. 96
6. CONCLUSION AND DISCUSSIONS..................................................................... 99
6.1 Limitations .................................................................................................... 103
6.2 Future Implications........................................................................................ 103
REFERENCES......................................................................................................... 105
APPENDICES
A. Transition and Emission Probabilities for *evdekiler* from the Corpus ...................... 124
B. Transition and Emission Probabilities for *evdekiler* from the Treebank ................... 136
C. Transition and Emission Probabilities for *affina* from the Corpus........................... 139
D. Transition and Emission Probabilities for *affina* from the Treebank ........................ 142
E. CURRICULUM VITAE .......................................................................................... 144
F. VITA...................................................................................................................... 147

# LIST OF TABLES

Table 1 Frequency sorted list of old and contemporary Turkish words ................................44

Table 2 Results for Intuitive Reduplication of Non-adjectives............................................65

Table 3 Distribution of Distinct Roots Ending with Linkers in Treebank.............................66

Table 4 Consonant Co-occurrence Statistics from Corpus..................................................67

Table 5 Results for Intuitive Reduplication of Nonce Words..............................................68

Table 6  Acceptability of nonce words results from method and participants.......................70

Table 7 Total numbers of observed types and tokens of *n*-grams (*N*<=4) in the Corpus.......72

Table 8 Most frequent *n*-grams (*N*<=4) in the METU corpus .............................................72

Table 9 Most frequent *n*-grams (*N*<=4) in the BOUN corpus .............................................73

Table 10 Most frequent Turkish in CHILDES uttered by children and other participants.....73

Table 11 Most frequent *n*-grams in CHILDES uttered by children (*N* <= 4)........................74

Table 12 Most frequent *n*-grams in CHILDES uttered by mothers (*N* <= 4)........................75

Table 13 Most frequent *n*-grams in CHILDES uttered by fathers (*N* <= 4)..........................75

Table 14 Most frequent *n*-grams in CDS portion of CHILDES (*N* <= 4).............................76

Table 15 Most frequent roots and suffixes in Treebank......................................................77

Table 16 Most frequent roots and distributions of attached suffixes in Treebank .................77

Table 17 Co-occurrence percentages of 10-fold selection from Treebank............................78

Table 18 Co-occurrence percentages of 10-fold selection from the child speech in CHILDES ..........................................................................................................................78

Table 19 Co-occurrence percentages of 10-fold selection of roots from the CDS in CHILDES ..........................................................................................................................79

Table 20 Recognition success rates with respect to rewards and punishments using CDSs ..80

Table 21 Recognition success rates with respect to rewards and punishments using Treebank ........ 81

Table 22 Transition probabilities of the *n*-grams ........ 85

Table 23 Emission probabilities of the *n*-grams (ε denotes empty string) ........ 86

Table 24 Root ending in Treebank ........ 88

Table 25 Achievement scores of HMM ........ 92

## LIST OF FIGURES

Figure 1 Representation of DISAGREE in Auslan ................................................................4

Figure 2 Constituent structure of *discontentedness*...............................................................4

Figure 3 Form-meaning mapping .........................................................................................18

Figure 4 Derivational and inflectional morphemic prototypes ............................................39

Figure 5 Three approaches for compound word representations in lexicon.........................41

Figure 6 Probability density graph for *çok-bilmiş* 'wiseacre'.............................................82

Figure 7 Probability density graph for *yer-çekimi* 'gravity' ...............................................82

Figure 8 The Representation of an HMM ...........................................................................83

Figure 9 Trellis diagram for *kedim* 'my cat'........................................................................85

# LIST OF ABBREVIATIONS

| | |
|---|---|
| **1** | First person |
| **2** | Second person |
| **3** | Third person |
| **ABL** | Ablative |
| **ABS** | Absolutive |
| **ACC** | Accusative |
| **ACT** | Active |
| **AGR** | Agreement |
| **AOR** | Aorist |
| **ART** | Article |
| **AUSLAN** | Australian Sign Language |
| **CAUS** | Causative |
| **CDS** | Child-directed speech |
| **CM** | Compound marker |
| **COM** | Comitative |
| **COP** | Copula |
| **DAT** | Dative |
| **DEM** | Demonstrative |
| **DER** | Derivative |
| **DUAL** | Dual |
| **EMPH** | Emphatic |
| **FRC** | Full relative clauses |
| **FSA** | Finite-state automate |
| **FST** | Finite-state transducer |
| **FUT** | Future |
| **GEN** | Genitive |
| **GM** | Generalizing modality |
| **HMM** | Hidden Markov Model |
| **IMP** | Imperfective |
| **INDIC** | Indicative |
| **INF** | Infinitive |
| **INST** | Instrumental |
| **IO** | Input/Output |
| **LF** | Logical Form |
| **LOC** | Locative |
| **LSA** | Latent semantic analysis |
| **MAS** | Masculine |
| **MDT** | Morphological Doubling Theory |
| **MEG** | Magnetoencephalography |
| **MLE** | Maximum-likelihood estimation |
| **M-RED** | M-reduplication |
| **NEG** | Negation |
| **NEUT** | Neutral |
| **NOM** | Nominal |
| **OPT** | Optative |
| **PASS** | Passive |
| **PAST** | Past tense |
| **PERF** | Perfective |
| **PF** | Phonological form |
| **PLU** | Plural |
| **POSS** | Possessive |
| **PRC** | Pronominal relative clauses |
| **PRE** | Prefix |
| **PRES** | Present |
| **PROG** | Progressive |
| **PSB** | Possibility |
| **QP** | Question particle |
| **REC** | Reciprocal |
| **REL** | Relativizer |
| **SBJ** | Subjective |
| **SD** | Simple doubling |
| **SG** | Singular |
| **SUB** | Subordinator |
| **SOV** | Subject-Object-Verb |
| **TER** | Turkish emphatic reduplication |
| **TRANS** | Transitive |

# CHAPTER 1

# INTRODUCTION

When I first decided to work on linguistic morphology, I was told that morphology was the most intertwined branch of linguistics; thus, it would be a steep learning curve for me. Now I know what my instructor actually meant. No matter which branch of linguistics they study, every theorist sooner or later has to deal with morphology in order to corroborate and avoid future breaches in the theories. As stated by Spencer and Zwicky (2001), it is not because morphology is the dominant sub-discipline but because morphology is the study of word structure, and words are at the interface between phonology, syntax and semantics.

Linguistic morphology refers to the mental system involved in word formation, the internal structure of words and how they are formed (Aronoff & Fudeman, 2011). Since the core element in morphology is the word, the central purpose of morphology is to map sound to meaning within the word and between words (Beard, 1995). There is no satisfactory answer to the question of how to define a word with necessary and sufficient conditions. According to Trask (2004), a word can be orthographic or phonological. He states that an "orthographic word is a written sequence which has a white space at each end but no white space in the middle". Yet, this definition is not sufficient to define words in languages like Chinese and Vietnamese which have little or no morphology. A phonological word is defined as "a piece of speech which behaves as a unit of pronunciation according to criteria which vary from language to language". Unfortunately, phonemes and syllables also fall into this definition. Allwood et al., (2010) mention gestural words, which are pieces of physical communication behaving as units of gesturing according to criteria which also vary from language to language. For example, a gestural word can occur in sign languages. Aronoff and Fudeman (2011) divide words into syntactical, phonological and grammatical categories. Some theorists define words as the smallest unit of syntax. If the syntax governs the ordering of items, then *-s* in [[*break*]*s*]$_{\mathrm{v}}$ must be a word. Thus, the word *my* cannot stand alone as a full sentence in English but *mine* can. Therefore, *my* cannot be distinguished as a word. Some theorists define words as the domain of phonological operations, such as stress assignment. However, clitics are words grammatically but not phonologically because they usually avoid stress assignments. The occurrences of *break* in the sentences "*I take a break*" and "*I break it*" are tokens of the same word but they are different grammatical words because of their morphosyntactic properties.

In this study, a word is assumed to be an orthographic form which can occur freely in corpora of a language. For example, the words *the*, *window-s* and *ev-ler-imiz* 'our houses' are free forms but *-s, -ler* PLU and *-imiz* 1.PLU.POSS are not. Yet, they are all morphemes, which are defined as the smallest unit of language that carries meaning (Fromkin & Rodman, 1993). The focus of this study is the statistical segmentation of Turkish orthographic words into morphs, which also includes segmentation of single-word compounds because compound words have more than one free morpheme. Morphs are the surface forms of morphemes.

Morphological awareness is crucial in Turkish because it is an agglutinating language with a considerably complex morphology (see Göksel & Kerslake, 2005; Lewis, 2000). When communicating, the word internal structure in Turkish is required to be segmented because Turkish morphosyntax is tortuous and it plays a central role in semantic analysis. For example, although Turkish is considered as an SOV language, sentences are usually in a non-rigid word order. Thus, the subject and object of a verb can only be determined by morphological markers as in (1) rather than the word order.

| | | |
|---|---|---|
| (1) | *Köpek adam-ı ısırdı.* | *Köpeğ-i adam ısırdı.* |
| | Dog man-ACC bit | Dog-ACC man bit |
| | The dog bit the man. | The man bit the dog. |

As given in (1), a single morph *-ı* or *-i*, which are allomorphs of the accusative case marker and altered due to Turkish vowel harmony, is an important determinant for the sentential meaning. Thus, it should be immediately distinguished by the listener. Distinguishing a sub-word unit actually means performing a morpheme segmentation task, which requires expertise via linguistic awareness and having been exposed to linguistic data. The precise modeling of morphological segmentation using statistical methods requires a set of data with frequencies as given in a corpus. This method is described in the following chapters.

## 1.1 From Communication to Morphology

If the definition of the language is narrowed down to a tool for communication, then the set of its users might include almost all organisms. Each biological form evolves its own way of communication: Single cell organisms communicate via protein molecules and receptors on their membranes. Prairie dogs use alarm calls to code the specifications of intruders, such as their color, speed and level of danger (Slobodchiko, Paske & Verdolin, 2009). Honeybees dance to show the location of flowers (von Frisch, 1967). Elephants use ultrasonic sounds to call relatives to help infants or pregnant females (Lee & Moss, 1999; Moss, 1988). Similarly, the wild chimpanzees use sounds to call for help or warn about intruders, and for example, they can learn to produce 4 words after 7 years of training (Hayes & Nissen, 1971) or use sign language to present about 122 words or phrases (Gardner, Gardner & van Cantfort, 1989). The communication abilities of other species cannot be compared to humans in fact; ability of many birds and whales for sound production is beyond human capability however, it is not the level of the human articulatory system or the production of distinct sound patterns that make humans superior. The difference lies in the productivity and creativity in the linguistic forms.

Among the set of communication mediums such as gestures, signs and postures, language is the ultimate complex product of human mind. From a finite set of units and resources such as phonemes, morphemes, words, rules, long and short term memory, humans can produce and understand theoretically infinite linguistic outputs. Indeed experiences, proficiencies, expectations and context play a central role in language production and understanding; also the reciprocal link between the two requires a set of invertible functional operations. The difficulty of the study of language lies in the fact that language is studied and understood via itself. In other words, it acts as both subject and tool of the study. It might seem odd but computer science already uses some languages to *understand* other languages, i.e., parsers, analyzing strings of symbols based on predefined rules and forms.

The study of language involves phonology, morphology, lexicology, semantics, pragmatics, discourse as well as syntax. Yet, the application of linguistic knowledge on particular occasions (i.e., performance) is full of errors, uncertainties, pauses, incomplete and subjective variants. As claimed by nativist researchers, the focus of linguistics should be the user's knowledge of the grammatical rules of the native language (i.e., competence). The performance is more related to the faculty of language in broad sense while the competence is closer to the faculty of language in a narrow sense (Hauser, Chomsky & Fitch, 2002). Since the initiation of Generative Linguistics (Chomsky, 1957; Chomsky, 1965) many linguistic studies have focused on the study of the grammatical structure of word sequences, such as phrases and sentences. This grammar innately provides human with a basis for the reciprocal link between language production and understanding.

Linguistic morphology combines grammar with syntax. Yet, there are no cross-linguistic universals in morphology and there are some isolated languages with no morphology at all. Thus, the generativists are blamed for neglecting the importance of morphology. Actually, they consider syntax is more central than morphology and many aspects of morphology are more closely associated with the lexicon than syntax. Carstairs-McCarthy (2010) states that morphology exists because morphophonology exists. Consecutively, morphophonology exists because of the way language evolves. In other words, morphology exists because while the biological foundations of language in humans were evolving, certain random characteristics of the cognitive and communicative materials acted in a specific way as a process of natural selection. If human brains had operated differently in certain essential ways or if human bodies had evolved differently, some forms of language could have evolved, and there would have been grammar, but there would have been nothing like the current morphology (Carstairs-McCarthy, 2010). This claim is consistent with Chomsky's *extrinsic* explanation for phonology (which gives rise to morphophonology and then morphology) in which he (Chomsky, 2004, p.405) states:

> "... a large range of imperfections may have to do with the need to 'externalize' language. If we could communicate by telepathy, they would not arise. The phonological component is in a certain sense 'extrinsic' to language, and the locus of a good part of its imperfection, so one might speculate."

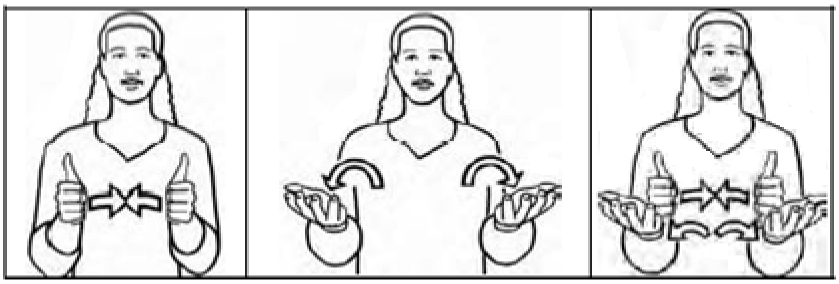
Speech is not the only system to convey language. Signed communication systems are complex, complete and grammatical languages, too. They are independent of, but equivalent to, spoken languages (Klima & Bellugi, 1979). In sign languages, despite some convergences, the morphology differs from the corresponding spoken language forms while syntax and word categories are highly related (Johnston & Schembri, 2007) as in the Australian Sign Language (Auslan) examples in (2).

(2) WOMAN STAY
N V
*The woman stayed.*

WOMAN BUY CAR D-A-R-W-I-N
N V N N
*The woman is buying a car in Darwin.*

The suffixes -*ed* and -*ing* are not represented in (2). Similarly, a subset of adverbs in English is easily recognized by the ending -*ly* (e.g., *happily)* but there is no adverbial ending in Auslan and the same signs may function as both adjectives and adverbs in many cases. On the other hand, in both English and Auslan sentences, adverbs can often occur next to the adjective, verb, or adverb they are modifying, but sometimes they may appear at the beginning or the end if they are modifying the sentence as a whole.

Word-grammar is often claimed to be determinant in syntax. Yet, the order of morphemes is determined by *common use*. Some morphemes, such as *dis*-, *un*-, and *non*- in English, are prefixes while others, like -*mA* in Turkish and -*less* in English, are postfixes for negation. An example is given in Figure 1 (taken from Johnston & Schembri, 2007).



AGREE NOT-DO DISAGREE (AGREE ^ NOT-DO)

**Figure 1** Representation of DISAGREE in Auslan

If the *common us*e for DISAGREE in Auslan was in accordance with 'prefix - verb', then it would be NOT-DO ^ AGREE form. Figure 2 shows an example given by Anderson (1992) for the constituent structure of *discontentedness.*

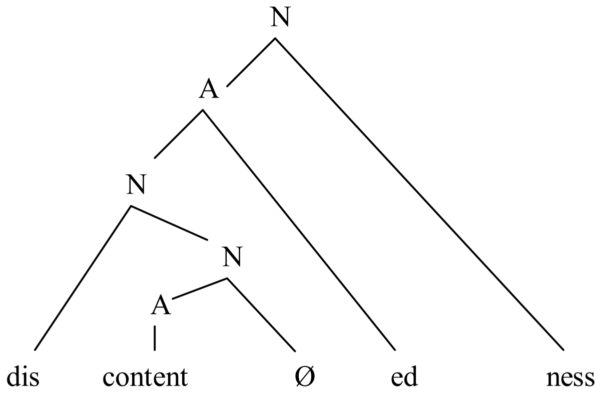


**Figure 2** Constituent structure of *discontentedness*

Anderson (1992) states that the morphemes in *discontentedness* cannot be combined as **ness-ed-content-dis* because of English morphotactics. Yet, if the adults in English speaking societies changed the daily use of this word to *nessedcontentdis* as well as all its related combinations, their newborn infants would acquire this new version. The orders of morphemes are not magical but random choices made in *common use* during the evolution of language. What children do is to deduce morphemes from *common use* and obtaining morphemes from *common use* requires statistical abilities.

Infants and adults use frequency information to segment language-like stimuli in different languages such as English, Japanese and Spanish (Batchelder, 1997). Actually, there are existing methods to solve some issues in morphology. For example, Clark (2007) described stochastic transducers and information geometry to model the Arabic broken plural. A linear method for Semitic nonconcatenative phonology was also given by Bird and Allison (1994). Stonham (1994) explained reduplication in Nitinaht language similar to the affixation process by using Semitic-like templates with vowel length constraints and melodic outputs.

Despite being disparaged and considered non-universal, morphology does exist. Even infants easily acquire the morphology of their languages and decompose words into morphemes. Therefore, there is no point in ignoring morphology by attributing it to syntax. The main question is whether the learning of morphology is innate or whether it is acquired via human cognitive skills, such as learning, memory capacity and computation.

During the second half of their first year, infants begin to show sensitivity to the sound organization of their native languages to build up their initial lexicon (Aslin, Jusczyk & Pisoni, 1997; Best, 1995). Infants rely on prosodic and statistical data to locate words in fluent speech. Stress patterns are effective in acquiring words and then attaching meaning to them (Jusczyk, 1999). A similar conclusion from an experimental study by Thiessen and Saffran (2003) stated that 7-month-old English learning infants attended more to statistical cues in patterns to determine word boundaries while 9-month-old infants were more sensitive to stress patterns. They also supported the findings of Mattys et al., (1999) that infants were sensitive to consonant clusters *within and across words* and were also very sophisticated in distinguishing such acoustic structures of clusters mixed with vowels. Alterations in the acquired clusters and structures speculatively resulted in morphemes, which were not necessarily concatenative. Similar experiments indicate the innateness of pattern recognition and statistics for word learning. The first lexical entries are not morphemes or not necessarily roots but are already inflected, derived, compounded or diacritic forms. As soon as adequate input is provided, the related patterns help infants to lexicalize the morphemes with corresponding combinatorial rules in a dynamic manner (for an application for combinatory lexicalized morphemes in Turkish, see Bozsahin, 2002).

### 1.2 Aim

The main aim of this study is to explore the ability of *n*-grams to close the gap between the poverty of stimulus argument and human behavior through a semi-supervised morphological word segmentation using frequencies. The cognitive plausibility of the model is another concern of the study.

### 1.3 Scope

The scope of this study is Turkish orthographical words. There is a close correspondence between phonotactics and orthotactics in Turkish, which means that orthographic morpheme segmentation resembles the actual task of phonological segmentation performed by the native speakers.

### 1.4 Research Questions

The research questions of this study are as follows:

- Is it possible to propose a semi-supervised statistical model with *n*-grams using frequencies in order to explain morphological word segmentation, acquisition of morphology and morphotactics?

- Will this be just a superfluous model fitting the existing linguistic data or will it be compatible with current cognitive empirical data?

- Is semi-supervision cognitively plausible?

## 1.5 Nature and Nurture

Before introducing material and method of the current study, it is informative to review the "nature versus nurture" discussion. The nativist perspective towards linguistics states that children achieve an adult-like and stable linguistic competency level with boundaries that are set by Universal Grammar (UG). In other words, there is some innate linguistic knowledge and this entails language acquisition. On the other hand, the empiricist approach states that language acquisition is inductive and language is learnable from input. The "native versus nurture" discussion depends on the poverty of stimulus argument. The poverty of stimulus argument can be summarized as follows:

- Premise 1: There are patterns in all natural languages that cannot be learned by children using positive evidence alone.
- Premise 2: Children are only presented with positive evidence for such particular patterns.
- Premise 3: Children learn the correct grammar for their native language.
- Conclusion: Therefore, human beings must have some form of innate linguistic capacity that provides additional knowledge to language learners.

Nativist scholars claim that some aspects of grammar such as the Binding Theory or Auxiliary Fronting cannot be learned from the positive data alone; thus, children must possess some linguistic knowledge motivating their language acquisition. For example, the declarative sentence "The man who is hungry is ordering dinner" is correct to front the main clause auxiliary as in (1a), but it is ungrammatical to front the subordinate clause auxiliary (1b) (Chomsky, 1965).

(1) a) Is the man who is hungry ordering dinner?
b) *Is the man who hungry is ordering dinner?

Children have two options for the example above: The first option is a structure-independent rule where the first *is* is relocated. The second is correct structure-dependent rule in which only the relocation of the *is* from the main clause is allowed (Chomsky, 1980). Strikingly, children do not go through a stage where they inaccurately move the first *is* to the front of the sentence (Crain & Nakayama, 1987), and they are not exposed to any explicit negative evidence to favor the structure-independent rule.

Crain and Pietroski (2001) stated that until empiricists show how specific principles can be learned on the basis of the primary linguistic data, the innateness hypotheses will continue to be the best available explanation for the gap between normal human experience and linguistic knowledge. In other words, empirical mechanisms should be introduced or improved in order to explain language acquisition and disprove the poverty of stimulus argument. However, Pullum and Scholz (2002) emphasized that the poverty of stimulus argument lacks empirical evidence, suggesting that the nativists work on empirical instances to support their argument. The authors defined a specification schema for the argument as follow:

- Acquirendum Characterization: Describe in detail what is supposed to be known.
- Lacuna (Gap) Specification: identify a set of sentences such that if the learner had access to them, the claim of data-driven learning of the acquirendum would be supported.

- Indispensability Arguments: give reason to think that if learning were data-driven then the acquirendum could not be learned without access to sentences in the lacuna.
- Inaccessibility Evidence: support the claim that tokens of sentences in the lacuna were not available to the learner during the acquisition process.
- Acquisition Evidence: give reason to believe that the acquirendum does in fact become known to learners during child-hood.

Although Pullum and Scholz (2002) analyzed some nativist arguments through this schema and concluded that the premises and corresponding reasoning schema for the poverty of stimulus argument were stated in an invalid way, they were ambivalent in terms of the native versus nurture debate. However, they emphasized that until data-driven learning is investigated in a more detailed way, linguists will remain ill-equipped and continue to fantasize and speculate about the issue.

In fact, empiricist researchers have attacked the premises of the poverty of stimulus argument. For example, Perfors et al., (2006) used context-free grammars to achieve auxiliary fronting by scoring the grammars, and concluded that structure dependence need not be part of innate linguistic knowledge. Computational linguistics provided many counter examples to the poverty of stimulus argument by various data-driven methods without or with minimum assumptions (e.g., Bod, 2009; Brent, 1993; Hsu, Chater & Vitanyi, 2013; Perfors, Tanenbaum & Regier, 2006; Reali & Christiansen, 2005; Pinker, 1989; Regier & Gahl; 2004; Schütze, 1995; Tomasello, 2000). Some researchers have tried to show that positive evidence itself is enough to learn language (contra Premise 1). Similarly, the researchers empirically suggest that negative evidence is abundant, at least in an indirect way (contra Premise 2). Finally, some researchers have speculated on the existence of a single correct grammar or a set of correct grammars allowing linguistic change and variance (contra Premise 3).

The joint generativist models, such as the HMM used in the current study, place probabilities over both the observed and the hidden data. The probabilities evaluated from the observed data can be employed as indirect negative evidence because such probabilities also indicate what is not attested in a language. For example, the forward evaluation algorithm for an HMM can accept or reject a genuine observation by using the probabilities of previously seen states as in the current study. Similarly, Reali and Christiansen (2005) employed a corpus and a statistical connectionist model to demonstrate that auxiliary fronting is possible even by only using the positive evidence in distributional information in the corpus. Moreover, syntactic acquisition is greatly facilitated when distributional information is integrated with other sources of probabilistic information (Christiansen & Monaghan, 2006; Monaghan et al., 2005; Morgan et al., 1987). Ramscar and Yarlett (2007) also designed a learning model that successfully simulated the learning of irregular plurals only based on positive evidence.

The discussions about the argument are very important because they incrementally contribute to not only the sides of the debate but also to linguistics. For example, Chomsky and his colleagues reviewed the empiricist studies and revisited the poverty of stimulus argument (Berwick et al., 2011). This study aims to show that it is possible to propose a semi-supervised HMM that can achieve a morphological segmentation of Turkish words. Moreover, it can also succeed in the acceptability decision for previously unseen observations by utilizing positive evidence. In other words, seen observations can provide indirect negative evidence as shown in Section 4.3.3.

## 1.6 Material and Method

A Hidden Markov Model (HMM) was utilized in this study to model morphological segmentation. The METU-Turkish Corpus (Say et al., 2002) was used to evaluate initial probabilities of the model. Semi-supervision was provided by the manual segmentation of the METU-Sabancı Turkish Treebank (Atalay et al., 2003). Moreover, the CHILDES database (MacWhinney, 2000) was also manually segmented into morphs in order to thoroughly understand the effects of *n*-gram frequencies and sequences in morphology. This was done to measure the plausibility of the semi-supervised HMM.

The METU Turkish Corpus is a collection of 2 million words from written Turkish samples post-1990. The corpus is XCES tagged at the typographical level. The words of the METU Turkish Corpus were taken from 10 different genres. At most 2 samples from one source are used; each sample consists of 2000 words or the sample ends when the next sentence starts. The METU-Sabanci Turkish Treebank is a morphologically and syntactically annotated corpus of 7,262 grammatical sentences. The sentences are taken from the METU Turkish Corpus. The similar percentages of different genres in METU-Sabanci Turkish Treebank and METU Turkish Corpus were maintained. The structure of the METU-Sabanci Turkish Treebank is based on XML.

# CHAPTER 2

# LINGUISTIC MORPHOLOGY AND ITS ASPECTS

The need for communication and learning for survival purposes drives children to a solution of language learning; grammar. If each word form in a language is composed of a single morpheme, then this language has distinct forms for the same lexeme in each different context and it is a grammatically perfect language. The main research issue in such a language will be the interactions of phonology, syntax and lexicon. Yet, this is not the case for great majority of human languages. Linguistic morphology basically refers to the mental system involved in word formation. Many morphologists state that the notions of morphology rest on the more basic notion of the lexeme (Bauer, 2003; Beard, 1995; Stump, 2001). A lexeme is a unit of linguistic analysis which belongs to a particular syntactic category, and has particular meaning and functions. Cross linguistic studies on morphology have argued that the lexicon should be morphemic.

In 1889, Baudine de Courtenay originally defined a new concept: the morpheme (Beard, 1995). He placed roots and affixes into this concept. Morphemes are the smallest linguistic elements that carry meaning, and the phonological realizations of the morphemes are called morphs. For example, the plural morpheme *-lAr* in Turkish has *-lar* and *-ler* morphs depending on Turkish vowel harmony in which allomorphs are created according to the roundness and backness of the preceding vowel. A root in a word formation constitutes the core meaning of the word. The operations involved in combining roots and affixes are presented in the Section 2.1 in relation to both concatenative and non-concatenative languages.

Saussure avoided the terminology, ‘morpheme’. He used *signifier* and *signified* and morphology was defined as phonological alternations of a *signifier*. Zero morphs are a contradiction to the Saussurean view because they indicate different concepts by preserving homonymy. In other words, zero morphs modify the meaning of a word without making any change in the surface form of this word. Bloomfield (1933) was the first person to place all morphemes in the lexicon, which had formerly considered being a storage component of words. He assumed a single grammatical function for each morpheme. Yet, morphological asymmetry, such as in Russian inflection, and portmanteau morphemes introduced a problem: the same morpheme could mark more than one grammatical function and one grammatical function can be marked by more than one morpheme. For example, *-s* in English can be used to mark *singular*, *present*, *3rd person* and *indicative* at the same time. Similarly, singular feminine nominal marking in Russian can be achieved by *-a*, *-ø*, or *-o*. This shift in lexical morpheme hypothesis (Beard, 1995) has led to the study of form and function in morphology. A language learner lexicalizes morphemic lexemes together with its functions. These functions can be deduced from semantic, syntactic and pragmatic analyses and the learning of corresponding language through linguistic experience. However, the forms must, initially, be identified by the listener, which is claimed to occur via statistical learning in this study.

There are languages without or with little morphology, such as Vietnamese and Mandarin, and languages with complex morphologies, such as K$^w$ak$^w$'ala, Hebrew, Tagalog, Hungarian and Turkish. Isolating languages lack inflection and systematic word derivation processes. On the other hand, polysynthetic or agglutinating languages can express in a single word what, in English, would be a sentence containing numerous words. The speakers of such languages must also learn a huge set of rules for morphological composition, since the number of forms that can be built from a small set of lexical stems can run into the millions (Hankamer, 1989). The striking differences and diversity in the morphologies of polysynthetic and isolating languages are mirrored by differences in grammatical organization extending to the deepest levels of how meaning is organized (Evans & Levinson, 2009)

Morphology is an interface problem; i.e., external. The expressiveness of a language can be achieved syntactically or morphologically. Although this language-specific orientation historically results from language evolution, morphology exists. Synchronic linguists concerned with the universals of natural language acquisition neglect morphology. However, there are some non-universal constraints in languages across the world indicating that the claims of a Universal Grammar are empirically false, misleading or non-disprovable (for a detailed discussion see Evans & Levinson, 2009). Perhaps the reduction of morphology to syntax might be an erroneous option to jettison the divergence but reducing morphology to a lexicon better mitigates the discussions about the status of morphology. From a lexicalist perspective, all morphemes reside in a lexicon reducing the problem to the interaction of syntax and lexicon. Due to this divergence, Aronoff (1993), and Aronoff and Fudeman (2011) state that linguists must consider *morphology by itself*, not merely as an appendage of syntax and phonology, and that linguistic theory must allow for a separate and autonomous morphological component to study the faculty of language.

Besides understanding the modularity and the processing of the human language faculty, studies concerning linguistic morphology also have technological implications. For example, in the field of machine translation, lexical gaps and the translations of phrases are problematic not syntactically but morphologically. In Turkish, the subject and the object of a verb are not determined by the word order but by the morphemes as in (1) in Chapter 1.

Mathews (1991) and Anderson (1992) argue for a separation of inflectional affixation from the grammatical feature inventory of lexical items, such as Number, Tense, Gender and Case features. The proponents of the Split Morphology Hypothesis (Anderson, 1992; Matthews, 1991; Perlmutter, 1988) state that affixation is a result of operations on roots, rather than listed items. They also consider that derivational morphology is too irregular to be combined with inflectional morphology. In the current study, inflectional and derivational morphemes are treated as the same element. The rationale for such a treatment is explained statistically in Section 2.3.4.

In level ordered morphology (Chomsky & Halle, 1968; Inkelas, 1993; Kiparsky, 1982b), morpheme and word boundaries are marked and different morpheme classes are introduced. Each class is allowed to operate within specific boundaries. This is also known as the item-and-arrangement approach in which morphology is considered as the arrangement of morphemes into a specific order. For example, the word *kediler* 'cats' is produced from the concatenation of two morphemes *kedi* 'cat' and *-ler* PLU whose positions of occurrences are predetermined. The other approach is the item-and-process, or word-and-paradigm morphology (Anderson, 1992; Matthews, 1991). In this approach, complex words are produced via the process acting on simpler words. For example, *kediler* ‘cats’ is retrieved if ‘make plural’ process is executed on *kedi* ‘cat’.

In his influential work, Amorphous Morphology, Anderson (1992) proposes that complex words are not incrementally built by concatenating morphemes whereas word structures are described by rule-governed relations among words. His famous examples are given in K$^{w}$ak$^{w}$’ala language. In this language, every sentence is verb-initial and some inflectional morphemes of noun phrases (NP) are not attached to constituents of the phrase but to the verb as in (4) (taken from Anderson, 1992).

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| (4) | *nanaqəsil-ida* | *iʔgəl’wat-i* | *əliwinux$^{w}$a-s-is* | *mestuwi* | *la-xa* |
|  | Guides-SBJ/ART | expert-DEM | hunter-INST-his | harpoon | PRE-OBJ/ART |

*mig$^{w}$at-i*
seal-DEM
An expert hunter guides the seal with his harpoon.

This is a quite striking example of morphology because the inflectional markers for case, possessor and deictic status of every NP are not within NP but on the preceding element. Anderson (1992) underlines that these strangely placed morphemes are definitely not for agreement but are grammatically subcomponents of the NP. Such peculiar formations require that phonological words are not actually the domain of morphology. Thus, morphology should not be about morpheme processing but about word processing.

Although Anderson buttresses a lexicalist and postsyntactic view, he rejects morphemic lexicons and advocates that morphology involves relations between forms, not simply the concatenation of primitive units of sound and meaning. He exemplifies his claim by presenting problems in the analyses of morphological forms and interactions of morphology with other linguistic domains. In this chapter, initially the operations in morphology and affix ordering are discussed, then, Morphology and its interactions are reviewed in Section 2.3 with the final section being concerned with the acquisition of morphology.

## 2.1 Operations in Linguistic Morphology

Although linguists may argue in support of other definitions of morphology, they mostly agree that morphology is the study of meaningful part of words: morphemes (McCarthy, 1991). It is easy to detect the morphemes in *haber-ler-de* ‘news-PLU-LOC’ but not in *men.* The English word *men* is a plural noun but the plural morpheme is in the vowel *-e-* as opposed to *-a-* in singular *man*. Therefore, the morpheme PLU is realized in various forms. The morpheme that gives the main meaning of the word is the *stem* or *root*. *Haber* ‘news’ is the root of the *haberlerde* ‘in the news’ and it is a free morpheme while *-ler* and *-de* are suffixes and bound morphemes. Yet, the *stem* and *affix* relation is not always easily captured. For example, the root of the word *nominee* is not *nomin* because the word *nomin* does not occur as a free form in English because it is the stem of Latin word *nomen* ‘a name’. Instead it is derived from the word *nominate* via truncation and suffixation processes.

Morphemes that precede the stem are prefixes while those that follow the stem being called suffixes. There are also circumfix morphemes whose first portion acts as a prefix and the second as a suffix. Kiraz (2001) gives the example of the Syriac word *neqtlun* ‘kill them’. The word *qtl* is a stem pattern for the verb “to kill” and *ne-un* is the circumfix for 3.PLU. MAS. The word formation rules in languages such as Hebrew, Syriac and Arabic usually follow a template style. For example, the Hebrew root *rkd* “dance” obeys a pattern *CaCCan* to produce *rakdan* “dancer” while the root *spr* ‘cut’ follows a template *miCCaCa* to form *mispara* “barbershop”. Indeed, there are some exceptions in such template approaches and infant speakers of such languages experience an overgeneralization of the templates on the exceptions.

A learner of a language has to identify morphological operations in their native language in order to comprehend and effectively use possible word formations. In the next section, the main morphological operations concerning concatenation and morphophonemic processes will be reviewed. The morphophonemic processes include addition, zero morpheme, epenthesis, vowel harmony, voicing, ablaut, umlaut and reduplication.

### 2.1.1 Concatenation

Concatenation occurs in compounding and affixation. At a very primitive level, it simply unifies two or more strings. For example, *fildişi* ‘ivory tusk’ is produced by concatenating *fil* ‘elephant’ *diş* ‘tooth’ and *-i* ‘CM-Compound marker’. Distinguishing single word compounds from single stem words and segmenting the compounds into their constituents are also topics of morphology. The recognition and segmentation of single compound words statistically in this study is modeled in Chapter 4.

Affixation has four types of morphological operation: suffixation, prefixation, infixation and circumfixation. The first three types can be deduced from their names. Suffixation is a process in which morphemes are concatenated at the end of roots as in (5).

| | | |
|---|---|---|
| (5) | *ayna-lar-a* | *mean-ing-ful-ly* |
| | Mirror-PLU-DAT | Mean-PROG-DER-DER |
| | to mirrors | |

Prefixation, on the contrary, is concatenation of morphemes at the beginning of roots as shown in (6).

| | | |
|---|---|---|
| (6) | *bi-haber* | *dis-locate* |
| | DER-News | DER-Locate |
| | unaware | |

Similarly, some morphemes are placed away from root boundaries and somewhere inside roots for infixation as in Tagalog in (7) (taken from McCarthy & Prince, 1993). Infixation does not occur in Turkish.

| | | | | |
|---|---|---|---|---|
| (7) | *sulat* | *s-**um**-ulat* | *asna* | *as-**ka**-na* |
| | teaching | to teach | clothes | his clothes |

Circumfixation is actually a hybrid process of suffixation and prefixation simultaneously as in Indonesian (8) (taken from Conner, 2003). In most cases of circumfixations, both prefix and suffix particles are independently attested but usually with different meanings and functions (Spencer, 2001). Circumfixation is a less common type of concatenation and it does not exist in Turkish.

| | | | | |
|---|---|---|---|---|
| (8) | *pátut* | *mem-(p)atút-kan* | *hántu* | *meng-hantú-i* |
| | proper | ACT-proper-CAUS | ghost | ACT-ghost-LOC |
| | | to correct | | to frighten/haunt |

Templatic languages, such as Arabic, Hebrew and Syriac, have words composed of consonant roots. The concatenation of morphemes usually occurs through the placing literals of the morphemes within the trilateral roots. Examples in Arabic and Hebrew are given in (9).

| | | | |
|---|---|---|---|
| (9) | Hebrew Pluralization: | *e**r**e**c*** ‘land’ | *a**r**a**c**ot* ‘lands’ |
| | | ***zimr**a* ‘melody’ | ***zimr**ot* ‘melodies’ |
| | Arabic Pluralization: | ***q**a**lb*** ‘hearth’ | ***q**u**l**u**b*** ‘hearths’ |
| | | ***k**i**t**a**b*** ‘book’ | ***k**u**t**u**b*** ‘books’ |

The boldface literals in (8) indicate templatic roots which do not occur as free forms but with vowels. Concatenation in templatic languages is generally quite systematic. For example, singular roots in forms of *CaCC* and *CiCaC* become plural as *CuCC* and *CuCuC* respectively in Arabic. Templatic concatenation differs from apophony in that it is quite regular and systematic. It may also collaborate with prefixation, suffixation, addition, epenthesis and deletion as in linearly concatenative languages.

### 2.1.2 Morphophonemic Operations

Affixes are often accompanied by morphophonemic processes. The most common case is *apophony* (ablaut and umlaut). In *ablaut*, the vowels in roots are altered in order to indicate grammatical changes as in (10). It is generally an Indo-European language process.

| | | | | |
|---|---|---|---|---|
| (10) | Tense: | *s**i**ng* | *s**a**ng* | *s**u**ng* |
| | Germanic Plural: | *d**o**m* ‘field’ | *d**u**m* ‘fields’ | |

The only vowel alternations in Turkish roots occur in first and second personal pronouns with a dative case as in (11). The regular dative case for these pronouns should have be **bene* and **sene*. They are exceptions with a historical basis in Turkish because there is no ablaut or umlaut in Turkish.

| | | | | | |
|---|---|---|---|---|---|
| (11) | *ben - A* → | *bana* | | *sen - A* → | *sana* |
| | I - DAT → | to me | | you - DAT → | to you |

Similarly, *umlaut* is a vowel change, too, but it is a Germanic effect in which *i* or *y* are degrades as in (12). Note that plural morpheme *-er* causes an apophony in the root *buch.*

| | | | |
|---|---|---|---|
| (12) | Tense: | *br**i**ng* | *br**ou**ght* |
| | Germanic Plural: | *buch* ‘book’ | *b**ü**ch**er*** ‘books’ |

In Turkish morphophonology, vowel harmony is quite important. While concatenating, morphs generally have to preserve the roundness and backness properties of the previously concatenated morph as in (13).

| | | | | | |
|---|---|---|---|---|---|
| (13) | *ev-ler.* | *araba-lar.* | *Sol-ü* | *çal-ma-yı* | *unut-tu-m.* |
| | house-PLU | car-PLU | Sol-ACC | play-SUB-ACC | forget-PAST-1.SG |
| | houses | cars | I forgot playing the G (note). | | |

An interesting application of vowel harmony is observed when a morph is attached to a root terminating with a post alveolar *l* instead of a regular dental alveolar *l* in Turkish. *Sol* in Turkish is pronounced with a post alveolar *l*. Thus, it makes the accusative case obey the harmony with itself instead of *-o-*. The expected form is *solu* not *solü.* The same case occurs with the word *hal-ler* ‘condition-PLU’ meaning conditions. These examples require a revision in the definition of Turkish vowel harmony.

Apophony in consonants is usually called consonant apophony or *C-mutation* in which the final consonants are altered as in (14).

| | | | | |
|---|---|---|---|---|
| (14) | *buil**d*** | *buil**t*** | *ağa**ç*** | *ağa**c**-ı* |
| | | | tree | tree-ACC |
| | | | | the tree |

Turkish consonant mutation is called *voicing*. If some of the strings terminating with the voiceless consonants, *p, t, k, ç,* are followed by the suffixes starting with vowels, then the consonants are voiced as *b, d, ğ, c* as in (15).

| | | | | |
|---|---|---|---|---|
| (15) | *sonu**ç*** | *sonu**c**-um* | *kana**t*** | *kanad-ı* |
| | result | result -1.SG.POSS | wing | wing-ACC |
| | | my result | | the wing |

*Consonant assimilation* is also important in Turkish morphophonology. The initial consonants of some morphemes undergo an assimilation operation if they are attached to the strings terminating in the voiceless consonants, *p*, *t*, *k*, *ç*, *f*, *s*, *ş*, *h*, *g*, as in the surface forms of the Turkish past tense *-DI* in (16).

| | | |
|---|---|---|
| (16) | *at-tı* | *konuş-tu* |
| | throw-PAST | speak-PAST |
| | threw | spoke |

Many languages make use of *tonal changes* to indicate grammatical categories or change in meaning. Spencer (2001) states that in DhoLuo language *ì* (decreasing tone) if for 2.SG imperfective while *í* (increasing tone) indicates 2.SG perfective. The classical tonal alternation changing word meaning is observed in Mandarin Chinese. *Mā* (stable tone) means *mother* while *mǎ* (decreasing and then increasing tone) is *horse* in Mandarin.

*Stress* assignment is also very common morphophonemic operation to mark lexemes. For example, *contrást* (verb) and *cóntrast* (noun) are different lexemes in English. Stress assignment can also be observed cross-linguistically in compounding. In English and Turkish, the majority of nominal compounds have two stresses and they are stressed more in the leftmost constituent. The stress change of each constituent in compounding helps the listeners to perceive the distinct constituents as a compound word, a single linguistic entity. There are some variations of the stress changes, such as, *köpékbalığı* 'shark' consisting of *köpek* 'dog' and '*balık*-CM' 'fish' which has the second stress in its second constituent.

*Metathesis* is reordering of phonemes. Turkish children often reorder *yumurta* 'egg' and *mutfak* 'kitchen' as **yuturma* and **muftak*. An example in Saanich is given in (17) (taken from Stonham, 1994).

| | | |
|---|---|---|
| (17) | *t'sə* | *t'əs* |
| | break something | break something |
| | (imperfective) | (perfective) |

*Deletion* or *subtraction* is another morphophonemic operation in which instead of adding or changing some portion of roots or affixes, they are deleted as given in (18).

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| (18) | *kar**ı**n* | - *ım* → | *karnım* | *vak**i**t* | - *i* | → *vakti* |
|  | abdomen | 1.SG.POSS | my abdomen | time | - ACC | the time |
|  | *de* | - *yor* → | *diyor* | *ye* | - *yor* | → *yiyor* |
|  | say | - PROG | saying | eat | - PROG | eating |

It is interesting that the accusative case marker *-I* agrees with the deleted vowel *i* in *vakit* to form *vakti*. It can be proposed that deletion occurs after concatenation and it is a postlexical operation.

*Epenthesis* occurs when one or more sounds interfere with concatenating morphemes. In Turkish, epenthesis usually occurs in loaned and monosyllabic words as in (19)

|  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|
| (19) | *hak* | - *ım* → | *hak**k**ım* | *af* - | *et* | → *a**f**fet* |
|  | right | 1.SG.POSS | my right | forgiveness | do/make | forgive |

The brief morphological operations and examples given above show that morphology is definitely intertwined with phonology, syntax and semantics. Before presenting the interactions of morphology in the following sections, first the modeling affix order is reviewed.

### 2.2 Ordering of Affixes

To express a larger set of meanings morphology possesses affixes which are a set of meaningful morphemes. The ordering of affixes is fairly strict in many languages and variations in the order results in drastic changes of meaning. The affix system of a language has a finite set of elements (morphemes and the operations such as emphatic reduplication) and a finite set of possible combinations. Manova and Aronoff (2010) state that affix ordering could be either *motivated* (rule-governed) or *unmotivated* (rote-learned):

- *Motivated* affix order obeys either
    - *Grammatical* principles reflecting the organization of grammar
        - *Formal grammatical* principles: phonological, morphological and syntactic principles.
        - *Semantic* principles.
    - *Extra-grammatical* principles
        - *Statistical*: there is a particular order because it prevails in all languages.
        - *Psycholinguistic*: related to the human way of processing and producing affix combinations.
        - *Cognitive*: related to the cognitive characterization of the world.
        - *Pragmatic*: the speech-act context affects the ordering.
        - *Other Principles*: Temporal, psychological, evolutionary, and such.
- *Unmotivated* affix ordering is inexplicable and it is in the following types:
    - *Templatic*: Inexplicable but ordered. It is only related to form.
    - *Arbitrary*: There is no affix ordering system.

Manova (2010) showed that even non-segmental morphological rules and subtractive formations, both among the crucial arguments of a-morphous morphology (Anderson, 1992), operated like the segmental affixations. She exemplified that addition, substitution, modification (such as emphatic reduplication and epenthesis), conversion (like zero derivation), and subtraction were all segmental affixations. Rice (2000) distinguished template and layered morphologies as follows:

- Zero morphemes are prevalent in template morphology but not in layered morphology.
- Layered morphology gives rise to a headed structure but template morphology does not.
- Layered morphology is constrained by some principle of adjacency, but template morphology is not.
- Layered morphology does not permit an inner morpheme to be effective in the selection of outer morphemes but template morphology can work in this way.

Layered morphology is semantically governed whereas the template one is form-dependent. A language can fall into either category and it will still have an affix ordering mechanism. This mechanism can be form-related or meaning-related. It can be driven by both grammatical and extra grammatical constraints. Whether templatic or layered, affix ordering is hierarchical. Although layered morphology and template morphology require different treatments in terms of morpheme segmentation and acquisition, a particular language does not have to utilize either of the two types of morphology or can benefit from both types, such as in the Athapascan language (Rice, 2000). Actually, inflectional morphology in general is both semantically and form governed because the slots where an inflectional morpheme is to be attached are predetermined and the selection of morphemes per slot is semantically organized at the same time. Rice (2009) stresses that syntactic affix ordering should be discussed in relation to semantic ordering because changing the inflectional affix order alters word meaning as in (20) (taken from Lewis, 2000) and (21) (taken from Mithun, 1999).

| | | |
|---|---|---|
| (20) | *Türk-ler-dir* | *Türk-tür-ler* |
| | Turkish-PLU-GM | Turkish-GM-PLU |
| | They are the Turks | They are Turkish |
| (21) | *yup-pag-cuar* | *yug-cuar-pag* |
| | person-big-little | person-little-big |
| | little giant | big midget |

The examples above show that the semantics of a composite word is related to relevance and scope. The scope of a morpheme to be attached to a word is all the morphemes previously utilized in the formation of that word. Bybee (1985) suggested that a meaning element was relevant to another meaning element if the semantic content of the first directly affected or modified the content of the latter. The ordering of affixes in combinatorial morphology is either encoded in the last affix of the base or in the base itself (Giegerich, 1999; Plag, 1996; Plag, 1999). This means that affix ordering is a step-by-step derivation. The closing suffixes (Aronoff & Fuhrhop, 2002; Manova, 2008) are examples for the effects of the most recently attached morpheme in affix ordering. The closing suffixes may attach to several morphemes and bases but they close the word to further derivation or inflection.

At first glance, some affix ordering might display a recursive processing. The Turkish suffix *-ki* is a pronominal relativizer that can only be attached to words with locative, genitive or temporal aspects. It also unlocks a word which has been closed for further suffixation. Although theoretically there is no upper bound in the number of *-ki* suffixes in a word, there are usually at most two occurrences of *-ki* in Turkish as in (22).

| | | |
|---|---|---|
| (22) | *ev*<br>house | *ev-de-ki*<br>house-LOC-REL.ki<br>the one in the house |
| | *ev-de-ki-ler*<br>house-LOC-REL.ki-PLU<br>the ones in the house | *ev-de-ki-ler-in-ki*<br>house-LOC-REL.ki-PLU-GEN-REL.ki<br>the one belongs to the ones in the house |

Another recursive production example that results in comprehension difficulties in English is provided by Plag and Baayen (2009) in (23).

(23) We must be fearless.
We must have fearlessness.
We must not be fearlessnessless.
We must not have fearlessnesslessness.
We must be fearlessnesslessnessless.

Thus, it is fair to assume that morphological processing is not recursive in principle. Indeed, there are some semi-recursive formations but they have upper bounds and less understandability. These formations still have to obey the affix ordering principles of the target language.

Greenberg (1963) proposed a universal constraint that if both the derivation and inflection followed the root, or they both preceded the root, the derivation was always between the root and the inflection. Yet, this is no longer a valid statement because statistically motivated affix ordering studies provide abundant information and counter examples on the topic as shown in (24). There is a tendency for derivational affixes to be closer to the root however, this not an obligation where the derivational suffix *-lik* is preceded by four inflectional suffixes. This example is taken from METU-Sabancı Turkish Treebank (Atalay et al., 2003).

(24) *anla-ş-ıl-abil-ir-**lik***
understand-REC-PASS-PSB-AOR-**DER**
understandability

Similar statistical analyses have shown that inflectional morphemes do not always close words to further derivation. Manova and Aronoff (2010) state that although statistical studies are quite successful in modeling affix ordering, a speaker can neither compare languages nor count the forms in a corpus. However, the statistical studies do not propose that speakers count forms or they immediately perform conditional probabilities to discover morphemes and such like. Instead, these studies stress that speakers are aware of the word and sub-word frequencies of their languages and this awareness is effective in developing their linguistic skills, learning and decision making. It is not grounding but modeling. Further discussions on this issue can be found in Chapter 5.

Dressler (1989), Dressler et al., (2009) and Manova (2005) used cognitive concepts, such as prototypes, to explain affix ordering. They offered prototypes for inflectional and derivational affixes according to their semantic impact on word meaning. Manova (2008) further argued that nouns, adjectives and verbs had a cognitive nature relevant to suffix order. For example, the derivational suffix *-Im* in Turkish can only be attached to verbs to create nouns as in (25).

| | | |
|---|---|---|
| (25) | *uy-um* | *yaz-ım* |
| | comply-DER | write-DER |
| | compliance | spelling |

Although Bickel et al., (2007) presented data from the Chintang language in which free permutations of prefixes were allowed to a certain degree, Manova and Aronoff (2009) commented that if a language was a system, completely arbitrary affix ordering was not possible. In this study, it is assumed that affix ordering is motivated and the statistical principles dominate the other extra-grammatical principles.

### 2.3 Morphology and Its Interfaces

An interface is encountered when there is a point of contact between two domains or there is a boundary phenomenon. It is a modern linguistic tendency to partition language phenomena into distinct description domains: phonetics, phonology, morphology, semantics, syntax, pragmatics and such. Although each of these domains is assumed to be discrete, they are constituents of a language system, and they have to interact to convey achieve the ultimate goal: communication and thinking.

Linguistic structures have meanings associated with forms as in Saussure mapping given in Figure 3 (taken from Déchaine, 2005).

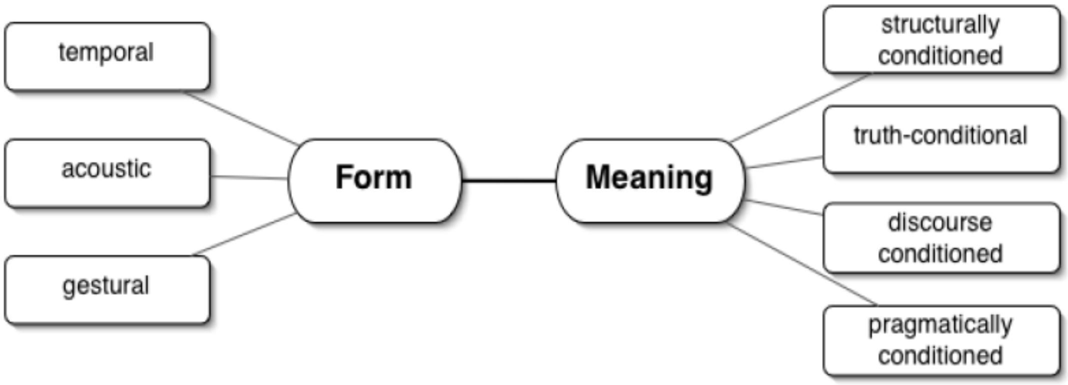


**Figure 3** Form-meaning mapping

Since morphology is fundamentally concerned with word formation and words are the main benchmarks for other linguistic domains, it is important to examine how morphology interacts with the other domains. In this section, morphology and its interactions are reviewed mainly in relation to the following questions related to the domains:

- Morphology and phonology (morphophonology): How do morphological formations affect the implementation of phonological rules?
- Morphology and syntax (morphosyntax): Do word internal structures reflect a syntactic process? Should derivational and inflectional morphological interactions with syntax be considered separately?
- Morphology and semantics (morphosemantics): Does morphological composition correspond to semantic composition?
- Morphology and Lexicon (morphemic lexicon): Does morphology operate on the lexicon?

### 2.3.1 Morphology and Phonology

There are two mainstream models of morphophonology: Lexical (Kiparsky, 1982b) and Prosodic Phonology (McCarthy & Prince, 1990). In lexical phonology, both morphological and phonological rules apply in the lexicon. Phonological rules can be lexical and postlexical. The lexical rules interact with the morphology in the lexicon while the postlexical ones occur once the syntactic rules have been satisfied. In other words, lexical rules are utilized at word level but postlexical rules apply to larger constituents than words.

Prosodic morphology emerges mainly for reduplication and templatic morphology. The roots and affixes form a labeled bracketing which heavily interacts with the foot, syllable and mora. Lappe (2007) focused on the truncation and clipping processes in English through prosodic morphology. These operations were assumed to be unpredictable but Lappe proposed a model derived from the framework constructed by Orgun and Sprouse (1999) to disprove this assumption.

Carstair-McCarthy (2001) indicates that phonology has a radical influence on morphological material because some morphological processes (such as affixation, reduplication) are restricted to bases with certain phonological characteristics, and they cannot be applied to bases without those characteristic even if the syntactic, morphological and semantic constraints are satisfied. This influence can be found in both derivation and inflection. For example, the English derivational suffix *-al* is restricted to bases with main stress on the final syllable with an exception in *burial* (Siegel, 1979). Similarly, English comparative and superlative suffixes *-er* and *-est* are allowed in short adjectives. In Turkish, consonant voicing in concatenation is usually not allowed when the root is monosyllabic.

Morphological operations exemplified in Section 2.1 show that hierarchical operations require local phonological modifications. For example, each morpheme concatenated to a word obeys Turkish vowel harmony stipulated by the previously attached morpheme. Even zero morphemes may require a tonal change or stress assignment in the surface forms. In order to further explore the topic, it is important to understand the terms *onset*, *nucleus* and *coda*. The onset of a syllable is made up of the first consonant or consonants. The nucleus is a simple diphthong vowel. The coda is the consonant(s) following the nuclei. All syllables must have a nuclei but the remainder is optional.

Morphs are surface forms of morphemes and there might be more than one surface form for a morpheme, these are called allomorphs. Allomorphs usually result from the phonological constraints of a language such as harmony, assimilation, voicing and epenthesis. Phonemes in morphemes must sometimes agree with the remaining the constituents on place, continuancy, or harmony. This agreement can be progressive (i.e., newcomers) or regressive (i.e., hosts). Sometimes extra phonemes that do not convey any morphological, semantical or grammatical information may occur between morphemes. For example, vowel-to-vowel contacts are usually not allowed in Turkish and French. Aronoff and Fudeman (2011) call this unwanted contact, *hiatus*.

However, there are many exceptions to the phonological constraints. For example, the Turkish progressive marker *-yor* never undergoes vowel harmony but it progressively forces the newcomer morphemes to obey the harmony. There is historical explanation for this phenomenon. Similarly, some of the monosyllabic roots assumed to avoid voicing; in fact some of them undergo voicing to prevent synonymy as in (26).

| | | | | | | |
|---|---|---|---|---|---|---|
| (26) | *kap* | - | *ı* | → | *kabı,* | **kapı* |
| | bowl | - | ACC | → | the bowl, | door |
| | *art* | - | *ı* | → | *ardı,* | **artı* |
| | back | - | ACC | → | the back, | plus |

The Optimality Theory (Prince & Smolensky, 1993) considers phonology as a universal set of constraints which are hierarchically ranked on a language-specific basis. It accounts for morphophonological operations by evaluating the phonological constraint(s) to simulate a morphophonological phenomenon. The Optimality Theory provides the winner constraint among the set of constraints but it cannot explain the morphophonological phenomenon in which the other constraints are also acceptable. Turkish emphatic reduplication is an example for which the Optimality Theory needs to employ lexical frequency to explain the phenomenon because in Turkish emphatic reduplications, alternatives do exist. Turkish emphatic reduplication, which is a doubling operation, needs further discussed. Turkish duplication can be seen to occur in the following three ways; m-reduplication, doubling and emphatic reduplication.

In Turkish a word or compound that undergoes m-reduplication and immediately follows its original form expresses a *broader meaning* than its simple form. If a word or compound to be m-reduplicated starts with a vowel, the original word is prefixed with *m-* and then duplicated as shown in (27a) below. If a word or a compound starts with a consonant other than *m-*, the consonant is replaced with *m-* then the new form is duplicated as shown in (27b). If the word or the compound already starts with *m-*, it is followed by the word *falan* meaning ‘and such like or so and so’. Although such constructions are considered to be informal, they are perfectly valid and common in colloquial usages of Turkish. The original form generally precedes the duplicated form.

(27) a. [*Çocuklar*]$_{\text{NP}}$ [[*akıcı makıcı*]$_{\text{ADV}}$ [ *konuşmazlar*]$_{\text{V}}$ ]$_{\text{VP}}$
Children do not speak fluently (and the like)
b. [*Çocuklar mocuklar*]$_{\text{NP}}$ [[*akıcı*]$_{\text{ADV}}$ [ *konuşmazlar*]$_{\text{V}}$ ]$_{\text{VP}}$
Children (and anyone) do not speak fluently

The doubling process *intensifies the meaning* of the duplicated words that are usually nouns, adjectives and adverbs. Similar to the m-reduplication, the doubled word succeeds the original form. The doubling occurs in two ways: simple doubling and doubling in lexical formations.

The *Simple doubling* (SD) process produces an exact copy of the word or the compound, and then places it immediately after the original form as in (28).

| | | |
|---|---|---|
| (28) | *tek tek* | *zaman zaman* |
| | one by one | time to time |

The m-reduplication (M-RED) and SD occur also in compounds and phrases as in (29a), (29b), and (29c). This indicates that M-RED and SD are post lexical formations.

(29) a. *duvar saat-i* *muvar saat-i* *al-ma*
wall clock-ACC M-RED buy-NEG
Do not buy a wall clock (or any sort of clock)

b. *yemek zaman-ı* *memek zaman-ı ara-ma*
eating time-ACC M-RED call-NEG
Do not call him at meal time (or any similar time)

c. *yemek zaman-ı* *yemek zaman-ı ara-ma*
eating time-ACC SD call-NEG
Do not call him at meal time (or any similar time)

Many idioms and phrases are effectively produced by the duplication of their first constituents. In this case, some additional morphemes, such as the plural suffix and the question particle (QP), are attached to the daughter constituents as in (30a) and (30b), or one of the daughter constituents undergoes some phonetic changes as in (30c).

(30)

a. *güzel-ler* *güzel-i* *bir kız*
beautiful-PLU beautiful-ACC a girl
a very beautiful girl

*ucuz-lar* *ucuz-u* *bir araba*
cheap-PLU cheap-ACC a car
a very cheap car

b. *güzel* *mi güzel* *bir kız*
beautiful QP beautiful a girl
a very beautiful girl

*ucuz* *mu ucuz* *bir araba*
cheap QP cheap a car
a very cheap car

*c. ufak* *tefek* *bir kutu*
little $\Phi_i$(little) a box
a tiny box

*çoluk* *çocuk* *duy-du*
$\Phi_j$(child) child hear-PAST
all the children (and families) heard it

The phonetically changed forms, such as *tefek* and *çoluk* in (30c), generally do not occur independently from the bases they are derived from, i.e., *ufak* 'little' and *çocuk* 'child' respectively. Similarly, although the question particle in Turkish is orthographically represented as a separate item as in (30b), it always follows a word and obeys the vowel harmony constraint with the word it follows, as in (31) Furthermore, if the question particle follows a verb, the agreement and intervening tense suffixes are attached to the question particle instead of the verb (Aygen, 2007; Kornfilt, 1996) as in (31a)

(31) a. *uyu-yor* *mu-ydu-nuz?*
sleep-PROG QP-PAST.COP-2.PLU
Were you sleeping?

b. kedi *mi?*
cat QP-3.SG
Is it a cat?

The question particles in (30b) are not necessary for the syntax but added for morphological reasons in the reduplicated phrase. Since the QP obeys the vowel harmony and morphotactic constraints, it acts as a morph-attached to the first word.

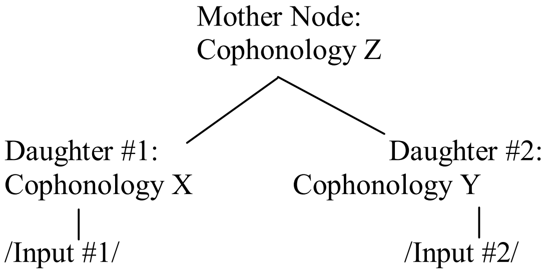
A further type of lexical doubling formation is the aorist verb doubling. The aorist verb and its negated duplicated aorist form constitute the meaning 'as soon as the verb occurs' as in (32).

(32) *ye-r* *ye-mez* *ilac-ın-ı* *al.*
eat-AOR eat-NEG.AOR medicine-2.POS-ACC take
As soon as you have eaten, take your medicine

*gör-ür* *gör-mez* *o-na* *sarıl-dı-m.*
see-AOR see-NEG.AOR he/she-DAT hug-PAST-1.SG
As soon as I saw him/her, I hugged him/her

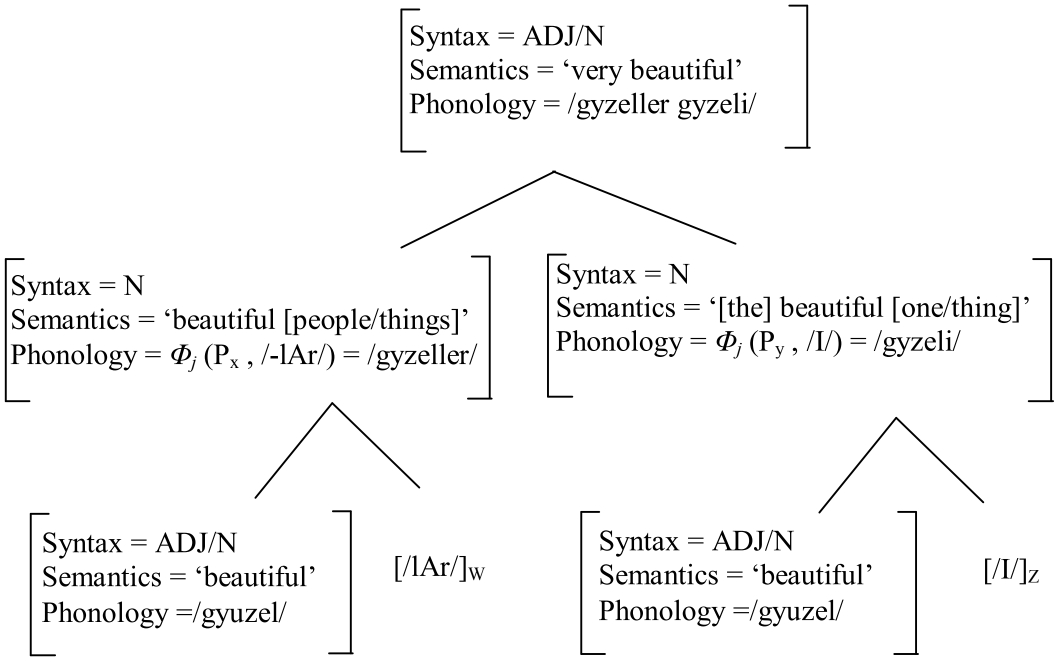
$\Phi_j$ and $\Phi_i$ in (34) below are cophonologies (Inkelas & Zoll, 2005; Orgun, 1996; Orgun, 1999), which are the morphological functions associated with particular morphological constructions to model morphologically conditioned phonology. They take words or morphemes as input, and perform operations such as constraint ranking, truncation, and velar deletion on the input to be sent to the phonology interface (Inkelas, 2000; Inkelas & Orgun, 1995) as in (33).

(33)



Inkelas and Zoll (2005) employ cophonologies in their Morphological Doubling Theory (MDT) and argue that this theory is morphologically motivated because it makes use of roots, morphs or affixes instead of mora, coda or foot. The model works in a binary manner, in which there are two inputs called daughter nodes, and the output in the root of the tree is called the mother node. In MDT, the reduplicant and base are both generated by the morphology as part of a construction that also embodies semantic and phonological generalizations concerning the output of reduplication (Inkelas, 2005). Inkelas and Zoll (2005) also state that Turkish reduplication is morphophonemic through the cophonologies and their representation as employed in the examples given in (34) below.

(34) Example: *güzel* → *güzeller güzeli*


Syntax = ADJ/N
Semantics = 'very beautiful'
Phonology = /gyzeller gyzeli/
Syntax = N
Semantics = 'beautiful [people/things]'
Phonology = $\Phi_j$ ($P_x$ , /-lAr/) = /gyzeller/
Syntax = N
Semantics = '[the] beautiful [one/thing]'
Phonology = $\Phi_j$ ($P_y$ , /I/) = /gyzeli/
Syntax = ADJ/N
Semantics = 'beautiful'
Phonology =/gyuzel/
[/lAr/]$_W$
Syntax = ADJ/N
Semantics = 'beautiful'
Phonology =/gyuzel/
[/I/]$_Z$

It should be noted that adjectives in Turkish can also be used as nouns when the nouns in adjectival phrases are dropped. In this case, the inflections on the nouns are suffixed to the adjectives: *güzel kız-lar-ın* 'beautiful girl-PLU-GEN' → *güzel-ler-in* 'beautiful-PLU-GEN'. This is a lexical operation on phrasal formations rather than a derivational operation.

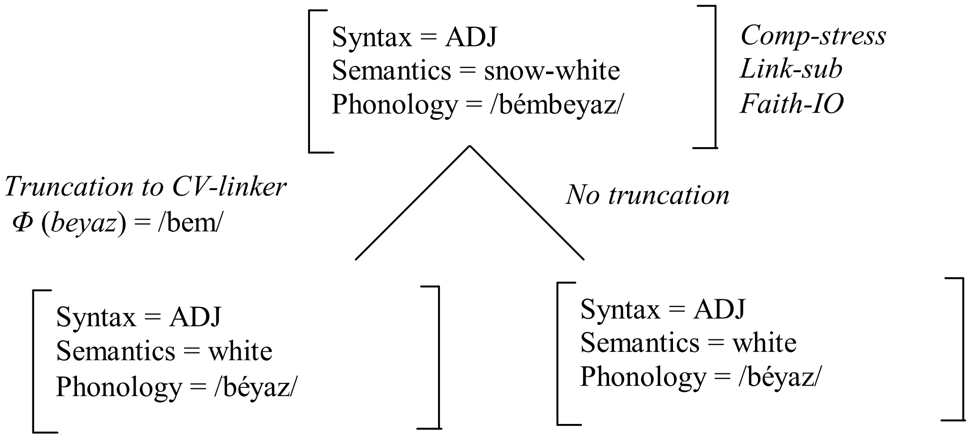
Turkish emphatic reduplication (TER) is used to accentuate the meaning of an adjective. It involves the duplication of the initial (C)V of the base then the addition of a prefix as a linker to the root, which is a consonant from the set "*-p, -s, -m, -r*" (Demircan, 1987; Dhillon, 2009; Kelepir, 2001; Kim, 2007; Wedel, 1999; Yu, 1998) as shown in the example in (35a) below. The output is a change in meaning. If the first letter of the base word is a vowel, then always *-p* is infixed as the linker. In some cases the (C)V-*linker* prefix is also followed by an additional infix from the set, *"-A, -Il, -Am*" as in (34b). Turkish emphatic reduplication is not a morpheme, but a morphological operation.

(35) $C_1V_1C_2\ldots$ → $C_1V_1$(*p, m, r, s*)(*A, Il, Am ,ε*) $C_1V_1C_2\ldots$
(*ε* denotes an empty string)

| | | | |
|---|---|---|---|
| a. | *be-m-beyaz*<br>TER white<br>snow-white | *ka-s-katı*<br>TER solid<br>hard as a rock | *te-r-temiz*<br>TER clean<br>very clean |
| b. | *ya-p-a-yalnız*<br>TER alone<br>all alone | *çı-r-ıl-çıplak*<br>TER *naked*<br>totally naked | *pa-r-am-parça*<br>TER piece<br>smashed to pieces |

The cophonologies of the MDT can operate in TER as well. Truncation and addition operations act on the word *beyaz* to produce *bem*. Then, the mother node links the subconstituent daughters faithful to the input and shifts stress to the truncated one to form */bémbeyaz/* as in (36).

(36) Example: *beyaz* → *bembeyaz*



Demircan (1987) and later, Wedel (1999; 2000) examined the Turkish E-RED as a phonological operation and summarized the linker selection constraints as:

1. The linker from the set {p, s, m, r} cannot be identical with the initial consonant ($C_1$) of the base: *pembe* 'pink' → **peppembe*, although *p* is in *{p, s, m, r}*. *Perpembe* is possible but not likely.
2. The linker cannot be identical to the second consonant ($C_2$) of the base: *pembe* → **pempembe/pespembe*, although *m* is in *{p, s, m, r}*.
3. The phonetic features {*coronal, sonorant, labial, continuant*} of the linker cannot be identical with those of the second segment of the base. The linker with the most contrasting features is selected for perceptual salience.
4. The linker is selected in a way that it can establish an optimization or balance among the features contributing to the featural contrast with respect to base.

The linker should be selected in such a way that it contributes features that can establish an optimization or balance among the featural contrast with respect to the base. In the MDT, the features given above can be ranked by the cophonologies to determine the linker of TER form. Yu (1998) argued that the allomorphy in Turkish reduplication could be accounted for by positing morphotactic constraints, which spell out the form of each of the allomorphs that dominate certain phonotactic constraints. The ultimate selection of the appropriate allomorph depends on satisfying the harmony rules of the lower-ranked phonotactic constraints of the linker. Demircan (1987) analyzed 121 emphatically reduplicated adjectives and concluded that the number of reduplicated adjectives shows the relationship; *-p* > *-m* > *-s* > *-r*. In another study, Wedel (1999; 2000) concluded that TER with the linker *-r* should be lexicalized in Turkish.

Despite these findings, there are constructions in everyday Turkish speech that ignore these constraints. For example, in addition to *çı-r-ıl-çıplak,* which is the expected reduplicated form of *çıplak* as in (34b), *çı-s-çıplak, çı-m-çıplak, çı-r-çıplak* and *çı-p-çıplak* can occur in informal settings[1] and they are acceptable. Furthermore, it can be seen that there is still regularity in these formations. The Optimality Theory cannot explain the situation where there is no winner but a ranking.

[1] Emphatically reduplicated forms of *çıplak* 'naked' do occur in web as *çı-s-çıplak, çı-r-çıplak, çı-m-çıplak* and *çı-p-çıplak,* other than *çı-r-ıl-çıplak.*

The phonological constraints of TER are derived from the set of adjectives that have already been reduplicated. Yet, it is possible to find a derived adjective that has never been emphatically reduplicated such *resim-siz kitap* ‘picture-DER book means a book without (a) picture(s)’. The acceptable TER formation of *resimsiz* should be constructed as in (37).

(37) *resimsiz* → *re* (*p, m, r, s*)(*A, Il, Am ,ε*) *resimsiz*

The ordering the selection rates of the linker type in Turkish emphatic reduplication were experimentally investigated and the detailed results are given in Chapter 4. The results indicate that Turkish speakers also make use of simple consonant co-occurrence frequencies to avoid synonymy and false-root deception in Turkish emphatic reduplication. In other words, the least frequent “*linker type-*$C_1$” co-occurrence is chosen to enhance communication because Turkish words usually have roots in the leftmost positions and a frequent “*linker type-*$C_1$” selection might deceive the listener as if there is a root instead of a prefixation process. This is an interesting finding because Turkish emphatic reduplication was thought to be driven by pure phonology but in fact it also makes use of lexical frequencies.

The speakers of a language do not have time to count co-occurrences or derive statistics but such results indicate that frequencies might explain phenomena such as emphatic reduplication. Next morphology-phonology interaction can be observed in the acceptability of nonce-words. These words are frequently employed in linguistic studies to evaluate areas such as well-formedness (Hammond, 2004), morphological productivity (Ansen & Aronoff, 1988) and development (Dąbrowska, 2006), judgment of semantic similarity (MacDonald & Ramscar, 2001), and vowel harmony (Pycha et al., 2003). Nonce words are also used to understand the process of adopting loan words. The majority of loaned words undergo certain phonetic changes to more closely resemble the lexical entries of the language into which they are to be adopted (Kawahara, 2012). For example, *television* in Turkish becomes *televizyon* /*televızjon*/ because /*jon*/ is more frequent than /*ʒɪn*/ in Turkish[2]. Similarly, *train* is adopted as *tren* /*tren*/ because, similar to diphthongs, vowel-to-vowel co-occurrences (*hiatus*) are not usually allowed in Turkish except some compound words. This phenomenon shows that the speakers of a language are aware of the possible sound frequencies and co-occurrences of their native languages, and they can make judgements on the naturalness of loan words, recently invented words and nonce words by using their knowledge of the existing Turkish lexis. Thus, the acceptability of nonce words can be a logical decision based on known-word statistics.

Previously, the acceptability of nonce words was investigated by experimental investigations through phonotactic properties or factor-based analysis (Albright, 2008). In the experimental investigations, it was observed that the participants accepted or rejected nonce words according to probable combinations of sounds (Albright, 2008; Hammond, 2004). In the factor-based analysis, the acceptability of nonce words was evaluated through the co-occurrences of syllables or consonant clusters locally (Hay et al., 2004) or non-locally (Coo & Callahan, 2011; Finley, 2012; Frisch & Zawaydeh, 2001) or through nucleus-coda combination probabilities (Treiman et al., 2000).

For this study, the acceptability of nonce words was assessed through a model using the conditional probabilities of the bigram co-occurrences of the orthographic representations locally and the pairwise co-occurrences of the vowels within the same word boundaries.

[2] In the METU-Turkish Corpus, there are 181 occurrences with the segment /*ʒɪn*/ of which only 30 are at the terminating word boundaries. On the other hand, there are 5,945 occurrences with the segment /*jon*/ of which 3,190 are at the terminating word boundaries, excluding the word *televizyon*.

Sliding the bigrams from left to right was chosen to mimic the effects of Turkish morphophonological changes, namely voicing, consonant assimilation, epenthesis, deletion and disallowance of hiatus. The co-occurrence probabilities of pairwise vowels were employed in the model to judge the effect of Turkish vowel harmony on the decision to accept a nonce-word. This study was undertaken to validate the cognitive plausibility of using conditional probabilities to simulate human level decision making. Similar methods within the context of phonotactic modeling were used for Finnish vowel harmony (Goldsmith & Riggle, 2012). However, in this study, the local bigram phonotactic modeling was used to evaluate Turkish nonce words. Two threshold values were set for the decision to reject, moderately accept and fully accept. The threshold values were computed according to the length of each input string.

For the evaluation of the conditional and co-occurrence probabilities, the METU-Turkish Corpus, containing about two million words, was employed (Say et al., 2002). The list of nonce words was created intuitively. The same list of nonce words evaluated by the method was also given to Turkish native speakers to judge the level of acceptability of each word. The results were relatively similar as explained in Chapter 4.

### 2.3.2 Morphology and Syntax

Traditionally, morphology is divided into inflectional and derivational domains. Leaving aside the necessity for such a division which will be presented in Section 2.3.4, definitions for these two domains need to be given. Derivational morphemes are assumed to create new lexical entries; thus, they more closely related to lexicon. On the other hand, inflectional morphemes are required by syntactic constraints, such as case, number and gender for nominal categories, and tense, aspect, mood, voice and agreement for verbal constraints.

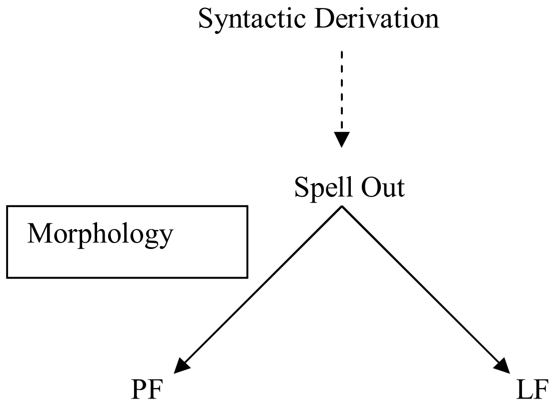
Generally, cases mark nominal forms for nominative, accusative, genitive, ablative, ergative and absolutive cases. Yet, some languages have more than 15 cases. Number indicates singular, plural, dual or trial aspects of nouns. The indication of gender varies greatly among languages. There are masculine, feminine, neutral, animate and inanimate genders but there is no universal way of determining the gender of a noun. It might depend on sex, phonetics and shape. In Russian, for example, verbs agree on gender of their subject only in the past tense. Arabic verbs agree with their subjects for number when the word order is Subject-Verb but not Verb-Subject. In German, gender is obligatory. In French, adjectives agree on gender and number of the nouns they modify.

Grammatical function change mainly occurs via morphology. For example, the suffix *-ed* in English is also for producing passive constructions. A verb is causative in Turkish if it is inflected with *-DIr*. Antipassivezation in ergative/absolutive languages, such as Greenlandic Eskimo, requires that the object of the verb is marked in an oblique case or becomes null. There are no universal constraints on morphosyntax but there is certainly a bidirectional interaction between the two. It can be seen, intuitively, that word order in phrasal or sentential structures (syntax) is related to word formations (morphology). Basically, a verb cannot be a subject of a sentence but a noun morphologically derived from the verb can be. Moreover, the cases relating objects and subjects to verbs, the genders determining the selection of adjectival forms, the agreements for grammatical cohesion also lie in the border of syntax and morphology for a plethora of languages.

In the 1960's through to 1970, disagreements on the nature of word formation led to the emergence of two trends in grammar, which are Generative Semantics and Lexicalism. The debate has continued eventually converging on the following questions about word structure: Is morphology an autonomous module? Is it subsumed under syntax? If it is an autonomous module, then the interaction between syntax and morphology needs to be explained. Although the generativists usually castigate the independence of morphology due to its extrinsic aspects, Border (2001) considers that the resolution of these questions is an empirical issue. On one hand there are morphological operations and constraints that cannot be reduced to syntactic conditions and on the other hand there are syntactically motivated operations resulting in rich word formations.

Lexicalists claim that words derived from a lexicon serve as the terminals in the syntactic derivation and they are special in a way that, for example, phrases are not (Embick & Noyer, 2007). The second line of research advocates that words are created by the rules of the syntax. This means that it requires the base elements of syntactic derivation, the principles of word assembling and the way of relating phonological forms to the assembled words. For example, Distributed Morphology (Halle & Marantz, 1993) proposes an architecture of grammar in which a single generative system is responsible for both word structure and phrase structure (Baker, 1988; Borer, 2004; Embick & Noyer, 2007; Pesetsky, 1995). In this non-lexicalist approach, the syntax consisting of a set of rules generative rules forms words by the syntactic operations, Merge and Move. Further operations are performed in the derivation of phonological form and logical form interface levels. Since morphology resides in phonological form interface, the morphological structure at phonological form level is simply the syntactic structure as in (38).

(38)

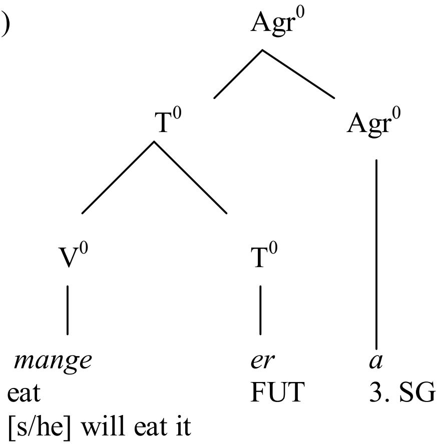


In this approach, the primitives of the syntax are an open-class of roots, such as √CAT, √OX and √FOOT, and a universal set of abstract morphemes, such as [PLU]. After the Move and Merge operations on these primitives, a hierarchically derived structure is, like √CAT.[PLU], √OX.[PLU] or √FOOT.[PLU] sent to spell out. The morphology in this model acts as the supplier of the phonological features including the zero element. This model also accepts the Separation Hypothesis (Beard, 1995), in which the components of the traditional morpheme do not contain syntax, semantics and phonology. The morphemes are rather underspecified according to the syntactico-semantic environment in which they are employed. The phonological form is imbued by the late insertion of a vocabulary item, such as [PLU] ↔ -en/{√OX ...}. The resulting phonological forms are derived according to the vocabulary item such as *cats, oxen* and *feet*. Embick and Noyer (2007) claim that the generation of all complex forms must be performed in the syntax because there is no lexicon in which complex objects can be assembled according to rules distinct from the syntactic rules. Cross-linguistic investigations easily show that the picture is far more complex. However, there are special cases called syncretism in which morphology lets down the syntax.

As Lieber (1992) stated, a simple theory of morphology will be one in which morphology is simulated as a part of syntax yet, so far, no one has succeeded in deriving all properties of words from the same principles of grammar. This is a similar approach to the Mirror

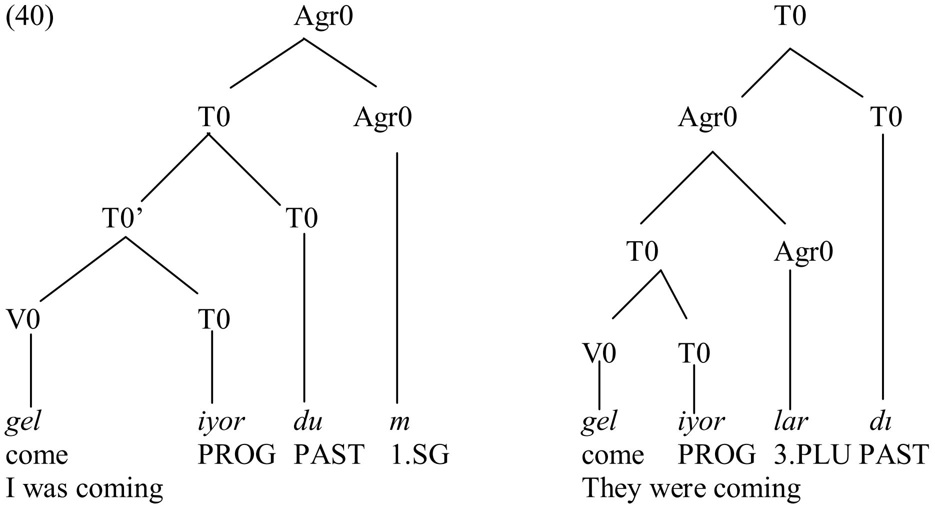
Principle (Baker, 1985) in which morphological derivations must directly reflect the syntactic derivations (and vice versa). If a morpheme denoting tense is closer to a stem than a morpheme representing person, then the syntactic node which dominates tense markers is lower in the tree (i.e., previously derived) than the person markers as shown in (39) (taken from Borer, 2001).

(39)



A counter example in Turkish is *geliyordum* and *geliyorlardı* in (40). Note that *?geliyordular* might be a practically acceptable formation but not **geliyorumdu*.

(40)



The examples in (40) differ only in the order of the tense and person morpheme which violates the assumption that morpheme order reflects the syntactic derivation order. After analyzing the verb, tense and aspects orders of 530 languages, Julien (2000; 2002) concluded that words were not produced by syntactic operations. Thus, a word is not a grammatical concept; instead, wordhood is a matter of distribution and it can be penetrated by a syntactic mechanism.

An example where the scope of an affix is not a word but a phrase is the suspended affixation phenomenon (Broadwell, 2008; Kabak, 2007). Suspended affixation is a counter example for the claim that morphology is only relevant for word formation as in (41).

| (41) | *çocuk* | *kitap,* | *silgi,* | *ve* | *kalem-i* | *unuttu.* |
|---|---|---|---|---|---|---|
| | child | book-NOM | eraser-NOM | and | pencil-ACC | forget-PAST |
| | child | [book, | eraser | and | pencil]-ACC | forget-PAST |

The child forgot the book, the eraser and the pencil.

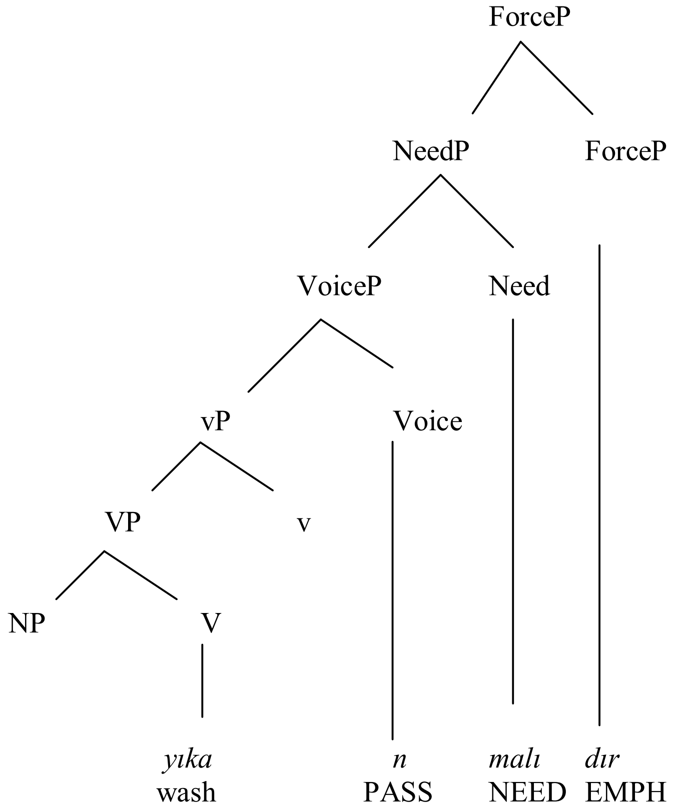
Kabak (2007) gives the example in (42) and states that the morphology limits suspended affixation. Broadwell (2008) claims that suspended affixation is created by syntax but filtered by morphology because not every affix can be used as a suspended affix.

| (42) | * [*Avşa-ya* | *git-ti* | *ve* | *deniz-e* | *gir-di*]-*y-di-k* |
|---|---|---|---|---|---|
| | Avşa-Dat | go-PAST | and | sea-DAT | enter-PAST.COP-PAST-1.PLU |

We went to Avşa and swam in the sea.

Broadwell (2008) further indicates that by providing a minimalist analysis tree as in (43), then nearly every morpheme in Turkish would head a separate phrase structure.

(43)



(It) should be washed

A tree of the sort given in (43) necessitates that every morpheme should license suspended affixation. In fact, only a few morphemes have this ability thus, an overgenerated minimalist tree should be stipulated by morphology. Göksel (2007) presents a similar conclusion about morphology and syntax interface by exemplifying that headless pronominal relative clauses in Turkish are ambiguous as in (44).

(44) *sev-dik*[3]*-ler*$_i$*-imiz*$_j$ *sev-en- ler*$_i$*-imiz*$_j$
like-REL-PLU-1.PLU.POSS like-REL-PLU-1.PLU.POSS
those who we like/liked those who like/liked us

(*köpek*) *sev-en- ler*$_i$*-imiz*$_i$ (*köpek*) *ısır-an- lar*$_i$*-ımız*$_i$
(dog) like-REL-PLU-1.PLU.POSS (dog) bite-REL-PLU -1.PLU.POSS
those among us who like/liked dogs. those among us who dogs bite/bit.

Göksel (2007) indicates three points as the basis of the claim that morphology and syntax are distinct components as suggested by Di Sciullo and Williams (1987) and Ackema and Neeleman (2004):

- Pronominal relative clauses in Turkish have a fixed ordering of affixes regardless of their syntactic role.
- They have a fixed maximal size independent of whether the expression of more functions is required syntactically.
- They use affixes from the nominal paradigm irrespective of the requirement that these fulfill grammatical functions.

Göksel (2007) compares full relative clauses (FRC) with pronominal relative clauses (PRC) and then underlines that pronominal relative clauses are not head-deleted versions of the full relative clauses. The suffix *-lAr* (and *-ø* for singular case) functionally differs in both relativizer as in (45). In PRC, it marks the third person plural pronoun while it just marks the plural in FRC.

(44-5) *Al-dık-lar-ımız* (PRC).
buy-REL-**PLU**-1.PLU.POSS
those we buy/bought.

*Al-dığ-ımız* *kitap-lar* (FRC).
buy-REL-1.PLU.POSS book-**PLU**
the books that we buy/bought.

**al-dık-lar-ımız* *kitap-lar* (FRC).
buy-REL-**PLU**-1.PLU.POSS book-**PLU**

A head noun in a full relative clause can be employed as a direct or oblique object by syntax. On the other hand, a relativized pronoun is coindexed with the direct object case in a pronominal relative clause. Moreover, genitive noun phrases in FRC cannot be overtly expressed in PRC as in (46) (taken from Göksel, 2007).

(46) *Tolstoy-**un** sık sık oku-duğ-**um** roman-lar-ı*
Tolstoy-**GEN** often read-REL-**1.SG.POSS** novel-PLU-**3.SG.POSS**
Tolstoy novel which I often read

* *Tolstoy-**un** sık sık oku-duk-lar-**ım-ı***
Tolstoy-**GEN** often read-Rel-Plu-**1.SG.POSS-3.SG.POSS**
Those of Tolstoy which I often read

[3] Note that Göksel (1997; 2001; 2007), Tekin (2001) and Kelepir (2007) assume that *-K-* is a separate morpheme attached to *-DI* as a relativizer. Yet, the remainder of the literature assumes that *-DIK* is an unanalyzable relativizer, which is also accepted in this study.

Göksel (2007) concludes that word formation is opaque to syntactic mechanism is a rejection of strong Lexicalism as suggested by Booij (2005a).

Another important phenomenon is clitics which are morphologically attached to words but are highly syntactical. Certain clitics are neither words nor affixes. Yet, they constitute a separate type of object whose behavior is partly governed by clitic-specific grammatical mechanism as in the Catalan clitics in (47) (taken from Hualde, 1992). They have special syntax (Anderson, 1992; Anderson, 2005; Zwicky, 1977).

| | | | | | |
|---|---|---|---|---|---|
| (47) | ***Ho****=vaig* | *fer* | *per* | *a* | *tu* |
| | 3.SG-NEUT-ACC=AUX-1.SG-PAST | do-INF | for | to | you |

I did it for you

*Rep=****ho****!*
Receive-IMP-SG=3.SG-NEUT-ACC
Receive it

Göksel and Kerslake (2005) summarize Turkish clitics as the particle *-mI*, the connectives *dA* and *-(y)sA/ise*, the copular markers *-(y)DI*, *-(y)mIş*, *-(y)sA*, the adverbial marker *-(y)ken*, the generalizing modality marker *-DIr*, some person markers and the comitative/instrumental and conjunctive marker *-(y)lA/ile*. Most are morphologically concatenated with the right most constituents of phrases. Anderson (1992; 2005) concludes that lexical morphophonology controls the distribution of affixes; syntax for affixes and postlexical morphophonology for the distribution of special clitics. Bermudez-Otero and Payne (2011) state that there are special clitics which cannot be accommodated in morphophonology because these recalcitrant elements interact with lexical morphology and phonological rules.

Finally, syncretism should be reviewed within the morphology-syntax interaction. Syncretism is the mismatch between morphology and syntax as exemplified in the previous section. There are different instances of syncretism such as: simple, nested, contrary and non-autonomous (Baerman, Brown & Corbett, 2005). Syncretism can occur in all types of morphosyntactic constraints (gender, number, case, tense, voice, and such). In simple syncretism, two or more different morphosyntactic paradigms are merged in equal numbers as in Yup'ik (48) (taken from Baerman et al., 2005)

| | | | | |
|---|---|---|---|---|
| (48) | *nunak* | *nunak* | *nunat* | *nunat* |
| | land-DUAL-ABS | land-DUAL-REL | land-PLU-ABS | land-PLU-REL |

Nested syncretism occurs if simple syncretism is compounded across different environments; i.e., an unequal number of paradigms is merged. In contrary syncretism, the pairing of paradigms is mutually exclusive. Their common elements are given by Baerman et al., (2005) as follows:

- There is a morphological distinction which is syntactically relevant (inflectional)
- There is a failure to make this distinction under particular morphological conditions.
- Thus, there is a resulting mismatch between syntax and morphology.

In Turkish, for example, verbs usually have to agree with numbers of the subjects in plural case. When a 3rd person plural subject is not expressed by an overt noun phrase, and the referents are animate, plural marking of the predicate is obligatory (Göksel & Kerslake, 2005). If the subjects are inanimate, then they do not have to agree with verbs in number as in (49). Similarly, *-lAr* can be utilized in both 3.PLU and PLU grammatical roles.

(49) *Ev-**ler*** *yan-dı.* *İtfaiyeci-**ler*** *yan-dı-**lar**.*
house-**PLU** be burned-PAST fireman-**PLU** be burned-PAST-**3.PLU**
The houses were burned. The firemen were burned.

Syncretism should not be characterized as the lexicon letting the semantics down such as the homonymy of *bow* (a weapon) and *bow* (to bend forward). It is instead a deviation between function and form because the morphology fails the syntax. Carstairs-McCarthy (2010) thinks that this is an imperfection and a short-sightedness in language due to evolutionary differences in morphology and syntax because morphology exists mainly for synonymy avoidance. Baerman et al., (2005) proposed that the solution for syncretism lied in the assumption that morphology was lexeme based but opaque to syntax and independent of meaning to some degree. In the example in (49), morphology fails the syntax in number but it should be semantically motivated because the filtering of 3.PLU occurs when the subject is inanimate.

Li (2005) uses the $X^o$ theory for the morphology syntax interface. He explains that the $X^o$ theory does not try to extend the mechanism of one component to another. Rather, it recognizes the separation of syntax and morphology and constructs their interface as the synthesis of both components. He admits that morphologically complex words are not formed non-syntactically as they are in the Lexicalist Hypothesis. The internal structures of words should be parametrically visible to syntax. Various types of syncretism show that morphology and syntax are sometimes mismatched. They have to interact and be opaque to each other. Considering that morphology also interacts with semantics, it should reside in the lexicon. Morphology cannot be subsumed into syntax due to such syncretism. Thus, the problem lies in the interaction of lexicon and grammar.

### 2.3.3 Morphology and Semantics

The aim of language is to convey meaning thus every constituent in each linguistic domain ultimately exists to compose and express meaning in a less ambiguous way. Although Aronoff (1976) claimed that morphemes were not necessarily associated with a constant meaning and that their nature was basically structural, morphemes do carry meaning. The assumption that words are the minimal units of the lexicon is too constraining (Jackendoff, 1997; Keenan & Stabler 1997). For example, derivational affixes, relativizers, phrasal affixes and clitics require semantic transparency. Similarly, the Turkish relativizer *-DIK,* which approximately stands for English relative pronoun *that,* requires a lexical representation and compositional semantics. Therefore, the interaction of morphology and semantics, namely morphosemantics, is an issue that needs to be reviewed.

The central question of morphosemantics is whether morphemes have the same relations as, predication, modification and other such elements belonging to sentential categories. Moreover, the semantic compositionality of words with multiple morphemes is also a topic of morphosemantics. The central element is the root in a multimorphemic word and the head in a compound word. Each morpheme incrementally and unequally contributes to semantics and syntactic functionality of the words they are combined with. In other words, every morpheme has its semantic and structural sides, which are the morphological realization (Aronoff, 1993; Zwicky, 1986). Kibort (2011) presents a contrary and a more radical view that unlike syntactic agreement and government, tense instances in the Kayardild language are not morphosyntactic but morphosemantics because syntax is insensitive to the tense value of the verb. Considering the fact that meanings need to be associated with a *name* and morphemes have meanings, the relation between lexical semantics and morphology should be investigated as associations between *name* and *meaning* (Levin & Rappaport; 2001). Gamback (2005) notes that unification-based grammars and computational semantics inherently utilize three strong trends:

1. Keeping most of the semantic information lexicalized.
2. Building structures in a compositional manner
3. Postponing decisions as much as possible

A fully articulated theory of lexical semantic representation should be a generative theory that allows for the characterization of all possible word meanings in a language (Carter, 1976; Pustejovsky, 1995) up to phrase and sentence level. Verbs can be, for example, ergative, unaccusative, transitive and intransitive (Pustevsky, 1995). Their requirements of nominal cases actually reflect their semantic representation in a language-specific manner. For example, the English verb *agree* requires a locative object while it is accusative for Turkish. The same situation exists for affixes, too.

A well-known English suffix with a lexicalized meaning is *-ism*. It has the meanings of 'doctrinal system of principles' and 'peculiarity of speech' (Aronoff & Fudeman, 2011). Yet, the meaning of the derivate *-er* is not so systematic. For example, a *gambler* is a person who gambles regularly however, a *driver* drives a car but not necessarily regularly. A company in Turkey has been producing a very popular drink, *limon-ata* 'lemon-DER' which is lemonade. It has been advertising the new products, *nar-ata* 'pomegranate-DER' and *mandalina-ta* 'mandarin-DER' which can be transferred to English as *mandarinade* and *pomegranatade*. There had been no other occurrences of the derivate *-ata* except the word *limonata* in Turkish. The company was actually doing morphemic wordplay but the audiences immediately comprehended the message in the new commercial.

The meanings of morphologically complex words are partially predictable from the meanings of their parts or the previously known words sharing the same parts. Indeed, there are ambiguities and non-systematic variations in the semantics of morphemes. It is mainly because of the conflict between contrast and efficiency; a user of a language needs to convey meaning as clearly as possible and in an efficient and economical way (Siddiqi, 2009). The optimization between the two gives rise to ambiguities in the non-regular semantic composition. Most of the derivational affixes and some of the inflectional ones are frequently polysemous. Moreover, the correspondence between form and meaning in word formation is not always one-to-one (Lieber, 2004). It is only through use in context that morphemes acquire specific meanings. A single word may acquire distinct lexicalized meanings and form a complex lexical entry.

Levin and Rappaport (2001) review the lexical conceptual structure of verbs as in (50) (taken from Pinker, 1989).

(50) [[x ACT] CAUSE [y BECOME [ ]$_{\text{STATE}}$]]
Walking causes people to become healthy
Boredom causes people to walk

They state that morphology provides support for the existence of a two level lexical representation, lexical conceptual structure and argument structure. There are morphemes that signal the relation between the verbs and related lexical conceptual structure and the verbs with common lexical conceptual structure but a distinct argument structure. Consequently it can be seen that semantic compositions and their lexical representations require morphemic representations and morphemes also possess semantic content.

### 2.3.4 Morphology and Lexicon

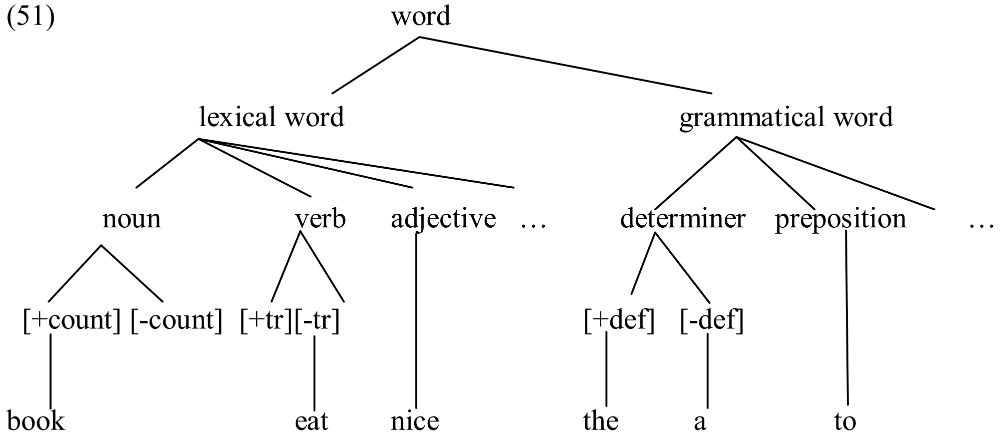
A lexicon as an inventory of words is considered to contain lexemes to serve as syntactic terminals. For example, *WALK* is a lexeme with two surface forms, the verb *walk* and the noun *walk*. Bauer (2003) excludes affixes from being lexemes and includes free words, which are stems and roots, because free words have more semantic substance than affixes. However, there are languages in which affixes have as much semantic substance as roots. Mithun (1999) provides some examples from Yup'ik, a Native American language, in which affixes bear meanings such as 'eat' and 'say' yet they do so by a discourse function (i.e., referring to previously introduced information in a discourse) rather than a semantic stem.

Traditionally, there is a lexical integrity hypothesis which states that no syntactic rule can refer to the elements of morphological structure (Lapointe, 1980), and that words are atomic at the level of syntax and phrasal semantics (Di Sciullo & Williams 1987). The lexical integrity hypothesis prevents syntax from intervening in the internal structure of words while the no phrase constraint prohibits morphology from interfering with syntax. They are historically the products of the Separationist Hypotheses. Liever and Scalise (2005) reviewed a set of data to propose a weaker lexical integrity hypothesis. They reported that syntax cannot manipulate the internal structure of words but can enter the internal structure of the words because syntactic rules have access to the internal structure of X$^{o}$ categories. Such claims require a review of the structure of the lexicon since it should be organized in a way that allows the interaction of syntax with the lexemes. If it is assumed that morphemes are lexemes, then the main point becomes the interaction of syntax and lexicon.

A widely known relation between morphology and lexicon is *blocking*. Aronoff (1976) defined the blocking as, "the nonoccurrence of one form due to the simple existence of another". For example, the existence of the irregular plural *women* for *woman* blocks the regular form **womans*. Similarly, there is no causative form of Turkish verb *git* 'go' because *git-tir* 'go-CAUS' is blocked by another verb *götür* 'take away'. The experimental study about Turkish emphatic reduplication (TER) performed in this research indicates that besides blocking, there is a frequency effect on a morphophonological operation, TER.

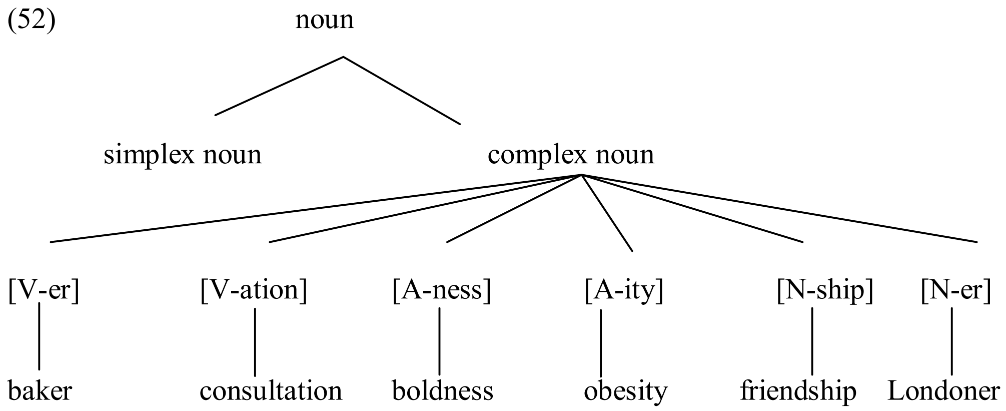
The lexicon has been represented as a hierarchy of types (Flickinger, 1987; Sag et al., 2003; Sag, 2007) as in (51) (taken from Booij, 2010). This hierarchy can be interpreted as an inheritance tree in which each node inherits the properties of its dominating nodes.

(51)



The inheritance hierarchy can also be used for morphological purposes (Hippisley, 2001; Riehemann, 2001) as in (52) (taken from Booij, 2010).

(52)



Each subclass in the morphological hierarchy tree may contain phonetic, syntactic and semantic properties. Schematic representations in the lexicon can be used for the unification of the complex structures. The same strategy can also be employed in compounding (see Krott, 2001). A hierarchical lexicon with different levels of abstractness and generalization, as outlined by Booij (2010), defines constituent families and sets of words sharing the same morphemes. The existence of constituent families has been confirmed psycholinguistically (Baayen, 2003; Schreuder & Baayen, 1997). The larger the size of a constituent family of a word, the faster it will be retrieved. This is valid for both morphemic constructions and compounding. Although the definition by Booij (2010) of construction morphology depends on the assumption that morphology is based on the paradigmatic relations of words and word forms, his studies on the hierarchical lexicon can be applied to morphemic lexicons. He only makes a distinction between inflectional and derivation morphology and focuses on the latter type.

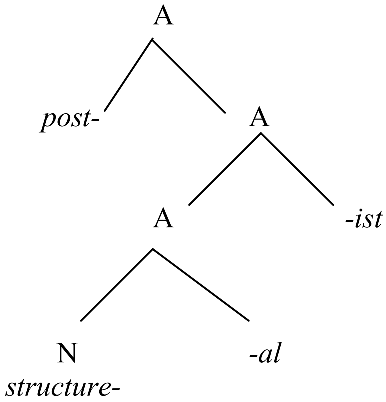
As an agglutinating language, Turkish can introduce brand new words such as *kitap-sız-lık-lar-ımız-dan* 'book-DER-DER-PLU-1.PLU-ABL' meaning 'because of our being without any book'. If it is searched in the web, it will be seen that there is no entry for it but this word is perfectly understood by native speakers of Turkish. Thus, it cannot be lexical item standing on its own, but it is rather a complex word with its semantics compositionally built up from the morphemes residing in the lexicon. There are psycholinguistic studies that suggest morphemic lexicon is required (Boudelaa & Marslen-Wilsow, 2001; Marslen-Wilsow, 1999; Marslen-Wilson, Zhou, & Ford, 1996). Some derived lexemes that are not perfectly compositional must be retained in the mental lexicon, such as *cranberry*, *boysenberry* and *raspberry*. Moreover, some frequently accessed compositional lexemes and their constituents should be stored in the lexicon to benefit from both computational and time efficiencies (Libben, 2006).

Aronoff and Fudeman (2011) give the following examples in (53) derived using a prefix *be-*.

| | | |
|---|---|---|
| (53) | *behead* | 'to remove someone's head' |
| | *befriend* | 'to make yourself a friend to someone' |
| | *besiege* | 'to lay siege to' |
| | *bewitched* | 'To be placed under the power of another as if by magic' |

In each case, the prefix *be-* produces a different meaning. The meanings of the stems in (53) are never lost but transformed. The forms are partially motivated. The derivational processes shown in (54) also provide clues about the structure of affixation. It shows that lexical categories are also opaque to affixes.

(54) [[*re*[*consider*]$_{V}$]$_{V}$ *-ation*]$_{N}$
[*post-*[[[*structure*]$_{N}$ *-al*]$_{A}$] *-ist*]$_{A}$]$_{A}$



Such affixes are called derivational affixes, which can create new lexemes. The inflectional affixes are for syntactic purposes such as number, gender, tense, aspect, modality and case marking. The other lexical morphological operation is compounding. It would be beneficial to review the conditions to distinguish inflectional suffixes from derivational ones. Actually, there are no necessary and sufficient conditions to distinguish inflection from derivation. Furthermore, compound words, their recognition and segmentation are discussed below.

Many morphologists support the Split Morphological Hypothesis (Beard, 1995) in which inflectional morphology is distinguished from compounding and derivational morphology. Since compounding and derivational morphology creates new lexical items and interact less with the syntax, they are exclusively assigned to the lexicon while inflectional morphology is attributed more to the syntax. Stump (2001) and Bauer (2003) summarized the criteria below to distinguish inflection from derivation:

- Derivation results in a change of lexical meaning.
- Derivation causes a change of word category.
- Inflectional affixes have regular meaning but derivations do not.
- Inflection is productive while derivation is semi-productive.
- Derivational affixes are nearer to the root than inflectional affixes. In other words, inflection closes words to further derivation.
- Derivatives can be replaced by monomorphemic words.
- Inflection uses a closed set of affixes.

However, even the category of a morpheme differs across languages. For example, the causative morpheme in Turkish is inflectional while it is considered to be derivational in Finnish. The distinction between the two mainly depends on the criteria given above. Yet, there are some conflicting examples and Turkish provides a valuable test-bed for these criteria. For this purpose, the statistics from the manual and morphological segmentation of the METU Turkish Corpus and the METU-Sabancı Turkish Treebank (Atalay et al., 2003) were examined.

Every morpheme results in a change in the meaning of stems. Even a noun with a plural morpheme does not have exactly the same meaning as the same noun in singular form. Derivational morphemes are considered to result from a change of lexical meaning. The locative and causative morphs in (55) and the ablative morphs in (56) and the plural morph in (57) are basically inflectional in Turkish yet they function as if they were derivational.

|  |  |  |  |  |
|---|---|---|---|---|
| (55) | *göz-de* | *yüz-de* | *öl-dür* | *an-dır* |
|  | eye-LOC | hundered-LOC | die-CAUS | recall-CAUS |
|  | favorite | percent | kill | resemble |

|  |  |  |  |
|---|---|---|---|
| (56) | *Soğuk-tan giyindim.* | *bir-den* | *yeni-den* |
|  | cold-ABL wear-PAST-1.SG | one-ABL | new-ABL |
|  | I wore [something] because of the cold. | suddenly | again |

(57) *Ali Bey-ler geldi.*
Ali mister-PLU come-PAST
Mr. Ali and his family have arrived. / The esteemed Mr. Ali has arrived.

The locative and causative morphs in (55) create new lexemes and they seem to be derivational. The ablative in *soğuk-tan* in (56) provides the meaning of 'because of' contextually because the verb *giy* 'wear' does not take an object in ablative case. The other two ablative morphs in (55) form adverbs. Similarly, the plural morph in (57) assigns either an honorific meaning to Mr. Ali or it means 'Mr. Ali and his family'.

Derivation is said to cause a change of word category. An apparent counter example is diminutive derivation. For example, *küçü(k)-cük* 'small-DER' meaning produces an adjective from another adjective meaning 'tiny'. There are also affixes that are usually inflectional. Yet, they change word categories and act like derivational morphemes and they can only be distinguished from their contexts. The locative in *göz-de* as in (55) and the ablatives in *bir-den* and *yeni-den* as in (56) form adjectival and adverbial words. Moreover, the reciprocal morpheme *-Iş* and the verbal noun marker *-mA* as in (58) change the word category.

|  |  |  |
|---|---|---|
| (58) | *Onu-nla bak-ış* | *Hızlı bir bak-ış.* |
|  | he-COM look-REC | quick a look-REC |
|  | Look at each other. | Have a quick look. |
|  | *hızlı bir don-dur-ma.* | *lezzetli bir don-dur-ma.* |
|  | quick a freeze-CAUS-VN | delicious a freeze-CAUS-VN |
|  | a quick freezing. | a delicious ice-cream. |

It is claimed that inflectional affixes have regular meanings but derivations do not. Some counter examples are observed in some word formations with inflectional tenses. For example, the Turkish future tense marker *-(y)AcAk*, and the perfective tense marker *-mIş* and the past tense marker *-DI* are required to have regular meaning when they are attached to verbs. Yet, they seem to be derivational in (59).

|  |  |  |  |  |  |
|---|---|---|---|---|---|
| (59) | *gel-ecek* | *y(e)i-yecek* | *geç-miş* | *gir-di* | *çık-tı* |
|  | come-FUT | eat-FUT | pass-PERF | enter-PAST | exit-PAST |
|  | future | food | past | input | output |

It is stated that inflection is productive while derivation is semi-productive. A T-test on the frequencies of the inflectional and derivational morphs discovered in the manual segmentation has shown that the number of the inflectional morphs is statistically higher than the derivational ones (p < .05). However, the manual segmentation task has also revealed that among the inflectional morphs, only about 20% are more frequent than the derivational morphs. The remainder have less than or equal frequency to the derivational morphs. Therefore, there are individual derivational morphs that are more productive than other inflectional ones.

Derivational affixes are assumed to be nearer to the root than inflectional affixes. In other words, inflection closes words to further derivation. A reasonable assumption made at first glance fails when the samples from the manual segmentation task are investigated. Indeed, the derivational morphemes tend to be closer to the stems. Yet, there are hundreds of counter examples obtained from the manual segmentation. For example, as shown in (23) *anla-ş-ıl-abil-ir-lik* has a derivational morph *-lik* after four inflectional morphs. Therefore, none of the inflections in this word closes the word to further derivation.

Another distinctive criterion is that derivatives can be replaced by monomorphemic words. It states that if a derived word in a sentence is replaced by a monomorphemic word from the lexicon, the sentence still makes sense. Yet, the same is not true for a word with inflections. A counter example is given in (60).

(60) *Çabuk konuş-tu-n.* *Çabuk ye*
fast speak-PAST-2.SG fast eat
You spoke in a fast way. Eat it fast.

The final criterion is that inflectional morphology uses a closed set of affixes. Yet, languages evolve and change over time. New morphs are formed from the existing ones or borrowed from other languages. For example, the feminine morpheme is borrowed from Arabic language as in *memur-e* 'civil servant - Fem' which means a female civil servant. In informal daily speech, an inflectional morpheme *-Ak* has been utilized as an optative suffix for a few decades. For example, *gör-ek* 'see-OPT' means 'let us see' whose formal form should have been *gör-elim*.

Considering the counter examples given above, there is no precise way of distinguishing inflectional morphemes from derivational ones. The conditions given are not necessary and sufficient conditions to make the distinction that derivational morphology is lexical and the inflectional morphology can be subsumed into the syntax. Bauer (2003) states that if we discard this distinction, then what remains are the stems and affixes and we must reconsider the definition of lexemes. Many languages cannot be studied if the words are assumed to be the only lexical items, thus, a morphemic lexicon is compulsory in linguistics of some languages.

Each stem in the lexicon has a semantic, and a functional content that determine the stem's meaning and its syntactic role. Every morpheme in the lexicon contributes to the semantic and functional contents of a stem to which the morpheme is attached. Prototypically, a derivational morpheme contributes more to the semantic content while an inflectional morpheme has more effect on the functionality as shown in Figure 4.

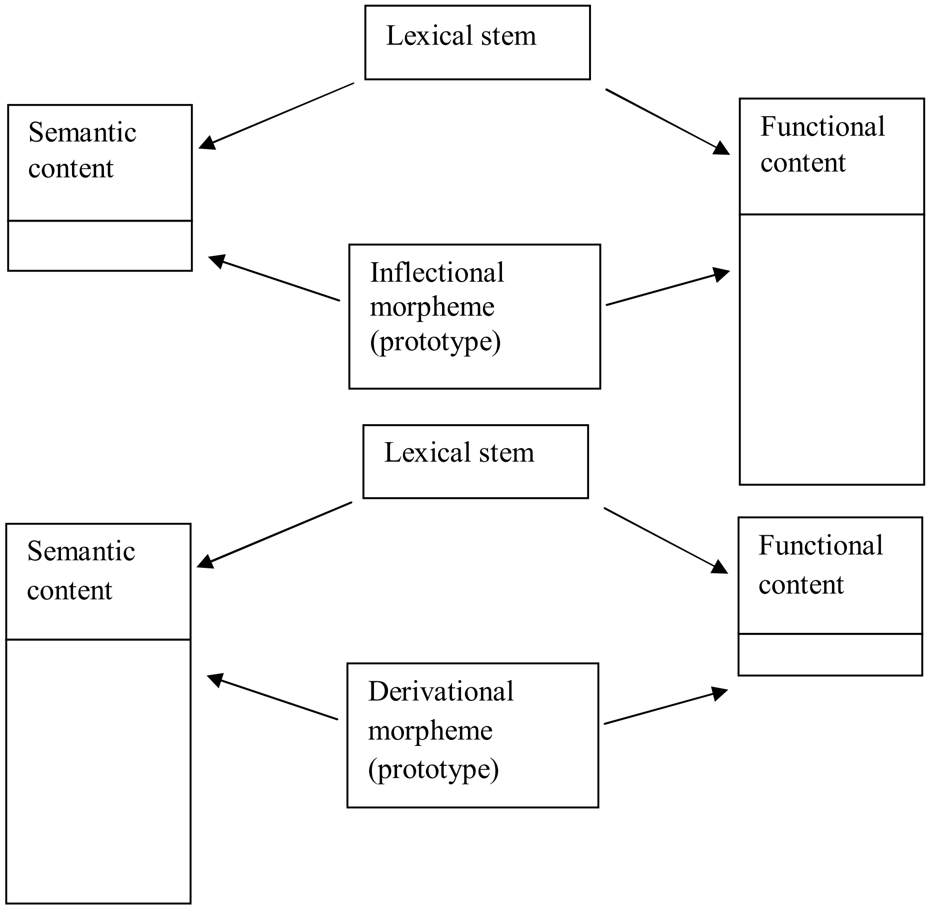


**Figure 4** Derivational and inflectional morphemic prototypes

A pure derivation in a language occurs in two situations. For the first case of pure derivation a stem tends to be extinct while its forms with derivational morphemes survive. For example, ***yuvar**lamak* 'rounding', ***yuvar**lak* 'rounded' and *al**yuvar*** 'erythrocyte' exist in Turkish while their stem *yuvar* 'round' does not exist anymore. Similarly, ***düğ**ün* 'wedding', ***düğ**üm* 'knot', ***düğ**me* 'button', ***bek**çi* 'guard' and ***bek**le* 'wait' are genuine lexemes while the root ***düğ*** 'to combine two things' and ***bek*** 'safe' do not exist anymore. In the second case a stem and its derived forms do not collocate in the same contexts. For example, all contributors in the manual segmentation task were able to conclude that ***yaz**ı* 'script' and ***yaz**ım* 'act of writing' were derived from ***yaz*** 'write' yet they failed to capture the derivation of ***yar**ı* 'half (adj)'and ***yar**ım* 'half (noun)' from ***yar*** 'split into two'. The reason is that the stem and its derived forms in the former example are often collocated in different sentences or contexts. On the other hand, the contextual co-occurrences of the latter example are quite low.

If the distinction is dropped and the morphemes are assumed to be lexemes, then the question concerns the lexicalization of the grammar because inflectional morphology is considered to provide an interface between the syntax and the lexicon. For example, Balogh and Kleiber (2003) discuss the computational benefits of a totally lexicalized grammar through a totally lexicalized morphology for Hungarian, which an inflectional language. They provide Prolog codes for the lexical representations of inflectional morphemes to capture Hungarian morphotactics. Moreover, they demonstrate how to check word order through morphological ranking parameters assigned to the morphemic lexemes.

For the distinction of the inflectional and derivational prototypical morphemes, all of the criteria given above can be assumed to be valid. Maintaining such a distinction in morphology is important to understand and analyze the morphology of a language. Yet, it should also be maintained that the distinction for particular occasions, contexts, or for particular languages can be discarded.

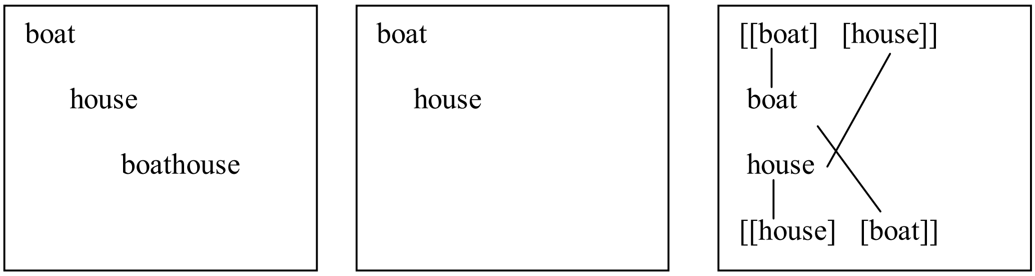
The next topic in the relation of morphology and lexicon is compounding. Compounding in linguistics is traditionally defined as the formation of a new lexeme by adjoining two or more lexemes (Bauer, 2003), such as the formation of *blackboard* from *black* and *board*. The involvement of a lexicon in compounding is a result of the loose grammatical combinations of words, and the inaccessibility of the internal structure of compounds to the syntax. Thus, Anderson (1992) excludes clitics and phrasal affixes from compounds, such as Spanish *darlo* ‘give it’, which consists of *dar* ‘give’ and *lo* ‘it’. Although compounding is widely accepted as a non-productive operation, it is, in fact frequent and productive in medical terminology, chemical compound naming, and in languages such as Dutch and German. Compounds are important objects of morphological investigations because they are present, globally, in all languages (Dressler, 2006). Libben (2006) speculates that the first word formation process in a language might have been compounding. Jackendoff (2002) also argues that compounds reveal the history of human language development and compounding precedes derivation. As pointed out by Dressler, there are languages with compounding and without affixation, but the reverse is not seen. Therefore, the study of compounding is also the subject of morphological processing and representation in language.

Inserting a new word or changing the linear order of the words in a compound result in a radical semantic change, or it is simply prohibited because one of the constituents of a compound assigns the semantic and syntactic properties of the compound. In other words, compounding heavily depends on the order of the constituents. Compounding also affects stress assignments. In English and Turkish, the majority of nominal compounds have two stresses and they are stressed more in the left-most constituent. The stress change of each constituent in compounding helps the listener to perceive the distinct constituents as a single linguistic entity.

In the discussion of compounding, Bauer’s definition of ‘lexeme’ needs to be clarified since he excludes affixes and includes free words, which are stems and roots, because free words have more semantic substance than affixes. A stem-only definition for compounding that excludes affixation is also problematic for Turkish. A Turkish compound, *ayakkabı* ‘shoe’ consists of *ayak* ‘foot’ and *kap*-CM ‘cover’ in which *kap* and the surface form of the compound marker *-(s)I* must be concatenated before the compounding because **ayakkap* is not a valid formation. The valid formation (*ayak* (*kap-ı*)) requires a pre-lexical and obligatory morphological operation. Furthermore, the free-word criterion is problematic, because, for example, it requires *outrun* and *underestimate* to be compounds rather than prefixed forms (Lieber and Stekauer, 2009). It also requires that Turkish verbs with the possibility suffix *-(y)Abil* as in *görebil* ‘[you] to be able to see’ is a compound because both *gör* ‘see’ and *bil* ‘know’ are free-forms in Turkish.

Booij (2005b) discusses afixoids, semi-affixes in compounding, such as -*like*, -*way* and -*wise* in *godlike*, *someway* and *clockwise*. He provides examples from Dutch to show that affixoids act as a bound lexical item on specific occasions although they are recognizable self-standing entries. He stresses that on some occasions it is difficult to distinguish compounding from derivation. Pre-lexical compounding operations, such as the use of the compound markers and post-lexical operations, such as inflectional and derivational operations regarding the head constituents, also indicate that compounding is intertwined with morphology. Further discussion on compounding lies in the approaches for the mental representations of compounds.

Libben and Jarema (2006) summarize the theories for the mental representations of compounds. In a full-listing hypothesis, compounds are listed in the mental lexicon as a single entity with compounds having no reference to their constituents. Another approach assumes that it is the constituents that are lexical items not the compounds. When needed, compounds and their semantics are computed from the constituents. The first approach provides computational efficiency while the second offers storage efficiency. The third approach is opportunity-based representation in which compounds and their constituents are stored in the lexicon as well as the corresponding references among them. All three approaches are summarized in Figure 5 (adapted from Libben, 2006, p.6). The properties of compounds, such as frequency, compositionality and semantics, determine the alternative, which is a choice of a full representation of a compound or temporarily computing it from its constituents.



a) Maximization of computational efficiency

b) Maximization of storage efficiency

c) Maximization of both

**Figure 5** Three approaches for compound word representations in lexicon

Clahsen and Almazan (2001) compared two groups of subjects with Williams Syndrome and Specific Language Impairment. They concluded that although the subjects were unable to syntactically produce irregular plurals, they were successful in producing a plural-inside-compounds. This finding makes the full lexical representation of compounds compulsory. Similarly, Fehringer (2012) performed an experimental study on subjects with aphasia and concluded that the opportunity based approach is more plausible. Fiorentino and Poeppel (2007) investigated the presence and neural instantiation of morphological decomposition in compounds using MEG and behavioral measures, demonstrating that compound processing involves decomposition, sub-served by a left-temporal component peaking around 300–400 msec post word-onset in the MEG signal, sensitive to morphemic, not whole-word properties.

Putting aside the discussions related to the mental representation of compounds and whether compounding is a subject of morphology, this leaves the question of how to differentiate single-stem words from single-word compounds and then how to segment compounds when no phonological information is at hand. In this study, simple bigram co-occurrence probabilities are employed to recognize and segment the single stem compound words in Turkish. The data for the statistical analysis was taken from the METU-Turkish Corpus (Say et al., 2002), the METU-Sabancı Turkish Treebank (Atalay et al., 2003) and the CHILDES database (MacWhinney, 2000). The method and findings are given in Chapter 4.

## 2.4 Acquisition of Morphology

Acquisition of morphology is part of human language acquisition. Niyogi (2006) formulated the process of language acquisition as having 4 key components as explained below:

1. Target grammar; $g_t \in G$ is a target grammar drawn from a class of possible target grammars (*G*) which are representational devices for generating languages. Languages are subset of $\Sigma^*$ where $\Sigma$ is a finite alphabet.
2. Example sentences; $s_i \in L_{g_t}$ are example sentences generated by the target grammar. $s_i$ is the $i^{th}$ example sentence in CDSs and $L_{g_t}$ is the target language of the corresponding grammar.
3. Hypothesis grammars; $h \in H$ are hypothesis grammars drawn from a class of possible grammars by learners (children). These grammars are used to both generate and comprehend the sentences of the target language.
4. Learning algorithm *A* is an effective procedure by which grammars from *H* are selected by the learner.

Many theories have been offered to explain how children acquire language. The behaviorist, nativist, social/cognitive and connectionist accounts have all been the most influential theories of language acquisition. The existing theoretical accounts of language learning would generally agree in the formal representation of language acquisition given above. They disagree mainly in the learning method itself and whether a hypothesis grammar exists or is there a single grammar with multiple parameters to be set or discovered.

Before and during the 1950s, the supporters of the behaviorist theory considered language learning as a behavior. Association, reward and punishment, reinforcement and imitation were the facilitators of language acquisition considered as ‘habits’. Chomsky’s (1959) review of Skinner’s Verbal Behavior (1957) started the nativist revolution in linguistics. The nativist perspective is that there is not sufficient information in the input to explain language acquisition. Instead, there needs to be innate syntactic knowledge and language-specific procedures. The theory of Universal Grammar (UG) provides a priori knowledge to acquire a language in terms of principles and parameters instead of a set of hypothesis grammars (H). Yet, the nativist account is not sufficient to explain how children acquire language.

The social/cognitive account mainly focuses on the learning procedure. In this account, an all-purpose learning mechanism is part of language learning. In other words, general learning mechanisms and abilities are effective in language acquisition. For example, a child can infer a word’s meaning by observing the gestures, movements and faces of others. Children do not speak like adults when they make utterances, ubiquitously, their language acquisition is developmental. The nativist view does not accept a developmental grammatical system but the social/cognitive account involving a changing grammatical system. The social/cognitive account also emphasize that there is sufficient information in child-directed speech for language acquisition. The child-directed speech is mainly slower, shorter, exaggerated, high-pitched and grammatically better-formed (*motherese*). Goldin-Meadow (2009) emphasized that motherese did not exist in different cultures because in some cultures children were not provided with such exaggerated and well-formed utterances, and they were not even addressees but indirect hearers during interactions. Goldin-Meadow (2009) further provides a stimulating view that children do not need a universally simplified input but they may undertake the simplification themselves. The limitations of a child’s memory may cause them to unable to recall long strings of words. As a result, they carry out the analytical work to discover linguistic regularities on a smaller or filtered database. This view is the ‘less-is-more’ hypothesis of Newport (1990). The connectionist account, on the other hand, is a movement within cognitive science with a goal of explaining human abilities through artificial neural networks composed of artificial analogs of neurons. From this perspective, language learning is a process of constantly adjusting the relative strengths and thresholds of the connections in the network until the linguistic output resembles linguistic input. Connectionism is more of a technique for exploring rather than explaining language acquisition. In this account (such as Plaut & Gonnerman, 2000), words are not single nodes but the co-activation of related phonological, orthographic and semantic features. For example, the word cat is the co-activation of its phonemes and orthographic form with semantic features, such as, “has tail”, “claws”, “runs”, and “meows”.

The nativist approach has a unitary view that every human language is motivated by an underlying UG. Yet, it is possible for children to arrive at language-like systems through other routes. Sign language used between deaf, and sometimes hearing, people differ marginally from spoken languages (Klima & Bellugi, 1979; Supall & Newport, 1978). Furthermore, deaf children can acquire sign language from their parents in the same way as hearing children acquire spoken language by achieving major milestones at approximately the same ages (Newport & Meier, 1985). However, when the isolated deaf children whose hearing parents did not know sign language were studied, it was discovered that each one of the isolated deaf children came to a language of his/her own (Feldman et al., 1978; ; Goldin-Meadow, 2003; Goldin-Meadow & Mylander, 1984). Each isolated deaf child presented a different language-like gesture system to their parents but received nonlinguistic co-speech gestures in return. More fascinatingly, when isolated deaf children were brought together, they developed a common set of signs such as in Nicaragua, from which, later, the Nicaraguan Sign Language (NSL) was born (Goldin-Meadow, 2009). When a new generation of deaf children acquired NSL, they began to make it more systematic and language-like. In other words, the need for communication motivated isolated deaf children to build a common set of signs while the need for transferring the set to the next generation resulted in a grammar. Thus, Goldin-Meadow (2006) indicated that innateness should not be defined as genetic encoding but that language was genetically resilient. In other words, it is a behavior likely to be developed by each member of the species under varying conditions. Resilience does not infer that language is a unitary phenomenon; instead, it suggests that language learning is gradual and varying, and uses human-specific and general abilities to acquire language.

The linguistic competence of a child gradually increases as growing up. However, it may also degrade in old age. Language not only changes during a person’s life but also across the generations. The requirement for a language to be teachable gives rise to the systematicity (i.e., grammar). Computational and robotic experiments attempt to explain the problem of language inventions (see Galantucci, 2005; Galantucci et al., 2010; Kirby et al., 2008; Smith and Kirby, 2008). There are two traditions in such experiments; the first approach assumes that linguistic structure arises as a solution to the problem of communication and the second does not take the communication pressure into account but considers language as a system transferred from generation to generation. The studies concerning the second approach concluded that a compositional system with recursion, word order and categories are the ultimate results of an unstructured communication system to be taught to the next generation. This conclusion is the same as the results of the research into the gesture system of deaf children (Golden-Meadow, 2006). This transfer process may result in the systematic change of grammar through generations. Similarly, word-frequency is a determinant in the lexical evolution. Pagel et al., (2007) showed that words with high frequencies have undergone a little lexical change through Indo-European history while less frequent words were more prone to phonetic changes. In order to test this hypothesis, 35 words from The Orkhon Inscriptions were randomly selected (Ergin, 2002). These words were used in ancient times between 7th and 8th centuries. The frequencies of their contemporary meanings were evaluated from the Corpus. When the words were sorted according to their frequencies, Table 1 was achieved.

**Table 1** Frequency sorted list of old and contemporary Turkish words

| **Old Turkish** | **Contemporary Turkish** | **Frequency** |
|---|---|---|
| *üçün* | *için* | 8332 |
| *bar* | *var* | 2820 |
| *yok* | *yok* | 1728 |
| *yer* | *yer* | 1427 |
| *kişi* | *kişi* | 850 |
| *kaltı* | *kaldı* | 394 |
| *ara* | *ara* | 332 |
| *oglı* | *oğlu* | 192 |
| *yiti* | *yedi* | 179 |
| *bay* | *zengin* | 152 |
| *bunça* | *bunca* | 117 |
| *olurup* | *oturup* | 115 |
| *bodunug* | *halkı* | 83 |
| *kagan* | *hakan* | 76 |
| *kök* | *gök* | 74 |
| *inim* | *kardeşim* | 59 |
| *yaşda* | *yaşta* | 58 |
| *ekin* | *ikisinin* | 54 |
| *olurtum* | *oturdum* | 37 |
| *teŋri* | *tanrı* | 32 |
| *çıgań* | *fakir* | 31 |
| *kergek* | *vefat* | 24 |
| *bodun* | *milleti* | 19 |
| *asra* | *altta* | 12 |
| *kazaganıp* | *kazanıp* | 11 |
| *üze* | *üstte* | 10 |
| *yagız* | *yağız* | 9 |
| *kuubratdım* | *toparladım* | 2 |
| *kıltım* | *kıldım* | 2 |
| *körür* | *bağlı* | 2 |
| *yarlıkadukin* | *lütüfkar* | 1 |
| *kuutum* | *talihim* | 1 |
| *apam* | *atam* | 1 |
| *törüg* | *töreler* | 1 |
| *udaçı erti* | *bozabilecekti* | 0 |

Table 1 indicates that Turkish lexical evolution is sensitive to the frequencies of words. The most frequent words, such as *için*, *var* and *yok*, will most probably undergo little change in near future. This finding is an indication that the frequent words are learned better and transmitted conservatively between generations. In other words, since the lexical evolution indicates that humans are sensitive to frequencies, these frequencies can also affect language learning through social interaction, such as the acquisition of morphology.

Lightbown and Spade (1999) state that for children their developmental sequences are related to their cognitive development and gradual mastery of the linguistic elements required for the expression of ideas. Experimental studies undertaken by Brown (1973), and Littlewood (1984) indicate that children acquire morphemes gradually in a sequence. They usually acquire the grammatical morphemes, i.e. inflectional ones, prior to the derivational morphemes. The distinction of inflection and derivation is claimed to be related with the internal configurations of mental lexicons. Yet, the internal characteristics of mental lexicons are unknown. Domínguez (1991) concludes that children learn complex and compound words first as an unanalyzable whole. They realize that some nouns follow patterns in the language, such as concatenating *-(e)s* to nouns for making plurals. The phonetic, contextual and semantic similarities and differences operate on the discovery and the lexicalization of the morphemes. Children first have to internalize lexical entries. Then, their linguistic experience leads them to understand that some words have a transparent internal structure that can be segmented. Thus, even morphemic lexicons are dynamic, incrementally improving and dependent on linguistic exposure.

In cognitive learning trajectory, the U-shaped learning of language has emerged as an area to be investigated (Ervin, 1964; Rumelhart & McClelland; 1986). U-shaped learning occurs when the learners first learn the correct forms of irregular verbs (such as *went and broke*), then overgeneralizing the rules regarding regular verbs they abandon the correct forms to produce *goed* and *breaked* and finally return to the correct forms. U-shaped learning seems to contradict the model of continuing cognitive development. The debates have continued on the connectionist and rule-based theories (Rumelhart & McClelland; 1986; Pinker & Prince, 1988; Plunkett & Marchman; 1991; McClelland & Patterson, 2002).

In the connectionist model, there are no symbols and rules to be learned but only a connectionist network to be trained through the linguistic input. There are input layers representing data and output layers reflecting the product. According to the activation parameters, learning involves a process whereby the network is fed by training input. A weighting procedure takes place to tune the network. The result is a pattern recognition task. Rumelharth and McClelland (1986) presented a connectionist network simulating children's overgeneralization in the U-shaped learning. This model predicts that acquisition takes time. In the rule-based theories, the learning is not a pattern recognition task but the acquisition of forms and corresponding rules.

Gordon (1985) reported that children typically produced 'rat-eaters' and not 'rats-eater'. Yet, this was not the case for irregular plurals because they produced 'mice-eater'. It supported the idea that inflection was universally represented on a separate level, which is ordered after all other morphological processes (Kiparsky, 1982a). Irregular plurals can be input to compounding because irregulars are stored in the lexicon. This suggests that morphology is not a pattern recognition but acquisition of rules, forms and exceptions.

Dual-processing models (Clahsen, 2006; Marcus et al., 1995; Pinker & Prince, 1988) are rule-based. These models take the differences in regular and irregular forms and propose that regular forms are produced by rules (such as concatenating *-ed* to verbs to form past tense) and irregular inflections are already in the memory. There is a filtering mechanism acquired by linguistic experience to block the regular formations when there is a corresponding irregular form in the memory, such as blocking of *goed* over *went*. U-shaped learning is a result of the process where children's morphological awareness is aroused when they acquire rules and form associatively. Beck (1995) reported a similar process for non-native speakers and broadened the overgeneralization phenomenon to second language learners. Blevins (2004) suggests that morphosyntactic templates and combinatorial rules are required to form words and a more specific rule can block a general one. For example, a realization rule <[V, 3.SG, PRES, IND], X+s> for English *-s* is less specific than a constant spell-out rule <[V, 3.SG, PRES, IND, BE], is>. Thus, *bes* is blocked and *is* is chosen.

Clahsen (2006) states that both build (i.e., rule-based) and frozen (i.e., memorized) forms are mentally represented. As in the compounding reviewed in the Section 2.3.4, a mental lexicon might be both built and frozen to maximize both storage and computational efficiencies. The optimization between the two during language acquisition can produce to the U-shaped learning.

Clahsen (1999) made a distinction between inflectional and derivational processes stating that regular inflectional processes were symbolic, whereas derivational forms were stored in the lexicon as whole forms. Laudanna, Badecker and Caramazza (1992) claimed that inflectional information was processed before derivational information. Although there are behavioral differences between inflectional and derivational morphology (Feldman, 1994), this division is prototypical as there are some cases where this categorization is not possible (Ford et al., 2003). Bertram et al., (2000) accepted that some words were processed as full forms while others were parsed. Yet, they neglected a distinction between inflection and derivation, but proposed that the factors determining whether a word was parsed were; the degree to which an affix cause a change in word's meaning, the productivity of affix form and the existence of homonymy affix.

Pinker (1991) and Clahsen (1999; 2006) advocated dual route models of lexical processing. They suggested that regular inflected forms were not stored but decomposed on-line because they were predictable. On the other hand irregular or novel forms were processed in a different manner and are stored. Schreuder and Baayen (1997) supported race models composed of two processing routes. Yet, in this version, both routes work in parallel with one route processing words as a whole and the other decomposing words into constituents. Only one of the routes wins the processing task. Highly frequent complex words are supposed to be stored as complex words while infrequent ones are processed via the decomposition route.

In word production, children may over-generalize morphological rules, such as, forming *goed* from *go* instead of *went* or producing *bene* from *ben* 'I' instead of *bana* 'I-DAT'. Lieven (2006) summarized the factors that are effective in how quickly children learn the morphology of their languages as: type frequency, token frequency, salience, transparency, formal complexity and the regularity and distributional consistency of the inflectional paradigm. While the type frequency indicates how many different lexemes are inflected in the same way, the token frequency is the relative frequency of different surface forms. The salience corresponds to audible and perceptible morphs. The transparency of morphemes is the degree of the acceptability of their semantics. Finally, the formal complexity of a morpheme is associated with the number of paradigms related to the morpheme. Portmanteau morphemes, for example, have more than one meanings or functions. These factors vary across the world's languages but children sooner or later master the morphology of their languages. Although different experimental studies indicated dual route models, there are some challenges from the supporters of the single route models (Lieven, 2006): Firstly, the connectionist networks successfully modeled some parts of the morphology learning task as a single mechanism. Secondly, these networks can also model the task with infrequent morphs. Thirdly, depending on frequency and phonological similarity factors of morphological markers, children show a long process of morphological development, and the overgeneralization problem might continue even after child's successful acquisition of the morphology.

Such challenges presented above have given rise to the race model of the dual route morphology (Frauenfelder & Schreuder, 1992). According to the race model, while parsing route gains the recognition of transparent low-frequent words, the direct route gains the recognition of high-frequency opaque words. Similarly, Gürel (1999) performed experiments and concluded that words with frequent suffixes seem to be accessed in a whole-word access procedure rather than through decompositional lexical access. In other words, morphemic frequencies take part in the race between the routes as well.

What lies at the heart of human language learning is segmentation and combination. Children need to segment what they hear; then, they need to discover patterns within and across words. These patterns eventually will lead to the syntax and morphotactics of their languages. When they acquire these patterns, children both understand the utterances they are exposed to and are successfully able to produce the novel linguistic forms of their own. It was shown experimentally that infants were sensitive to patterns not only within words and also across words through transitional probabilities, stress patterns and consonant clusters (Aslin et al., 1997; Aslin et al., 1998; Best, 1995; Jusczyk, 1999; Mattys et al., 1999; Saffran et al., 1996; Thiessen & Saffran; 2003). Therefore, a computational study modeling the acquisition of morphology from a corpus can make use of statistical cues in the corpus as in the current study.

# CHAPTER 3

## COMPUTATIONAL MORPHOLOGY and ITS ASPECTS

In this chapter, computational aspects of morphology, needs and models for computational morphology are discussed. Hauser, Chomsky and Fitch (2002) hypothesized that the faculty of language, in a narrow sense, included recursion and this was the only uniquely human component of the language. Pinker and Jackendoff (2005), on the other hand, considered that the innate language faculty contained more than recursion. Yet, they and the other authors all agree that the human cognitive system has a computational module and intelligence. These two aspects lead to computational abilities, followed by decision-making.

A morphological word can be defined in a recursive manner as in (61).

(61) Word = [stem]
Word = [Word + morpheme]

Mithun (2010) exemplifies that some languages show recursion in the morphological structures as in the Yup'ik and the Khalkha languages in (62).

(62) *Aya*___llru___*ni*___llru___*at*
go-PAST-say-PAST-TRANS.INDIC-3.PLU/3.SG
they said he had left

*jav*___uul___ *jav*___uuluul___
go-CAUS go-CAUS-CAUS
cause to go cause to cause to go

Turkish displays recursive morphological constructions in multiple causative words as in (63), similar to the relativizer *-ki* following genitive or locative cases in (22).

(63) *ye*___dirtdirt___
eat-CAUS-CAUS-CAUS-CAUS
make cause to make cause to eat

As in the Turkish relativizer *-ki*, the recursive morphological example in (63) is limited to two or at most three occurrences; otherwise it loses its comprehensiveness. Yet, it shows that there is systematicity in morphological constructions. Systematicity requires mathematical models of segmentation and acquisition because it is assumed that the human mind is a computational device that systematically interprets the received input and systematically transmits the output. Morphology might form a bridge between the world we conceive and the structured information in the lexicon. Many languages have a quite complex morphology requiring an extensive amount of work to define; however, morphological analyzers can provide researchers with a much needed short-cut. Thus, it can be said that computational approaches to morphology are required for machine translation (Goldwater & McClosky, 2005; Oflazer & El-Kahlout, 2007), speech recognition (Creutz, 2006) and information retrieval (Kurimo & Turunen, 2008).

Computational approaches to morphology are generally concerned with formal devices, such as grammars, stochastic models, tagging and parsing. Finite-state automata (FSA) and transducers (FST) are used as formal devices for encoding morphological grammars and analyzers. As stated by Monson (2008), most morphological computational approaches are generally hand-built. They require a savant in a specific language to define its rules in advance and consequently they are rule-based. Other approaches are statistical involving running algorithms on corpora to deduce morphemes or tags. These statistical approaches which as much as possible avoid domain-specific rules and knowledge have been gaining importance.

There is need for computational morphology because if one uses an electronic dictionary that depends only on words, he will fail to take advantage of regularities. For example, Hankamer (1989) reported that a word-only Turkish dictionary might contain hundreds of millions of words due to the highly productive suffixation even from a single root. Sproat (1992) stated that computational morphology is required for:

- Natural language application;
    - parsing, text generation, machine translation, dictionary tools and lemmatization.
- Speech application;
    - text-to-speech systems, speech recognition.
- Word processing applications;
    - spelling checkers, text input.
- Document retrieval.

Although the rule-based and statistical approaches act in a quite different manner, they share the same aim: producing a machine-readable morphological output. In this study, it is assumed that the morphological segmentation and the acquisition of words can be modeled and they are within the domain of humans' statistical abilities. Therefore, it is not innate but limited to humans because of the restrictions on the memory and computational capacity of other animals.

All computational models of morphology assume that language acquisition is a problem of induction, which is the creation of an internal representation of language that allows the learner to generalize the observed linguistic input, interpret and produce novel linguistic forms (Goldwater, 2006). In Chomskian nativism, general learning mechanisms are not sufficient to acquire a language, and humans inherit specific endowments to do this. Nativist approaches accept Gold's Theorem (Gold, 1967), in which any formal language that has hierarchical structure capable of infinite recursion is unlearnable from positive evidence alone. There are certain features in the input claimed to be specified by the Universal Grammar that allow children to set the value of some particular parameter (Dresher & Kaye, 1990). Empiricists, who oppose nativism support the idea that language acquisition is based on the statistical properties of the input, and language acquisition occurs in a similar way to other associative learning (Elman et al., 1996). Some scholars are apt to amalgamate two opposing views into a claim in which the language acquisition mechanisms of the Universal Grammar are assumed to gain benefit from statistics (Yang, 2004).

Despite being castigated by the nativists, computational approaches to linguistics support the empiricist view. Borrowing Marr's terminology (Marr, 1982), computational morphology primarily aims to be either at the *computational* level or the *algorithmic* level of the information processing performed by humans. Initially, Harris (1954) proposed that algorithms with statistical information could be used to mark morpheme boundaries. Later on, Koskenniemi (1983) proposed the two-level morphology approach in which the surface level is to describe the word form as they occur in written text and the lexical level is to encode lexical units such as stem and suffixes. This approach is a rule-based model using FST.

In this chapter, the rule-based and then the statistical morphological analyses are reviewed. Then, the need for and the use of statistical morphological analyzers are discussed.

### 3.1 Rule-based Morphological Analyses

Rule-based morphological analyzers have previously well-defined set of rules and corresponding finite-state machines to employ the rules. Although Koskenniemi's (1983) two-level morphology model was not the first rule-based model, it has been the most cited for two reasons. It used very standard machinery and it was the most successful model ever. Its machinery consisted of a finite-state transducer, which is actually a finite-state automaton, and language-specific morphotactics embedded in the transducer.

Karttunen and Beesley (2005) state that two-level morphology is based on three ideas:

- Rules are symbol-to-symbol constraints that are applied in parallel.
- The constraints can refer to the lexical context, to the surface context, or both.
- Lexical lookup and morphological analyses are performed in a cycle manner.

All morphological operations including concatenation, deletion, reduplication, epenthesis, subtraction, and infixation can be modeled using rule-based morphological analyzers (Roark & Sproat, 2007). Even the root-pattern morphology of Semitic languages was successfully modeled using rule-based methods (Kiraz, 2001; McCarthy, 1979).

#### 3.1.1 Finite State Automata and Transducers

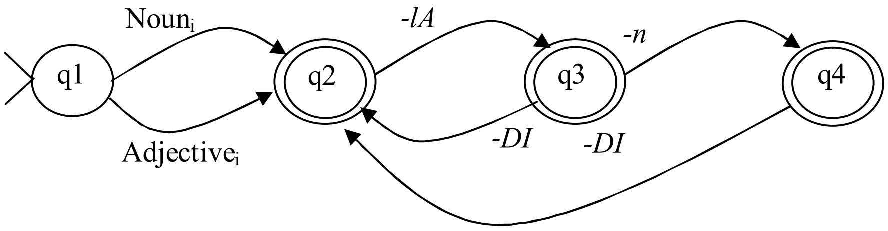
Finite-state automata (FSA) are quite well known in language theory (Hopcroft & Ullman, 1979; Lewis & Papadimitriou, 1997). We can define an FSA, namely *M*, over language *L* as in (64).

(64) $M = (Q, \Sigma, \delta, q_1, F)$ where

- $Q$ is the set of states $q$ of $M$.
- $\Sigma$ is the set of alphabet of morphemes $\sigma$ of $L$.
- $\delta$ is the set of transition states from $Q$ x $\Sigma$ to $Q$, such that for each $q_1 \in Q$ and $\sigma \in \Sigma$ there is $q_i \in Q$ such that $\delta(q_{i,} \sigma) = q_j$ where $q_j$ is non-final state unless morpheme $\sigma$ is at state $q_i$
- $q_1$ is the initial state of $M$
- $F$ is the set of final states of $M$.

An example FSA for the Turkish derivative *-lA* and the reflexive/passivizer *-n,* and the past tense *-DI* is given in (65). The derivative *-lA* is used for constructing transitive verbs from nouns and adjectives, and the reflexive morpheme is used to make the corresponding verbs intransitive. The automaton will accept bare nouns and adjectives, derived verbs from nouns and adjectives by *-la*, the reflexive forms of the derived verbs and corresponding past tenses.

(65)

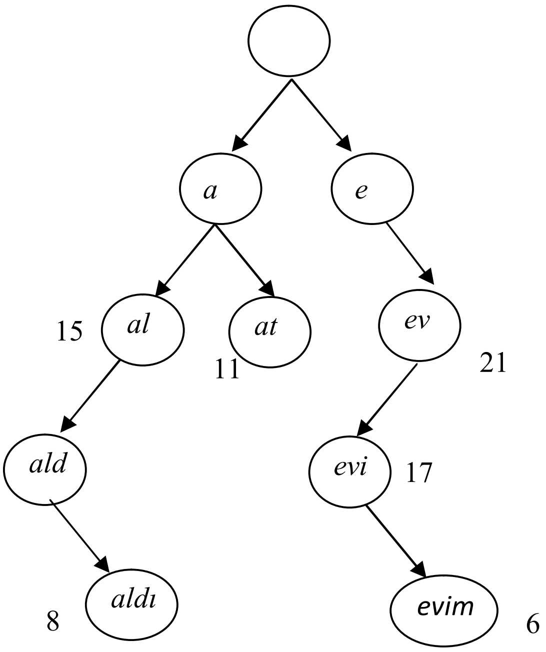


*M = (Q, Σ, δ, q₁, F)* where

- $Q = \{q_1, q_2, q_3, q_4\}$
- $\Sigma = \{Noun_i, Adjective_i, \text{-}lA, \text{-}n, \text{-}DI\}$
- $\delta = \{(q_{1,}\ Noun_i) \rightarrow q_2,$
  $(q_{1,}\ Adjective_i) \rightarrow q_2,$
  $(q_{2,}\ \text{-}lA) \rightarrow q_3,$
  $(q_{3,}\ \text{-}DI) \rightarrow q_2,$
  $(q_{3,}\ \text{-}n) \rightarrow q_4,$
  $(q_{4,}\ \text{-}DI) \rightarrow q_2\}$
- $q_1$ is the initial state
- $F = \{q_2, q_3, q_4\}$.

The automaton in (65) accepts the adjective *kuru* ‘dry’, the noun *av* ‘prey’, the verbs *kuru-la* ‘to dry’, *av-la* ‘to hunt’, *kuru-la-n* ‘to be dried’, *av-la-n* ‘to be hunted’ and their forms in past tense, *kuru-la-dı* ‘She/He dried something’, *av-la-dı* ‘She/He hunted something’, *kuru-la-n-dı* ‘She/He was dried’*, av-la-n-dı* ‘It was hunted’. Yet, it rejects **kuru-n* and **av-dı-la*. An apparent disadvantage is that morpheme boundaries must be predetermined to be employed in the automaton. The letter tree, which is also called the discrimination network or *trie*, is composed of the nodes that represent possible morphs and word formations of a lexicon (Knut, 1973) as in (66).

(66)



The trie in (66) is for a lexicon containing Turkish words *al* 'buy', *at* 'throw', *ev* 'house', *aldı* 'bought', *evi* 'house-Acc' and *evim* 'hous-1.SG.POSS'. The numbers indicate the frequencies of nodes in a sample text. It is possible to propose an FSA, which accepts or rejects the words according to the trie. Yet, the assumption that lexicons in human minds are tries is cognitively implausible (Forster, 1976). This statement depends on the psycholinguistic experiments in which native speakers evaluate nonwords slower than native words. However, an automaton for the above trie would reject a nonword **aldn* faster than accepting the word *aldı* 'bought' because the automaton will get stuck at earlier nodes of the trie. In other words, FSA approaches using trie fail to capture the frequency effect (Bradley, 1978), which states that there is a negative correlation between the frequency of a word and its retrieval time.

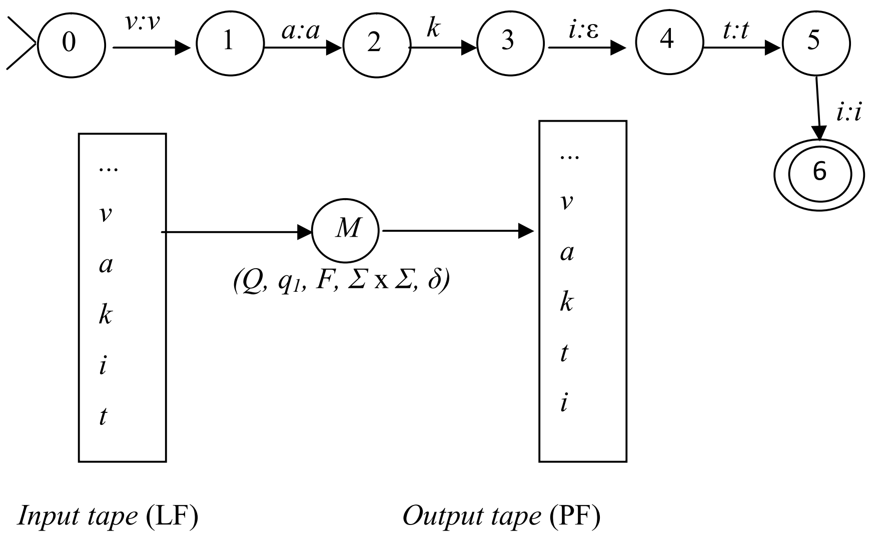
An FST is simply an FSA with two tapes instead as *M* in (67). It can compare lexical/logical forms (LF) with surface/phonologic (PF) forms to produce or accept a word.

(67) $M = (Q, q_1, F, \Sigma \times \Sigma, \delta)$ where

- $Q$ is the set of states q of $M$.
- $q_1$ is the initial state of $M$
- $F$ is the set of final states of $M$.
- $\Sigma$ is the set of alphabet of morphemes $\sigma$ of $L$.
- $\delta$ is the set of transition relation from $Q \times (\Sigma \cup \varepsilon \times \Sigma \cup \varepsilon)$ to $Q$.

The transition function causes the FST to both change states and write to the output tape. An FST that deletes *i* in Turkish word *vakit* 'time' concatenated with the accusative *-i* to produce the surface form *vakti* in (68).

(68)



$M = (Q, q_1, F, \Sigma \times \Sigma, \delta)$ where

- $Q = \{0, 1, 2, 3, 4, 5, 6\}$
- 0 is the initial
- 6 is the final state
- $\Sigma = \{v, a, k, i, t, \varepsilon\}$
- $\delta = \{(0, v, v) \rightarrow 1,$
  $(1, a, a) \rightarrow 2,$
  $(2, k, k) \rightarrow 3,$
  $(3, i, \varepsilon) \rightarrow 4,$

(4, *t*, *t*) → 5,
(5, *i*, *i*) → 6}, where ε denotes the empty string.

As FSA, FSTs can be concatenated to form larger constituents. The idea behind this concatenation is that larger constructions such as phrases and sentences can also be processed by FSTs. Yet, ambiguities create problems even in word formations. Simple FSTs are gullible because of the roots and the morphs containing morph-like sub-constituents. Weighted and probabilistic FSTs are used to improve the rule-based morphological implementations.

### 3.1.2 Implementations and Improvements for Rule-Based Models

The implementations of rule-based models in morphology usually require a large set of morphosyntactic rules, a lexicon of morphemes and corresponding tags. They can be implemented for both *item-and-arrangement* and *item-and-process* approaches. These different approaches are motivated by the properties of different languages. For Indo-European languages, morphological rules are more important than morphemes. Some morphemes represent more than one syntactic function and they are sometimes realized through alternation of stems. Thus, these phenomena yielded to the *item-and-arrangement* approach. Yet, for agglutinating languages like Finnish and Turkish, individual morphemes can indicate very systematic syntactic roles and they are arranged in a particular linear order. This caused to the *item-and-process* approach.

Two approaches later led to Stump's (2001b) distinction of *lexical* and *inferential* realization of morphology. In the *lexical* theories, morphs are not lexical entries but morphemes whereas morphs are lexicalized in the *inferential* one. Karttunen (2003) and Roark and Sproat (2007) discussed and showed that these two approaches can be reduced to finite-state operations. In other words, in terms of computational morphology, these distinctions are not quite informative but they are enlightening for cognitive science.

After Koskenniemi's analyzer for Finnish, Antworth (1990) proposed PC-KIMMO, a two-level morphological analyzer in which the user can build their own set of rules and lexicons. Later different rule-based analyzers offered for various languages such as Turkish (Çöltekin, 2010; Hankamer, 1986; Oflazer, 1994; Solak & Oflazer, 1993), English (Black et al., 1987; Ritchie et al., 1987), Hindi (Kumar et al., 2012) and Quechua (Weber et al., 1988). Since defining each rule was time-consuming, autosegmental methods lacking the distinction of rules and representations emerged.

Kay (1987) combined McCarthy's (1981) nonconcatenative morphology for Arabic with autosegmental phonology in a finite-state model. He used an FST that read *four* tapes to model Arabic causative word formation similar to the model developed by Wiebe (1992). Bird and Ellison (1994) proposed autosegmental representations and rules to form a one-level phonology. They claimed that the nonlinear phonology can be incorporated into constraint-based grammars such as HPSG (Pollard & Sag, 1987). They compared their model with Kay's (1987) and Wiebe's (1992) models and concluded that the automata would be less restrictive if it had no rules. Johnson and Martin (2003) suggested that morpheme boundaries could be identified by examining the properties of the minimal finite state automaton accepting the word types of a corpus. The automaton was, in a cyclic manner, accepting trie-like representations of the frequent morphs called *hubs*. Yet, the world's languages are full of exceptions, ambiguities and irregularities requiring the unions of plethora of the automata.

Clark (2007) used statistical stochastic transducers to compare the supervised and unsupervised learning of Arabic morphology. Stochastic transducers are FST with transition probabilities among the states. He made use of forward and backward probabilities on tapes to compare Arabic templates. The probabilities employed in the stochastic transducers are usually derived in a supervised manner. For example, Altun and Johnson (2001) proposed a probabilistic FSA approach for English auxiliaries and Turkish morphology. The training set they used to deduce probabilities was annotated meaning that the initial probabilities of the stochastic transducers are evaluated and then assigned by experts in advance. These experts can be human annotators or the pre-existing algorithms with high success rates. Whether stochastic or not, rule-based models require disambiguation among the alternant parses of a word.

Recently, Sak et al., (2011) proposed a set of complete electronic language processing resources and a morphological analyzer in Turkish. The analyzer is based on the two-level description given by Oflazer (1994) and the FST from Mohri (1997). Morphological analyses require disambiguation due to ambiguities and multi-functional morphs as in (69) (taken from Sak et al., 2011).

(69) Morphological parses of *alın*
(+ indicates inflectional suffixes and - is for derivational ones)
*alın*[Noun]+[A3sg]+[Pnon]+[Nom]
*al*[Noun]+[A3sg]+*Hn*[P2sg]+[Nom]
*al*[Adj]-[Noun]+[A3sg]+*Hn*[P2sg]+[Nom]
*al*[Noun]+[A3sg]+[Pnon]+*NHn*[Gen]
*al*[Adj]-[Noun]+[A3sg]+[Pnon]+*NHn*[Gen]
*alın*[Verb]+[Pos]+[Imp]+[A2sg]
*al*[Verb]+[Pos]+[Imp]+*YHn*[A2pl]
*al*[Verb]-*Hn*[Verb+Pass]+[Pos]+[Imp]+[A2sg]

Sak et al., (2011) used the average perceptron algorithm described in Sak et al., (2007) which was a trigram model previously employed in the implementation of morphological disambiguation in Turkish by Hakkani-Tür et al., (2002). Sak et al., (2011) reported that they achieved 97.81% accuracy in disambiguation. Yüret and Türe (2006) reported 96% accuracy in the same task using the Greed Prepend Algorithm adapted from Webb and Brkic (1993). In fact, the first morphological disambiguation task in Turkish was reported by Hakkani-Tür (2002) with 95.07% success rate. Yatbaz and Yüret (2009) used unsupervised methods running on morphologically annotated data set and achieved 64.5% accuracy in the morphological disambiguation task.

Karttunen and Beesley (2005) reviewed twenty-five years of finite-state morphology. They indicated that finite-state morphology supported the Optimality Theory (OT) (McCarthy, 2002; Prince & Smolensky, 1993). The OT was considered to be a two-level theory with *ranked* parallel constraints. The main difference between the OT and the finite-state morphology indicated by Karttunen and Beesley was that two-level rules were not universal while the OT proposed universal constraints. Roark and Sproat (2007) gave an interesting example of a local morphological dependency in Kanuri. In this language, person and number are marked with either prefixes or suffixes but not both. A finite-state network must record the fact that a prefix is given when a suffix is not given. It doubles the size of the corresponding network. Another way to model this local dependency is to use a push-down automaton in which previous states are registered.

The improvements in rule-based models vary with respect to the language being modeled. Enlarging the network model, enriching the rule-set and the lexicons, employing a register, and, more popularly, embedding statistics into the models have been among the common methods for improving rule-based models. Roark and Sproat (2007) indicated that the morphological analysis paradigm was shifted to statistical methods in the early 1990s; this is reviewed in the following section.

### 3.2 Statistical Morphological Analyses

Statistical morphological analyses are in the domain of machine learning. Recently, there has been an increasing interest in statistical modeling of morphology with *unsupervised*, *semi-supervised* or *supervised* algorithms running on corpora. In particular, agglutinating languages such as Turkish and Korean are good candidates for statistical morphological methods because words in these languages can consist of long morpheme sequences driven by morphosyntactic constraints.

The most influential and laudable statistical approaches to morphology were developed by Goldsmith (2001), Schone and Jurafsky (2001), Yarowsky and Wicentowski (2000), Baroni et al., (2002), Creutz and Lagus (2002; 2005), Sharma et al., (2002), Wicentowski (2002), Johnson and Martin (2003), Goldwater (2007), and Monson (2008). Before reviewing these models in the following subsections, an important aspect of the statistical methods, supervision, needs a brief explanation.

Supervision in statistical methods refers to the degree of labeled data or feedback on which the probabilities of the method depend. Consider a machine receiving a sequence of inputs $x_1, x_2, x_3,$... where $x_i$ is the sensory input at time $t$. In supervised learning, the machine is also given a sequence of expected output $y_1, y_2, y_3$... It enhances the probabilities it produced with the probabilities it is delivered. A subclass of supervised learning is *reinforcement learning*. In reinforcement learning, the machine is punished or rewarded in terms of its parameters according to its output. The machine aiming to improve the rewards for next trials tunes the probabilities and parameters. In unsupervised learning, the machine simply receives unlabeled inputs but no rewards or expected outputs are provided.

In the following sections, the most influential methods are briefly reviewed. The technical details are mentioned shortly.

#### 3.2.1 Supervised Approaches to Morphology

Most word segmentation methods involve many language-specific heuristics, hand-segmented training data, and/or lexicon of morphs. I consider the FST models reviewed in Section 3.1 as supervised models because they all had hand-built lexicons or morphotactic rules embedded in their automata. The initial ideas about supervised morphological learning were provided by Wothke (1985; 1986) and Klenk (1985a; 1985b). Then, Chang and Chen (1993) used a supervised morph segmentation method in which the training data was also used to evaluate the probabilities of different segmentations. Chang and Su (1997) used an iteratively growing dictionary while a corpus was being segmented with the help of statistics from previously segmented data. All of these studies were also aimed to be cognitively plausible in terms of language acquisition. Children are not exposed to utterances in which morpheme boundaries and types are labeled; therefore, due to their cognitive implausibility, supervised methods are out of focus of this study.

#### 3.2.2 Unsupervised and Semi-supervised Approaches to Morphology

Learning in artificial machines might be considered as mysterious with the dearth of concrete feedback or expected output. Yet, the actual task undertaken in unsupervised models is to uncover patterns in the data (Ghahramani, 2004). The pattern can be later used in decision making, feature detection, providing other machines with elements such as appropriate inputs. Semi-supervised learning falls into a category between supervised and unsupervised ones since it makes use of both labeled and unlabeled dataset. Usually, the labeled set is much smaller than the unlabeled one. Labeling, for example, can be carried out in advance by human annotators or by other successful methods. Then, statistics from the unlabeled raw data are combined with statistics from the labeled data. Employing partially labeled data sets results from the fact that creating sufficient labeled data can be very time-consuming, and the assumption that a sample inherits the patterns of its parent (Abney, 2008). As well as a small set of labeled data, heuristics are also used in semi-supervised learning to enhance the model. Children are exposed to *annotated* data early, but later they rely on self-discovery.

Almost every method in machine learning such as; the Hidden Markov Models, Naïve Bayes and *k*-nearest-neighbor classifiers, graphs, Support Vector Machines, clustering algorithms, Gibbs sampling, propagations and co-training can be revised as a semi-supervised learning model in computational linguistics (for detailed reviews see Abney, 2008). The ways of combining statistics from two sets vary across the models. Semi-supervised learning methods use unlabeled data to either modify or reprioritize hypotheses obtained from labeled data alone (Xiaojin, 2005). For the joint distribution of probabilities *p(x, y)*, *p(x)* of the labeled set and *p(y|x)* of the unlabeled set can be used. Another approach is weighting or averaging the probabilities from both sets to calculate the joint probabilities. It depends on the task and the intuition about the level of representativeness of the labeled data as regards the pattern in the unlabeled set. In other words, if it is assured that the labeled data quite successfully represents the pattern hidden in the unlabeled set, then it is wise to assign greater weights to the statistics of the labeled one. If it is explorative or unknown, it is better to undertake averaging or run a series of trials to determine weights.

Goldsmith (2001) developed *Linguistica,* based on the Minimum Description Length (MDL) (Risanen, 1989) which is a method of statistical inference. It views learning process as an inference from data compression. For a given set of hypotheses, *H* and data set, *D*, it tries to find the hypothesis in *H* that compresses *D* most (Grünwald, 2007). If *L(H)* is the length, in bits, of the description of the hypothesis; and *L(D|H)* is the length, in bits, of the description of the data when encoded with the help of the hypothesis, then the MDL aims to minimize *L(H)* + *L(D|H)*. Goldsmith's system works on an unannotated corpus then it derives *signatures* of the target language. *Null.er.ing.s* is an example *signature* that accepts *drink*, *drinker, drinking* and *drinks*. By using the MDL principle, the system first generates candidate signatures and then evaluates them.

The MDL is a reaction to maximum-likelihood estimation (MLE). In the MLE, the goal of learning is to select the hypothesis *H* with the highest likelihood as in (70).

(70) $H = \arg\max_h P(d \mid h)$

The MLE is simply a forward implementation of expectation maximization (Neal & Hinton, 1998). Therefore, it suffers from two problems, local maxima and predetermined parameters. In this method, the algorithm can possibly converge to a local maximum instead of a global one. Furthermore, the number of parameters to be evaluated must be known in advance. The optimization between the two might give rise to the *overfitting* problem when an unwanted hypothesis might dominate and model all data.

Goldsmith's candidate generation formula starts by producing all possible substrings with the length six. Then for each substring, the formula in (71) is computed. The substrings are ranked according to the corresponding values evaluated by the formula.

(71)
$$\frac{freq(n_1,n_2,...,n_k)}{N_k}\log\frac{freq(n_1,n_2,...,n_k)}{\prod_{1}^{k} freq(n_i)}$$

The first 100 top ranking substrings are selected as possible candidates. Then, a *take-all-splits* approach is employed. This splits the words in the Corpus in a probabilistic way according to the top candidate list and eliminates the non-optimal candidates. Then the remaining list is exposed to candidate evaluation procedure as in (72) according to the MDL principle. Both the candidates and the wordlist, i.e., the lexicon, are used in the formula.

(72)

a) $\lambda < T > + \lambda < F > + \lambda < \Sigma >,$

b) $\sum_{t \in T} (\log(26) * length(t) + \log\frac{[W]}{[t]})$

c) $\sum_{f \in suffixes} (\log(26) * length(f) + \log\frac{[W_A]}{[f]})$

d) $\sum_{\sigma \in \Sigma} \log\frac{[W]}{[\sigma]}$

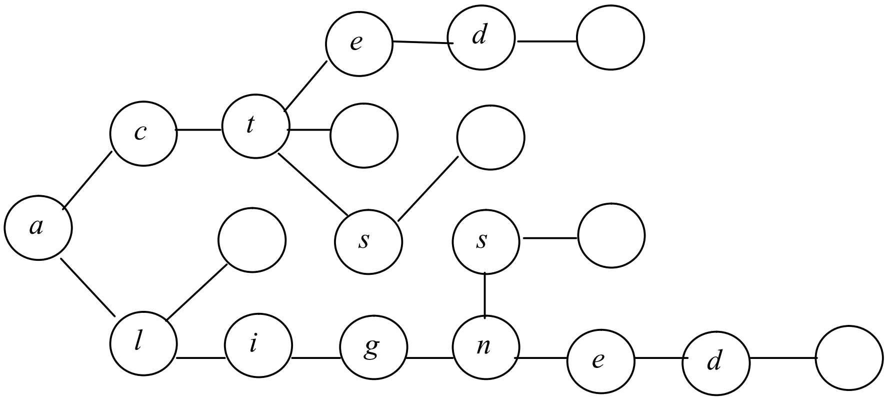
where *T* represents the set of stems, *F* is the set of affixes, *Σ* stands for the set of signatures, 26 is the cardinality of English alphabet, *W* is all word tokens in the corpus, *t* is the number of individual stems and σ refers to individual signatures.

The formula in (72a) is the main one that evaluates the compressed length of the model in bits. The other formulas in (72b), (72c) and (72d) compute each term in the main one. Then the overall maximum likelihood estimation for the compression model is computed for the whole corpus. Goldsmith (2001) reported a success rate of 82.9% percent in English (Roark and Sproat computed 81.8% *f*-score for *Linguistica*). This method is criticized for neglecting the syntactic and semantic information that children have access to. Yet, no method has been created that is able to model the kind of syntactic and semantic information that a child is considered to have access to (Roark & Sproat, 2007).

Demberg (2007) extracted suffix clusters resembling Goldsmith's signatures by employing forward and backward tries. She measured the distance between suffixes in clusters and roots. She considered that suffixes with a small distance result morphophonological changes. Schone and Jurafsky (2001) used semantic, orthographic and syntactic information derived from an unannotated corpus. They focused on the problems of Goldsmith's approach, which solely relied on orthographic representations. Without semantic information and morphophonological changes, it would be hard to distinguish *hate-d* from *hat-ed* because of the one of the best candidates, *Null.ed*. Their system was advantageous because it also considered circumfixes, incorporated frequency information, and used distributional information to identify syntactic properties and transitive closures to find semantic variants. They marked the potential affixes as the branches (excluding leaf nodes) of a trie.

The Schone and Jurafsky model starts with stripping off prefixes from stems according to predetermined thresholds. Then, it takes the original lexicon and the potential stems from the stripping process to build a trie. The branching in the trie represents a suffix as in (73) (taken from Schone & Jurafsky, 2001).

(73)



In order to find prefixes, the words in the corpus were reversed and a reverse trie was built. Again, branching in the reverse trie represented prefixes. Schone and Jurafsky used a version of the Latent Semantic Analysis (LSA) introduced by Schütze (1998) in order to compute semantic similarities among the words in the corpus. In the LSA, a matrix of *N* x *2N* dimensions is used to evaluate the semantic similarities among the words with respect to sentential co-occurrences with other words. If two words whose lists of co-occurrences with other words resembled each other, then they are assumed to be similar semantically. Then, a semantic transitive closure net among orthographically similar words was built. A threshold value was set in Schone and Jurafsky's model to accept that two words shared high semantic similarity. A combination of semantic and orthographic similarities was used for morphological segmentation together with the suffixes and prefixes determined from the tries. The method achieved 88.1% *f*-score value for English.

Yarowsky and Wicentowski (2000) proposed a semi-supervised modeling of inflectional morphology. Their method consisted of two stages; aligning roots and inflected forms from data, such as *take/took*, and training a supervised morphological analyzer over a subset of the aligned data. They created a table of inflectional categories for the language. For instance, -*ed*, -*t* and -*Ø* were input as past tense inflectional markers. Furthermore, they had a large corpus of text, a list of candidate noun, verb and adjective roots. They also made use of lists containing consonants, vowels and common function words selected in advance.

Yarowsky and Wicentowski used suffixes to distinguish roots. For example, the existence of *announce* and *announced* in the corpus and *-ed* in the suffix list aided the system in learning that *announce* was the root and *announced* is its past tense form. For irregular cases, such as *sing/sang*, they made use of the ratio of their frequencies in the corpus. As a third heuristic, they used weighted contextual vectors similar to those employed by Schone and Jurafsky (2001). The detailed trials can be found in Wicentowski (2002).

Besides producing a table of paired roots and their inflected forms, they also aimed to perform part-of-speech tagging. By using an interpolated backoff model iteratively on (74), they assigned the tags to words in the corpus.

(74) $$P(change \mid root, suf, POS) = P(a \rightarrow \beta \mid root, suf, POS)$$

where *a*→*β* represents the context-independent stem change.

The method proposed by Yarowsky and Wicentowski benefited from many heuristics and reported 99.2% success rate for the morphological segmentation of English verbs and their inflected forms. Johnson and Martin (2003) proposed the use of *hubs,* which were the nodes with words and in/out probabilities in automata in order to detect morpheme boundaries.

They combined the ideas of Harris (1951) with those of Schone and Jurafsky (2000). Similarly, Baroni et al., (2002) proposed a method similar to that of Schone and Jurafsky (2000). They used the edit-distance measure of orthographic similarities together with semantic similarities to determine morphologically related words. Snover et al., (2002) reported a very successful method using probabilities for morphological analysis. They used the frequencies of the stems and suffixes, their lengths, joint probabilities and number of paradigms. Then a novel search algorithm was run to determine the optimum analysis.

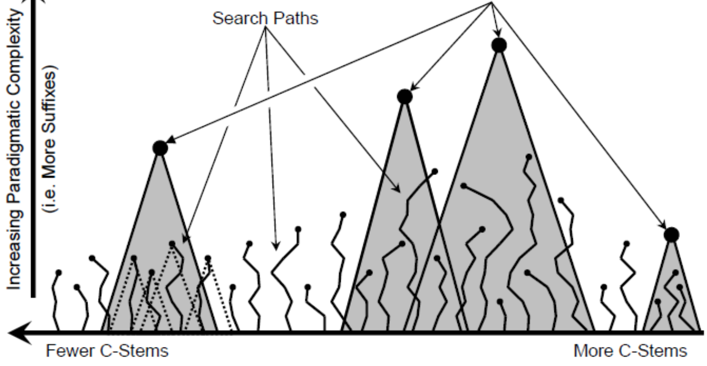
Creutz and Lagus (2002; 2005) introduced and improved the *Morphesor;* an unsupervised morphological analyzer with MDL principles, and it was specifically designed for agglutinating languages. They changed the MDL method so that it consisted of just two parts; a list of morphs including stems, prefixes and suffixes and a list of morph sequences resulting in valid word forms. They defied the single-suffix-per-word restriction observed in many of the computational morphology induction studies. In agglutinating languages, the search space of morphological models is quite immense. Each string can contain a number of morphological paradigms that is even larger than or equal to the cardinality of the string. In other words, each character in a word can represent a distinct morphological type. Their improvement in the traditional MDL framework was to employ a generative probability model through a greedy recursive search strategy through the probabilistic *maximum a posteriori* parameters as in (75).

(75)
$$\arg\max_M P(M \mid corpus) = \arg\max_M P(corpus \mid M)P(M)$$
$$where\ P(M) = P(lexicon \mid grammar)$$

The model acted recursively as long as the segmentation lowered the global description length. The lexicon was composed of morphs and the grammar was of the form *(prefix* stem suffix*)+*. The model used the probabilities in the Hidden Markov Model over the corpus. In other words, the *lexicon* had types and the *corpus* had tokens. The grammar was used in the model to find the optimum segmentation of the tokens with respect to the types. Moreover, in the model was the assumption that the prefix morphs were not allowed to occur in the rightmost position and suffixes could not be in the head positions of any word. Moreover, the models considered only the words with single stem. The authors reported 0.74 *f*-score for Finnish.

Monson (2008) improved *ParaMor,* which used the methods similar to the *Morphesor* but was paradigm-based. The model also made use of *schemas* that were actually the same as the *signatures* introduced by Goldsmith (2001). Clustering algorithms determined schemas each of which represented a morphological paradigm and was registered in the lexicon. The paradigms were composed of clusters as in (76) where the *c-stem* represents the candidate stems.

(76)



The segmentation algorithm of Manson (2008) runs in a bottom-up manner. In other words, starting from the most probable ending of *schemas* and performed searching among the *c-stem* space. The model achieved 56.3% success rate for English. Although Manson (2008) reported relatively successful experiments, the main weakness of his model is that it does not consider derivational affixes but focuses on inflectional paradigms.

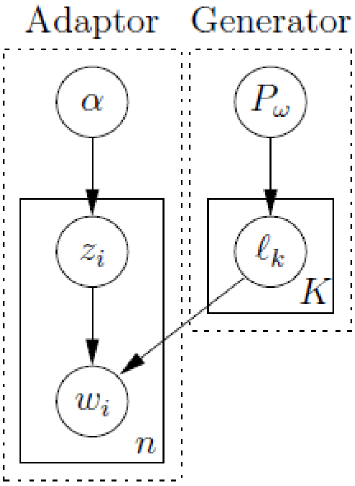
Goldwater (2007) reviewed the statistical morphology analyses models and claimed that the *Morphesor* contained no morpheme boundaries at all. She stated that the models worked because the number of parameters in the model was limited in an ad hoc way. Since Creutz and Lagus (2002; 2005) assumed that the probability of a suffix was independent of the stem, the probability distributions were different from the empirical data because stems were indeed the determiners in the selection of suffixes. In her doctoral dissertation, Goldwater (2007) proposed the use of non-parametric Bayesian inference to overcome the problems in the *Morphessor* in which the parameters could not be tuned to observe changes. She stated that non-parametric Bayesian inference was more appropriate when the number of parameters, their distributions and probabilities were unknown. Moreover, she emphasized that morphological analyzers attached more importance to the patterns over word types rather than word tokens. Thus, she claimed that the current explanations of word segmentation based on transition probabilities may be oversimplified since they generally assume independence between words. On the other hand, her segmentation experiments shown that bigram dependencies between words might result in better segmentation.

Bayesian learning depends on Bayes' rule in (77) which defines the probability of hypothesis *h* given data *d*.

(77)
$$P(h|d) = \frac{P(d|h)P(h)}{P(d)}$$
$$which\ \ is\ \alpha\ P(d|h)P(h)$$

From a cognitive point of view, the prior distribution, *P(h)*, is a combination of the innate learning biases and previous experience, and serves as a constraint on learning (Goldwater, 2007). The non-parametric model of Goldater (2007) had two components; a lexicon generator and an adaptor grammar that assign probabilities to lexical items as in (78) (taken from Goldwater, 2007). The number of parameters in her two-stage model was not specified in advance; yet, it tended to grow incrementally with the size of the data.

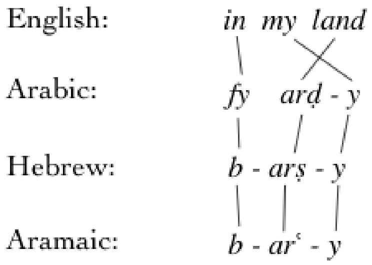
(78)



The generator acted on a corpus of tokens to produce morphs in a way to similar to the Chinese restaurant process (see Aldous, 1985). She used a *K*-dimensional and a Hierarchical Dirichlet distribution to return a set of parameters. Then, she used a CRP adaptor (Griffiths, 2005; Pitman, 1995) to receive the probability distributions of the parameters from the corpus and the generated tokens. Although she employed quite complicated statistical formulas, her method was more unsupervised compared to that improved by Crista and Lagus (2002; 2005) in which a lexicon of morphs was been utilized. Yet, it is cognitively less implausible. In a similar way, Johnson (2008) showed that a nonparametric model for Sesotho is better for word segmentation when it takes morphology into account.

A similar approach with non-parametric Bayesian learning was reported by Snyder and Barzilay (2008). They used parallel multilingual data to enhance the training of a morpheme segmentation model by aligning different morphs as in (79).

(79)



This study is interesting because it combined the hierarchical Dirichlet distribution and the Gibb's sampler with aligned multilingual data. It achieved 72.20 *f*-score for Hebrew and Arabic alignment. Thus, they concluded that related languages provided a better result. The practical implications of this study are that if there is a language with detailed available electronic resources, it can be used for the morphological analyses of another language provided that they are morphologically related.

## 3.4 Need for Statistical Models of Morphology

The practical reason for employing statistical models of morphology usually in an unsupervised manner is that full natural language lexicons are far too gigantic to fit in the working memory. Moreover, Hammarström and Borin (2011) review the history of statistical morphological analyzes. Then, he indicates that the problem of word segmentation is ever-present in speech processing, and the computational tools for taking on this problem are becoming increasingly sophisticated and increasingly available. Thus, researchers in linguistics and computational linguistics began revisiting the problems of word segmentation and learning of morphology with speech processing tools.

Another reason for employing statistical analyzers is that they can contribute to answering various questions in language acquisition (Batchelder, 1997; Brent, 1999; Goldwater 2007). Considering that children have access to semantics and pragmatics in addition to simple utterances, it would be implausible to assume that they only use statistics to acquire language. Moreover, statistical methods use large corpora while this does not necessitate that children need a specific size of utterances for language acquisition. Yet, statistical models propose that frequencies provide children with the clues necessary for language acquisition.

Ostler (2008) states that there are approximately 7,000 spoken languages in the world with 80% of the world's languages having fewer than 100,000 speakers. In other words, most of the languages are at risk of extinction. There is neither the time nor sufficient experts to document these endangered languages. Unsupervised or semi-supervised models could be utilized in order to rapidly and cheaply record the data for these languages that have not been studied yet. This needs to be undertaken quickly because some languages are disappearing at a rapid rate (Krauss, 2007).

In addition to documenting endangered languages Hammerström and Borin (2011) list the motivations for statistical learning of morphology as;

- linguistic theory
- elimination of the large lexicon for morphs
- modeling child language acquisition
- speech recognition
- machine learning
- morphological engine for the models in other linguistic domains
- language description and documentation bootstrapping for unstudied languages

In terms of Cognitive Science, statistical models attenuate the nativism. In particular, semi-supervised learning is appropriate for language acquisition because a small portion of the linguistic data to which children are exposed is supervised and the remainder of the data is unlabeled and massive. In other words, the quantity of parental or educational feedback that children receive is much less than the unlabeled data they meet in their daily life. Engaging interactions between parents and children is crucial for child language acquisition (Kuhl, 2004). These interactions provide children with required feedback, informative linguistic data and offer the intentional teaching of language. Therefore, statistical models for morphology are needed not just for practical purposes but also cognitive reasons.

Yet, it should be noted by the researchers with cognitive motivations that the primary instinct for a child is to communicate. Since this instinct is satisfied by first-language acquisition, adults are not as good as children in language learning even though their computational capacities (*memory* and *mathematical abilities*) are higher. Bearing this fact in mind together with the aim developing a computationally and cognitively acceptable model for morphological analysis, this study focuses on the Hidden Markov Model with a semi-supervised learning algorithm and minimum language specific heuristics, such as the average root length in Turkish. The model first tries to decide whether a given word is a compound or not. If the word is not a compound word, it is directly sent to a segmentation module. If it is a compound word, it is split then sent to the segmentation module. Since it is assumed that emphatic reduplication, epenthesis, voicing and deletion are morphophonological *operations* rather than morphemes, they are treated as special rules embedded into the model. The detailed method and findings are given in the next chapter.

# CHAPTER 4

# SEMI-SUPERVISED MORPH SEGMENTATION in TURKISH

In this study, a semi-supervised Hidden Markov Model (HMM) was employed. The formal definition of the HMM is given in Section 4.3.3. The HMM learned its initial emission and transition probabilities from a dataset. The dataset was composed of two subsets: the METU-Turkish unannotated corpus of about 1.7 million words, and a manually and morphologically segmented treebank of approximately 45 thousand words, the METU-Sabancı Turkish Treebank. The corresponding probabilities from both sources were averaged while inducing overall morph segmentations.

METU Turkish Corpus is a collection of 2 million words of post-1990 written Turkish samples. It is XCES tagged at the typographical level. The words of METU Turkish Corpus were taken from 10 different genres. At most 2 samples from one source is used; each sample is 2000 words or the sample ends when the next sentence ends. METU-Sabanci Turkish Treebank is a morphologically and syntactically annotated treebank corpus of 7262 grammatical sentences. The sentences are taken form METU Turkish Corpus. The percentages of different genres in METU-Sabanci Turkish Treebank and METU Turkish Corpus were kept the similar. The structure of METU-Sabanci Turkish Treebank is based on XML.

Before describing the morph segmentation method, the methods and findings of two partially independent studies concerning Turkish emphatic reduplication and the acceptability of nonce words in Turkish are explained in order to provide insights into the power of simple frequencies, co-occurrences and transition probabilities. Then, the method and findings of the semi-supervised morph segmentation task will be given in Section 4.3. The process of compound word identification and segmentation and the findings will also be explained in that section.

## 4.1 Method and Findings for Turkish Emphatic Reduplication

Emphatic reduplication is a morphophonological operation performed to accentuate the meaning of an adjective as formulated in (80).

(80) $C_1V_1C_2 \ldots \rightarrow C_1V_1(p, m, r, s)(A, Il, Am, \varepsilon)C_1V_1C_2 \ldots$
($\varepsilon$ denotes an empty string)

In order to thoroughly understand the linker selection choices available to native speakers of Turkish, a questionnaire was prepared which consisted of 31 non-adjectival words composed of Turkish nouns and verbs. The word list was given to the 25 male and 25 female participants whose native language is Turkish. Non-adjectives were deliberately selected to guarantee that the participants had never applied Turkish E-RED to the words. The participants were asked that if the words were adjectives how would they emphatically reduplicate them. They were allowed to give single word answers. In addition, they were asked; whether they had ever reduplicated the words, to explain how they knew to reduplicate the words and give the average time in seconds it took them to reduplicate of each word.

All the participants had at least a university first degree and their average age was 34.20 (*SD*=2.60). They reported that they had never reduplicated any of the words. The participants responded that they reduplicated the words 'intuitively' and each word required 5 seconds or less for reduplication. All the participants only used *-p, -m, -s,* and *-r* for the linker position none used *"-A, -Il, -Am"* as an additional prefix as shown in Table 2. It indicates that the existing emphatically reduplicated words with additional linkers are actually lexicalized.

**Table 2** Results for Intuitive Reduplication of Non-adjectives

| WORDS | # of p-linker | # of m-linker | # of s-linker | # of r-linker |
|---|---|---|---|---|
| *bıçak* | 11<br>(*bıpbıçak*) | 10<br>(*bımbıçak)* | 28<br>(*bısbıçak*) | 1<br>(*bırbıçak*) |
| *böcek* | 13 | 8 | 25 | 4 |
| *cevap* | 30 | 0 | 16 | 4 |
| *cami* | 28 | 8 | 14 | 0 |
| *çorba* | 29 | 0 | 16 | 5 |
| *dilek* | 32 | 11 | 7 | 0 |
| *davet* | 26 | 3 | 21 | 0 |
| *duvar* | 27 | 7 | 16 | 0 |
| *eğlen* | 50 | 0 | 0 | 0 |
| *fırın* | 17 | 2 | 23 | 8 |
| *felek* | 12 | 3 | 30 | 5 |
| *getir* | 32 | 0 | 18 | 0 |
| *götür* | 37 | 0 | 13 | 0 |
| *hüzün* | 43 | 2 | 5 | 0 |
| *jilet* | 36 | 6 | 8 | 0 |
| *kıble* | 18 | 2 | 30 | 0 |
| *kemir* | 23 | 8 | 14 | 5 |
| *leğen* | 36 | 3 | 11 | 0 |
| *laf* | 26 | 7 | 17 | 0 |
| *masal* | 14 | 5 | 29 | 2 |
| *nizam* | 32 | 8 | 10 | 0 |
| *pırasa* | 14 | 9 | 17 | 10 |
| *resim* | 38 | 4 | 8 | 0 |
| *surat* | 43 | 6 | 0 | 1 |
| *seçim* | 32 | 9 | 0 | 9 |
| *şerit* | 26 | 19 | 0 | 5 |
| *tutkal* | 37 | 5 | 8 | 0 |
| *tekerlek* | 24 | 7 | 18 | 1 |
| *vazo* | 29 | 2 | 19 | 0 |
| *yutkun* | 38 | 3 | 9 | 0 |
| *zarf* | 40 | 10 | 0 | 0 |
| **Distribution (%)** | **57%** | **11%** | **28%** | **4%** |

Many formations obey the constraints as previously defined in the literature (Demircan, 1987; Wedel, 1999; Wedel, 2000) but some formations such as *böpböcek, fırfırın, mammasal, kerkemir* and *kemkemir* violate the given constraints. Unlike the study by Demircan (1987), the linker order of this study is *-p* > *-s* > *-m* > *-r*. Moreover, the E-REDs with *r*-linkers also violate the conclusion by Wedel (1999; 2000) that the *r*-forms should be lexicalized; rather, they are just less-frequently formed.

The explanations for these irregularities might lie in the statistics. When the METU-Sabancı Turkish Treebank was examined, it was concluded that there are 43,574 roots of which 5,544 are distinct. The linker order found in the study is exactly opposite to that of the number of words with roots ending with a linker as shown in Table 3.

**Table 3** Distribution of Distinct Roots Ending with Linkers in Treebank

| Root ending | Number of distinct roots |
|---|---|
| *-p* | 100 |
| *-s* | 128 |
| *-m* | 281 |
| *-r* | 470 |

The average root length in the Treebank is 4.09 (Kılıç & Bozşahin, 2012). In a similar study, the average was reported as 4.02 (Güngör, 2003). Considering that Turkish is a suffixing language, an emphatic reduplication with prefixation might result in the degrading of the success of the root detection in communication. Besides the phonological constraints, selecting an appropriate linker so that the first segment of the reduplicated word has less resemblance to an existing root-word is additionally effective in the process of reduplication. For example, *dar-davet*, *dur-duvar*, *göm-götür*, *gör-götür*, *ger-getir*, *gem-getir, hür-hüzün, var-vazo* and *zar-zarf* are not produced (see Table 2) because *dar* 'tight', *dur* 'stop', *göm* 'bury', *gör* 'see', *ger* 'stretch', *gem* 'curb', *hür* 'free', *var* 'exist' and *zar* 'die' are already meaningful stems in Turkish.

The outputs from the participants that violated some of the constraints might also result from the consonant co-occurrences in Turkish. For a word $C_1V_1C_2$…, the consonant co-occurrence on the boundary of the prefix and the base will be taken from the set "*p*$C_1$, *m*$C_1$, *s*$C_1$, *r*$C_1$". The linker is selected so that the resemblance between the known word and the consonant co-occurrence is minimized. In order to test this hypothesis, statistics from the METU Turkish Corpus were examined. Table 4 shows the number of distinct words containing the consonant co-occurrences composed of one linker and the initial letter of the non-adjectival word derived from the Corpus.

**Table 4** Consonant Co-occurrence Statistics from Corpus

| Letters | *p-* | *m-* | *s-* | *r-* |
|---|---|---|---|---|
| *b* | 46 | 482 | 101 | 633 |
| *c* | 44 | 435 | 136 | 705 |
| *ç* | 112 | 13 | 48 | 602 |
| *d* | 106 | 1599 | 148 | 9958 |
| *f* | 25 | 28 | 66 | 191 |
| *g* | 11 | 92 | 54 | 1519 |
| *h* | 189 | 114 | 168 | 200 |
| *j* | 7 | 1 | 7 | 90 |
| *k* | 275 | 134 | 845 | 2575 |
| *l* | 1799 | 3171 | 1655 | 8005 |
| *m* | 340 | 257 | 519 | 5156 |
| *n* | 90 | 82 | 140 | 559 |
| *p* | 100 | 447 | 404 | 499 |
| *r* | 952 | 201 | 139 | 277 |
| *s* | 529 | 926 | 612 | 3119 |
| *ş* | 10 | 90 | 10 | 624 |
| *t* | 820 | 109 | 4338 | 3321 |
| *v* | 25 | 25 | 61 | 61 |
| *y* | 132 | 122 | 719 | 346 |
| *z* | 36 | 161 | 16 | 195 |
| **Distribution (%)** | **6%** | **16%** | **17%** | **61%** |

The distribution of the consonant co-occurrences is inversely correlated with the distribution of the participants' responses. For example, the number of words with the consonant co-occurrences using *-p* is very low while the number of emphatically reduplicated words using *-p* is very high.

The Mann-Whitney non-parametric test was employed to compare the distributions of the two sets given in Table 2 and Table 4. The word starting with a vowel, *eğlen*, was excluded from the calculation because it was reduplicated by using only the linker *-p*. It shows that the two sets were significantly different ($U_p$ = 190, $U_m$=27, $U_s$=58, $U_r$=0, $p$ < .05). For the participants' responses, the number of the answers were significantly and negatively correlated with the linker types changing from *-p* to *-r* (Spearman's $r$(120) = -.64, $p$ < .01). On the other hand, the number of the consonant co-occurrences in the Corpus significantly and positively correlate with the linker types changing from *-p* to *-r* (Spearman's $r$(120) = .34, $p$ < .01).

In order to corroborate the findings, another list was prepared containing 31 nonce words with initial $C_1V_1C_2$ patterns that were identical to the ones in Table 2 was prepared. The nonce words were evaluated as *acceptable* or *moderately acceptable* by the method proposed by Kılıç (2012). The new list of nonce words was given to another group of 25 male and 25 female Turkish native speakers (age $M$ = 28.00, $SD$ = 2.40). The participants were told to assume that these words were the names of recently invented colors. They were asked how they would emphatically reduplicate these new colors if they were in adjective phrases, such as *davlar* in a phrase *davlar araba* 'car'. The list of the words and the corresponding responses are given in Table 5.

**Table 5** Results for Intuitive Reduplication of Nonce words

| WORDS | # of p-linker | # of m-linker | # of s-linker | # of r-linker |
|---|---|---|---|---|
| *bıçır* | 10 | 9 | 31 | 0 |
| *böcele* | 15 | 10 | 24 | 1 |
| *ceverek* | 29 | 0 | 20 | 1 |
| *camat* | 30 | 5 | 14 | 1 |
| *çortu* | 31 | 0 | 17 | 2 |
| *dilit* | 29 | 13 | 8 | 0 |
| *davlar* | 28 | 2 | 20 | 0 |
| *duvsu* | 28 | 7 | 15 | 0 |
| *eğet* | 50 | 0 | 0 | 0 |
| *fırıtya* | 17 | 4 | 20 | 9 |
| *felipez* | 10 | 2 | 34 | 4 |
| *getelli* | 32 | 0 | 18 | 0 |
| *göttüp* | 37 | 1 | 12 | 0 |
| *hüziş* | 45 | 2 | 3 | 0 |
| *jiler* | 37 | 2 | 11 | 0 |
| *kıbut* | 17 | 1 | 31 | 1 |
| *kemtir* | 27 | 7 | 12 | 4 |
| *leğlef* | 38 | 2 | 10 | 0 |
| *lafut* | 26 | 7 | 17 | 0 |
| *mastun* | 16 | 7 | 26 | 1 |
| *nizeri* | 30 | 7 | 12 | 1 |
| *pırlaka* | 18 | 5 | 15 | 12 |
| *reser* | 40 | 3 | 6 | 1 |
| *surnup* | 43 | 7 | 0 | 0 |
| *seçper* | 37 | 11 | 0 | 2 |
| *şerleti* | 30 | 15 | 0 | 5 |
| *tuttarı* | 36 | 6 | 8 | 0 |
| *teken* | 24 | 7 | 18 | 1 |
| *vazar* | 25 | 1 | 24 | 0 |
| *yutnas* | 37 | 1 | 12 | 0 |
| *zartan* | 41 | 7 | 2 | 0 |
| **Distribution (%)** | **59%** | **10%** | **28%** | **3%** |

The Mann-Whitney non-parametric test was employed to compare the distributions of the two sets given in Table 4 and Table 5. The nonce word starting with a vowel, *eğet*, was excluded from the calculation because it was reduplicated by using only the linker *-p*. It shows that the two sets were significantly different ($U_p$ = 202, $U_m$=26, $U_s$=56, $U_r$=0, $p < .05$). The number of answers were significantly and negatively correlated with the linker types changing from *-p* to *-r* (Spearman's $r(120)$ = -.61, $p < .01$). On the other hand, the frequencies of consonant co-occurrences in the Corpus significantly and positively correlate with the linker types changing from *-p* to *-r* (Pearson's $r(120) = .31, p < .01$).

The results indicate that speakers are aware of the consonant co-occurrences in their native language. These co-occurrences are effective in the selection of the linker type in Turkish emphatic reduplication. This study (excluding the nonce words section) will be presented as a poster in the 35th annual meeting of the Cognitive Science Society, to be held in Berlin, Germany, Wednesday, July 31 - Saturday, August 3, 2013 (Kılıç & Bozşahin, 2013)

### 4.2 Method and Findings for Acceptability of Nonce Words in Turkish

Another study indicating the awareness and effects of orthographic co-occurrences and transitions was performed on the acceptability of nonce words in Turkish. In order to explain the model used for this purpose, let $s$ be a string such that $s = u_1 u_2 \ldots u_n$, where $u_i$ is a letter in the Turkish alphabet. The string $s$ is unified with the empty strings $\sigma$ and $\varepsilon$ such that $s = \sigma u_1 u_2 \ldots u_n \varepsilon$, where $\sigma$ denotes the initial word boundary and $\varepsilon$ denotes the terminal word boundary. The overall transition probability of the string $s$ is evaluated from the METU-Turkish Corpus using the following formula.

$$(81) \qquad P_t(s) = \prod_{1}^{\varepsilon} P(u_i \mid u_{i-1})$$

For example, using the formula in (81), $P(a|\sigma)$ gives the probability of the strings starting with the letter $a$, and $P(b|a)$ estimates the probability of the substring *ab* in the corpus. Now let $v$ be a subset of the string $s$ such that $v = v_{i1} v_{j2} \ldots v_{km}$ where $v_{km}$ is the $m^{th}$ vowel in the $k^{th}$ location of the string $s$. The overall probability of vowel co-occurrences of the string $s$ is estimated from the substring of vowels $v$ using (82).

(82)

$$P_c(v) = \prod_{2}^{j} \frac{g(v_{i-1}v_i)}{f(v_{i-1})} \qquad \textit{if the length of v is greater than 1}$$

$$P_c(v) = \frac{f(v_i)}{CorpusSize} \qquad \textit{if the length of v is 1}$$

In formula (82) the function $f(v_i)$ gives the frequency of the words that contain the vowel $v_i$ as a substring in the Corpus. The function $g(v_{i-1}\ v_i)$ gives the frequency of words in which the vowels $v_{i-1}$ and $v_i$ are co-occurring not necessarily in immediately consecutive positions but within the same word boundaries. The acceptability probability of the string $s$ is calculated by $P_a(s) = P_t(s)P_c(v)$. The acceptability decision of the string $s$ in the method is achieved using (83).

(83) *Accept* if $P_a(s) \geq 10^{-(t+v)}$
*Moderately Accept* if $10^{-(t+v+1)} \leq P_a(s) < 10^{-(t+v)}$
*Reject* if $10^{-(t+v+1)} > P_a(s)$

where $t$ is the number of transitions (which is *the length of the string* + 1) and $v$ is the number of the vowel co-occurrences (which is *the number of the vowels* - 1) in the string. If the string $s$ has only one vowel, then $v = 1$.

The method was applied to the list of nonce words given Table 4. The same list was also given to a group of 50 native Turkish speakers to evaluate the acceptability of each item. The nonce word *talar*, for example, was evaluated as in (84).

(84) $P_a(talar)$

$$= P_t(\sigma talar\varepsilon)\ P_c(aa)$$
$$= P(t|\sigma)\ P(a|t)\ P(l|a)\ P(a|l)\ P(r|a)\ P(\varepsilon|r)\ \ P_c(aa)$$
$$= 7.66\text{e-}06\ P_c(aa) = 7.66\text{e-}06 \ \ \text{x} \ \ 4.75\text{e-}01$$
$$= 3.63\text{e-}06$$

Since $P_a(talar) \geq 10^{-(6+1)}$, in which 6 conditional probability estimations and 1 vowel co-occurrences are evaluated, the nonce word *talar* is *accepted.*

The word list was evaluated by the 50 Turkish speaker participants. The distribution of the responses from the 50 native speakers and the results of the method are given in Table 6 (The bold text indicates a strong similarity of the results).

**Table 6** Acceptability of nonce words results from method and participants

| | | **Responses of the Participants** | | |
|---|---|---|---|---|
| **WORD** | **Results of the Method** | **Reject** | **Moderately Accept** | **Accept** |
| *öğtar* | **Reject** | 96% | 4% | |
| *söykıl* | **Reject** | 96% | 4% | |
| *talar* | **Accept** | | | 100% |
| *telüti* | **Reject** | 64% | 28% | 8% |
| *prelüs* | **Reject** | 84% | 14% | 2% |
| *katutak* | **Moderately Accept** | 8% | 50% | 42% |
| *par* | **Accept** | | 14% | 86% |
| *öçgöş* | **Reject** | 100% | | |
| *jeklürt* | **Reject** | 100% | | |
| *böşems* | **Reject** | 88% | 12% | |
| *trüğat* | **Reject** | 96% | 4% | |
| *cakeyas* | **Reject** | 92% | 8% | |
| *çörottu* | **Reject** | 74% | 16% | 10% |
| *döyyal* | **Reject** | 78% | 22% | |
| *efföl* | **Reject** | 92% | 8% | |
| *aznı* | Reject | 32% | 60% | 8% |
| *fretanit* | **Reject** | 64% | 30% | 6% |
| *erttiçe* | **Moderately Accept** | 36% | 64% | |
| *goytar* | Reject | 38% | 52% | 10% |
| *hekkürük* | Reject | 40% | 48% | 12% |
| *henatiya* | **Moderately Accept** | 36% | 64% | |
| *taberarul* | **Reject** | 84% | 16% | |
| *gövük* | Reject | 30% | 44% | 26% |
| *sör* | **Moderately Accept** | | 78% | 22% |
| *perolus* | **Reject** | 84% | 16% | |
| *kletird* | **Reject** | 98% | 2% | |
| *ojuçı* | **Reject** | 100% | | |
| *ürtanig* | **Reject** | 94% | 6% | |
| *lezğaji* | **Reject** | 100% | | |

| | | | | |
|---|---|---|---|---|
| *lamafi* | **Moderately Accept** | | 64% | 36% |
| *nort* | Reject | 38% | 42% | 20% |
| *netik* | **Accept** | | 18% | 82% |
| *meşipir* | Moderately Accept | | 24% | 76% |
| *oblan* | **Moderately Accept** | | 58% | 42% |
| *öftik* | **Reject** | 62% | 34% | 4% |
| *özola* | **Moderately Accept** | 32% | 60% | 8% |
| *ayora* | Accept | | 72% | 28% |
| *sengri* | **Moderately Accept** | 32% | 68% | |
| *sakkütan* | **Reject** | 58% | 34% | 8% |
| *şepilt* | **Reject** | 78% | 22% | |
| *şür* | **Moderately Accept** | | 78% | 22% |
| *puhaptı* | **Moderately Accept** | 38% | 44% | 18% |
| *upapık* | **Reject** | 54% | 28% | 18% |
| *ülü* | Reject | 28% | 52% | 20% |
| *yukta* | **Moderately Accept** | | 74% | 26% |
| *zerafip* | **Reject** | 54% | 34% | 12% |
| *upgur* | **Reject** | 70% | 16% | 14% |
| *kujmat* | **Reject** | 90% | 10% | |
| *lertic* | **Reject** | 94% | 6% | |
| *düleri* | Accept | | 64% | 36% |

For 82% of the words, the participants responses were in agreement with the results from the method, however, in 18% of the results the method failed to simulate the responses from the participants. This study was previously presented in the Students Session of the European Summer School in Logic, Language and Information 2012 (Kılıç, 2012).

These two experimental studies using statistical models with values evaluated from electronic resources have been described in order to provide the basis for the cognitive plausibility of statistical morph segmentation. Now we can return to the focus of this study, compound word recognition and segmentation and morph segmentation.

### 4.3 The Morph Segmentation Method and Findings

Before explaining the HMM model and its improvements through the manual segmentation task for the Turkish morph segmentation, it is relevant to review the raw frequencies evaluated from the Corpus and the CHILDES database, which were the starting point of this thesis. These frequencies show the extent to which ranking the frequencies of exhaustively generated orthographic representations can show the initial morphemes that have been acquired. After the power of frequencies, compound word recognition and segmentation tasks are explained, this section will conclude with the method, improvements and findings of the study.

#### 4.3.1 Power of Raw Frequencies

I started with the naive method of the exhaustive generation of possible *n*-grams from the Turkish alphabet, which consists of 29 letters. The possible unigrams ranged from *a* to *z*; bigrams were listed from *aa* to *zz*; trigrams from *aaa* to *zzz* and so on. Then, the frequencies of the *n*-grams were evaluated from the METU-Turkish Corpus. No phonological filtering was applied to the *n*-grams before evaluating their frequencies. The frequencies speak for themselves, for example, the most frequent *n*-grams in this group are the inflectional morphemes, as well as some connectives and frequent function words, such as *-lar* (Plu), *ve* 'and' and *bir* 'a/one'. The least frequent *n*-grams are usually rare stems, onomatopoeic words and nonce words, such as *ihya* 'enliven', *zzzt* and *ğaü*. A summary is provided in Table 7.

**Table 7** Total numbers of observed types and tokens of *n*-grams (*N*<=4) in the Corpus

| | Unigram | Bigram | Trigram | Quadragram |
|---|---|---|---|---|
| **Total Types** | 29 | 779 | 8,948 | 35,628 |
| **Total Tokens ~** | 20 million | 7.5 million | 6.5 million | 5 million |

Most frequent 15 tokens and their percentages from approximately 1.7 million words in the Corpus are given in Table 8.

**Table 8** Most frequent *n*-grams (*N*<=4) in the METU corpus

| Rank | Unigram | Bigram | Trigram | Quadragram |
|---|---|---|---|---|
| 1 | *a* | *ar* | *lar* | *ları* |
| 2 | *e* | *la* | *ler* | *leri* |
| 3 | *n* | *an* | *eri* | *erin* |
| 4 | *r* | *er* | *arı* | *ında* |
| 5 | *i* | *le* | *bir* | *arın* |
| 6 | *l* | *in* | *ara* | *inde* |
| 7 | *k* | *de* | *nda* | *iyor* |
| 8 | *d* | *en* | *yor* | *nlar* |
| 9 | *ı* | *ın* | *ini* | *anal* |
| 10 | *m* | *da* | *ını* | *asın* |
| 11 | *t* | *ir* | *ile* | *için* |
| 12 | *y* | *ma* | *rin* | *inin* |
| 13 | *s* | *ka* | *ası* | *ıyor* |
| 14 | *b* | *ya* | *anı* | *iler* |
| 15 | *o* | *bi* | *nde* | *alar* |
| **Percent of the Total Tokens** | 78% | 22% | 8.9% | 5.1% |

When the same method was applied to the BOUN Corpus of about 490 millions of words (Sak et al., 2011), the results in Table 9 were achieved. Table 9 indicates that 490 millions of words are biased because they are mostly collected from the websites of Turkish newspapers. For example, because of a highly frequent word, *Türkiye* 'Turkey', the trigrams *tür*, *iye* and the quadragrams *türk*, *kiye* amd *ürki* dominate the rest of the *n*-grams. Even though the METU Turkish Corpus has much less words; it is a well-balanced corpus, and suitable for use in a plausible statistical model.

**Table 9** Most frequent *n*-grams (*N*<=4) in the BOUN corpus

| Rank | Unigram | Bigram | Trigram | Quadragram |
|---|---|---|---|---|
| 1 | *a* | *an* | *bir* | *için* |
| 2 | *e* | *ar* | *ara* | *türk* |
| 3 | *i* | *er* | *baş* | *öyle* |
| 4 | *r* | *ya* | *iye* | *konu* |
| 5 | *l* | *ka* | *yap* | *iste* |
| 6 | *k* | *ir* | *ile* | *daha* |
| 7 | *n* | *de* | *kar* | *deği* |
| 8 | *t* | *bi* | *ama* | *başk* |
| 9 | *y* | *ra* | *ist* | *kiye* |
| 10 | *s* | *en* | *kon* | *ürki* |
| 11 | *o* | *et* | *tür* | *rkiy* |
| 12 | *b* | *la* | *kan* | *ilgi* |
| 13 | *m* | *al* | *gör* | *aşka* |
| 14 | *d* | *le* | *son* | *kara* |
| 15 | *u* | *il* | *ver* | *endi* |
| **Percent of the Total Tokens** | 81% | 21% | 6% | 4% |

The Child Language Data Exchange System (CHILDES) is a corpus launched in 1984 by Brian MacWhinney and Catherine Snow to serve as a central repository for first language acquisition data (MacWhinney, 2000). It also contains Turkish data consisting of the collected speech of children, their parents, relatives, friends and expert investigators (Slobin, 1982). The ages of the children varied from 8 months to 2 years. Of the 70,867 Turkish words in CHILDES, 32,272 were uttered by children, and the remaining 38,595 words were considered to be a sample of child-directed speech (CDSs). The most frequent 15 Turkish words[4] in CHILDES uttered by the children and the CDS in the study are given in Table 10.

**Table 10** Most frequent Turkish in CHILDES uttered by children and other participants

| Rank | Children | Others |
|---|---|---|
| 1 | *bu* | *ne* |
| 2 | *bak* | *sen* |
| 3 | *da* | *bu* |
| 4 | *ben* | *kim* |
| 5 | *var* | *nasıl* |
| 6 | *sonra* | *peki* |
| 7 | *bir* | *niye* |
| 8 | *de* | *var* |
| 9 | *o* | *o* |
| 10 | *ne* | *senin* |
| 11 | *işte* | *zaman* |
| 12 | *böyle* | *başka* |
| 13 | *anne* | *mi* |
| 14 | *yok* | *özge* |
| 15 | *ama* | *yapıyor* |
| **Percent in Total Utterances** | 17.7% | 18.2% |

[4] The 6th most frequent transcription was actually *xxx* which meant incomprehensible words uttered by the child which was subsequently discarded.

While segmenting the words in CHILDES into morphs, some transcription errors or usages specific to Turkish spoken language were not corrected or transferred into regular written forms deliberately in order to preserve coherence with the original database. The CDSs have 377 morphs and 1,584 stems. For example, Turkish question particle and the connector *dA* were erroneously written as concatenated to the previous words. Similarly, spoken form of Turkish future tense *-(y)AcAk* suffix is sometimes informally shortened as in *oynıcam* or *oynucam* instead of *oynayacağım* 'I will play'. All erroneous or colloquial forms were left as they were in the database. In total, 59,766 morphs were suffixed to the words. The words uttered by the children have 24,352 suffixes. Thus, the ratios of the morphs per word are .75 and .92 for the children and the others respectively. The frequency tables are given below and also include the overall numbers of distinct roots and morphs per speakers.

The words with highest number of suffixes in the Childes have 6 morphs. These are:

- *bin-dir-e-mi-yor-um-ki*[5] (uttered two times by the children)
- *süs-le-n-mi-yor-mu-sun* (spoken two times by the interviewers)
- *geç-ir-e-mi-yor-um-ki* (uttered once by a child)

The Turkish children's utterances in CHILDES were usually collected through question-directed interactions. Thus, the question word *ne* 'what' was the most frequent word uttered by the others. The exhaustive generation of *n*-grams without any phonological filtering and calculating corresponding frequencies from CHILDES database resulted in a meaningful ranking. Most frequent *n*-grams were grouped separately by the children, the mothers and the fathers as shown in Tables 11[6], 12 and 13 respectively. Table 14 shows the most frequent *n*-grams for all the CDS.

**Table 11** Most frequent *n*-grams in CHILDES uttered by children ($N$ <= 4)

| Rank | Unigram | Bigram | Trigram | Quadragram |
|---|---|---|---|---|
| 1 | ***a*** | ***ar*** | ***yor*** | ***iyor*** |
| 2 | ***r*** | *yo* | ***lar*** | *yoru* |
| 3 | ***e*** | *or* | *iyo* | ***ıyor*** |
| 4 | ***n*** | ***an*** | *oru* | *orum* |
| 5 | *y* | ***la*** | *ıyo* | ***anne*** |
| 6 | ***i*** | *ba* | *rum* | ***uyor*** |
| 7 | ***m*** | ***ya*** | ***ben*** | ***öyle*** |
| 8 | *k* | ***er*** | ***bir*** | *onra* |
| 9 | *b* | ***ne*** | *aba* | *sonr* |
| 10 | *o* | ***da*** | ***bak*** | ***ları*** |
| 11 | *l* | ***en*** | *arı* | ***nlar*** |
| 12 | *d* | ***le*** | ***yap*** | *rlar* |
| 13 | ***u*** | ***ra*** | ***ler*** | ***işte*** |
| 14 | ***ı*** | ***de*** | *ann* | *böyl* |
| 15 | ***t*** | *iy* | *nne* | ***üyor*** |

[5] *-ki* suffixes in both examples are not the relativizer *-ki* but the repudiative discourse connective *ki*. They should not have been concatenated to the previous words but they were in the original database.

[6] The bold items indicate morph-like *n*-grams uttered by children.

**Table 12** Most frequent *n*-grams in CHILDES uttered by mothers (*N* <= 4)

| Rank | Unigram | Bigram | Trigram | Quadragram |
|---|---|---|---|---|
| 1 | *a* | *ne* | *yor* | *kızı* |
| 2 | *e* | *ım* | *ama* | *özge* |
| 3 | *n* | *an* | *bak* | *ızım* |
| 4 | *m* | *ak* | *ızı* | *anne* |
| 5 | *i* | *bu* | *kız* | *ecim* |
| 6 | *r* | *in* | *zge* | *ıyor* |
| 7 | *k* | *ya* | *özg* | *hadi* |
| 8 | *ı* | *ma* | *zım* | *iyor* |
| 9 | *y* | *ge* | *ann* | *alım* |
| 10 | *b* | *en* | *nne* | *baka* |
| 11 | *l* | *ka* | *sen* | *neci* |
| 12 | *u* | *or* | *yap* | *nnec* |
| 13 | *d* | *di* | *aka* | *yoru* |
| 14 | *s* | *ba* | *cim* | *erin* |
| 15 | *o* | *er* | *alı* | *öyle* |

**Table 13** Most frequent *n*-grams in CHILDES uttered by fathers (*N* <= 4)

| Rank | Unigram | Bigram | Trigram | Quadragram |
|---|---|---|---|---|
| 1 | *a* | *or* | *yor* | *iyor* |
| 2 | *n* | *yo* | *bur* | *burç* |
| 3 | *r* | *en* | *iyo* | *rçak* |
| 4 | *e* | *ur* | *çak* | *urça* |
| 5 | *i* | *ne* | *rça* | *ıyor* |
| 6 | *y* | *ak* | *urç* | *peki* |
| 7 | *k* | *in* | *sen* | *ormu* |
| 8 | *u* | *an* | *ıyo* | *yorm* |
| 9 | *l* | *ar* | *pek* | *orsu* |
| 10 | *s* | *bu* | *eki* | *rsun* |
| 11 | *m* | *ya* | *sun* | *yors* |
| 12 | *o* | *iy* | *yap* | *asıl* |
| 13 | *ı* | *er* | *lar* | *nası* |
| 14 | *b* | *ni* | *rmu* | *diyo* |
| 15 | *d* | *un* | *ası* | *için* |

**Table 14** Most frequent *n*-grams in CDS portion of CHILDES (*N* <= 4)

| Rank | Unigram | Bigram | Trigram | Quadragram |
|---|---|---|---|---|
| 1 | *n* | *ne* | *yor* | *iyor* |
| 2 | *a* | *ar* | *lar* | *ıyor* |
| 3 | *e* | *or* | *yap* | *rsun* |
| 4 | *r* | *yo* | *sen* | *ormu* |
| 5 | *i* | *an* | *sun* | *yorm* |
| 6 | *y* | *en* | *iyo* | *yors* |
| 7 | *m* | *er* | *ıyo* | *orsu* |
| 8 | *k* | *in* | *ler* | *anne* |
| 9 | *l* | *ya* | *ama* | *yapı* |
| 10 | *ı* | *la* | *bak* | *apıy* |
| 11 | *s* | *ba* | *eni* | *uyor* |
| 12 | *o* | *un* | *rmu* | *pıyo* |
| 13 | *u* | *le* | *rsu* | *nası* |
| 14 | *b* | *im* | *kim* | *yapa* |
| 15 | *d* | *ak* | *alı* | *usun* |

Tables 11 to 14 show that the frequency-sorted *n*-grams uttered by children include morph-like units; and they mostly resemble the frequency-sorted *n*-grams uttered by the people interacting with the children. Therefore, it can be claimed that children's language acquisition is mediated by the people that they interact with. This claim is made for semi-supervised models because parents and other people in contact with the children are sources of supervised learning, and children are also exposed to a plethora of unstructured and unlabeled data. The combination of two is a form of semi-supervised learning.

Similarly, in the manual segmentation of the Treebank, 5,544 distinct roots, 240 distinct inflectional suffixes and 289 distinct derivational suffixes were observed. Initially, Çöltekin (2010)'s morphological analyzer was employed to make the segmentation task easier. This analyzer was also prone to false segmentations and oversegmentations and it also neglected derivational suffixes. Thus, cross checks and corrections on the data were performed by three people. There are 15,772 bare roots or some idioms, numbers, proper nouns, and compounds that are not segmented. A total of 49,451 segmentations was made, which constitute the total number of morphs in the Treebank. In other words, the average morph per word is 1.14. The words with the maximum number of morphs have 7 morphs attached, such as *yanlış-la-n-abil-ir-liğ-i-nden* 'from its being falsifiable' where -*la* and -*liğ* are derivational morphs. The most frequent 15 inflectional and derivational suffixes and roots are listed in Table 15.

**Table 15** Most frequent roots and suffixes in Treebank

| Rank | Root | Inflectional Suffix | Derivational Suffix |
|---|---|---|---|
| 1 | *bir* | *-lar* | *-lı* |
| 2 | *de* | *-i* | *-li* |
| 3 | *ol* | *-ı* | *-la* |
| 4 | *bu* | *-ler* | *-ce* |
| 5 | *o* | *-ma* | *-lik* |
| 6 | *ve* | *-m* | *-im* |
| 7 | *ben* | *-di* | *-lık* |
| 8 | *da* | *-a* | *-k* |
| 9 | *ne* | *-sı* | *-le* |
| 10 | *yap* | *-ın* | *-ı* |
| 11 | *bil* | *-me* | *-m* |
| 12 | *gel* | *-dı* | *-ık* |
| 13 | *iç* | *-u* | *-den* |
| 14 | *gör* | *-in* | *-lığ* |
| 15 | *için* | *-du* | *-ra* |

When the most frequent roots and the morphs attached to them were investigated the results given in Table 16 were observed.

**Table 16** Most frequent roots and distributions of attached suffixes in Treebank

| Rank | Root (from 5544 roots) | Number Inflectional Suffixes Attached (from 240 suffixes) | Number Derivational Suffixes Attached (from 289 suffixes) |
|---|---|---|---|
| 1 | *bir* | 24 | 10 |
| 2 | *de* | 51 | 3 |
| 3 | *ol* | 83 | 17 |
| 4 | *bu* | 23 | 3 |
| 5 | *o* | 33 | 6 |
| 6 | *ve* | 0 | 0 |
| 7 | *ben* | 12 | 4 |
| 8 | *da* | 0 | 0 |
| 9 | *ne* | 22 | 7 |
| 10 | *yap* | 80 | 6 |
| 11 | *bil* | 55 | 14 |
| 12 | *gel* | 51 | 6 |
| 13 | *iç* | 43 | 12 |
| 14 | *gör* | 78 | 8 |
| 15 | *için* | 3 | 0 |
| **Percent of the Total** | **0.27%** | **80.83%** | **24.91%** |

For example, the most frequent root in the Treebank was *bir* ‘a/one’ which co-occurred with 24 of all inflectional suffixes and 10 of all the derivational suffixes in the Treebank. When a unique list of derivational and inflectional suffixes co-occurring with the most frequent 15 roots in the Treebank was compiled, it was observed that only 0.27% of the roots covered 80.83% of all the distinct inflectional morph and 24.91% of all distinct derivational morphs. In order to understand the effect of ranking in the morph coverage percentages, 15 random roots were selected 10 times from the Treebank and their morph co-occurrence frequencies were observed as shown in Table 17.

**Table 17** Co-occurrence percentages of 10-fold selection from Treebank

| Trial | Percent for Inf. | Percent for Der. |
|---|---|---|
| 1 | 21.25% | 1.38% |
| 2 | 37.92% | 3.46% |
| 3 | 39.58% | 2.08% |
| 4 | 40.42% | 1.38% |
| 5 | 26.25% | 1.73% |
| 6 | 26.67% | 7.27% |
| 7 | 22.08% | 3.46% |
| 8 | 21.25% | 1.38% |
| 9 | 10.42% | 0.00% |
| 10 | 15.42% | 6.23% |
| **Avg. (10-fold)** | **26.13%** | **2.84%** |
| **Avg. (Top 15 words)** | **80.83%** | **24.91%** |

In none of the trials did the co-occurrence percentages of the selected 15 roots exceed 40.42% of the inflectional and 7.27% of the derivational suffixes in the Treebank.

When the distinct morph co-occurrences per the most frequent 15 roots in both the child speech and the CDS in the CHILDES database were observed, it was seen that the most frequent roots had a higher morph co-occurrence ratio than the less frequent ones. In order to verify this finding, an additional 15 roots were randomly selected 10 times. In each selection, the co-occurrence percentages were evaluated as summarized in Table 18 for the child speech and Table 19 for the CDS. The root-morph coverage percentages of 10-fold selections (on average) and the most frequent 15 words for the child speech and the CDS were almost identical in the inflectional morphs while they were highly parallel in the derivational ones.

**Table 18** Co-occurrence percentages of 10-fold selection from **the child speech** in CHILDES

| Trial | Percent for Inf. | Percent for Der. |
|---|---|---|
| 1 | 11.57% | 2.78% |
| 2 | 5.79% | 0.00% |
| 3 | 20.25% | 1.85% |
| 4 | 9.09% | 2.78% |
| 5 | 22.73% | 4.63% |
| 6 | 24.79% | 0.93% |
| 7 | 9.92% | 3.70% |
| 8 | 10.33% | 1.85% |
| 9 | 9.09% | 2.78% |
| 10 | 4.55% | 0.93% |
| **Avg. (10-fold)** | **12.81%** | **2.22%** |
| **Avg. (Top 15 words)** | **42.98%** | **4.63%** |

**Table 19** Co-occurrence percentages of 10-fold selection of roots from **the CDS** in CHILDES

| Trial | Percent for Inf. | Percent for Der. |
|---|---|---|
| 1 | 11.79% | 5.26% |
| 2 | 18.25% | 1.75% |
| 3 | 7.60% | 2.63% |
| 4 | 6.84% | 4.39% |
| 5 | 10.65% | 1.75% |
| 6 | 3.04% | 0.00% |
| 7 | 22.05% | 1.75% |
| 8 | 9.13% | 2.63% |
| 9 | 20.53% | 0.88% |
| 10 | 11.03% | 3.51% |
| **Avg. (10-fold)** | **12.09%** | **2.46%** |
| **Avg. (Top 15 words)** | **41.83%** | **6.14%** |

It can be possibly claimed that when children acquire the most frequent roots and the corresponding attached morphs, they have mastered and acquired an immense proportion of the morphological paradigms in their native language. This and the previous data represent the power of frequencies in children's morpheme acquisition processes. In other words, children might benefit from frequencies more often than expected.

### 4.3.2 Compound Word Recognition and Segmentation

The minimum description length based algorithms, such as the one used in Linguistica (Goldsmith, 2001; Goldsmith, 2005), cannot handle compounding because they have an assumption that each word contains a single stem (Creutz & Lagus, 2007). Therefore, it is necessary to assume that every word might have more than one stem. In that perspective, Qu et al., (2008) employed several supervised models to identify Chinese noun compounds in corpora achieving 90% recognition precision. Zhang et al., (2000) made use of the mutual information of Chinese characters in order to segment compound words. They achieved a 94% precision rate in segmentation. However, they had many parameters and thresholds needing to be set specific to Mandarin Chinese in which most of the characters were morpho-syllabic. Weller and Heid (2012) employed a German corpus with POS tags in splitting compound words for machine translation. They used word frequencies according to the German compound word patterns derived from the POS tags. They reported a 94% precision value for compound splitting. In a similar study, Adda-Decker (2003) made use of a German corpus for decompounding. She considered successor letters for each possible splitting point within a word with a length that was twice the minimum split length *L* for German compound words.

The existing models address each one of the issues with language specific parameters or compounding patterns and POS tags. The model in this study considers both of the tasks with minimum Turkish-specific assumptions and without any compounding patterns. The recognition task depends on the results of two manual segmentation processes while the segmentation employs two corpora with no annotation as explained below.

There are many single-stem words with lengths greater than the compound words in Turkish. Therefore, the first task is to differentiate compound words from non-compounds. For this purpose, the manual segmentation statistics of the words in the METU-Sabancı Turkish Treebank were used. In manual segmentation, the inflectional and derivational affixes were treated as the same and each morph was accepted as a distinct item. Some orthographic representations are highly frequent in word stems while others are not attested in the morphs. For example, *f, h, j* and *ö* are never allowed in any of the morphs in Turkish.

Two similarity functions, $f_s$ stem-similarity function and $f_a$ affix-similarity function, are used for the recognition of compounds. The functions correspondingly calculate the average pairwise conditional probabilities for a given word from the stem and affix types previously identified by the manual segmentation. The Treebank contains compounds, idioms and numeric representation and these are excluded from the similarity calculations.

In order to understand the function, let $s$ be a string such that $s = u_1\, u_2 \ldots u_n$, where $u_i$ is a letter in the Turkish alphabet. The string s unified with the empty strings $\sigma$ and $\varepsilon$ such that $s = \sigma u_1 u_2 \ldots u_n \varepsilon$, where $\sigma$ denotes the initial word boundary and $\varepsilon$ denotes the terminal word boundary. The $f_s$ and $f_a$ values of the string $s$ are evaluated comparatively from the Treebank and CHILDES using (85) where $n$ is the length of $s$.

(85)

$$f_s(s) = \frac{\sum P_s(u_i \mid u_{i-1})}{n+1} \quad \text{where} \quad P_s(u_i \mid u_{i-1}) = \frac{freq(u_{i-1}u_i)}{SizeOfStemList}$$

$$f_a(s) = \frac{\sum P_a(u_i \mid u_{i-1})}{n+1} \quad \text{where} \quad P_a(u_i \mid u_{i-1}) = \frac{freq(u_{i-1}u_i)}{SizeOfAffixList}$$

$$\text{if } \; f_s > f_a, \;\; \text{then s is a compound word.}$$

The method tested on 1,524 compound (Oflazer, 1994) and 1,524 random non-compound words. The average character lengths were 9.96 and 9.02 respectively. Initially, 51% of the non-compound and 70.4% of the compound words were successfully recognized by the method by using the affixes and stems in the Treebank. The success rates were 54.8% for non-compound and 68.7% for compound words using the CDSs. It was reported that the average stem length in Turkish was about 4 characters (Güngör, 2003; Kılıç and Bozşahin, 2012). Therefore, the bigram probabilities in $f_a$ before 4th orthographic position were decreased as a punishment while the probabilities in $f_s$ after 4th orthographic position were increased as a reward by fractions of .2, .3 or .5. The motivation lying beneath the application of the fractions as rewards or punishments was that it was not expected to find that an affix-like bigram co-occurrence before the 4th character and it was unlikely that a stem-similar co-occurrence would appear after the 4th character of a string with a single-stem. The success rates are summarized in Table 20 and 21.

**Table 20** Recognition success rates with respect to rewards and punishments using CDSs

| Reward/punishment fraction | Non-compounds | Compounds |
|---|---|---|
| 0 fraction | 54.8% | 68.7% |
| .2 fraction | 71.2% | 81.9% |
| .3 fraction | 68.4% | 85.5% |
| .5 fraction | 64.2% | 93.9% |

**Table 21** Recognition success rates with respect to rewards and punishments using Treebank

| Reward/punishment fraction | Non-compounds | Compounds |
|---|---|---|
| 0 fraction | 51.0% | 70.4% |
| .2 fraction | 69.5% | 85.2% |
| .3 fraction | 67.8% | 89.9% |
| .5 fraction | 66.0% | 95.7% |

Considering the recognition success rates for the non-compound words, .2 ratio was chosen as appropriate one to apply while calculating the probabilities before or after 4th orthographic position to be utilized in the recognition task. The tables also indicate that the CDS data was more successful in the recognition of non-compound words but less successful in recognizing compound words than the Treebank. This implies that the number of known morphs is effective in the recognition of non-compound words while the high number of known stems increases the recognition of compound words in Turkish.

When a word is identified as a compound word, the next stage is to segment it because during communication the internal structure of compounds must be accessed through a hierarchical segmentation by the hearers. For the segmentation task in this study, it was assumed that all members of a compound occur as a free form, except for the cranberry morphs (Aronoff, 1976). In order to understand the segmentation task, let *w* be a compound word with length *n* with boundaries marked by the empty string *ε*. There are $k = n\text{-}1$ candidate points in the string for segmentation. For each point, the mean of the probabilities of the substrings that either start or end at the point are evaluated from the Corpus using the formula in (86) adapted from Bernhard (2006). The points above a determined threshold are selected as compound splitting points. The threshold formula is given in (87). Only the substrings occurring as free forms were used in probability estimations.

(86)

$$f(k) = \frac{\sum_{i=0}^{k-1} \sum_{j=k+1}^{n} p(s_{i,k}) + p(s_{k,j})}{k}$$

where $s_{i,j}$ is a substring of *w* starting from the $i^{\text{th}}$ position to the $j^{\text{th}}$ position. It is a free form in the Corpus with the probability of

$$p(s_{i,j}) = \frac{freq(s_{i,j})}{corpus - size}$$

(87)

$$If \quad 2 * \frac{\sum f(k)}{k} < f(k_i) \text{ , then } k_i \text{ is a segmentation point}$$

Two single-word Turkish compounds, *çokbilmiş* 'wiseacre' = *çok* 'many' - *bilmiş* '[he has] known' and *yerçekimi* 'gravity' = *yer* 'land, earth' - *çekim*-CM 'attraction' are successfully segmented as *çok-bilmiş* and *yer-çekimi* as in Figures 6 and 7, respectively.

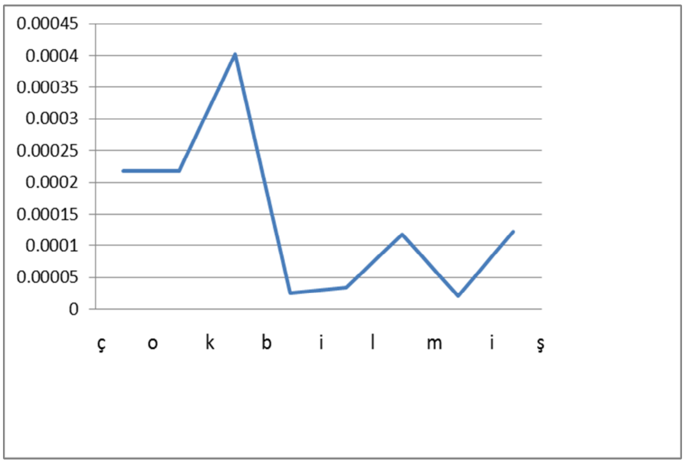


**Figure 6** Probability density graph for *çok-bilmiş* ‘wiseacre’

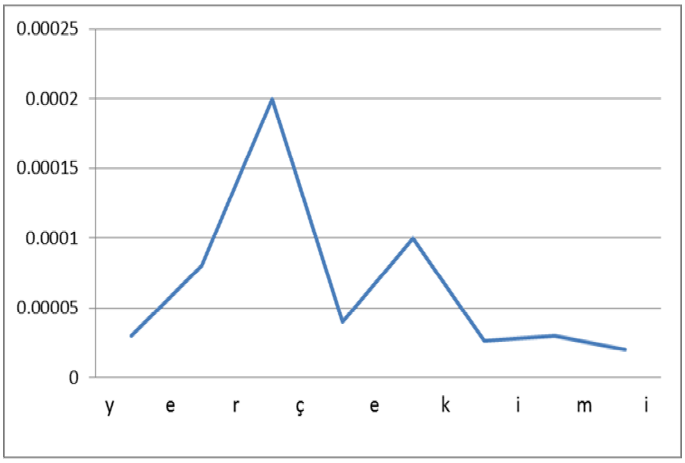


**Figure 7** Probability density graph for *yer-çekimi* ‘gravity’

The method was tested on the 1,524 words used in the recognition task. The probabilities were evaluated from the METU-Turkish Corpus and the BOUN web corpus. The method failed to recognize the segmentation of the compounds with constituents having substrings that resemble frequent free-forms in Turkish. For example, *devetüyü* ‘light brown’ = *deve* ‘camel’ - *tüy*-CM ‘hair’ was segmented as *de-ve-tüyü* because of the substrings *de* and *ve* were highly frequent connectives in Turkish. Similarly, *ayakkabı* ‘shoe’ = *ayak* ‘foot’ - *kap*-CM ‘cover’ was segmented into *ay-ak-kabı* because *ay* ‘moon’ and *ak* ‘white’ were more frequent than *ayak* in the corpora. *Kediotu* ‘valerian’ = *kedi* ‘cat’ - *ot*-CM ‘grass’ were to be segmented as *kedi-o-tu* due to a very frequent word *o* ‘that’ in both corpora. Since there was no free word as *tu* in both corpora, the free-word assumption was not satisfied; therefore, *kediotu* was not segmented into its constituents. Furthermore, the web corpus had a worse success rate despite its tremendous size, which indicated that global statistics failed if the sample size increased.

The probabilities of the most frequent 100 words were decreased by.8 in both corpora. These words are usually pronouns, connectives and some frequent verbs, which rarely occur in compounding. The method achieved (.87, .82, .84) precision, recall and *f*-score measures for the METU-Turkish Corpus and (.83, .80, .81) precision, recall and *f*-score measures for the web corpus. When the inputs were *görebil* ‘[you] to be able to see’ and *gidiver* ‘[you] go suddenly/swiftly', they were segmented as *gör-e-bil* and *gidi-ver*. This is because of the resemblance of *-(y)Abil* and *-(y)Iver* to free-forms *bil* 'know' and *ver* 'give'. These two formations are actually not compounding but morphological operations with *affixoids*.

Compounding is a cross-linguistically wide phenomenon occurring semantically in a similarly way to a derivational process with constituents that are free-forms unlike derivational affixes. Compounds are stored in the lexicon but their constituents need to be accessed during both learning and producing the compounds. The recognition method indicates that a morphemic lexicon is useful in identifying a word as a compound or non-compound word. The segmentation method shows that probability densities can be utilized in identifying constituents in a compound word. However, as the sizes of corpora increase, global statistics fails in segmentation.

### 4.3.3 The HMM and Morph Segmentation

The HMM is a statistical Markov model introduced by Baum and Petrie (1966) to evaluate the probability of a sequence of observations. The representation of an HMM is $HMM = (S, A, B, \pi)$ with the following elements:

- A set of states $S = s_1, ..., s_n$
- A set of transition probabilities $A = a_{11}, a_{12}, ... a_{nn}$ in which $a_{ij}$ represents the probability of transitioning from the state $s_i$ to $s_j$.
- Emission probabilities: A set $B$ of functions of the form $b_i(o_t)$ which gives the probability of observation $o_t$ being emitted by $s_i$.
- Initial state distribution: $\pi$ is the probability that $s_i$ is the initial state.

Figure 8 shows a representation of the HMM.

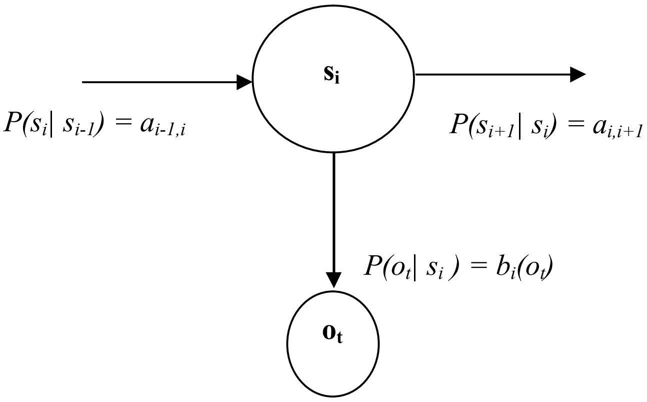


**Figure 8** The Representation of an HMM

HMMs have the Markov chain property:

- $P(s_{ik} \mid s_{i1}, s_{i2}, \ldots, s_{ik-1}) = P(s_{ik} \mid s_{ik-1})$

  this means that the probability of each subsequent state depends only on the previous single state.

An HMM posits 3 basic problems:

1. The Learning Problem: Given an HMM *M*, and a set of observations $O = \{o_1, o_2, ..., o_n\}$, how should we adjust model parameters $(A, B, \pi)$?

2. The Decoding Problem: Given an HMM *M*, and a set of observations $O = \{o_1, o_2, ..., o_n\}$, what is the most likely state sequence in the model that produced the observations?

3. The Evaluation Problem: Given an HMM *M*, and a set of observations $O = \{o_1, o_2, ..., o_n\}$, what is the probability that the observations are generated by the model, $P(O \mid M)$?

The solution to the learning problem is the evaluation of the initial probabilities from the Corpus and the Treebank. The solution to the decoding problem actually results in morph segmentation at the same time. Finally, solving the evaluation problem means checking if a word obeys morphotactics of the target language.

In the current study, the set of states were *n*-grams starting from unigrams to the longest word. The emission probabilities represent the likelihood of the *n*-grams for emitting possible orthographic co-occurrences. For example, the total number of types for the trigram is $29^3$, of which only 8,948 occur in the Corpus with 6,374,844 tokens, and the possible emissions are estimated by exhaustively searching through the Corpus. Similarly, for a given word, all possible *n*-grams are produced. Then, the initial probabilities of transitions and emissions for the corresponding *n*-grams were estimated through training the model on the Corpus; thus, the learning problem was solved.

Then, the Viterbi Algorithm chooses the best path and solves the decoding problem. The algorithm is modified to select the top 3 possible paths. The paths return segmentations. Additionally, the Markov Chain property is morphologically plausible because in Turkish morphotactics the concatenation possibility of a morpheme is determined by the recent suffix attached to a stem or by the stem itself. For example, the suffix *-ki* to form pronominal expressions can only be attached to words with the latest suffix that is either GEN or LOC.

The Viterbi algorithm is a recursive optimal solution to the problem of estimating the state sequence of a discrete time finite-state Markov process observed in a memoryless way (Forney, 1972; Jurafsky & Martin, 2000) as given in (88).

(88)
$$\delta_t(j) = \left[\max_{1 \le i \le N} \delta_{t-1}(i) a_{ij}\right] b_j(o_t)$$

In (88), $a_{ij}$ and $b_i(o_t)$ parameters are same as those in the HMM, and the  function gives a path through states. The evaluation algorithm works in a forward recursive manner and requires backtracking to find the best path of states. The HMM described above is totally unsupervised without any assumption except that Turkish is concatenated from left to right.

In Figure 9, a very simple trellis diagram shows the possible state transitions for the word *kedim* 'my cat', which also corresponds to the possible segmentations. The transition probabilities among the corresponding *n*-grams and emission probabilities of each *n*-gram are also given in Tables 22 and 23 respectively.

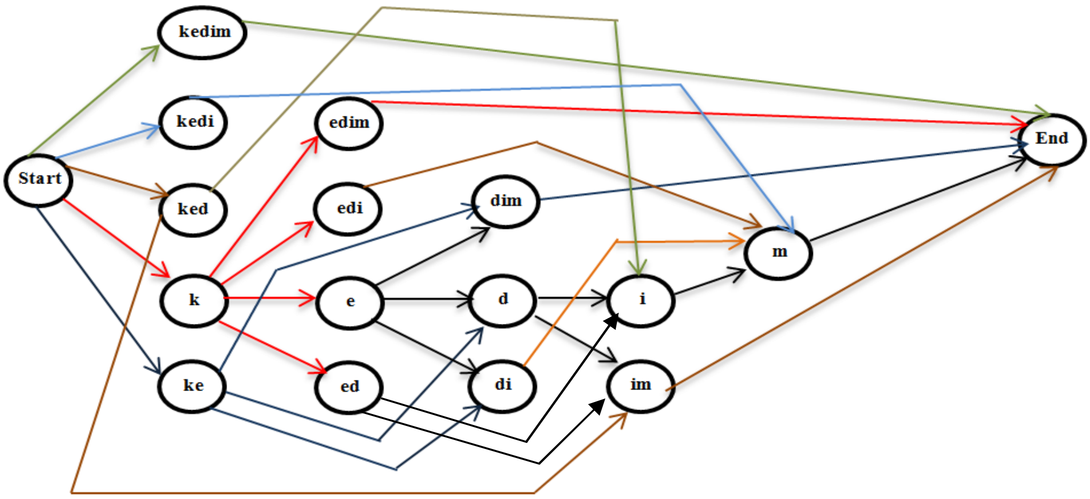


**Figure 9** Trellis diagram for *kedim* ‘my cat’

**Table 22** Transition probabilities of the *n*-grams

| | **Start** | *kedim* | *kedi* | *edim* | *ked* | *edi* | *dim* | *ke* |
|---|---|---|---|---|---|---|---|---|
| **Start** | | 1.82E-05 | 8.66E-03 | | 1.87E-03 | | | 1.34E-01 |
| *kedim* | | | | | | | | |
| *kedi* | | | | | | | | |
| *edim* | | | | | | | | |
| *ked* | | | | | | | | |
| *edi* | | | | | | | | |
| *dim* | | | | | | | | |
| *ke* | | | | | | | 1.19E-04 | |
| *ed* | | | | | | | | |
| *di* | | | | | | | | |
| *im* | | | | | | | | |
| *k* | | | | 1.03E-05 | | 6.12E-04 | | |
| *e* | | | | | | | 2.05E-03 | |
| *d* | | | | | | | | |
| *i* | | | | | | | | |
| *m* | | | | | | | | |
| **End** | | | | | | | | |

**Table 22 (**cont'd.**)** Transition probabilities of the *n*-grams

| **Start** | *ed* | *di* | *im* | *k* | *e* | *d* | *i* | *m* | **End** |
|---|---|---|---|---|---|---|---|---|---|
| *kedim* | | | | 4.02E-01 | | | | | |
| *kedi* | | | | | | | | | 2.50E-01 |
| *edim* | | | | | | | | 1.67E-01 | |
| *ked* | | | | | | | | | 2.10E-02 |
| *edi* | | | 6.02E-03 | | | | 3.60E-01 | | |
| *dim* | | | | | | | | 5.54E-02 | |
| *ke* | | | | | | | | | 2.65E-01 |
| *ed* | | 7.10E-03 | | | | 1.97E-02 | | | |
| *di* | | | 3.05E-02 | | | | 5.51E-01 | | |
| *im* | | | | | | | | 4.90E-02 | |
| *k* | | | | | | | | | 4.26E-01 |
| *e* | 1.70E-03 | | | | 8.62E-02 | | | | |
| *d* | | 3.70E-02 | | | | 6.73E-02 | | | |
| *i* | | | 9.12E-03 | | | | 1.86E-00 | | |
| *m* | | | | | | | | 7.14E-02 | |
| **End** | | | | | | | | | 1.75E-01 |

**Table 23** Emission probabilities of the *n*-grams (ε denotes empty string)

| **Output** | **N5** | **N4** | **N3** | **N2** | **N1** | **Start** | **End** |
|---|---|---|---|---|---|---|---|
| *kedim* | 3.81E-06 | | | | | | |
| *kedi* | | 4.73E-05 | | | | | |
| *edim* | | 2.31E-04 | | | | | |
| *ked* | | | 1.04E-04 | | | | |
| *edi* | | | 3.30E-03 | | | | |
| *dim* | | | 5.43E-04 | | | | |
| *ke* | | | | 4.40E-03 | | | |
| *ed* | | | | 4.99E-03 | | | |
| *di* | | | | 9.22E-03 | | | |
| *im* | | | | 4.93E-03 | | | |
| *k* | | | | | 5.23E-02 | | |
| *e* | | | | | 7.62E-02 | | |
| *d* | | | | | 5.08E-02 | | |
| *i* | | | | | 7.06E-02 | | |
| *m* | | | | | 4.12E-02 | | |
| ε | | | | | | 1.00E00 | 1.00E00 |

The Viterbi algorithm chooses the path as (*Start, N4, N1, End*) = (*ε, kedi, m, ε*), in which ε denotes the empty string. This is the correct sequence of the morphs in the word. The second most probable path, which is slightly closer to the first path, is (*Start, N2, N2, N1, End*) = (*ε, ke, di, m, ε*) because of high number of occurrences of the past tense suffix *-di* in the Corpus. This is an incorrect segmentation. In a corpus containing significantly more verbs than nouns the second path would be a correct alternative.

Finally, the evaluation problem needs to be handled. Given a set of observations, a typical HMM can evaluate whether this observation is probable or not, through the forward algorithm, the backward algorithm or the forward-backward algorithm. Since Turkish greatly employs suffixation, the forward algorithm was selected to solve the evaluation problem given in (89).

(89) $$P(O \mid M) = \sum_{i=1}^{N} \alpha_T(i) \ \ where\ \alpha_T(i) = p\{o_1, o_2, \ldots o_t, q_t = i \mid M\}$$

For example, two words *evlerdekilerinkiler* 'the ones belong to the ones in the houses' and **uyudumyor* can easily be segmented as *ev-ler-de-ki-ler-in-ki-ler* and **uyu-du-m-yor* by Turkish native speakers although the valid formation of the second one would be *uyuyordum* 'I was sleeping'. Since there was no transition from *-m* to *-yor* in the Treebank, a backoff algorithm was implemented to smooth the value. The details of smoothing are explained in Section 4.3.5. The HMM in the current study successfully performed the segmentation although these words did not occur in the Treebank and the Corpus. However, when the observations $O_i$ = (*ev, ler, de, ki, ler, in, ki, ler*) and $O_j$ = (*uyu, du, m, yor*) were evaluated by the forward algorithm, it was seen that $P(O_i \mid M) > 0$ and $P(O_j \mid M) = 0$. Since there was no transition between *-m* 1.SG and *-yor* PROG in the Treebank, $O_j$ = (*uyu, du, m, yor*) was not an acceptable observation by the HMM given the current history. In other words, existing positive evidence was enough to judge the acceptability of an unseen observation whether it was valid or not. Similarly, another observation $O_k$ = *(gör, ece, m)* for *?görecem* which was a colloquial usage of *göreceğim* 'I will see [it]' was evaluated by comparatively using the data from manual segmentations of the Treebank and CHILDES. Since the CHILDES database had colloquial usages of the future tense -(*y*)*AcAK* and its corresponding transitions, it was evaluated as $P(O_k \mid M_C) > 0$ by CHILDES but rejected by the Treebank as $P(O_k \mid M_T) = 0$.

### 4.3.4 Assumptions and Improvements for the HMM

A subset of the METU-Sabancı Turkish Treebank containing 5,010 words was manually segmented. The Treebank originally consisted of 7,262 sentences with 43,574 words. In this task, both derivational and inflectional affixes were segmented. The allomorphs, such as the plural suffixes *-lar* and *-ler* or derivational suffixes *-lik* and *-liğ* were treated as different morphs and they are the emissions of trigrams.

With respect to their orthographic length, the segments corresponded to *n*-grams in the HMM. Similarly, the orthographic distribution of the segments with respect to their length corresponded to the emission probabilities in the HMM. The statistics from the manual segmentation were used to improve the model by attempting to reduce the number of false segmentations and oversegmentations. The average root length of the subset from the Treebank was about 4 characters. There were 150 derivational and 214 inflectional morphemes in the subset and this became the gold standard for the subset in the current study. The inflectional suffixes were very frequent; however, the derivational suffixes were not nearly so frequent. For example, in the segmentation of the first 100 words, 59 new morphemes were discovered, of which only 6 were derivational. To understand the reason for the oversegmentation of roots by the HMM, the statistics of the 5,544 distinct roots with endings identical to morphemes in the gold standard had were from the Treebank, as shown in Table 24. For example, the most frequent root termination had the ending *-n* (10.82%).

**Table 24** Root ending in Treebank

| Root Ending Segment | Percent in the Treebank |
|---|---|
| *n* | 10.82% |
| *k* | 10.13% |
| *t* | 9.59% |
| *a* | 9.56% |
| *e* | 8.25% |
| *r* | 7.69% |
| *i* | 6.02% |
| *et* | 4.71% |
| *m* | 4.60% |
| *an* | 3.86% |
| *ş* | 3.18% |
| *ı* | 3.04% |
| *ol* | 2.41% |
| *la* | 2.32% |
| *u* | 2.32% |
| *er* | 2.14% |
| *le* | 2.00% |

These edge statistics were incorporated in the model as follows; if the sum of the indices of visited states (a measure of length) was close to the calculated average root length 4.09, and if in the current state a symbol identical to one of our morpheme endings *x* from Table 24 was observed, then the state's transition probability was multiplied by (1- *percentage-of-x*), which gave the probability of *x* not being an edge of one the roots from the Treebank. For example, if a unigram was in the 4th orthographic position of a word and it emitted *-n*, then its transition probability was multiplied by (1 - 0.1082). This was a simple way to check the effect of the edge statistic on the oversegmentation of roots, because it forced the Viterbi algorithm to favor the likely endings of roots and morphemes. Next, the false segmentation problem of morphemes was tackled. The statistics from the segmented subset were used for this purpose to look at the structure extending beyond the average root length. For example, *-lArI* (3.PLU.POSS) and *-lar-I* (PLU-ACC) are identical orthographically, hence they are prone to false segmentation. Manual segmentation showed that there were 190 occurrences of the latter, of which 59% had at least one more segment before the word boundary. On the other hand, 3Plu.Poss occurred in 40 words of which 30% were in word boundaries. The statistics of such problematic cases were part of the experiments in the current study. Their (1- 'edge probabilities') were multiplied with the transition probabilities of the HMM considering the locations and emission types of the states. For example, if *-lArI* had the transition probability of .085, and *-lAr* .075, and if 70% of *-lArI* were not at the word boundary compared to 59% for *-lAr*, determined from supervision, the numbers (1-.7)x.085 and (1-.59)x.075 would be the contenders. By doing so, the Viterbi algorithm was partially directed to a path starting with a 3-gram (PLU) instead of a 4-gram (3.PLU.POSS) for *-lArI* representations occurring before the word boundaries.

As a further improvement, the Treebank was completely segmented. The transition and emission probabilities from the Treebank and the Corpus were averaged to obtain the best segmentations. For example, the top three segmentations of the word *evdekiler* 'house-LOC-REL.Ki-PLU means the ones in the house' were *ev-de-ki-ler, *ev-de-kiler* and **ev-de-ki-le-r*. The segmentation **ev-de-kiler* was given as the second most probable because it contained a substring resembling a free word *kiler* ‘pantry’. The corresponding transition and emission values from the Corpus are given in Appendix A. The transition and emission probabilities were also evaluated from the manual segmentation of the Treebank as in shown Appendix B.

In this study, when the HMM failed to find a root placed in the leftmost position, the existence of 4 important morphophonological operations in the roots, Turkish emphatic reduplication, deletion, epenthesis and voicing, were checked and reversed. For example, the word *affina* ‘forgive-2.SG-DAT meaning to your forgiveness’ could not be successfully segmented because of the epenthesis of the extra *f*. Since there was no root in the form of **aff* and none of the Turkish bounded morphs contained *f,* the model produced **affi-n-a, *affi-na* and **affin-a.* Thus, the extra *f* should be deleted before the segmentation. The corresponding transition and emission probabilities for the word *affina* from the Corpus are given in Appendix C and the same probabilities from the Treebank are presented in Appendix D. The reduplication, deletion, epenthesis and voicing operations are not morphemes but phonologically driven root modifications. The following actions were taken to improve the model:

- The existence of the emphatic reduplication is checked via the template in (79). The substring following the linker is sent to the HMM. Yet, there are Turkish words which appear to be emphatically reduplicated. For example, *çerçeve* ‘frame’ is in the form of $C_1V_1$(*p, m, r, s*) $C_1V_1$… Since there is no Turkish word like **çeve*, the word is re-sent to the model as a whole by undoing the emphatic reduplication modification.
- The application of deletion is validated via the Corpus. In all deleted stems, the fourth and the third leftmost orthographic representations are consonants. The substring consisting of the four leftmost orthographic representations of the word is not a freeform in the Corpus. Turkish vowels are sorted according to the roundedness and backness properties of the succeeding vowel, not the preceding one because the deletion occurs after concatenation as in the *vakti* example in (66). Then, the candidate vowels are inserted between the third and fourth orthographic representations one by one. The existence of the new substring in the Corpus is checked at each step. Whenever a possible free form is found, the new substring is assumed to be the root and the string is sent to the HMM.
- The possibility of epenthesis is investigated by the leftmost and immediate double co-occurrences of Turkish fricatives, *s* and *f*, nasals, *n* and *m,* and the post-alveolar *l.* One of the doublets is deleted, then the word is sent to the HMM.
- The voicing in root is reversed if a voiced consonant exists in or later than the third leftmost position. Unvoicing is performed once and the word is resent to the HMM.

Then, the model produced *af(f)-ın-a* as the top segmentation, the transition and emission probabilities from the Corpus are given in Appendix C. In Appendix D, the transition and emission probabilities of the word from the Treebank are represented.

To summarize, the final model was composed of the HMM using the unannotated corpus (1.7 million words) and the supervision from the manually segmented treebank (43,574 words), compound word recognition and segmentation, and the morphophonological root alternations. The smoothing and backoff for the model is discussed below.

### 4.3.5 Smoothing and Backoff for the HMM

The statistical morphological language model in this study involves frequencies of the substrings in a corpus of words to explain morph decomposition and acquisition. The underlying proposition is that the substring frequencies and co-occurrences within a corpus offer morphological patterns. These patterns can be captured by statistical models (such as, Support Vector Machines, HMM, entropy, Bayesian, and such models) to be used in morphological segmentation. One important aspect of the statistical models is *smoothing*. Since there are combinations of possible morph sequences which are rare or never occur in a corpus (sparse data), the machine learning models assign a zero probability to them. If a new combination of morphs is seen during testing or a new morph is perceived, they will not be segmented. Therefore, model parameters are smoothed and the probability mass is reassigned to unseen morphs and their co-occurrences.

There are many smoothing techniques and their combinations are employed in Machine Learning. The simplest smoothing technique is the *add-one* or *Laplace smoothing* in which the frequencies of unseen morphs are increased by adding one while the frequent morphs are decreased by one. Yet, there might be a large number of unseen morphs and their combinations, in which case too much weight would be assigned to the unseen *n*-grams. The other common method is the Good-Turing discount (Good, 1953), this re-estimates the amount of probability mass for zero (or low count) *n*-grams by looking at *n*-grams with higher counts. In particular, this method reallocates the probability mass of *n*-grams that were seen once (twice, three times or more) to the *n*-grams that have never been seen. For each count *r*, we compute an adjusted count $r^*$ as;

(90) $$r^* = (r+1)\frac{n_{r+1}}{n_r}$$ where $n_r$ is the number of *n*-grams seen *r* times.

There might be a problem in the Good-Turing smoothing if $n_{r+1}=0$ in that there might be zero counts in the next *n*-grams. The backoff model proposed by Katz (1987) offers a solution if the conditional probability of an *n*-gram is zero, its probability is re-evaluated from lower level *n*-grams iteratively. The Katz's backoff algorithm, which is given in (91), has influenced other methods.

(91)

$$P_{bo}(w_i \mid w_{i-n+1}...w_{i-1}) = \begin{cases} d_{w_{i-n+1}...w_i} \dfrac{C(w_{i-n+1}...w_{i-1}w_i)}{C(w_{i-n+1}...w_{i-1})} \; if \; C(w_{i-n+1}...w_i) > k \\ \alpha_{w_{i-n+1}...w_{i-1}} P_{bo}(w_i \mid w_{i-n+2}...w_{i-1}) \; otherwise \end{cases}$$

In the formula given in (91), *C* is the counting function. α and *d* are used to normalize probability mass so that it still sums to 1, and to smooth the lower order probabilities that are used. For example, if there are no counts for a trigram model, then there is a backoff to the bigram estimation. If the bigram estimation is zero, then the unigram estimation α and *d* are evaluated (see Katz, 1987). If a constant value, *λ*, is evaluated and involved throughout the model, this is called *linear interpolation* (Jelinek and Mercer, 1980). This is a mixture of backoff models with add-*x* models because it interpolates unseen *n*-grams with values evaluated from seen ones. Finally, Kneser and Ney (1995) proposed an interpolated version of a backoff algorithm. They calculated the probability of an *n*-gram by computing the raw probability of the *n*-gram following a context (seen *n*-grams) and subtracting a discounting amount. This discounting amount was then re-added equally to all *n*-gram probabilities following the same context as the continuation values given in (92).

(92)
$$P_{KN} = \frac{C(w_{i-1}w_i) - D}{C(w_{i-1})} + \beta(w_i)P_{CONTINUATION} \text{ } where$$

$$P_{CONTINUATION} = \frac{|\{w_{i-1} : C(w_{i-1}w_i) > 0\}|}{\sum_{w_i} |\{w_{i-1} : C(w_{i-1}w_i) > 0\}|}$$

The *C* function in (92) gives the number of occurrences in the Treebank. The numerator in $P_{CONTINUATION}$ is composed of the number of morph types seen to precede $w_i$. The denominator is the number of morphs preceding all morphs. *β(w$_i$) = C(w$_i$)/(Total Number of Transitions)*. *D* is a number (e.g., .50) subtracted from every count. The Kneser-Ney algorithm pays attention to the number of contexts in which a morph occurs. In other words, it makes use of previous transition frequencies to a morph to assign a value to some unseen transitions to this morph. Zhai and Lafferty (2004) compared smoothing techniques in the information retrieval field and concluded that interpolation models are better than the backoff and add-one models. Similarly, Chen and Goodman (1999) compared smoothing techniques for different *n*-grams (for words) on corpora of different sizes. They concluded that Katz smoothing was better for trigrams on larger data sets while it was acceptable for bigrams in an average size corpora (with few hundred thousand words). Chen and Goodman (1999), then, reported that Kneser-Ney smoothing consistently outperforms all the other algorithms evaluated in a speech recognition task.

Previous studies have shown that smoothing depends on the task, algorithm and corpus size. The idea beneath all smoothing strategies is that known data can provide clues about unseen data. It is also cognitively plausible that from known patterns arise expectations for newcomer data. In this study, a combination of smoothing techniques is used because there were two kinds of problems. Firstly, the training corpora might have no emission probability for an *n*-gram. In other words, a word might have a substring (or morph) which does not occur in the corpora. Secondly, the training corpora might lack a transition probability for an *n*-gram in a context. That is, the corpora have no immediate co-occurrences for two *n*-grams.

In this study, two cases needed smoothing: probability mass was reassigned to unseen morphs (emissions) and unseen co-occurrences (transitions). The Good-Turing discount (Good, 1953) was employed for unseen emissions while the Kneser-Ney algorithm (Kneser & Ney, 1995) was utilized in unseen transitions.

For example, a nonce word *üj* which did not occur in both the Treebank and the Corpus was employed as a root in the formation of **üjlerimin* 'of my *üjs*'. The Good-Turing smoothing reallocated the existing bigram probabilities in the Treebank and the Corpus to smooth the emission probability of *üj* as *.20e-6*, which had initially been zero. Then, the Kneser-Ney's interpolation was used to smooth the transition probabilities from *üj-* to *-ler*, *-le* and *-l* which were annotated as morphs in the Treebank. The smoother transition values were *.03*, *.004* and *.20e-6* respectively. Eventually, the segmentation was *{Start, N2, N3, N2, N2, End}* = *{üj, ler, im, in}*.

### 4.2 Results

The results of the model were received before and after the incorporation of the subset containing 5,010 words. Then, the statistics from the manual segmentation of 43,574 words, compound word handling and the morphophonological operations checking were used. In the final model, the corresponding transition and emission probabilities evaluated from both the Corpus and the Treebank were averaged because, despite their huge fragmental difference, the semi-supervision was assumed to be equally effective as in the case of children's language acquisition. The details of the results are given in the next sections.

### 4.2.1 Results of the Method before and after 5,010 Words

The initial unsupervised HMM model achieved the precision, recall and f-measure values of .51, .72 and .59 respectively, which were not satisfactory. To reduce oversegmentations of roots and false segmentations of affixes, the co-occurrences of root endings and morpheme starts discovered by the manual segmentation of 5,010 words were incorporated into the model. Employing this much semi-supervision from a very small fragment (0.29%) of the database successfully increased the precision, recall, *f*-measure values to .72, .87, and .79 (precision, recall, *f*-measure respectively), from the unsupervised method with over 1.7 million unlabeled words. Considering the knowledge-poor strategies employed, and the fact that nothing had been undertaken to compensate for overfitting in advance, this was quite striking, and showed more avenues moving towards unsupervised and semi-supervised segmentation. Of 5,010 there were 1,838 words that were either roots or compounds, which seems to be a representative percentage. It should also be noted that the model obtained a .79 *f*-measure of correct segmentation into morphemes. Thus, the model delivered the morphs, not just the overall tag for the word, without the morphological analysis. What manual segmentation provided was syntactic and semantic disambiguation in an indirect way, hence some semantic-phonological cues (such as intonation, stress) and some limited syntactic knowledge (e.g. for compounds), were next targets that the model addressed.

### 4.2.2 Results of the Method after Assumptions, Improvements and 43,574 Words

Compound word recognition and segmentation were incorporated into the model. Moreover, the manual segmentation of 43,574 words was completed and utilized in the model as semi-supervision. The statistics of the manual segmentation of 43,574 words was used as the information from the segmentation of the 5,010 words. This time, the fragment of the manually segmented set of words (43,574) and the unannotated words (1.7 million) was 2.5%. Finally, handling the morphophonological operations on roots was embedded into the model as described in section 4.3.4. When the top 3 segmentations were considered, the method achieved .88, .92 and .90 precision, recall and *f*-score measures respectively. Table 25 summarizes the achievements.

**Table 25** Achievement scores of HMM

| The Method | Precision | Recall | *F*-score |
|---|---|---|---|
| Unsupervised | .51 | .72 | .59 |
| Semi-supervision from 5,010 words[7] | .72 | .87 | .79 |
| Semi-supervision from 43,574 words | .88 | .92 | .90 |

[7] This part of the work was presented in LREC 2012, Istanbul (Kılıç & Bozşahin, 2012).

# CHAPTER 5

## COGNITIVE ASPECTS OF SEMI-SUPERVISED METHODS

The semi-supervised method was successful in morphological segmentation of Turkish words. However, another concern of this study is the cognitive plausibility of semi-supervision. Tenenbaum et al., (2011) reviewed the construction of the human mind focusing on the following three central questions:

- How does abstract knowledge guide learning and inference from sparse data?
- What forms of abstract knowledge take across different domains and tasks?
- How is abstract knowledge acquired?

They proposed that cognition and its origin could be understood in terms of Bayesian inference over richly structured and hierarchical generative models with abstractions. The Bayesian formula is recalled in (93) where *D* is a data set (search space) and *h* stands for a hypothesis.

(93) $$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

The Bayesian approach provides a unification of mathematical language for inductive learning and cognitive models with minimum parameters and assumptions. The main assumption of Bayesian learning is learning from *prior* information. The prior information in Bayesian learning is not necessarily UG. Instead, it is previously exposed linguistic data. In this study, the probability estimations were performed as in the Bayesian formula. Various Bayesian models have been proposed in cognitive science and language acquisition (Frank et al., 2009; Goldwater et al., 2009; Griffiths & Tenenbaum, 2006; Perfors et al., 2011a; Xu & Tenenbaum, 2007). For example, the probability of a Turkish word starting with a vowel was estimated as in (94) as given below.

(94) $$P(V \mid C) = \frac{P(C \mid V)P(V)}{P(C)}$$

$$where\ V = \{a,e,ı,i,o,ö,u,ü\}\ in\ initial\ position$$

$$C = Metu - Turkish\ Corpus$$

The prior information is the Corpus, the list of Turkish vowels and the knowledge that the starting point of Turkish words is the leftmost position. If the probability of a Turkish word starting with a vowel succeeded by a consonant is questioned, then the corresponding search space will have the probability estimated in (94) and the formula will be evaluated over

Turkish consonants in the second leftmost position. In other words, statistical algorithms act incrementally or iteratively as in the HMM because it is also a dynamic Bayesian Network.

As children construct their mental representations of concepts and languages, they carry out a statistical analysis. If I was designing a natural language, I would almost ignore morphology and rely heavily on syntax in order to have a morphology-free grammar. However, natural languages are determined by evolution and are flooded with exceptions, irregularities and ambiguities in formations of both words and phrases. Communication demands a reciprocal way of transferring ideas that are generally turned into audible utterances and requires the mental skills to capture the systematicity in the utterances which allows understanding the ideas. From a very early age, children gain prior information from their environment in the form of words, utterances, visual input and feedback, which helps them in the vital task of communicating their needs and wants. Yang (2002; 2004) proposed that parameter-setting in the UG occur in a probabilistic way over the input children received. To summarize, a statistical learning algorithm with a Bayesian rule can be applied to model human-level segmentation and the acquisition of morphs in Turkish.

In terms of this being a cognitive science study, this study aims to investigate not only a computational model that segments Turkish words into morphs, but also its cognitive properties. Thus, in this chapter, the cognitive plausibility of statistical learning and the degree of supervision are reviewed.

## 5.1 Cognitive Reflections on Statistical Learning

The methods for the acquisition and segmentation of Turkish compound words, evaluating the acceptability of nonce words and the HMM used in the present study make use of the simple transition probabilities evaluated from raw text data as simply shown in (95).

(95) $P(B \mid A) = frequency\_of(AB) / frequency\_of(A)$

The methods in this study have biases in that cognitive models are probabilistic and inductive. As underlined by Griffiths et al., (2010), cognitive science aims to reverse-engineer the mind. Mental processes are modeled using algorithms approximating to the human-level success rate. It requires a top-down analysis of the processes but bottom-up modeling process. This is relevant to the emergentist philosophy since higher level explanations do not have independent validity but are approximations to the truth because they are described by a lower-level mechanism. Yet, the probabilistic models pursue a top-down or function-first strategy with the initial abstraction of the cognitive agent. Then, these models implement bottom-up inductions to undertake the abstraction.

Opponents of statistical learning methods support the innateness hypothesis. For example, children appear to favor hierarchical rules that operate on grammatical constructs such as phrases and clauses over linear rules that operate only on the sequence of words, even in the apparent absence of direct evidence supporting this preference such as auxiliary fronting (Chomsky, 1965; Chomsky; 1980). However, Perfors et al., (2011b) proposed and tested a model in which given typical child-directed speech and certain innate domain-general capacities an unbiased ideal learner could recognize the hierarchical phrase structure of language in a Bayesian way without having this knowledge innately specified as part of the language faculty. A similar study was also performed by Reali and Christiansen (2005).

The opponents of emprricisim also claim that neither infants nor adults can continuously count occurrences and calculate corresponding transition probabilities; and even if they could, they would not be able to perform these evaluations in such a short period of time.

However, Xu and Garcia (2008) showed that very young infants could make inferences from samples of populations. Infants of 11 and 12.5 months can integrate psychological and physical knowledge in probabilistic reasoning (Teglas et al., 2007; Xu & Denison, 2009). Kirkham et al., (2002) studied visual statistical learning in infancy observing that infants viewed familiar patterns alternating with novel sequences of identical stimulus components. At all ages the children displayed significantly greater interest in the novel sequence. Similarly, Arciuli and Simpson (2011) showed that statistical learning was effective in visual reading. This raises another question: Even if infants can count and carry out statistical analysis, can they employ these skills in linguistic domains? Many experimental studies indicate that the answer might be yes. As well as the experimental studies referred to above, there are linguistic studies (Aslin et al., 1998; Safran et al., 1996; Saffran & Thiessen, 2003) indicating the likelihood of domain general statistical learning in infancy. This means that there are structures for detecting inherent patterns in the environment and they may play an important role in cognitive development. As shown in Section 4.3.3, **uyu-du-m-yor* is an invalid formation. Although it was not observed in the Corpus, the Treebank and CHILDES, it was successfully evaluated as an unacceptable formation by the HMM using the forward evaluation algorithm. It is an implication that seen observations can be employed as indirect negative evidences. Thus, semi-supervised models, such as the current one, support the empricisim in the discussion of nature versus nurture.

Infants are reported to successfully discriminate speech segments using the transitional probabilities of syllable pairs (Aslin et al., 1998; Gomez, 2002). Saffran et al., (1996) stated that even 8-month-old infants pursued statistical learning. They concluded that adjacent sounds co-occurring with a high probability in language are usually found within words; yet, low probability sound pairs span word boundaries. This inverse ratio provides the potential information for word boundaries and it may further contribute to language acquisition by reinforcing the ability of segmenting speech into units.

Saffran and Thiessen (2003) showed that infants acquire the phonotactics of their languages statistically and became sensitive to these phonotactic patterns. In another study by Thiessen and Saffran (2004), they showed that frequency as a source of information could be used in a language-independent way, and seemed to be used by infants earlier than most of the other cues, by the age of 7 months. Similarly, Jusczyk (1999) experimentally showed that 7.5 and 10.5-month-old infants are able to segment words according to stress patterns, statistical regularities, allophonic cues and phonotactic patterns. He suggested that language learners might draw upon multiple cues to determine word boundaries in fluent speech and the task of having to attach meanings to sound patterns affect infants' abilities to segment words. Similarly, Saffran et al., (2008) performed comparative experiments on human and tamarin-monkey infants and concluded that the infants rapidly acquired complex grammatical structures by using statistically predictive patterns, failing to learn structures that lacked such patterns. Alishahi and Stevenson (2008) proposed a probabilistic model in which associations enabled the model to learn general conceptions of roles, based only on exposure to individual verb usages, and without requiring explicit labeling of the roles in the input. Similarly, Kwiatkowski et al., (2012) implemented an incremental probabilistic learner that models the acquisition of syntax and semantics from a corpus of child-directed utterances paired with possible representations of their meanings. They also explained that lexical items can be acquired on a single exposure and word order is learnt suddenly rather than gradually. Sudden learning can occur as rule-learning and can be determined by pragmatics.

Although infants cannot count hundreds of occurrences, they might have mental representations to keep track of types and token counts. Moreover, it does not have to be an intentional or conscious behavior. For example, Turkish native speakers in the emphatic reduplication task given in the previous chapter instantaneously selected the appropriate linker types that have co-occurrence frequencies with the initial consonants of words were comparatively lower. Similarly, for the nonce word acceptability task, they evaluated the words in seconds but the same task was modeled by counts and transitions in the Corpus.

Furthermore, through experiments on children with language impairment, Evans et al., (2009) found that IQ did not mediate the relationship between statistical learning and known vocabulary. Similarly, Newman et al., (2006) discovered that the relationship between infants' ability to segment speech into individual words via statistical learning and their later proficiency with natural language was not brought about by IQ but through memory. In other words, human beings have a statistical learning ability and it is not specific to linguistics but to general cognition. Although it is experimentally proven that statistical learning takes place in the linguistic domain, it is not an appropriate tool for every question in cognitive science. Statistical approaches address inductive problems, for example, they cannot model *attention* and *IQ* without appropriate extensions. The other limitation of statistical learning using the Bayesian method is that the model will be wrong if the underlying assumptions about prior knowledge are incorrect. Moreover, it cannot make computational level (Marr, 1982) assumptions about the human mind; but it can devise the specification of the problem or the goal of the learning (Perfors et al., 2011a). Yet, it is the best fit explanation of the experimental data in the inductive problems of cognitive science.

To conclude, statistical learning has been shown to be cognitively plausible both experimentally and computationally. As a cognitive concern, the degree of supervision needs to be briefly discussed.

## 5.2 Degree of Supervision in Morphology Learning

A child growing up under normal circumstances incrementally builds his or her mental lexicon and acquires the grammar of his or her mother tongue. The child displays an astonishing competence in language given utterances, meanings, contextual information, and responding to environmental and parental feedback. However, language acquisition is a life-long process. There is no hypothetical upper limit on the capacity of learning new vocabulary and the acquisition of new grammatical constructions because languages can change even in the life-time of a person. Thus, the lexicon must be dynamic.

There is virtually no way of knowing the exact structure of mental lexicons including the representations of form and meaning fed by linguistic, visual, auditory, olfactory, gustatory, and tactile input. Practically, it is possible to simulate lexicons as written and spoken corpora for the linguistic work in this study. Marquis and Shi (2012) performed experiments on 11-month-old French learners. The infants learned the new suffix and used it to interpret novel affixed words that had never occurred during training. These findings demonstrate that the initial learning of sub-lexical functions and morphological alternations is frequency-based, and does not rely on word meaning. This study validates the use of the corpus as a source of morpheme acquisition in this study. The next issue is the degree of supervision that is either inherited or learned.

The critical period hypothesis states that language is linked to biological age. It is claimed that if a child grows up in an isolated environment and misses the critical period of linguistic exposure, he or she will never be able to gain full competence in the target language. Similarly, adults suffer more difficulties in learning new languages. Ioup et al., (1994) set a context that lacked formal instruction and more closely resembled the environment in which the first language acquired. After the experiments conducted with adults, they concluded that a native-like competence was possible. Friederici et al., (2001) observed event-related brain activations of adults and concluded that adults who learned a miniature artificial language displayed a similar real-time pattern of brain activation when processing this language as native speakers do when processing natural languages. In other words, the difference between L1 and L2 learning may not lie in biology but motivation. Even a well motivaed L2 learner has a limit compared to L1.

During first language acquisition, infants desperately need to communicate to survive. They also need to construct their identities and express themselves. Since these motivations are extinguished or mitigated in time, they lead to differences in competences. Feral children achieve proficiency in learning a language but not as successful as the children who grew up under normal conditions. The reason could be that they have already somehow achieved a way of interacting, expressing themselves, surviving and constructing their identities.

For children who grow up normally, supervision is provided by the environment in which the child lives. It mainly includes the sentences parents uttered in context and the corrections made by parents on the child's utterances. There is abundant evidence that parents can provide children with finely adjusted and sensitive input (Snow, 1995). Studies concerning child language acquisition depend either on socialization theory, learning theory, or nativist theory. As expected, the nativist theory claims that a genetic module is programmed to acquire language whether feedback exists or not. The socialization theory states that language is acquired through social interactions, and the learning theory is based on empiricism; claiming that language acquisition occurs by learning from data. There are connectionist implementations with artificial networks that learn language with or without feedback (MacWhinney & Leinbach, 1991; Plunkett & Marchman, 1991; Rumelhart & McClelland, 1986).

In this study, learning theory is assumed to be more plausible in terms of morpheme acquisition. Yet, it should be noted that supervision does not solely depend on parental feedback. Children being able or unable to attend to objects and actions in context are a form of feedback. An erroneously uttered word might not receive a correction but at the same time the owner of the word might be deprived of the object or the action he or she intended to carry out. Children can even set up their own virtual environment. Easterbrooks and Baker (2002) asserted that play skills aid children in developing pre-linguistic skills. They claimed that as children manipulated toys and they developed the ability to represent objects, actions, their descriptions, and their relationships as a precursor to representing these through language. This is also a form of feedback. Thus, children are exposed to partial supervision while acquiring language and the portion and magnitude of this supervision might be greater than expected.

The sorted raw frequencies from the CHILDES Database show that there is a high parallelism between the morphs the parents uttered and those that their children uttered. In other words, the morphs children acquired were mostly what the parents uttered. It can be considered as an implication that parental supervision is an important aspect of child morphology acquisition. Yet, most of the data to which children are exposed is unstructured, unbiased, unlabeled, raw and, eventually, unsupervised. Combining this raw data and parental supervision results in the semi-supervision used in this study. The METU-Turkish Corpus acts as the raw data source. While evaluating the transition and emission probabilities, no assumption was made concerning the morphological structure of about 1.7 million words in the Corpus. Simply, the *n-grams* sliding from left-to-right were used in frequency and transition estimations. However, the supervision was achieved by the hand-made segmentation of about 47,000 words in the METU-Sabancı Turkish Treebank. Instead of character by character sliding *n-grams*, the transition and emission probabilities were evaluated from the morphs whose boundaries had been marked by minus signs or word boundaries.

In conclusion, implementing a semi-supervised HMM in this study is plausible not only practically but also cognitively. It opens more paths to the implementation of semi-supervised methods.

# CHAPTER 6

# CONCLUSION AND DISCUSSIONS

Morphology is the dominant subdiscipline in linguistics because it is the study of word structure, and words are at the interface between phonology, syntax and semantics. Although it is ignored and considered non-universal, morphology does exist. Even infants easily acquire the morphology of their languages and decompose words into morphemes. For example, Turkish infants have to distinguish word internal structure because Turkish morphosyntax is tortuous and it plays a central role in semantic analysis.

Many psycholinguistic studies showed that infants use statistics to build their initial lexicons (Aslin, Jusczyk & Pisoni, 1997; Best, 1995). They can identify speech segments using transitional probabilities (Aslin et al., 1998; Gomez, 2002; Saffran et al., 1996). Such statistical abilities are not only innate to language acquisition but to other domains as well (Kirkham et al., 2002; Teglas et al., 2007; Xu & Denison, 2009; Xu & Garcia, 2008) because accumulating and analyzing statistics requires general cognitive skills such as memory capacity and mathematics.

Morphology was selected as the topic of this study because morphology has a more central role than the syntax in Turkish grammar. Besides the success rate of the current model, it is also worth considering the premises of the poverty of stimulus arguments under the consequences of the current study. Only the first two premises given in Section 1.5 are relevant to the current study. Firstly, this study emphasizes that the probability mass in seen observations can provide children with necessary clues for not only morphological word segmentation but also indirect negative evidence for language acquisition. In other words, positive evidence in Turkish is sufficient for morphological word segmentations. Secondly, children are also provided with negative evidence as well. The CHILDES database has speech errors and self-corrections indicating negative evidences. A morphotactic rule in one dialect can be considered as an error in another dialect. The following examples in (96) are morphologically incorrect in formal Turkish, but they were discovered in the manual segmentation of CHILDES and they might be acceptable in some dialects:

(96) *bırak-aca-n mı?* → *bırak-acak mı-sın?*
(leave-*FUT-2.SG QP) (let-FUT QP-2.SG)
Will you leave [it]?

*mayo-n-lan* → *mayo-n-la*
swimming suit-2.SG.POSS-*COM swimming suit-2.SG.POSS-COM
with your swimming suit

In this study, the Hidden Markov Model (HMM) with semi-supervision was used for the morphological segmentation of Turkish words. The METU-Turkish Corpus (Say et al., 2002) and the manual segmentation of the METU-Sabancı Turkish Treebank (Atalay et al., 2003) were employed in the model to calculate the transition and emission probabilities. The model eventually achieved the scores of .88, .92 and .90 for precision, recall and *f*-measure respectively. It also checked the existence of Turkish voicing, emphatic reduplication, epenthesis and deletion in roots. In addition, the model contained a module for Turkish compound word recognition and segmentation. There are some rule-based morphological analyzers in Turkish as cited above, but they are unable to cope with language changes because these analyzers employ finite-state approaches with a previously compiled lexicon of morphemes. They use a set of rules for language-specific morphotactics and morphophonological constraints. Such methods are language-specific, and require their lexicons and rule sets to be updated. Furthermore, they cannot handle changes in languages over time whereas statistical models can do this. The rule-based approaches usually employ tries; yet, using tries in lexicons is cognitively implausible (Forster, 1976) because native speakers assess native words faster than non-words. The finite-state approaches using tries fail to capture the frequency effect (Bradley, 1978). However, the *n*-gram frequencies alone are not adequate in compounding. The frequencies within words cannot be used either in morphological ambiguity resolution. For these tasks they require additional information, such as semantic and phonetic knowledge, frequency distribution over lexical categories and so forth.

The research questions explored in this study were as follows:

- Is it possible to propose a semi-supervised statistical model with *n*-grams using frequencies in order to explain morphological word segmentation, the acquisition of morphology and morphotactics?

- Will this be just a superfluous model fitting the existing linguistic data or will it be compatible with current cognitive empirical data?

- Is semi-supervision cognitively plausible?

Firstly, the semi-supervised HMM successfully performed morphological word segmentation. Most frequent *n*-grams in both the Corpus and the CHILDES database were morphs and functional words. When the transition probabilities among the *n*-grams are investigated, it can be seen that the probabilities indicate Turkish morphotactics. In other words, given history of words, conditional probabilities are sufficient to capture the morphotactics and judge the morphological plausibility of a word rather than a set of rules. This can be mimicked by different machine learning models provided that there is a balanced corpus and supervision.

Secondly, as discussed in Chapter 5, the model is compatible with current cognitive empirical data. Experimental studies on compound words, emphatic reduplication and nonce words buttress the compatibility of using frequencies and probabilities in modeling of human activities. Infants count and undertake statistical analysis and they can also employ these skills in linguistic domains. They have the cognitive skills for detecting inherent patterns in the environment and the speech they are exposed to.

Thirdly, semi-supervised learning is appropriate for language acquisition because children receive a small portion of supervised linguistic data or feedback but the greater portion is unlabeled and massive. The frequent *n*-grams and the results of manual segmentation from CHILDES show that the morphs uttered by parents are fairly parallel to what children acquire. Thus, in reality, children are exposed to untagged data and parental supervision. The combination of these two elements results in the semi-supervised learning which was achieved in this study via the Corpus and the manually segmented Treebank.

The study also indicates that there is a probability mass in child-directed speech (CDS) and it is skewed toward possible word forms and unlikely morph sequences. This mass can be handled by different statistical methods, such as an HMM as in this study, and Support Vector Machines, entropy models or Bayesian learning models in other studies. The frequencies and sequences in the CDS are also indirect negative evidence because infrequent formations are sources of morphologically 'ill-formations' for children. In other words, children do not always need to receive negative evidence or corrections directly from adults. The skewed probability mass in the CDS instead, indirectly, tells children what is not to be done in a way that is similar to the forward evaluation in the HMM. In other words, known frequencies and the corresponding probability mass are types of indirect negative evidences because they cannot only segment words into morphs but also indirectly show what morphological formations are not allowed for a language. Pinker (1984) indicated the use of positive evidences as indirect negative evidences, and implemented it in the acquisition of the affix *-s* attached to the verbs in declarative sentences. He stated that a child is not a completely rational hypothesis-tester; however, a child progressively abandons a hypothesis contradicting by some input. In other words, the hypothesis (or morphological segmentation) which is more probable overwhelms the rest.

Suppose a genuine word $w$ has a genuine morph segmentation $s_1$ as $m_1$-$m_2$-...-$M_N$ which cannot be acquired by the positive observations of a language $L$. The corresponding morphotactic rule $R$ in this segmentation is induced and acquired by a learner without any evidence about the truth of the segmentation $s_1$. In other words, $R$ is acquired by innate knowledge. Is it possible to judge the plausibility of $R$? If the plausibility of this morphotactic rule $R$ is evaluated by some available observations in $L$, then the learner could have acquired $R$ from the available observations. That is a contradiction. If it is proposed that the plausibility of $R$ can be judged natively, then nativism is assumed to be an argument of nativism. The premises and the corresponding reasoning for the strong poverty of stimulus argument have been erroneously stated (Pullum & Scholz; 2002). Instead, the frequencies and probability mass in available observations can be used to decide on the *invalidity* of $s_2$ as $m_1$-$m_2$-...-$m_e$, and direct the learner to acquire $s_1$. It should be noted that low probability alone cannot always imply morphotactic invalidity. A very long word with a substring which occurs once in a corpus can have a very low probability while a very short but invalid word with substrings resembling frequent morphs in a corpus can have a high probability. In other words, what directs the learner to $s_1$ and acquire $R$ is the *comparative probability* between $s_1$ and $s_2$. Therefore, seen observations can be employed as indirect negative evidences because they can show language learners "what is not allowed" in an indirect way.

The manual segmentations of the Treebank and the CHILDES database also shows that when children acquire the most frequent roots and their morphological sequences and co-occurrences, they have learned most of the morphological paradigms in their language. In order to corroborate the use of frequencies in the cognitive studies, experimental studies and the corresponding statistical models in Turkish emphatic reduplication and the acceptability of nonce words were also discussed in this study. To the best of the author's knowledge, no similar approaches exist on this topic. The overall method and experiments suggest empiricism instead of nativism because human beings have a statistical learning ability that is not specific to linguistics but general cognition.

For Turkish emphatic reduplication, selecting a linker which has a frequent (admittedly orthographic) representation in the corpus would seem to direct the speaker to consider whether there was a root instead of a prefix. This points to more ways of looking at morphology-lexicon relation, rather than just the *blocking*, such as *went/*goed* and *git/*gittir/götür* 'leave/cause to leave/take away'. It seems that the speakers are putting the co-occurrence frequencies in their language to immediate use. Turkish emphatic reduplication, an apparently phonological operation, depends on global lexical knowledge to select an appropriate linker whose co-occurrence with the initial consonant of the reduplicated word is infrequent. Yavas (1980) was first to point out the lexical source of the linker type. It can be concluded that the TER base form paradigm is not consistent. In other words, something other than phonological ranking is also effective in TER. It is clear that the reduplicated "prefix", the linker types {*p, s, m, r*}, or the "infix" from {*-A, -Il, -Am*} are not morphological objects, affixes or morphemes. It also seems clear that the process is not purely lexical or phonological. There seems to be no discernible TER morpheme, or a purely morphophonological process. It should be noted that TER is co-determined by morphology and the lexicon. Moreover, its semantics depend on lexical properties, and it cannot be repeated. Thus, TER is morpholexical (Kılıç & Bozşahin, 2013).

Statistical models of morphology learning are useful for linguistic theory, the elimination of the large lexicon of morphs, modeling child language acquisition, speech recognition, machine learning and documentation of unstudied or endangered languages. The findings in this study and the uses of statistical models are striking, and show researchers how to move toward semi-supervised segmentation.

This study also emphasizes that the lexicon should be morphemic. However; there is no precise way of distinguishing inflectional morphemes from derivational morphemes. Bauer (2003) states that if this distinction is discarded then what remains are the roots and affixes and thus, the definition of lexemes must be reconsidered. Many languages cannot be studied in depth if words are assumed to be the only lexical items. Thus, a morphemic lexicon is compulsory in the linguistics of some languages. Bozsahin (2002) provides an account of Turkish combinatory morphemic lexicon. Balogh and Kleiber (2003) discuss the computational benefits of a totally lexicalized grammar. The frequencies and sequences are cues for the acquisition of forms. Yet, the forms are useful only if they are connected with meanings. As stressed by Marantz (2013), morphological and syntactic processing involves the central exploitation of a grammar of morphemes. Contemporary linguistics should focus on the notion of morphemes and how they are learned through the acquisition of *form/meaning connections*. These connections can be studied if the morphemes are lexicalized.

Finally, in the current study, it is not claimed that the frequencies are the solution to every problem in linguistics. Instead, it is emphasized that the frequencies are clues, and they can solve some of the problems in linguistics, such as morphological word segmentation, emphatic reduplication, and the acceptability of nonce-words, because by utilizing frequencies some algorithms have learned much more from corpora than most linguists would have thought possible. Furthermore, in this study it is not claimed that nativism is wrong. Instead, it is argued that frequencies and probability mass can inform a learner in the form of both positive and indirect negative evidences. Despite the fact that the probabilistic nature of language learning in UG has been also proposed by various nativist researchers (Yang, 2004), probabilities are only considered as a way to set the parameters of UG. Crain and Pietroski (2001) have stated that the innateness hypotheses will continue to be the best available explanation for the gap between normal human experience and linguistic knowledge until empiricists show how specific principles can be learned on the basis of the primary linguistic data. This study shows that frequencies and sequences can narrow down this gap, thus, it supports empiricism.

## 6.1 Limitations

The main limitation of the study is that it is based on orthographic representation; however, young children are exposed to speech rather than texts. Auspiciously, there is a close correspondence between phonotactics and orthotactics in Turkish and this allows the method to mimic children's acquisition of morphology to some extent. The model still suffers from the local maxima problem. The local maxima problem is serious because of the nature of the method. For example, *deler* 'pierces' → **de-ler* is an over-segmentation problem due to the high number of occurrences of *-de* LOC and *-ler* PLU in both the Treebank and the Corpus. Although the manual segmentations aid the model in avoiding the local maxima problem, more clues are required. For example, phonological and semantic information is necessary to improve the model.

There is about 1:36 ratio between the number of words in the Corpus and the Treebank, but the corresponding probabilities were averaged as if they had equal effects on morph segmentation. This was undertaken in order to simulate the importance of environmental feedback in language acquisition. For emphatic reduplication, deletion and epenthesis checked in the model, the corresponding rules were input in a supervised manner. Morph segmentation is an ambiguous task, but in real life, children have access to phonological information, such as, stress and prosody, visual and contextual information to overcome ambiguities. Thus, the most probable three segmentations were selected for *f*-score evaluations to compensate for this lack of regularity in the task.

The method delivers morphs without morphological analysis, but it requires enhancements due to ambiguities. Semi-supervision provides the method with the required improvement to cope with the ambiguities. However, phonology and semantics are strongly determinative in the acquisition of morphology and the disposal of ambiguities. Without such information and cues, the success rate of the method is limited. It is possible to improve the success rate; however, this would necessitate more language-specific assumptions, rules and exceptions. There should be an optimum level of supervision in order to ensure that such models are kept as little language-specific as much as possible.

## 6.2 Future Implications

During the development of the method and implementation of the corresponding experiments, it was evident that electronic language resources were crucial for computational linguistics. Researchers working in development and dissemination of such resources should be better funded. A corpus with phonological, semantic and syntactic annotations will not only improve the model in this study but also help researchers in other linguistic domains.

The segmentation method should be tested with speech data when a spoken corpus becomes available. In this study, it is not claimed that the whole language acquisition process is frequency-based and statistical. Rather, it emphasizes that frequencies and probabilities play a crucial role in morph segmentation and acquisition. It is strongly believed that if frequent phonemes are collected from a speech corpus, they will certainly be composed of morphs and functional words.

Another future implication is about the acceptability of nonce words. The method requires improvements in terms of the morphophonological properties of target languages. The threshold values for the acceptability decisions depend on word lengths. They also need to be improved with respect to the target languages. The method is successful for Turkish, a linearly concatenative language. However, it needs to be tested and adapted for the languages with ablaut or umlaut phenomena such as English and German, and the templatic languages such as Arabic and Hebrew. Furthermore, the participants who had a knowledge of a foreign language responded differently than those who only spoke their native Turkish language. Thus, future researchers should bear this in mind when studying nonce words.

For the compound word recognition and segmentation task, without the phonological and contextual information, the success of the statistical segmentation is limited by the distribution of the constituents in the corpus. Intonation and stress are quite important in this task; thus, a phonologically annotated corpus will very useful in similar studies. A further enhancement can be made by including language-specific compounding rules in the model, such as compound types and allowed head vs. modifier types as global statistics. Although the assumption that the lexicon is morphemic was useful in the identification and segmentation of compound words, the failure rates, because of too frequent words, indicated that *n*-grams can be informative for *in-word* paradigms while it requires global statistics for *across-word* paradigms.

The final task to be addressed for future research, especially studies of Turkish syntax will require morphosyntactic tags rather than morphs themselves. A combinatory approach unifying the tags with the corresponding morphs and a disambiguation process over the tags will be supportive for future research. This is crucial because if it is claimed that the lexicon has a morphemic structure, as in this study, then it is important to represent morphs with corresponding functional content. Ultimately, lexicalizing the morphs with semantic and syntactic content will turn the research questions of morphology towards the interaction of lexicon and syntax.

# REFERENCES

Abney, S. (2008). *Semisupervised learning for computational linguistics.* Boca Raton, FL: Taylor & Francis Group.

Ackema, P., & Neeleman, A. (2004). *Beyond morphology*. Oxford: Oxford University Press.

Adda-Decker, M. (2003). A corpus-based decompounding algorithm for German lexical modeling in LVCSR. *Proceedings of the European Conference on Speech Communication and Technology*, 257-260.

Albright, A. (2008). From clusters to words: Grammatical models of nonce word acceptability. Handout of *talk presented at 82nd LSA*, Chicago, January 3, 2008.

Aldous, D. J. (1985). Exchangeability and related topics. In P. L. Hennequin (ed.), *Ecole d'Ete de probabilitrs de Saint-Flour XII.*, *Lect. Notes Math., 1117.* New York: Springer.

Alishahi, A., & Stevenson, S. (2008). A computational model for early argument structure acquisition. *Cognitive Science*, *32*(5), 789-834.

Allwood, J., Hendrikse, A. P., & Ahlsén, E. (2010). Words and alternative basic units for linguistic analysis. In P. J. Henrichsen (ed.), Linguistic theory and raw sound. *Copenhagen Studies in Language, 40*, (pp. 9-26). Copenhagen: Samfundslitteratur.

Altun, Y., & Johnson, M. (2001). Inducing SFA with ϵ-transitions using minimum description length. *Finite State Methods in Natural Language Processing Workshop at ESSLLI*, Helsinki, Finland.

Anderson, S. R. (1992). *A-morphous morphology.* New York: Cambridge University Press.

Anderson, S. R. (2005). *Aspects of the theory of clitics.* Oxford: Oxford University Press.

Anshen, F., & M. Aronoff (1988). Producing morphologically complex words. *Linguistics, 26*, 641–655

Antworth, E. L. (1990). PC-KIMMO: a two-level processor for morphological analysis. *Number 16 in Occasional publications in academic computing*. Summer Institute of Linguistics, Dallas.

Arciuli, J., & Simpson, I. C. (2011). Statistical learning is related to reading ability in children and adults. *Cognitive Science, 36*(2), 286-304.

Aronoff, M. (1976). *Word formation in generative grammar.* Cambridge, MA: MIT Press.

Aronoff, M. (1993). *Morphology by itself. Stems and inflectional classes*. Cambridge, MA: MIT Press.

Aronoff, M., & Fudeman, K. (2011). *What is morphology? 2nd Edition*. Oxford: Wiley Blackwell.

Aronoff, M., & Fuhrhop, N. (2002). Restricting suffix combinations in German and English: Closing suffixes and the monosuffix constraint. *Natural Language and Linguistic Theory, 20*, 451-490.

Aslin, R. N., Jusczyk, P. W., & Pisoni, D. B. (1997). Speech and auditory processing during infancy. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology*, (pp. 147-198). John-Wiley & Sons.

Aslin, R. N., Safran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by human infants. *Psychological Science, 9*, 321-324.

Atalay, N. B., Oflazer, K., & Say, B. (2003). The annotation process in the Turkish Treebank. In *Proc. of the EACL Workshop on Linguistically Interpreted Corpora - LINC*, April 13-14, 2003, Budapest, Hungary.

Aygen, G. (2007). Q-Particle. *Journal of Linguistics and Literature, 4*(1), 01-30.

Baayen, R. H. (2003) Probabilistic approaches to morphology. In R. Bod, J. Hay & S. Jannedy (Eds.), *Probabilistic linguistics*, (pp. 229-287). Cambridge, MA: The MIT Press.

Baerman, M., Brown, D., & Corbett, G. G. (2005). *The syntax-morphology interface: A study of Syncretism*. NY: Cambridge University Press.

Baker, M. (1988). *Incorporation: A theory of grammatical function changing*. Chicago: University of Chicago Press.

Balogh, K., & Kleiber, J. (2003). Computational benefits of a totally lexicalized grammar. In V. Matouek & P. Mautner (Eds.), *Text, speech and dialogue, Proceedings of TSD2003*, (pp. 114-119). Springer-Verlag, Berlin Heidelberg New York.

Baroni, M., Matiasek, J., & Trost, H. (2002). Unsupervised discovery of morphologically related words based on orthographic and semantic similarity. *ACL Special Interest Group in Computational Phonology in Cooperation with the ACL Special Interest Group in Natural Language Learning (SIGPHON/SIGNLL)*, Philadelphia, Pennsylvania, 48-57.

Batchelder, E. O. (1997). *Computational evidence for the use of frequency information in discovery of the infant's first lexicon*. Unpublished Dissertation, Department of Linguistics, the City University of New York.

Bauer, L. (2003). *Introducing morphology 2^nd^ Edition.* Edinburg: Edinburg University Press.

Baum, L. E., & Petrie, T. (1966). Statistical inference for probabilistic functions of finite state Markov chains. *The Annals of Mathematical Statistics, 37*(6), 1554–1563.

Beard, R. (1995). *Lexeme-morpheme base morphology*. NY: SUNY Press.

Beck, M. (1995). Tracking down the source of NS–NNS differences in syntactic competence. *Unpublished manuscript*, University of North Texas.

Bermudez-Otero, R., & Payne, J. (2011). There are no special clitics. In A. Galani, G. Hicks & G. Tsoulas (Eds.), *Morphology and its interfaces,* (pp. 57-96). Amsterdam: John Benjamins.

Bernhard, D. (2006). Unsupervised morphological segmentation based on segment predictability and word segments alignment. In *Proceedings of the 2^nd^ Pascal Challenges Workshop*, 19-24, Venice, Italy, 2006.

Bertram, R., Schreuder, R., & Baayen, R. H. (2000). The balance of storage and computation in morphological processing: The role of word formation type, affixal homonymy, and productivity. *Journal of Experimental Psychology: Learning, Memory and Cognition, 26*(2), 489-511.

Berwick, R. C., Pietroski, P., Yankama, B., & Chomsky N. (2011). Poverty of the stimulus revisited. *Cognitive Science, 35,* 1207-1242

Best, C. T. (1995). Learning to perceive the sound patterns of English. In C. Rovee-Collier & L. P. Lipsitt (Eds.), *Advances in infancy research*, (pp. 217-304). Ablex.

Bickel, B., Banjade, G., Gaenszle, M., Lieven, E., Paudyal, N. P., Rai, P. I., et al. (2007). Free prefix ordering in Chintang. *Language, 83*, 1-31.

Bird, S., & Ellison, T. M. (1994). One-level phonology: Autosegmental representations and rules as finite automata. *Computational Linguistics*, *20*(1), 55-90.

Black, A., Ritchie, G., Pulman, S., & Russell, G. (1987). Formalisms for morphographemic description. In *Proceedings of the Third Conference of the European Chapter of the Association for Computational Linguistics*, 11-18.

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

Bod, R. (2009). From exemplar to grammar: A probabilistic analogy-based model of language learning. *Cognitive Science, 33*, 752-793.

Booij, G. (2005a). Construction-dependent morphology. *Lingue e Linguaggio, 2*, 163-178.

Booij, G. (2005b). Compounding and derivation: Evidence for construction morphology. In W. U. Dressler, F. Rainer, D. Kastovsky & O. Pfeiffer (Eds.), *Morphology and its demarcations,* (pp. 109-132). Amsterdam / Philadelphia: John Benjamins.

Booij, G. (2010). *Construction morphology.* Oxford: Oxford University Press.

Border, H. (2001). Morphology and syntax. In A. Spencer, & A. M. Zwicky (Eds.), *The handbook of morphology*, (pp. 151-190). Oxford: Blacwell Publishers.

Borer, H. (2004). *Structuring sense*. Oxford: Oxford University Press.

Boudelaa, S., & Marslen-Wilson, W. D. (2001). Morphological units in the Arabic mental lexicon. *Cognition, 81,* 65-92.

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

Bradley, D. C. (1978). *Computational distinctions of vocabulary type.* Unpublished PhD Dissertation, Cambridge, MA: MIT Press.

Brent, M. R. (1993). Minimal generative explanations: A middle ground between neurons and triggers. *Proceedings of the 15th Meeting of the Cognitive Science Society*, 28-36, Hillsdale, NJ: LEA.

Brent, M. R. (1997). Toward a unified model of lexical acquisition and lexical access. *Journal of Psycholinguistic Research, 26*, 363-375

Broadwell, G. A. (2008). Turkish suspended affixation is lexical sharing. In M. Butt & T. H. King (Eds.), Proc. of the LFG08 Converence.

Brown, R. (1973). *A first language: The early stages*. Cambridge: Harvard University Press.

Bybee, J. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam: J. Benjamins.

Carstairs-McCarthy, A. (2001). Phonological constraints on morphological rules. In A. Spencer & A. M. Zwicky (Eds.), *The handbook of morphology*, (pp. 144-148). Oxford: Blacwell Publishers.

Carstairs-McCarthy, A. (2010). *The evolution of morphology*. New York: Oxford University Press.

Carter, R. J. (1976). Some constraints on possible words. *Semantikos, 1*, 27-66.

Chang, C. H., & Chen, C. D. (1993). A study on integrating Chinese word segmentation and part-of-speech tagging. *Communications of the Chinese and Oriental Languages Information Processing Society, 3*(2), 69–77.

Chang, J. S., & Su, K. Y. (1997). An unsupervised iterative method for Chinese new lexicon extraction. *International Journal of Computational Linguistics and Chinese Language Processing,* 97-148.

Chen, S. F., & Goodman, J. (1999). An empirical study of smoothing techniques for language modeling. *Computer Speech and Language, 13*, 359–394.

Chomsky, N. (1957). *Syntactic structures.* The Hague: Mouton.

Chomsky, N. (1959). A review of B. F. Skinner's verbal behavior. *Language, 35*(1), 26-58.

Chomsky, N. (1965). *Aspect of the theory of syntax.* Cambridge, MA: MIT Press.

Chomsky, N. (1980). The linguistic approach. In M. Piatelli-Palmarini (ed.), *Language and learning: The debate between Jean Piaget and Noam Chomsky*. Cambridge, MA: Harvard University Press.

Chomsky, N. (2004). Language and mind: current thoughts on ancient problems. In L. Jenkins (ed.), *Variation and universals in biolinguistics*, (pp. 379-405). Amsterdam: Elsevier.

Chomsky, N., & Halle, M. (1968). The sound pattern of English. *Studies in Language.* New York: Harper & Row.

Christiansen, M. H. & Monaghan, P. (2006). Discovering verbs through multiple-cue integration. In K. Hirsh-Pasek & R.M. Golinkoff (Eds.), *Action meets words: How children learn verbs,* (pp. 88-107). New York: Oxford University Press.

Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of German inflection. *Behavior and Brain Sciences, 22*(6), 991-1003.

Clahsen, H. (2006). Dual-mechanism morphology. In K. Brown (ed.), *Encyclopedia of language and linguistics,* (pp. 1-5). Oxford: Elsevier.

Clahsen, H., & Almazan, M. (2001). Compounding and inflection in language impairment: Evidence from Williams Syndrome (and SLI). *Lingua*, *111*, 729-757.

Clark, A. (2007). Supervised and unsupervised learning of Arabic morphology. In A. Soudi, A. van den Bosch & G. Neumann (Eds.), *Arabic Computational Morphology*, (pp. 181-200). Springer.

Conner, T. J. (2003). Circumfixation: An unnoticed problem for Indonesian stress. In Proceedings of AFLA'03, March 2003, Hawaii, 28-30.

Crain, S., & Nakayama, M. (1987). Structure dependence in grammar formation. *Language, 63,* 522-543.

Crain, S., & Pietroski, P. (2001). Nature, nurture and universal grammar. *Linguistics and Philosophy, 24,* 139-186.

Creutz, M. (2006). *Induction of the morphology of natural language: Unsupervised morpheme segmentation with application to automatic speech recognition*. Ph.D. Thesis, Computer and Information Science, Report D13, Helsinki, University of Technology, Espoo, Finland.

Creutz, M., & Lagus, K. (2002). Unsupervised discovery of morphemes. *ACL Special Interest Group in Computational Phonology in cooperation with the ACL Special*

*Interest Group in Natural Language Learning (SIGPHON/SIGNLL)*, Philadelphia, Pennsylvania, 21-30.

Creutz, M., & Lagus, K. (2005). Inducing the morphological lexicon of a natural language from unannotated text. In *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning*, Espoo, 106-113.

Creutz, M., & Lagus, K. (2007). Unsupervised models for morpheme segmentation and morphology learning. *ACM Transaction on Speech and Language Processing*, *4*(1), 1-34.

Çöltekin, Ç. (2010). A freely available morphological analyzer for Turkish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010)*, Valletta, Malta, May 2010.

Dąbrowska, E. (2006). Low-level schemas or general rules? The role of diminutives in the acquisition of Polish case inflections. *Language Sciences, 28*, 120-135.

Déchaine, R. M. (2005). Grammar at the borderline: A case study of P as a lexical category. In J. Alderete et al. (ed.), Proc.of the *24th West Coast Conference on Formal Linguistic*, (pp. 1-18). Somerville, MA: Cascadilla Proceedings Project

Demircan, O. (1987). Emphatic reduplication in Turkish. In H. E. Boeschoten & L. Th. Verhoeven (Eds.), *Studies on modern Turkish: Proc. of the 3rd Conference in Turkish Linguistics* , (pp. 24-41). Tilburg: Tilburg University Press.

Dhillon, R. (2009). Turkish emphatic reduplication: Balancing productive and lexicalized forms. *GLS, 71*, 3-20.

Di Sciullo, A. M., & Edwin W. (1987). *On the definition of word*. Cambridge, MA: MIT Press.

Domínguez, J. A. (1991). The role of morphology in the process of language acquisition and learning. *Revista Alicantina de Estudios Ingleses, 4,* 37-47.

Dresher, B. E., & Kaye, J. (1990). A computational learning model for metrical phonology. *Cognition, 34*, 137–195.

Dressler, W. U. (1989). Prototypical differences between inflection and derivation. *Zeitschrift für Phonetik, Sprachwissenschaft and Kommunikationsforschung, 42*, 3–10.

Dressler, W. U. (2006). Compound types. In G. Libben & G. Jarema (Eds.), *The representation and processing of compound words,* (pp. 23-44). Oxford: Oxford University Press.

Dressler, W. U., Aksu-Koc¸, A., Laalo, K., & Pfeiler, B. (2009). *Early phases in the acquisition of affix order*. Paper presented at the Second Vienna workshop on affix order, Vienna, June 5–6, 2009.

Easterbrooks, S. P., & Baker, S. (2002). *Language learning in children who are deaf and hard of hearing: Multiple Pathways*. Boston: Allyn & Bacon.

Elman, J., Bates, E., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press/Bradford Books.

Embick, D., & Noyer, R. (2007). Distributed morphology and the Syntax/Morphology interface. In G. Ramchand & C. Reiss (Eds.), *The Oxford handbook of linguistic interfaces*, (pp. 289-324). New York : Oxford University Press.

Ergin, M. (2002). *Orhun Abideleri*. İstanbul: Boğaziçi Yayınları.

Ervin, S. (1964). Imitation and structural change in children's language. In E. Lennenberg (ed.), *New directions in the study of language.* Cambridge, MA: MIT Press.

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research, 52*(2), 321–335.

Evans, N., & Levinson, S. (2009). The myth of language universals: language diversity and its importance for cognitive science. *Behavioural and Brain Science, 32* (5), 429-448.

Eyüboğlu, İ. Z. (1998). *Türk dilinin etimoloji sözlüğü.* İstanbul: Sosyal Yayınlar.

Fehringer, C. (2012). The lexical representation of compound words in English: Evidence from aphasia. *Language Sciences, 34*, 65-75.

Feldman, L. B. (1994). Beyond orthography and phonology - differences between inflections and derivations. *Journal of Memory and Language, 33*(4), 442-470.

Feldman, H., Goldin-Meadow, S., & Gleitman, L. R. (1978). Beyond Herodotus: The creation of language by linguistically deprived deaf children. In A. Lock (ed.), *Action, symbol, and gesture*. New York: Academic Press.

Finley, S. (2012). Testing the limits of long-distance learning: Learning beyond a three-segment window. *Cognitive Science, 36,* 740–756.

Fiorentino, R., & Poeppel, D. (2007). Processing of compound words: An MEG study. *Brain and Language*, 103, 18–19.

Flickinger, D. P. (1987). *Lexical rules in the hierarchical lexicon.* Stanford: Stanford University Press.

Ford, M. A., Marslen-Wilson, W. D., & Davis, M. H. (2003). Morphology and frequency: contrasting methodologies. In E. H. Baayen et al. (Eds.), *Morphological structure in language processing.* Berlin: De Gruyter.

Forney, D. (1973). The Viterbi algorithm. In *Proceedings IEEE 61*, 268–278.

Forster, K. I. (1976). Accessing the mental lexicon. In R. J. Wales & E. W. Walker (Eds.), *New approaches to language mechanisms.* Amsterdam: North-Holland.

Frank, M., Goodman, S., & Tenenbaum, J. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science, 20*(5), 578–585.

Frauenfelder, U. H., & Schreuder, R. (1992). Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In G. E. Booij & J. van Marle (Eds.), *Yearbook of morphology*. Dordrecht: Kluwer.

Friederici A. D., Steinhauer, K., & Pfeifer, E. (2001). Brain signatures of artificial language processing: Evidence challenging the critical period hypothesis. *PNAS, 99*(1), 529–534.

Frisch, S. A., & Zawaydeh, B. A. (2001). The psychological reality of OCP-Place in Arabic. *Language, 77*, 91–106.

Fromkin, V., & Rodman, R. (1993). *An introduction to language.* Harcourt Brace Collage Publishers, 5th edition.

Galantucci, B. (2005). An experimental study of the emergence of human communication systems. *Cognitive Science, 29*(5), 737-767.

Galantucci, B., Kroos, C., & Rhodes, T. (2010). The effects of rapidity of fading on communication systems. *Interaction Studies, 11*(1), 100-111.

Gamback, B. (2005). Semantic morphology. In A. Arpp et al. (Eds.), *Inquiries into words, constraints and context*, (pp. 204-213). Stanford: CSLI Publication.

Gardner R. A., Gardner B. T., & van Cantfort T. E. (1989). *Teaching sign language to a chimpanzee,* Albany: State University of New York Press.

Ghahramani, Z. (2004). Unsupervised learning. In O. Bousquet, G. Raetsch & U. von Luxburg (Eds.), *Advanced lectures on machine learning. LNAI 317, 6*, 72-112.

Giegerich, H. J. (1999). *Lexical strata in English. Morphological causes, phonological effects*. Cambridge: Cambridge University Press.

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*, 447–474.

Goldin-Meadow, S. (2006). The resilience of language: What gesture creation in deaf children can tell us about how all children learn language? In J. Werker & H. Wellman (Eds.), *The essays in developmental psychology series*. New York: Psychology Press, 2003.

Goldin-Meadow, S. (2009). Language acquisition theories. In J. B. Benson & M. M. Haith (Eds.), *Language, memory, and cognition in infancy and early childhood.* Oxford, UK: Elsevier Inc.

Goldin-Meadow, S., & C. Mylander (1984). Gestural communication in deaf children: Non-effects of parental input on early language development. *Science, 221,* 372-374.

Goldsmith, J. (2001). Unsupervised acquisition of the morphology of natural language. *Computational linguistics, 27*(2), 153-198.

Goldsmith, J. (2005). *An algorithm for the unsupervised learning of morphology*. Technical report, TR-2005-06, Department of Computer Science, University of Chicago, 2005. http://humfs1.uchicago.edu/~jagoldsm/Papers/Algorithm.pdf.

Goldsmith, J., & Riggle, J. (2012). Information theoretic approaches to phonological structure: the case of Finnish vowel harmony. *Natural Language and Linguistic Theory, 30* (3), 859-896.

Goldwater, S. (2007). *Nonparametric bayesian models of lexical acquisition.* PhD Dissertation, Department of Cognitive and Linguistic Sciences, Brown University, Rhode Island.

Goldwater, S., Griffiths, T. L., & Johnson, M. (2009). A Bayesian framework for word segmentation: Exploring the effects of context. *Cognition, 112*(1), 21–54.

Goldwater, S., & McClosky. D. (2005). Improving statistical MT through morphological analysis. In *Proceedings of Empirical Methods in Natural Language Processing*, Vancouver, Canada, 676–683.

Gomez, R. L. (2002). Variability and detection of invariant structure. *Psychological Science, 13*, 431-436.

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*, 237-264.

Gordon, P. (1985). Level ordering in lexical development. *Cognition, 21,* 73-93.

Göksel, A. (1997). Morphological asymmetries between Turkish and Yakut. In K. İmer & N. E. Uzun (Eds.), *Proceedings of the VIIIth International Conerence on Turkish Linguistics.* Ankara: Ankara Üniversitesi Basımevi.

Göksel, A. (2001). The auxiliary verb ol at the morphology-syntax interface. In E. E. Erguvanlı-Taylan (ed.), *The verb in Turkish,* (pp. 151-181). Amsterdam: John Benjamins.

Göksel, A. (2007). Morphology and syntax inside the word: Pronominal participles of headless relative clauses in Turkish. In G. Booij, L. Ducceschi, B. Fradin, E. Guevara, A. Ralli & S. Scalise (Eds.), *Proceedings of the Fifth Mediterranean Morphology Meeting (MMM5)*, Fréjus, 15-18 September 2005.

Göksel, A., & Kerslake, C. (2005). *Turkish: A comprehensive grammar*. Routledge: London and New York.

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (ed.), *Universals of language*. Cambridge: MIT Press.

Griffiths. T, L. 2006. *Power-law distributions and nonparametric Bayes*. Unpublished manuscript.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbabum J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Science, 14*, 357-364.

Griffiths, T. L., & Tenenbaum, J. B. (2006). Optimal predictions in everyday cognition. *Psychological Science, 17*, 767-773

Grünwald, P. D. (2007). *The minimum description length principle: Adaptive computation and machine learning*. Cambridge, MA: MIT Press.

Güngör, T. (2003). Lexical and morphological statistics for Turkish. In *Proc. of TAINN 2003*, 409-412.

Gürel, A. (1999). Decomposition: To what extent? The case of Turkish. *Brain and Language, 68,* 218-224.

Hakkani-Tür, D. Z., Oflazer, K., & Tür, G. (2002). Statistical morphological disambiguation for agglutinative languages. *Computers and the Humanities, 36*, 381-410.

Halle, M., & Marantz, A. (1993). Distributed morphology and the pieces of inflection. In K. Hale & S. Keyser (Eds.), *The view from building 20: Essays in linguistics in honor of Sylvain Bromberger*, (pp. 111–176). Cambridge, MA : MIT Press.

Hammarström, H., & Borin, L. (2011). Unsupervised learning of morphology. *Computational Linguistics 37* (2), 309–350.

Hammond, M. (2004). Gradience, phonotactics and the lexicon in English phonology. *International Journal of English Studies, 4*, 1–24.

Hankamer, J. (1986). Finite state morphology and left to right phonology. *Proceedings of the West Coast Conference on Formal Linguistics 5*, 41-52.

Hankamer, J. (1989). Morphological parsing and the lexicon. In W. Marslen-Wilson (ed.), *Lexical representation and process*, (pp. 392-408). The MIT Press, Cambridge, Mass.

Harris, Z. (1954). Distributional structure. *Word, 10*, 146–162.

Hauser, M., Chomsky, N., & Fitch, W. T. (2002). The language faculty: What is it, who has it, and how did it evolve? *Science, 298*, 1569–1579.

Hay, J., Pierrehumbert, J., & Beckman, M. (2004). Speech perception, well-formedness and the statistics of the lexicon. In J. Local, R. Ogden & R. Temple (Eds.), *Phonetic* *interpretation: Papersbin laboratory phonology VI.* Cambridge: Cambridge University Press.

Hayes K. J., & Nissen C. H. (1971). Higher mental functions of a home-raised chimpanzee. In AM Schrier & F Stollnitz (Eds.), *Behavior of nonhuman primates*, *4,* (pp. 59-115). New York: Academic Press.

Hippisley, A. (2001). Word formation rules in a default inheritance framework: A network morphology account of Russian personal nouns. In J. van Marle & G. Booij (Eds.), *Yearbook of morphology 1999*, (pp. 221-261). Dordrecht: Kluwer.

Hopcroft, J. E., & Ullman, J. D. (1979). *Introduction to automata theory, languages and computation*. Menlo Park: Addison-Wesley.

Hsu, A. S., Chater, N., & Vitanyi, P. (2013). Language learning from positive evidence, reconsidered: A simplicity-based approach. *Topics in Cognitive Science, 5*, 35-55.

Inkelas, S. (1993). Nimboran position class morphology. *Natural Language & Linguistic Theory, 11*, 559-624.

Inkelas, S. (2000). Phonotactic blocking through structural immunity. In B. Stiebels & D. Wunderlich (Eds.), *Lexicon in focus. Studia Grammatica, 45*, (pp. 7-40). Berlin: Akademia Verlag.

Inkelas, S. (2005). Morphological doubling theory: evidence for morphological doubling in reduplication. In B. Hurch (ed.), *Studies on reduplication*, (pp. 63-86). Berlin: Mouton de Gruyter.

Inkelas, S., & Orgun, C. O. (1995). Level ordering and economy in the lexical phonology of Turkish. *Language, 71*, 763-793.

Inkelas, S, & Zoll, C. (2005). *Reduplication: Doubling in morphology*. Cambridge: Cambridge University Press.

Ioup, G., Boustaguia, E., El Tigi, M., & Moselle, M. (1994). Reexamining the critical period hypothesis. *Studies in Second Language Acquisition, 16*(1), 73-98.

Jackendoff, R. (1997). *The architecture of the language faculty*. Cambridge, MA: MIT Press.

Jackendoff, R. (2002). *Foundations of knowledge*. Oxford: Oxford University Press.

Jelinek, F., & Mercer, R. (1980). Interpolated estimation of Markov source parameters from sparse data. In E. S. Gelsema & L. N. Kanal (Eds.), *Pattern recognition in practice*, (pp. 381-402).

Johnson, H., & Martin, J. (2003). Unsupervised Learning of Morphology for English and Inuktitut. *Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, Edmonton, Canada.

Johnson, M. (2008). Unsupervised word segmentation for Sesotho using adaptor grammars. In *Proceedings of the Tenth Meeting of ACL Special Interest Group on Computational Morphology and Phonology*, 20-27, Columbus, Ohio, June 2008.

Johnston, T., & Schembri, A. (2007). *Australian sign language (Auslan): An introduction to sign language linguistics.* New York: Cambridge University Press.

Julien, M. (2000). *Syntactic heads and ford formation: A study of verbal inflection.* PhD Dissertation, University of Tromsø.

Julien, M. (2002). Inflectional morphemes as syntactic heads. In B. Sabrina, W. U. Dressler, O. E. Pfeiffer & Maria D. Voeikova (Eds.), *Morphology 2000: Selected papers from the 9th Morphology Meeting*, (pp. 175-184). Vienna, 24–28 February 2000.

Jurafsky, D. S., & Martin, J. H. (2000). *Speech and language processing*. Englewood: Prentice Hall.

Jusczyk, P. W. (1999). How infants begin to extract words from speech. *Trends in Cognitive Science*, vol.3 (9), 323-328.

Kabak, B. (2007). Turkish suspended affixation. *Linguistics, 45*(2), 311-347.

Karttunen, L. (2003). Computing with realizational morphology. In A. Gelbukh (ed.), Computational linguistics and intelligent text processing. *Lecture Notes in Computer Science, 2588*, (pp. 205-216). Heidelberg: Springer Verlag.

Karttunen, L., & Beesley, K. R. (2005). Twenty -five years of finite-state morphology. In A. Arpple et al. (Eds.), *Inquiries into words, constraints and contacts*, (pp. 71-83), Festschrift for Kimmo Koskenniemi, Aalto, Finland.

Katz, S. M. (1987). Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustic, Speech and Signal Processing, 35*(3), 400-401.

Kawahara, Shigeto (2012) Lyman's Law is active in loanwords and nonce words: Evidence from naturalness judgment studies. *Lingua, 122* (11), 1193-1206.

Kay, M. (1987). Nonconcatenative finite-state morphology. In *Proceedings of the 3$^{rd}$ Meeting of the European Chapter of the Association for Computational Linguistics,* 2-10.

Keenan, E. L., & Stabler, E. (1997). Bare grammar. *Ninth European Summer School on Logic, Language and Information*, Aix-en-Provence.

Kelepir, M. (2001). To be or not to be faithful. In A. Göksel & C. Kerslake (Eds.), *Proc. of the 9thInternational Conference on Turkish Linguistics*, 12-14 August 1998, Oxford, 11-18.

Kelepir, M. (2007). Copular forms in Turkish, Turkmen and Noghay. In M. Kelepir & B. Öztürk (Eds.), *Proceedings of the 2nd Workshop on Altaic Formal Linguistics*, 11-13 October 2004, Boğaziçi University. Cambridge (Mass.): MIT Working Papers in Linguistics.

Kibort, A. (2011). The feature of tense at the interface of morphology and semantics. . In A. Galani, G. Hicks & G. Tsoulas (Eds.), *Morphology and its interfaces,* (pp. 171-193). Amsterdam: John Benjamins.

Kim, H. (2007). The Full-to-Partial reduction in Korean and Turkish reduplication. *Linguistic Research, 26*(2), 121-148.

Kiparsky, P. (1982a). From cyclic phonology to lexical phonology. In H. v.d. Hulst & N. S. H. Smith (Eds.), *The structure of phonological representations,* (pp. 131-175). Dordrecht: Foris.

Kiparsky, P. (1982b). Lexical morphology and phonology. In I. S. Yang (ed.), *Linguistics in morning calm*, (pp. 3-91). Seoul: Hanshin.

Kiraz, G. A. (2001). *Computational nonlinear morphology: With emphasis on Semitic languages*. Cambridge, UK: Cambridge University Press.

Kirby, S., Cornish, H., & Smith, K. (2008). Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language. *PNAS, 105*(31), 10681-10686.

Kirkham, N., Slemmer, J., & Johnson, S. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83*, B35–B42.

Kılıç, Ö. (2012). Using conditional probabilities and vowel collocations of a corpus of orthographic representations to evaluate nonce words. In *Proc. of the Student Session in ESSLLI 2012*, 6-16 August, Opole, Poland.

Kılıç, Ö. & Bozşahin, C. (2012). Semi-supervised morpheme segmentation without morphological analysis. In Ş. Demir, İ. El-Kahlout & M. U. Doğan, *Proceedings of the Workshop in Turkic Languages, LREC 2012*, İstanbul, 52-56.

Kılıç, Ö. & Bozşahin, C. (2013). Selection of linker type in emphatic reduplication: Speaker's intuition meets corpus statistics. (to appear) In *Proceedings of the 35th annual meeting of the Cognitive Science Society*, Berlin, Germany, July 31 - August 3, 2013.

Klenk, U. (1985a). Ein nicht-lexikalisches Verfahren zur Erkennung spanischer Wortstämme. In U. Klenk (ed.), *Strukturen und Verfahren in der maschinellen Sprachverarbeitung*, (pp. 47-65). Dudweiler: AQ-Verlag.

Klenk, U. (1985b). Recognition of Spanish inflectional endings based on the distribution of characters. In *Computers in Literary and Linguistic Computing: Proceedings of the Eleventh International Conference*, 246-253,

Klima, E. S., & Bellugi, U. (1979). *The sign language.* Cambridge, MA: Harvard University Press.

Kneser, R., & Ney, H. (1995). Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 494-497.

Knuth, D. E. (1973). *The art of computer programming*. Reading, MA: Addison-Wesley.

Koo, H., & Callahan, L. (2011). Tier-adjacency is not a necessary condition for learning phonotactic dependencies. *Language and Cognitive Processes*, 1-8.

Koskenniemi, K. (1983). Two-level morphology: A general computational model for wordform recognition and production. *Technical Report 11*, University of Helsinki, Department of General Linguistics.

Krauss, M. E. (2007). Mass language extinction and documentation: The race against time. In O. Miyaoka, O. Sakiyama & M. Krauss (Eds.), *Vanishing Languages of the Pacific Rim*, (pp. 3-24). Oxford: Oxford University Press.

Krott, A. (2001). *Analogy in morphology: The selection of linking elements in Dutch compounds.* Nijmegen: Max Planck Institut für Psycholinguistik.

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

Kumar, D., Singh, M., & Shukla, S. (2012). FST based morphological analyzer for Hindi language. *IJCSI International Journal of Computer Science Issues*, 9(4), No 3, 349-353.

Kurimo, M., & Turunen, V. (2008). Unsupervised morpheme analysis evaluation by IR experiments – Morpho Challenge 2008. *Working Notes for the CLEF 2008 Workshop*, Aarhus, Denmark.

Kwiatkowski, T., Goldwater, S., Zettelmoyer, L., Steedman, M. (2012). A probabilistic model of syntactic and semantic acquisition from Child-directed utterances and their meanings. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France.

Laudanna, A., Badecker, W., & Caramazza, A. (1992). Processing inflectional and derivational morphology. *Journal of Memory and Language, 31*(3), 333-348.

Lapointe, S. (1980). *A theory of grammatical aggrement.* PhD Dissertation. Amherst, University of Massachusetts.

Lappe, S. (2007). *English prosodic morphology.* Dordrecht: Springer.

Lee, P. C., & Moss, C. J. (1999). The social context for learning and behavioural development among wild African elephants. In H. O. Box & K. R. Gibson (Eds.) *Mammalian social learning: Comparative and ecological perspectives*, (pp. 102-125). Cambridge: Cambridge University Press.

Levin, B., & Rappaport M. H. (2001). Morphology and lexical semantics. In A. Spencer & A. M. Zwicky (Eds.), *The handbook of morphology*, (pp. 248-271). Oxford: Blackwell Publishing.

Lewis, G. (2000). *Turkish grammar, Second edition.* Oxford: University Press.

Lewis, H. R., & Papadimitriou, C. H. (1997). Elements of the theory of computation, Second Edition. Englewood Cliffs, NJ: Prentice-Hall

Li, Y. (2005). *$X^o$ A theory of the syntax-morphology interface.* MA: MIT Press.

Libben, G. (2006). Why study compound processing? An overview of the issue. In G. Libben & G. Jarema (Eds.), *The representation and processing of compound words*, (pp. 1-22). Oxford: Oxford University Press.

Libben, G., & Jarema, G. (2006). *The representation and processing of compound nouns*. New York: Oxford University Press Inc.

Lieber, R. (1992). *Deconstructing morphology: word formation in syntactic theory.* Chicago: University of Chicago Press.

Lieber, R. (2005). *Morphology and lexical semantics.* Cambridge: Cambridge University Press.

Lieber, R., & Scalise, S. (2005). The lexical integrity hypothesis in a new theoretical universe. In G. Booij et al. (Eds.), *On-line proceedings of the fifth Mediterranean morphology meeting (MMM5),* Frejus, 15-18 September, Bologna.

Lieber, R., & Stekauer, P. (2009). Introduction: Status and definition of compounding. In R. Lieber & P. Stekauer (Eds.), *The Oxford handbook of compounding*, (pp. 3-18). Oxford: Oxford University Press.

Lieven, E. (2006). Language development: Overview. In K. Brown (ed.), *Encyclopedia of language & linguistics*, *2nd edition, vol.13*, (pp. 376-391). Oxford, UK: Elsevier.

Lightbown, P. M., & Spada, N. (1999). *How language is learned*. Oxford: Oxford University Press.

Littlewood, W. (1984). *Foreign and second language learning: Language acquisition research and its implications for the classroom*. Cambridge: Cambridge University Pres

MacDonald, S., & Ramscar, M. (2001). Testing the distributional hypothesis: The influence of context on judgements of semantic similarity. *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, University of Edinburgh.

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

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition, 29*, 121-157.

Manova, S. (2005). Derivation versus inflection in three inflecting languages. In W. U. Dressler, D. Kavtovsky, O. Pfeiffer & F. Rainer (Eds.), *Morphology and its* *demarcations. Selected papers from the 11th International Morphology Meeting,* (pp. 233-252). Vienna, Feb. 2004.

Manova, S. (2008). Closing suffixes and the structure of the Slavic word: Movierung. In *Austrian contributions to the fourteenth international congress of Slavists*, Ohrid, Macedonia, September 2008, Wiener Slavistisches Jahrbuch, 54, 91–104.

Manova, S. (2010). *Understanding morphological rules. With special emphasis on conversion and subtraction in Bulgarian, Russian and Serbo-croatian*. Dordrecht: Springer.

Manova, S., & Aronoff, M. (2010). Modeling affix order. *Morphology, 20,* 109-131.

Marantz, A. (2013). No escape from morphemes in morphological processing. *Language and Cognitive Processes*, 2013. http://dx.doi.org/10.1080/01690965.2013.779385

Marcus, G., Vijayan, S., Rao, S., & Vishton, P. M. (1999). Rule-learning in seven-month-old infants. *Science, 238,* 77-80.

Marquis, A., & Shi, R. (2012). Initial morphological learning in preverbal infants. *Cognition, 122*, 61–66.

Marr, D. (1982). *Vision: A computational approach*. San Francisco: Freeman & Co Publications.

Marslen-Wilson, W. (1999). Abstractness and combination: The morphemic lexicon. In S. Garrod & M. J. Pickering (Eds.), *Language processing*, (pp 101–119). East Sussex, UK: Psychology Press.

Marslen-Wilson, W. D., Zhou, X. L., & Ford, M. (1996). Morphology, modality, and lexical architecture. In G. Booij & J. Van Mark (Eds.), *Yearbook of morphology*, (pp. 117-134). Dordrecht: Kluwer.

Matthews, P. H. (1991). *Morphology. 2nd ed.* Cambridge: Cambridge University Press.

Mattys, S. L., Juszcyk, P. W., Luce, P. A., & Morgan, J. L. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive Psychology*, *38*, 465-494.

McCarthy, J. J. (1979). *Formal problems in Semitic phonology and morphology*. Unpublished PhD Dissertation, MIT, Cambridge, MA.

McCarthy, J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry, 12,* 373-418.

McCarthy, J. J. (2002). *The foundations of optimality theory.* Cambridge, England: Cambridge University Press.

McCarthy, J. J., & Prince, A. (1990). Foot and word in prosodic morphology. *Natural Language and Linguistic Theory, 8,* 209-283.

McCarthy, J. J., & Prince, A. (1993). Generalized alignment. In G. Booij & J. van Marle (Eds.), *Yearbook of morphology.*

McCarthy, M. (1991). Morphology. In K. Malmkjær (ed.), *The linguistic encyclopedia*. Routledge, (pp. 314-323). London & New York: Routledge

McClelland, J. L., & Patterson, K. (2002). Rules or connections in past tense inflections: what does the evidence rule out? *Trends in Cognitive Science, 6*(11), 465–472.

Mithun, M. (1999). *The languages of native North America.* Cambridge: Cambridge University Press.

Mithun, M. (2010). The fluidity of recursion and its implications. In H. v.d. Hulst (ed.), *Recursion and human language,* (pp. 17-41). Berlin: De Gruyter Mouton.

Mohri, M. (1997). Finite-state transducers in language and speech processing. *Computational Linguistics, 23*(2), 269–311.

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition, 96,* 143-182.

Monson, C. (2008). *Paramor: From paradigm structure to natural language morphology induction*. PhD Dissertation, Language Technologies Institute, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA.

Morgan, J. L., Meier, R. P., & Newport, E. L. (1987). Structural packaging in the input to language learning: Contributions of prosodic and morphological marking of phrases to the acquisition of language. *Cognitive Psychology, 19,* 498-550.

Moss C. (1988). *Elephant memories: Thirteen years in the life of an elephant family*. New York: William Morrow and Company.

Neal, R., & Hinton, G. (1998). A new view of the EM algorithm that justifies incremental and other variants. In M. I. Jordan (ed.), *Learning in Graphical Models*, (pp. 355–368). Dordrecht: Kluwer.

Newman, R., Ratner, N. B., Jusczyk, A. M., Jusczyk, P. W., & Dow, K. A. (2006). Infants' early ability to segment the conversational speech signal predicts later language development: A retrospective analysis. *Developmental Psychology, 42*(4), 643–655.

Newport, E. L. (1990). Maturational constraints on language learning. *Cognitive Science, 34*, 11-28.

Newport, E. L., & R. Meier (1985). The acquisition of American sign language. In D. I. Slobin (ed.), *The crosslinguistic study of language acquisition*. Hillsdale, NJ: Erlbaum.

Nişanyan, S. (2009). *Sözlerin soyağacı: Çağdaş Türkçe'nin etimolojik sözlüğü.* İstanbul: Everest.

Niyogi, P. (2006). *The computational nature of language learning and evolution*. Cambridge, MA: MIT Press.

Oflazer, K. (1994). Two-level description of Turkish Morphology. *Literary and Linguistic Computing, 9*(2), 137-148.

Oflazer, K., & El-Kahlout, İ. D. (2007). Exploring different representational units in English-to-Turkish statistical machine translation. *Statistical Machine Translation Workshop at ACL*, Prague, Czech Republic, 25-32.

Orgun, C. O. (1996). *Sign-based morphology and phonology: with special attention to Optimality Theory*. Unpublished PhD Dissertation, University of California, Berkeley.

Orgun, C. O. (1999). Sign-based morphology: A declarative theory of phonology-morphology interleaving. In B. Hermans & M. von Oostendorp (Eds.), *The derivational residue in phonological Optimality Theory*, (pp. 247-267). Amsterdam: John Benjamins.

Orgun, C. O., & Sprouse, R. L. (1996). From MParse to control: Deriving ungrammaticality. *Phonology, 16,* 191-224.

Ostler, N. (2008). Is it globalization that endangers languages? In *UNESCO/UNU Conference: Globalization and Languages: Building our Rich Heritage*, Paris, 206–211.

Pagel, M., Atkinson, Q. D., & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature, 449*(11), 717-721.

Perfors, A., Tenenbabum, J. B., Griffiths, T. L., & Xu, F. (2011a). A tutorial introduction to Bayesian models of cognitive development. *Cognition, 120*(3), 302-321.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2006) Poverty of the stimulus? A rational approach. *Proceedings of the 2006 Cognitive Science conference*, Vancouver, BC, Canada.

Perfors, A., Tenenbaum, J. B., & Regier, T. (2011b). The learnability of abstract syntactic principles. *Cognition, 118*(3), 306 – 338.

Perlmutter, D. (1988). The Split-morphology hypothesis: Evidence from Yiddish. In M. Hammond & M. Noonan (Eds.), *Theoretical morphology: Approaches in modern linguistics*, Orlando, Academic Press.

Pesetsky, D. (1995). *Zero syntax: Experiencers and cascades.* Cambridge, MA: MIT Press.

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

Pinker, S. (1989). *Learnability and cognition: the acquisition of argument structure.* Cambridge, MA: MIT Press.

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

Pinker, S., & Jackendoff, R. (2005). The faculty of language: What's special about it? *Cognition, 95,* 201-236.

Pinker, S., & Prince, A. (1988). On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition, 28,* 73–193.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields, 102*, 145–158.

Plag, I. (1996). Selectional restrictions in English suffixation revisited. A reply to Fabb (1988). *Linguistics, 34*, 769–798.

Plag, I. (1999). *Morphological productivity. Structural constraints in English derivation*. Berlin: Mouton de Gruyter.

Plag, I., & Baayen, H. (2009). Suffix ordering and morphological processing. *Language, 85*(1), 109-152.

Plunkett, K., & Marchman, V. (1991) U-shaped learning and frequency effects in a multi-layered perceptron: implications for child language acquisition. *Cognition, 38*(1), 43–102.

Pollard, C., & Sag, I. A. (1987). *Information-based syntax and semantics. Volume 1. Fundamentals*. CLSI Lecture Notes 13. Stanford, CA: Center for the Study of Language and Information.

Prince, A., & Smolensky, P. (1993). *Optimality Theory: constraint interaction in generative grammar.* New Brunswick: Rutgers University.

Pullum G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *The Linguistic Review, 19*, 9-50

Pustejovsky, J. (1995). *The generative grammar.* Cambridge, MA: MIT Press.

Pycha, A., Novak, P., Shosted, R., & Shin, E. (2003). Phonological rule-learning and its implications for a theory of vowel harmony. In *Proceedings of WCCFL 22*, G. Garding & M. Tsujimura (Eds.), 423-435.

Qu, W., Ringlstetter, C., & Goebel, R. (2008). Targeting Chinese nominal compounds in corpora. *Proceedings of the LREC2008.*

Ramscar, M., & Yarlett, D. (2007). Linguistic self-correction in the absence of feedback: A new approach to the logical problem of language acquisition. *Cognitive Science, 31*, 927-960.

Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. *Cognitive Science, 29,* 1007-1028.

Regier, T., & Gahl, S. (2004). Learning the unlearnable: the role of missing evidence. *Cognition, 93*, 147–155.

Rice, K. (2000). *Morpheme order and semantic scope*. Cambridge: Cambridge University Press.

Rice, K. (2009). *Principles of affix ordering: An overview.* Paper presented at the 2nd Vienna Workshop on Affix Order, Vienna, June 5-6, 2009.

Riehemann, S. Z. (2001). *A constructional approach to idioms and word formation*. PhD Dissertation, Stanford University, Stanford.

Ritchie, G., Black, A., Pulman, S., & Russell, G. (1987). The Edinburgh/Cambridge morphological analyser and dictionary system (version 3.0) user manual. *Technical Report Software Paper No. 10*, Department of Artificial Intelligence, University of Edinburgh.

Roark, B., & Sproat, R. (2007). *Computational approaches to morphology and syntax.* Oxford: Oxford University Press.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J.L. McClelland & D.E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructures of cognition, Vol. II*, (pp. 216–271). Cambridge, MA: MIT Press.

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

Saffran, J. R., Hauser, M., Seibel, R., Kapfhamer, J., Tsao, F., & Cushman F. (2008). Grammatical pattern learning by human infants and cotton-top tamarin monkeys. *Cognition, 107*(2), 479-500.

Saffran, J. R., & Thiessen, E. D. (2003). Pattern induction by infant language learners. *Developmental Psychology, 39*, 484-494.

Sag, I. A. (2007). *Sign-based construction grammar: an informal synopsis*. MS. thesis, Stanford University, Stanford, CA.

Sag, I. A., Wason, T., & Bender, E. M. (2003). *Syntactic theory: a formal introduction.* Stanford: CSLI.

Sak, H., Güngör, T., Saraçlar, M. (2007). Morphological disambiguation of Turkish text with perceptron algorithm. (2007). In A. Gelbukh (ed.) *CICLing 2007. LNCS, vol. 4394*, (pp. 107– 118). Springer, Heidelberg.

Sak, H., Güngör, T., Saraçlar, M. (2011). Resources for Turkish morphological processing. *Language Resources and Evaluation, 45* (2), 249–261.

Say, B., Zeyrek, D., Oflazer, K., & Ozge, U. (2002). Development of a corpus and a treebank for present-day written Turkish. *Proc. of the Eleventh International Conference of Turkish Linguistics*, Famagusta, Northern Cyprus.

Schone, P., & Jurafsky, D. (2991). Knowledge-free induction of morphology using latent semantic analysis. In *Proceedings of the 4th conference on computational linguistics,* 67-72.

Schreuder, R., & Baayen, R. H. (1997). How complex simplex words can be. *Journal of Memory and Language, 37,* 118-139.

Schütze, H. (1995). *Ambiguity in language learning: computational and cognitive models*. Unpublished PhD Dissertation, Stanford University.

Schütze, H. (1998). Automatic word sense discrimination. *Computational Linguistics, 24*(1), 97-124.

Shademan, S. (2007). *Grammar and analogy in phonotactic well-formedness judgments*. Ph. D. thesis, University of California, Los Angeles.

Sharma, U., Jugal, K., & Das, R. (2002). Unsupervised learning of morphology for building lexicon for highly inflectional language. In *Proceedings of the ACL-02 workshop on morphological and phonological learning,* 1-10.

Siddiqi, D. (2009). *Syntax within words: Economy, allomorphy, and argument selection in distributed morphology.* Amsterdam; Philadelphia: John Benjamins.

Siegel, D. (1979). *Topics in English morphology.* New York: Garland.

Skinner , B. F. (1957). *Verbal behavior*. New York: Appleton-Century-Crofts.

Slobin, D. (1982). Universal and particular in the acquisition of language. In E. Wanner & L. Gleitman (Eds.), *Language acquisition: The state of the art*, (pp. 128-172). New York: Cambridge University Press.

Slobodchiko, C. N., Paseka, A., & Verdolin, J. L. (2009). Prairie dog alarm calls encode labels about predator colors. *Animal Cognition*, *12*, 435–439.

Smith, K., & Kirby, S. (2008). Cultural evolution: implications for understanding the human language faculty and its evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363(*1509), 3591-3603.

Snow, C. (1995). Issues in the study of input: Finetuning, universality, individual and developmental differences, and necessary causes. In P. Fletcher & B. MacWhinney (Eds.), *The handbook of child language,* (pp. 180-193). Oxford: Blackwell Publishing.

Solak, A., & Oflazer, K. (1993). Design and implementation of a spelling checker for Turkish. *Literary and Linguistic Computing, 8*(3), 113-130.

Spencer, A. (2001). Morphophonological operations. In A. Spencer & A. M. Zwicky (Eds.), *The handbook of morphology*, (pp. 1-10). Oxford: Blackwell Publishing.

Spencer, A., & Zwicky, A. M. (2001). Introduction. In A. Spencer & A. M. Zwicky (Eds.), *The handbook of morphology*, (pp. 123-143). Oxford: Blackwell Publishing.

Sproat, R. (1992). *Morphology and computation*. Cambridge, MA: MIT Press.

Stonham, J. T. (1994). *Combinatorial morphology.* Amsterdam: John Benjamin Publishing Co.

Stump, G. T. (2001). Inflection. In A. Spencer & A. M. Zwicky (Eds.), *The handbook of morphology*, (pp. 13-43). Oxford: Blacwell Publishers.

Stump, G. T. (2001b). *Inflectional morphology: A theory of paradigm structure*. Cambridge: Cambridge University Press

Supalla, T., & E. L. Newport (1978). How many seats in a chair? The derivation of nouns and verbs in American Sign Language. In P. Siple (ed.), *Understanding language through sign language research*. New York: Academic Press.

Teglas, E., Girotto, V., Gonzalez, M., & Bonatti, L. (2007). Intuitions of probabilities shape expectations about the future at 12 months and beyond. In *Proceedings of the National Academy of Sciences of the United States of America, 104*, 19156–19159.

Tekin, Ş. (2001). *İştikakçının köşesi*. Istanbul: Simurg Yayınları.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279-1285.

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

Thiessen, E. D, & Saffran, J. (2004). Spectral tilt as a cue to word segmentation in infancy and adulthood. *Perception and Psychophysics, 66*(5), 779–791.

Tomasello, M. (2000). The item-based nature of children's early syntactic development. *Trends in Cognitive Science, 4*, 156-163.

Trask, R.L. (2004). What is a word? *Working Papers 11*. Department of Linguistics and English Language, University of Sussex.

Treiman, R., Kessler, B., Knewasser, S., Tincoff, R., & Bowman, M. (2000). English speakers' sensitivity to phonotactic patterns. In M. B. Broe & J. Pierrehumbert (Eds.), *Papers in laboratory phonology V: Acquisition and the lexicon*, (pp. 269–282). Cambridge: Cambridge University Press.

von Frisch, K. (1967). *The dance language and orientation of bees*. New York: Harvard University Press.

Webb, G. I. & Brkic, N. (1993). Learning decision lists by prepending inferred rules. In *Proceedings of the AI93 Workshop on Machine Learning and Hybrid Systems*, 6-10, Melbourne

Weber, D. J., Black, H.A., & McConnel, S.R. (1988). *AMPLE: a tool for exploring morphology*. Dallas, TX: Summer Institute of Linguistics.

Wedel, A. (1999). *Turkish emphatic reduplication*. Ms. Thesis, UC Santa Cruz.

Wedel, A. (2000). Perceptual distinctiveness in Turkish emphatic reduplication. In *Proc. of the 19th West Coast Conference on Formal Linguistics*, 546-559.

Weller, M., & Heid, U. (2012). Analyzing and aligning German compound nouns. *Proceedings of the LREC2012*, 2395-2400.

Wicentowski, R. (2002). *Modeling and learning multilingual inflectional morphology in a minimally supervised framework*. PhD Dissertation, John Hopkins University, Baltimore, MD.

Wiebe, B. (1992). *Modeling autosegmental phonology with multi-tape finite state transducers*. Master' Dissertation, Simon Fraser University.

Wothke, K. (1985). *Maschinelle Erlernung und Simulation morphologischer Ableitungsregeln*. PhD Dissertation, Rheinische Friedrich-Wilhelms-Universität zu Bonn.

Wothke, K. (1986). Machine learning of morphological rules by generalization and analogy. In *Proceedings of the 11$^{th}$ Conference on Computational Linguistics*, 289–293, Morristown, NJ.

Xiaojin, Z. (2005). Semi-supervised learning literature survey. *Technical Report 1530*, Department of Computer Sciences, University of Wisconsin, Madison.

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition, 112*, 97–104.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. In *Proceedings of the National Academy of Sciences of the United States of America, 105*, 5012–5015.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review, 114*, 245–272.

Yang, C. (2002). *Knowledge and learning in natural language*. New York: Oxford University Press.

Yang, C. D. (2004). Universal grammar, statistics or both? *Trends in Cognitive Science, 8*(10), 451-456.

Yarowsky, D., & Wicentowski, R. (2000). Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of the 39$^{th}$ annual meeting of the ACL,* 207-216.

Yatbaz, M. A., Yüret, D. (2009). Unsupervised morphological disambiguation using statistical language models. In *Proceedings of the NIPS 2009 Workshop on Grammar Induction, Representation of Language and Language Learning*.

Yavas, M. (1980). *Borrowing and its implications for Turkish phonology*. Unpublished PhD Dissertation, University of Kansas.

Yu, A. (1998). *On the Origin of the Turkic emphatic reduplication*. Ms. Thesis, University of California, Berkeley.

Yüret, D., & Türe, F. (2006). Learning morphological disambiguation rules for Turkish. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, 328-334.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems, 22*(2), 179-214.

Zhang, J., Gao, J., & Zhou, M. (2000). Extraction of Chinese compound words - An experimental study on a very large corpus. *Proceedings of the second workshop on Chinese language processing*, 132-139.

Zwicky, A. M. (1977). *On clitics.* Bloomington IN: Indiana University Linguistics Club.

Zwicky, A. M. (1986). The general case: basic form versus default form. *In Proceedings of the 12th annual meeting of the Berkeley Linguistics Society*, 305–314.

# APPENDIX A

## Transition and Emission Probabilities for *evdekiler* from the Corpus

| | ***START*** | ***evdekiler*** | ***evdekile*** | ***vdekiler*** | ***evdekil*** | ***vdekile*** | ***dekiler*** | ***evdeki*** |
|---|---|---|---|---|---|---|---|---|
| **START** | | 3.8E-05 | 2.7E-05 | | 2.0E-05 | | | 6.6E-05 |
| ***evdekiler*** | | | | | | | | |
| ***evdekile*** | | | | | | | | |
| ***vdekiler*** | | | | | | | | |
| ***evdekil*** | | | | | | | | |
| ***vdekile*** | | | | | | | | |
| ***dekiler*** | | | | | | | | |
| ***evdeki*** | | | | | | | | |
| ***vdekil*** | | | | | | | | |
| ***dekile*** | | | | | | | | |
| ***ekiler*** | | | | | | | | |
| ***evdek*** | | | | | | | | |
| ***vdeki*** | | | | | | | | |
| ***dekil*** | | | | | | | | |
| ***ekile*** | | | | | | | | |
| ***kiler*** | | | | | | | | |
| ***evde*** | | | | | | | | |
| ***vdek*** | | | | | | | | |
| ***deki*** | | | | | | | | |
| ***ekil*** | | | | | | | | |
| ***kile*** | | | | | | | | |
| ***iler*** | | | | | | | | |
| ***evd*** | | | | | | | | |
| ***vde*** | | | | | | | | |
| ***dek*** | | | | | | | | |
| ***eki*** | | | | | | | | |
| ***kil*** | | | | | | | | |
| ***ile*** | | | | | | | | |
| ***ler*** | | | | | | | | |
| ***ev*** | | | | | | | 6.9E-04 | |
| ***vd*** | | | | | | | | |
| ***de*** | | | | | | | | |
| ***ek*** | | | | | | | | |
| ***ki*** | | | | | | | | |

| | | | |
|---|---|---|---|
| ***il*** | | | |
| ***le*** | | | |
| ***er*** | | | |
| ***e*** | 2.3E-05 | 2.3E-05 | |
| ***v*** | | | 1.5E-04 |
| ***d*** | | | |
| ***k*** | | | |
| ***i*** | | | |
| ***l*** | | | |
| ***r*** | | | |
| **END** | | | |

| | ***vdekil*** | ***dekile*** | ***ekiler*** | ***evdek*** | ***vdeki*** | ***dekil*** | ***ekile*** | ***kiler*** | ***evde*** |
|---|---|---|---|---|---|---|---|---|---|
| **START** | | | | 5.2E-05 | | | | | 4.3E-04 |
| ***evdekiler*** | | | | | | | | | |
| ***evdekile*** | | | | | | | | | |
| ***vdekiler*** | | | | | | | | | |
| ***evdekil*** | | | | | | | | | |
| ***vdekile*** | | | | | | | | | |
| ***dekiler*** | | | | | | | | | |
| ***evdeki*** | | | | | | | | | |
| ***vdekil*** | | | | | | | | | |
| ***dekile*** | | | | | | | | | |
| ***ekiler*** | | | | | | | | | |
| ***evdek*** | | | | | | | | | |
| ***vdeki*** | | | | | | | | | |
| ***dekil*** | | | | | | | | | |
| ***ekile*** | | | | | | | | | |
| ***kiler*** | | | | | | | | | |
| ***evde*** | | | | | | | | 2.1E-02 | |
| ***vdek*** | | | | | | | | | |
| ***deki*** | | | | | | | | | |
| ***ekil*** | | | | | | | | | |
| ***kile*** | | | | | | | | | |
| ***iler*** | | | | | | | | | |
| ***evd*** | | | 1.1E-02 | | | | 1.1E-02 | | |
| ***vde*** | | | | | | | | 1.6E-02 | |
| ***dek*** | | | | | | | | | |
| ***eki*** | | | | | | | | | |
| ***kil*** | | | | | | | | | |
| ***ile*** | | | | | | | | | |
| ***ler*** | | | | | | | | | |
| ***ev*** | | 6.9E-04 | | | | 6.9E-04 | | | |
| ***vd*** | | | 9.0E-03 | | | | 9.0E-03 | | |
| ***de*** | | | | | | | | 9.4E-04 | |
| ***ek*** | | | | | | | | | |
| ***ki*** | | | | | | | | | |
| ***il*** | | | | | | | | | |
| ***le*** | | | | | | | | | |
| ***er*** | | | | | | | | | |
| ***e*** | 2.3E-05 | | | | 9.7E-05 | | | 3.1E-04 | |
| ***v*** | | 1.5E-04 | | | | 1.5E-04 | | | |

| | | |
|---|---|---|
| ***d***<br>***k***<br>***i***<br>***l***<br>***r*** | 2.8E-04 | 2.8E-04 |
| **END** | | |

| | ***vdek*** | ***deki*** | ***ekil*** | ***kile*** | ***iler*** | ***evd*** | ***vde*** | ***dek*** | ***eki*** |
|---|---|---|---|---|---|---|---|---|---|
| **START** | | | | | | 4.0E-04 | | | |
| ***evdekiler*** | | | | | | | | | |
| ***evdekile*** | | | | | | | | | |
| ***vdekiler*** | | | | | | | | | |
| ***evdekil*** | | | | | | | | | |
| ***vdekile*** | | | | | | | | | |
| ***dekiler*** | | | | | | | | | |
| ***evdeki*** | | | | | | | | | |
| ***vdekil*** | | | | | | | | | |
| ***dekile*** | | | | | | | | | |
| ***ekiler*** | | | | | | | | | |
| ***evdek*** | | | | | 2.4E-01 | | | | |
| ***vdeki*** | | | | | | | | | |
| ***dekil*** | | | | | | | | | |
| ***ekile*** | | | | | | | | | |
| ***kiler*** | | | | | | | | | |
| ***evde*** | | | | 2.1E-02 | | | | | |
| ***vdek*** | | | | | 2.3E-01 | | | | |
| ***deki*** | | | | | | | | | |
| ***ekil*** | | | | | | | | | |
| ***kile*** | | | | | | | | | |
| ***iler*** | | | | | | | | | |
| ***evd*** | | | 1.1E-02 | | | | | | 4.7E-02 |
| ***vde*** | | | | 1.6E-02 | | | | | |
| ***dek*** | | | | | 2.2E-02 | | | | |
| ***eki*** | | | | | | | | | |
| ***kil*** | | | | | | | | | |
| ***ile*** | | | | | | | | | |
| ***ler*** | | | | | | | | | |
| ***ev*** | | 2.9E-03 | | | | | | 2.9E-03 | |
| ***vd*** | | | 9.0E-02 | | | | | | 3.9E-02 |
| ***de*** | | | | 9.4E-04 | | | | | |
| ***ek*** | | | | | 3.2E-03 | | | | |
| ***ki*** | | | | | | | | | |
| ***il*** | | | | | | | | | |
| ***le*** | | | | | | | | | |
| ***er*** | | | | | | | | | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| ***e*** | 9.7E-05 | | | 4.9E-04 | | | 1.1E-03 | | |
| ***v*** | | 6.3E-04 | | | | | | 6.3E-04 | |
| ***d*** | | | 2.8E-04 | | | | | | 1.1E-02 |
| ***k*** | | | | | 3.2E-03 | | | | |
| ***i*** | | | | | | | | | |
| ***l*** | | | | | | | | | |
| ***r*** | | | | | | | | | |
| **END** | | | | | | | | | |

| | *kil* | *ile* | *ler* | *ev* | *vd* | *de* | *ek* | *ki* | *il* | *le* |
|---|---|---|---|---|---|---|---|---|---|---|
| **START** | | | | 3.2E-03 | | | | | | |
| ***evdekiler*** | | | | | | | | | | |
| ***evdekile*** | | | | | | | | | | |
| ***vdekiler*** | | | | | | | | | | |
| ***evdekil*** | | | | | | | | | | |
| ***vdekile*** | | | | | | | | | | |
| ***dekiler*** | | | | | | | | | | |
| ***evdeki*** | | | 2.4E-01 | | | | | | | 2.4E-01 |
| ***vdekil*** | | | | | | | | | | |
| ***dekile*** | | | | | | | | | | |
| ***ekiler*** | | | | | | | | | | |
| ***evdek*** | | 2.4E-01 | | | | | | | 2.4E-01 | |
| ***vdeki*** | | | 2.3E-01 | | | | | | | 2.3E-01 |
| ***dekil*** | | | | | | | | | | |
| ***ekile*** | | | | | | | | | | |
| ***kiler*** | | | | | | | | | | |
| ***evde*** | 2.1E-02 | | | | | | | 9.0E-02 | | |
| ***vdek*** | | 2.3E-01 | | | | | | | 2.3E-01 | |
| ***deki*** | | | 2.6E-02 | | | | | | | 2.6E-02 |
| ***ekil*** | | | | | | | | | | |
| ***kile*** | | | | | | | | | | |
| ***iler*** | | | | | | | | | | |
| ***evd*** | | | | | | | 4.7E-02 | | | |
| ***vde*** | 1.6E-02 | | | | | | | 0.07 | | |
| ***dek*** | | 2.2E-02 | | | | | | | 0.02231 | |
| ***eki*** | | | 1.6E-02 | | | | | | | 2.6E-02 |
| ***kil*** | | | | | | | | | | |
| ***ile*** | | | | | | | | | | |
| ***ler*** | | | | | | | | | | |
| ***ev*** | | | | | | 3.2E-02 | | | | |
| ***vd*** | | | | | | | 3.9E-02 | | | |
| ***de*** | 9.4E-03 | | | | | | | 3.6E-02 | | |
| ***ek*** | | 5.1E-03 | | | | | | | 4.6E-02 | |
| ***ki*** | | | 2.6E-02 | | | | | | | 3.8E-02 |

|  |  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|
| ***il*** |  |  |  |  |  |  |  |  |  |  |
| ***le*** |  |  |  |  |  |  |  |  |  |  |
| ***er*** |  |  |  |  |  |  |  |  |  |  |
| ***e*** | 4.4E-03 |  |  |  | 2.1E-03 |  |  | 1.9E-02 |  |  |
| ***v*** |  |  |  |  |  | 9.0E-03 |  |  |  |  |
| ***d*** |  |  |  |  |  |  | 1.3E-02 |  |  |  |
| ***k*** |  | 4.8E-03 |  |  |  |  |  |  | 1.6E-02 |  |
| ***i*** |  |  | 2.0E-02 |  |  |  |  |  |  | 5.1E-02 |
| ***l*** |  |  |  |  |  |  |  |  |  |  |
| ***r*** |  |  |  |  |  |  |  |  |  |  |
| **END** |  |  |  |  |  |  |  |  |  |  |

|  | ***er*** | ***e*** | ***v*** | ***d*** | ***k*** | ***i*** | ***l*** | ***r*** | **END** |
|---|---|---|---|---|---|---|---|---|---|
| **START** |  | 3.0E-02 |  |  |  |  |  |  | 1.5E-01 |
| ***evdekiler*** |  |  |  |  |  |  |  |  |  |
| ***evdekile*** |  |  |  |  |  |  |  | 1.0E00 |  |
| ***vdekiler*** |  |  |  |  |  |  |  |  | 1.5E-01 |
| ***evdekil*** | 1.0E00 | 1.0E00 |  |  |  |  |  |  |  |
| ***vdekile*** |  |  |  |  |  |  |  | 1.0E00 |  |
| ***dekiler*** |  |  |  |  |  |  |  |  | 3.8E-01 |
| ***evdeki*** |  |  |  |  |  |  | 2.4E-01 |  |  |
| ***vdekil*** | 1.0E00 | 1.0E00 |  |  |  |  |  |  |  |
| ***dekile*** |  |  |  |  |  |  |  | 1.0E00 |  |
| ***ekiler*** |  |  |  |  |  |  |  |  | 3.7E-01 |
| ***evdek*** |  |  |  |  |  | 1.0E00 |  |  |  |
| ***vdeki*** |  |  |  |  |  |  | 2.3E-01 |  |  |
| ***dekil*** | 1.0E00 | 1.0E00 |  |  |  |  |  |  |  |
| ***ekile*** |  |  |  |  |  |  |  | 6.4E-01 |  |
| ***kiler*** |  |  |  |  |  |  |  |  | 2.9E-01 |
| ***evde*** |  |  |  |  | 9.0E-02 |  |  |  |  |
| ***vdek*** |  |  |  |  |  | 1.0E00 |  |  |  |
| ***deki*** |  |  |  |  |  |  | 2.6E-02 |  |  |
| ***ekil*** | 7.0E-02 | 1.1E-01 |  |  |  |  |  |  |  |
| ***kile*** |  |  |  |  |  |  |  | 6.7E-01 |  |
| ***iler*** |  |  |  |  |  |  |  |  | 3.0E-01 |
| ***evd*** |  | 5.2E-01 |  |  |  |  |  |  |  |
| ***vde*** |  |  |  |  | 7.0E-02 |  |  |  |  |
| ***dek*** |  |  |  |  |  | 8.4E-01 |  |  |  |
| ***eki*** |  |  |  |  |  |  | 2.3E-01 |  |  |
| ***kil*** | 2.1E-01 | 3.1E-01 |  |  |  |  |  |  |  |
| ***ile*** |  |  |  |  |  |  |  | 3.9E-01 |  |
| ***ler*** |  |  |  |  |  |  |  |  | 2.5E-01 |
| ***ev*** |  |  |  | 6.2E- |  |  |  |  |  |

|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
|  |  |  |  | 02 |  |  |  |  |  |
| ***vd*** |  | 5.5E-01 |  |  |  |  |  |  |  |
| ***de*** |  |  |  |  | 4.2E-02 |  |  |  |  |
| ***ek*** |  |  |  |  |  | 2.0E-01 |  |  |  |
| ***ki*** |  |  |  |  |  |  | 1.3E-01 |  |  |
| ***il*** | 1.4E-01 | 3.5E-01 |  |  |  |  |  |  |  |
| ***le*** |  |  |  |  |  |  |  | 4.6E-01 |  |
| ***er*** |  |  |  |  |  |  |  |  | 2.1E-01 |
| ***e*** |  |  | 3.3E-02 |  | 9.6E-02 |  |  | 2.4E-01 |  |
| ***v*** |  |  |  | 1.6E-02 |  |  |  |  |  |
| ***d*** |  | 3.0E-01 |  |  |  |  |  |  |  |
| ***k*** |  |  |  |  |  | 1.2E-01 |  |  |  |
| ***i*** |  |  |  |  |  |  | 1.5E-01 |  |  |
| ***l*** | 1.1E-01 | 2.4E-01 |  |  |  |  |  |  |  |
| ***r*** |  |  |  |  |  |  |  |  | 2.6E-01 |
| **END** |  |  |  |  |  |  |  |  |  |

| OUTPUT | N9 | N8 | N7 | N6 | N5 | N4 | N3 | N2 | N1 | START | END |
|---|---|---|---|---|---|---|---|---|---|---|---|
| ***evdekiler*** | 3.8E-05 | | | | | | | | | | |
| ***evdekile*** | | 2.7E-05 | | | | | | | | | |
| ***vdekiler*** | | 2.7E-05 | | | | | | | | | |
| ***evdekil*** | | | 2.0E-05 | | | | | | | | |
| ***vdekile*** | | | 2.0E-05 | | | | | | | | |
| ***dekiler*** | | | 1.6E-04 | | | | | | | | |
| ***evdeki*** | | | | 6.6E-05 | | | | | | | |
| ***vdekil*** | | | | 1.6E-05 | | | | | | | |
| ***dekile*** | | | | 1.3E-04 | | | | | | | |
| ***ekiler*** | | | | 2.1E-04 | | | | | | | |
| ***evdek*** | | | | | 5.2E-05 | | | | | | |
| ***vdeki*** | | | | | 5.3E-05 | | | | | | |
| ***dekil*** | | | | | 1.0E-04 | | | | | | |
| ***ekile*** | | | | | 2.6E-04 | | | | | | |
| ***kiler*** | | | | | 1.2E-03 | | | | | | |
| ***evde*** | | | | | | 5.1E-04 | | | | | |
| ***vdek*** | | | | | | 4.7E-05 | | | | | |
| ***deki*** | | | | | | 3.4E-03 | | | | | |
| ***ekil*** | | | | | | 2.1E-03 | | | | | |
| ***kile*** | | | | | | 1.6E-03 | | | | | |
| ***iler*** | | | | | | 8.7E-03 | | | | | |
| ***evd*** | | | | | | | 8.8E-04 | | | | |
| ***vde*** | | | | | | | 6.0E-04 | | | | |
| ***dek*** | | | | | | | 3.6E-03 | | | | |
| ***eki*** | | | | | | | 8.1E | | | | |

| | | | | | |
|---|---|---|---|---|---|
| | -03 | | | | |
| ***kil*** | 4.6E-03 | | | | |
| ***ile*** | 2.0E-02 | | | | |
| ***ler*** | 4.3E-02 | | | | |
| ***ev*** | | 1.3E-02 | | | |
| ***vd*** | | 1.0E-03 | | | |
| ***de*** | | 8.0E-02 | | | |
| ***ek*** | | 3.8E-02 | | | |
| ***ki*** | | 3.4E-02 | | | |
| ***il*** | | 5.4E-02 | | | |
| ***le*** | | 8.8E-02 | | | |
| ***er*** | | 9.5E-02 | | | |
| ***e*** | | | 3.3E-01 | | |
| ***v*** | | | 5.2E-02 | | |
| ***d*** | | | 2.2E-01 | | |
| ***k*** | | | 2.3E-01 | | |
| ***i*** | | | 3.1E-01 | | |
| ***l*** | | | 3.0E-01 | | |
| ***r*** | | | 3.2E-01 | | |
| ***ε*** | | | | 1.0E00 | 1.0E00 |

# APPENDIX B

## Transition and Emission Probabilities for *evdekiler* from the Treebank

| | ***ler*** | ***ev*** | ***de*** | ***ek*** | ***ki*** | **END** |
|---|---|---|---|---|---|---|
| **START** | | 3.7E-03 | | | | |
| ***evdekiler*** | | | | | | |
| ***evdekile*** | | | | | | |
| ***vdekiler*** | | | | | | |
| ***evdekil*** | | | | | | |
| ***vdekile*** | | | | | | |
| ***dekiler*** | | | | | | |
| ***evdeki*** | | | | | | |
| ***vdekil*** | | | | | | |
| ***dekile*** | | | | | | |
| ***ekiler*** | | | | | | |
| ***evdek*** | | | | | | |
| ***vdeki*** | | | | | | |
| ***dekil*** | | | | | | |
| ***ekile*** | | | | | | |
| ***kiler*** | | | | | | |
| ***evde*** | | | | | | |
| ***vdek*** | | | | | | |
| ***deki*** | | | | | | |
| ***ekil*** | | | | | | |
| ***kile*** | | | | | | |
| ***iler*** | | | | | | |
| ***evd*** | | | | | | |
| ***vde*** | | | | | | |
| ***dek*** | | | | | | |
| ***eki*** | | | | | | |
| ***kil*** | | | | | | |
| ***ile*** | | | | | | |
| ***ler*** | | | | | | 2.6E-01 |
| ***ev*** | | | 2.8E-02 | | | |
| ***vd*** | | | | | | |
| ***de*** | | | | | 7.1E-01 | |
| ***ek*** | | | | | | |
| ***ki*** | 4.4E-02 | | | | | |
| ***il*** | | | | | | |
| ***le*** | | | | | | |

| |
|---|
| ***er*** |
| ***e*** |
| ***v*** |
| ***d*** |
| ***e*** |
| ***k*** |
| ***i*** |
| ***l*** |
| ***e*** |
| ***r*** |
| **END** |

| **OUTPUT** | **N5** | **N3** | **N2** | **N1** | **START** | **END** |
|---|---|---|---|---|---|---|
| ***evdekiler*** | | | | | | |
| ***evdekile*** | | | | | | |
| ***vdekiler*** | | | | | | |
| ***evdekil*** | | | | | | |
| ***vdekile*** | | | | | | |
| ***dekiler*** | | | | | | |
| ***evdeki*** | | | | | | |
| ***vdekil*** | | | | | | |
| ***dekile*** | | | | | | |
| ***ekiler*** | | | | | | |
| ***evdek*** | | | | | | |
| ***vdeki*** | | | | | | |
| ***dekil*** | | | | | | |
| ***ekile*** | | | | | | |
| ***kiler*** | 2.0E-04 | | | | | |
| ***evde*** | | | | | | |
| ***vdek*** | | | | | | |
| ***deki*** | | | | | | |
| ***ekil*** | | | | | | |
| ***kile*** | | | | | | |
| ***iler*** | | | | | | |
| ***evd*** | | | | | | |
| ***vde*** | | | | | | |
| ***dek*** | | | | | | |
| ***eki*** | | | | | | |
| ***kil*** | | | | | | |
| ***ile*** | | 1.2E-02 | | | | |
| ***ler*** | | 9.7E-02 | | | | |
| ***ev*** | | | 2.1E-02 | | | |
| ***vd*** | | | | | | |
| ***de*** | | | 2.8E-02 | | | |
| ***ek*** | | | 1.0E-03 | | | |
| ***ki*** | | | 2.9E-02 | | | |
| ***il*** | | | 1.8E-02 | | | |
| ***le*** | | | 2.4E-02 | | | |
| ***er*** | | | 1.1E-02 | | | |
| ***e*** | | | | 6.3E-02 | | |
| ***v*** | | | | 2.0E-04 | | |
| ***d*** | | | | | | |
| ***k*** | | | | 3.7E-02 | | |
| ***i*** | | | | 2.4E-01 | | |
| ***l*** | | | | 3.3E-03 | | |
| ***r*** | | | | 2.0E-02 | | |
| ***ε*** | | | | | 1.0E00 | 1.0E00 |

# APPENDIX C

## Transition and Emission Probabilities for *affına* from the Corpus

| | ***affına*** | ***affın*** | ***ffına*** | ***affı*** | ***ffın*** | ***fına*** | ***aff*** | ***ffı*** |
|---|---|---|---|---|---|---|---|---|
| **START** | 4.8E-06 | 2.2E-05 | | 5.1E-05 | | | 1.1E-04 | |
| ***affına*** | | | | | | | | |
| ***affın*** | | | | | | | | |
| ***ffına*** | | | | | | | | |
| ***affı*** | | | | | | | | |
| ***ffın*** | | | | | | | | |
| ***fına*** | | | | | | | | |
| ***aff*** | | | | | | | | |
| ***ffı*** | | | | | | | | |
| ***fın*** | | | | | | | | |
| ***ına*** | | | | | | | | |
| ***af*** | | | | | | 4.8E-04 | | |
| ***ff*** | | | | | | | | |
| ***fı*** | | | | | | | | |
| ***ın*** | | | | | | | | |
| ***na*** | | | | | | | | |
| ***a*** | | | 6.0E-06 | | 3.5E-05 | | | 9.2E-05 |
| ***f*** | | | | | | 9.6E-05 | | |
| ***ı*** | | | | | | | | |
| ***n*** | | | | | | | | |
| **END** | | | | | | | | |

|  | fın | ına | af | ff | fı | ın | na | a | f | ı | n | END |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **START** |  |  | 4.4E-04 |  |  |  |  | 6.3E-02 |  |  |  |  |
| **affına** |  |  |  |  |  |  |  |  |  |  |  | 1.0E+00 |
| **affın** |  |  |  |  |  |  |  | 1.7E-01 |  |  |  |  |
| **ffına** |  |  |  |  |  |  |  |  |  |  |  | 1.0E+00 |
| **affı** |  |  |  |  |  |  | 6.6E-02 |  |  |  | 3.8E-01 |  |
| **ffın** |  |  |  |  |  |  |  | 1.7E-01 |  |  |  |  |
| **fına** |  |  |  |  |  |  |  |  |  |  |  | 9.8E-01 |
| **aff** |  | 1.5E-02 |  |  |  | 8.6E-02 |  |  |  | 2.3E-01 |  |  |
| **ffı** |  |  |  |  |  |  | 6.6E-02 |  |  |  | 3.8E-01 |  |
| **fın** |  |  |  |  |  |  |  | 5.5E-02 |  |  |  |  |
| **ına** |  |  |  |  |  |  |  |  |  |  |  | 8.7E-01 |
| **af** | 2.8E-03 |  |  |  | 7.4E-03 |  |  |  | 3.2E-02 |  |  |  |
| **ff** |  | 9.7E-03 |  |  |  | 5.6E-02 |  |  |  | 1.5E-01 |  |  |
| **fı** |  |  |  |  |  |  | 2.8E-02 |  |  |  | 5.0E-01 |  |
| **ın** |  |  |  |  |  |  |  | 1.3E-01 |  |  |  |  |
| **na** |  |  |  |  |  |  |  |  |  |  |  | 4.9E-01 |
| **a** |  |  |  | 4.0E-04 |  |  |  |  | 1.2E-02 |  |  | 2.6E-01 |
| **f** | 5.5E-04 | 2.2E-03 |  |  | 1.5E-03 | 4.1E-02 |  |  | 9.9E-03 | 8.1E-02 |  |  |
| **ı** |  |  |  |  |  |  | 4.0E-02 |  |  |  | 3.1E-01 |  |
| **n** |  |  |  |  |  |  |  | 7.8E-02 |  |  |  |  |
| **END** |  |  |  |  |  |  |  |  |  |  |  |  |

| OUTPUT | N6 | N5 | N4 | N3 | N2 | N1 | Start | End |
|---|---|---|---|---|---|---|---|---|
| *affina* | 4.8E-06 | | | | | | | |
| *affin* | | 5.7E-06 | | | | | | |
| *ffina* | | 3.8E-06 | | | | | | |
| *affi* | | | 5.1E-05 | | | | | |
| *ffin* | | | 1.9E-05 | | | | | |
| *fina* | | | 7.8E-05 | | | | | |
| *aff* | | | | 2.0E-04 | | | | |
| *ffi* | | | | 4.6E-05 | | | | |
| *fin* | | | | 1.3E-03 | | | | |
| *ına* | | | | 9.6E-03 | | | | |
| *af* | | | | | 5.8E-03 | | | |
| *ff* | | | | | 2.9E-04 | | | |
| *fi* | | | | | 2.4E-03 | | | |
| *ın* | | | | | 7.0E-02 | | | |
| *na* | | | | | 3.0E-02 | | | |
| *a* | | | | | | 3.9E-01 | | |
| *f* | | | | | | 2.4E-02 | | |
| *ı* | | | | | | 1.8E-01 | | |
| *n* | | | | | | 3.2E-01 | | |
| *ε* | | | | | | | 1.0E-00 | 1.0E-00 |

## APPENDIX D

### Transition and Emission Probabilities for *affına* from the Treebank

| | *af* | END |
|---|---|---|
| START | 9.2E-05 | |
| *affına* | | |
| *affın* | | |
| *ffına* | | |
| *affı* | | |
| *ffın* | | |
| *fına* | | |
| *aff* | | |
| *ffı* | | |
| *fın* | | |
| *ına* | | |
| *af* | | |
| *ff* | | |
| *fı* | | |
| *ın* | | |
| *na* | | 5.4E-02 |
| *a* | | 1.7E-01 |
| *f* | | |
| *ı* | | |
| *n* | | |
| END | | |

| OUTPUT | N2 | N1 | Start | End |
|---|---|---|---|---|
| *affına* | | | | |
| *affın* | | | | |
| *ffına* | | | | |
| *affı* | | | | |
| *ffın* | | | | |
| *fına* | | | | |
| *aff* | | | | |
| *ffı* | | | | |
| *fın* | | | | |
| *ına* | | | | |
| *af* | 2.5E-04 | | | |
| *ff* | | | | |
| *fı* | | | | |
| *ın* | 5.2E-03 | | | |
| *na* | 6.3E-05 | | | |
| *a* | | 8.1E-02 | | |
| *f* | | | | |
| *ı* | | 2.1E-01 | | |
| *n* | | 5.2E-02 | | |
| *ε* | | | 1.0E+00 | 1.0E+00 |

# APPENDIX E

## CURRICULUM VITAE

### PERSONAL INFORMATION

| | |
|---|---|
| Surname, Name: | Kılıç, Özkan |
| Nationality: | Turkish (TR) |
| Date and Place of Birth: | 28 November 1981, Urfa |
| Marital Status: | Single |
| Phone: | +90 537 492 35 42 |
| email: | ozkan.kilic@gmail.com |

### EDUCATION

| **Degree** | **Institution** | **Year of Graduation** |
|---|---|---|
| MS | METU Cognitive Science | 2007 |
| BS | METU Computer Engineering | 2005 |
| BS | METU Computer Ed. & Inst. Tech. | 2004 |
| High School | Buca EML, İzmir | 1999 |

### WORK EXPERIENCE

| **Year** | **Place** | **Enrollment** |
|---|---|---|
| 2008- Present | Turkish Treasury | Treasury Expert |
| 2006-2008 | Atılım University, CENG | Instructor |
| 2004-2006 | Atılım University, CENG | Teaching Assistant |

### FOREIGN LANGUAGES

Advanced English

### PUBLICATIONS

1. **Kılıç, Ö.**, & Bozşahin, C. (2013). Selection of linker type in emphatic reduplication: Speaker’s intuition meets corpus statistics. To appear in *Proceedings of the CogSci 2013*, Berlin, Germany.
2. **Kılıç, Ö.** (2013). Tracking scrambled word order while reasoning with diagrams. To appear in *Proceedings of the CogSci 2013*, Berlin, Germany
3. Çagıltay, N., Tokdemir, G., **Kılıç, Ö.**, & Topallı, D. (2013). Performing and analyzing non-formal inspections of entity relationship diagram (ERD). *Journal of Systems and Software*. http://dx.doi.org/10.1016/j.jss.2013.03.106

4. Seyidov, I., Kükrer, S., & **Kılıç, Ö.** (2013). Individual and social barriers of Turkish people to climate change and pollution. To appear in *Proceedings of the ICOEST2013*, Cappadocia, Turkey.
5. **Kılıç, Ö.** (2012). Using conditional probabilities and vowel collocations of a corpus of orthographic representations to evaluate nonce words. Proceedings of the Student Session of the *24th European Summer School in Logic, Language, and Information (ESSLLI 2012)*, 6-17 August 2012, Opole, Poland.
6. **Kılıç, Ö.,** & Bozşahin, C. (2012). Semi-supervised morpheme segmentation without morphological analysis. *Proceedings of the LREC 2012*, 21-27 May 2012, İstanbul, Turkey.
7. Tokdemir, G., Cagiltay, N., & **Kilic, O**. (2012). How engineers understand entity relationship diagrams: Insights from eye tracker data. *Proceedings of the IADIS International Conference Information Systems 2012*, Berlin, Germany, 10-12 March 2012.
8. **Kilic, O.**, Say B., & Demirörs, O. (2011). An experimental study on the cognitive characteristics of modeling notations. In F. V. C., Ficarra et al. (Eds.), *Advances in Dynamic and Static Media for Interactive Systems: Communicability, Computer Science and Design*, *11*:145-155. Blue Herons Editions: Italy. ISBN: 978-88-96471-08-1.
9. **Kilic, O.**, & Solak, A. (2008). Recent improvements in satellite networks for search and rescue: MEOSAR. *Proceeding of 4th ASMS 2008*, pp. 317-319, Bologna, Italy.
10. **Kilic, O.**, Say, B., & Demirors, O. (2008). Cognitive aspects of error finding on a simulation conceptual modeling notation. *Proocedings of Computer and Information Sciences 2008, ISCIS '08*, pp.1-6. doi: 10.1109/ISCIS.2008.4717930.
11. **Kılıç, Ö**., & Akman, İ. (2007). Türkiye'de ve yurt dışındaki bilgisayar bilimleri yüksek lisans programlarının karşılaştırması. *III. Lisansüstü Eğitim Sempozyumu: Lisansüstü Eğitimde Sorunlar ve Çözüm Önerileri*, Eskişehir. (Comparing Computer Science Graduate Programs in Turkey and Abroad. Proceedings of the 3rd Symposium on Graduate Education: The Problems in Graduate Education and Solution Proposals, Eskişehir, Turkey)
12. Kalem, G., **Kılıç, Ö.**, & Akman, İ. (2007). Lisansüstü öğrencilerinin proje ve tez yürütme sorunları. *III. Lisansüstü Eğitim Sempozyumu: Lisansüstü Eğitimde Sorunlar ve Çözüm Önerileri*, Eskişehir. (The Project and the Thesis Progress Problems of Graduate Students. Proceedings of the 3rd Symposium on Graduate Education: The Problems in Graduate Education and Solution Proposals, Eskişehir, Turkey).
13. **Kılıç, Ö.**, Say, B. & Demirörs, O. (2007). Kavramsal modelleme diyagramlarının bilişsel incelemesi. 2007 Ulusal Yazılım Mühendisliği Sempozyumu, Ankara. (Cognitive Inspection of Conceptual Modeling Diagrams. Proceedings of 2007 National Software Engineering Symposium, Ankara, Turkey)
14. **Kılıç, Ö.**, Akman, İ. (2007). Türkiye'de bilişim teknolojileri meslekleri ve eğitimi. *Ulusal Teknik Eğitim Mühendislik ve Eğitim Bilimleri Genç Araştırmacılar Sempozyumu*, Kocaeli, Bildiri Kitabi, s. 777-781, Kocaeli, Türkiye, 2007. (Information Technology Jobs and Training in Turkey. Proceedings of The National Symposium on Technical Education, Engineering and Education Sciences: Young Researchers, pp. 77-781, Kocaeli, Turkey).
15. Mishra, A., Cagiltay, N. E., **Kilic, O**. (2007). Software engineering education: Some important dimensions. *European Journal of Engineering Education*, Volume 32, Issue 3 June 2007, pp. 349 – 361. (http://dx.doi.org/10.1080/03043790701278607)
16. Öney, M. U., Çevik, A., Çağıltay, N.E., **Kılıç, Ö.** (2007). Topluluk zekası yönetimi ve optimizasyonu. *Akademik Bilişim Konferansı*, 2007, Kütahya (Collective Behavior Management and Optimization. Proceedings of the Conference of Academic Information, 2007, Kütahya, Turkey)
17. **Kılıç, Ö.,** Çağıltay, N. E., Tokdemir, G. (2006). Yazılım mühendisliği diyagramlarının kullanımındaki bilişsel ve davranışsal özellikler. *II. Mühendislik Kongresi*, Zonguldak,

Bildiri ve Poster Kitabı, s. 349-355, Türkiye, 2006. (Cognitive and Behavioral Properties of Using Software Engineering Diagram. Proceeding of the 2nd Engineering Conference, pp.349-355, Zonguldak, Turkey)

## HOBBIES

Animated Movies, Computer Technologies, Mangas, Movies

# APPENDIX F

## VITA

Özkan Kılıç was born in Urfa, on November 28, 1981. He received his B.S. (high honors) degree in Computer Education and Instructional Technologies from Middle East Technical University in June 2004. He received his B.S. (double major) in Computer Engineering from Middle East Technical University in June 2005. He also received M.S. degree in Cognitive Science, Graduate School of Informatics from the same university. He worked as a research assistant and instructor in the Department of Computer Engineering, Atılım University in Ankara. He is currently a treasury expert at the Department of System Development in Turkish Treasury. His main areas of interest are Computational Morphology, Diagrammatic Reasoning, Eye Tracking and Natural Language Processing.