A BAYESIAN MODEL OF TURKISH DERIVATIONAL MORPHOLOGY

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

UTKU CAN KUNTER

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

JULY 2023

A BAYESIAN MODEL OF TURKISH DERIVATIONAL MORPHOLOGY

submitted by **UTKU CAN KUNTER** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Cognitive Science Department, Middle East Technical University** by,

Prof. Dr. Banu Günel Kılıç
Dean, **Graduate School of Informatics**

Dr. Ceyhan Temürcü
Head of Department, **Cognitive Science**

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

Examining Committee Members:

Prof. Dr. Özgür Aydın
Department of Linguistics, Ankara University

Prof. Dr. Cem Bozşahin
Department of Cognitive Science, METU

Assoc. Prof. Dr. Burcu Can
Department of Computing Science and Mathematics, University of Stirling

Assist. Prof. Dr. Umut Özge
Department of Cognitive Science, METU

Assoc. Prof. Dr. Barbaros Yet
Department of Cognitive Science, METU

Date: 21.07.2023

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

Name, Surname: Utku Can Kunter

Signature :

ABSTRACT

A BAYESIAN MODEL OF TURKISH DERIVATIONAL MORPHOLOGY

Kunter, Utku Can
Ph.D., Department of Cognitive Science
Supervisor: Prof. Dr. Cem Bozşahin

July 2023, 306 pages

Building on an extensive review of the psycholinguistics literature and Turkish Derivational Morphology (DM), we propose a novel structure for representing DM in three hierarchical layers: segmentation, lexical selection and derivation. This proposal involves laying a belief structure over the traditional morphological structure of DM. We call this novel structure the Conventionalized Structure (CdS). We develop a computational model of morphology processing based on CdS using Bayesian Belief Networks (BBN). We present an algorithmic implementation for this model that learns and accurately represents new lexical items, recognizes affixes and tracks the salience of each item probabilistically. We carry out trials on this model with realistic observation lists and observe that model predictions are in line with the findings in studies in psycholinguistics. To support our claims and methodology, we carry out an extensive study of Turkish DM, looking into both Modern Turkish and Orkhon Turkic. We also look into the distributional semantics of derivational affixes and observe a high degree of regularity. In order to represent the complex semantics arising from interactions between morphemes, we use the categorial grammar framework. We build a baseline grammar, based on which we construct observation lists for exploration trials. While we focus on Turkish DM, we do not make any language-specific assumptions, our methods and results should be generalizable to other languages with segmental morphology.

Keywords: Derivational Morphology, Categorial Grammar, Bayesian Belief Networks

ÖZ

TÜRKÇE TÜRETİM MORFOLOJİSİNİN BAYES AĞLARI İLE MODELLENMESİ

Kunter, Utku Can
Doktora, Bilişsel Bilimler Bölümü
Tez Yöneticisi: Prof. Dr. Cem Bozşahin

Temmuz 2023, 306 sayfa

Türkçe yapım eklerinin ve psikodilbilim literatürünün geniş bir incelemesine dayanarak, yapım eklerinin 3 hiyerarşik seviyede temsil edileceği yeni bir yapı önerilmektedir. Bu seviyeler bölümleme, sözcük seçimi ve türetmedir. Bu öneri, türetim morfolojisinin geleneksel biçimbilgisel yapısının üzerine uyumlamaya izin veren bir yapının yerleştirilmesini kapsamaktadır. Bu yeni yapı esas alınarak ve Bayes ağları yöntemi kullanılarak bir hesaplama modeli geliştirilmektedir. Modelin hesaplamalı bir uygulaması oluşturulmakta, yeni sözcükleri isabetli şekilde temsil ettiği, yeni ekleri beklenen şekilde öğrendiği ve tüm unsurların belirginliğini istatistiksel olarak takip ettiği gösterilmektedir. Gerçekçi gözlem listeleri üzerinde yapılan denemelerde modelin psikodilbilim alanındaki gözlemlerle uyuşan tahminler yaptığı ortaya konmaktadır. İddiaları ve yöntemi desteklemek için Türkçe türetim morfolojisinin ayrıntılı bir incelemesi yapılmakta, hem Modern Türkçe, hem de Orhon Türkçesi üzerine çalışılmaktadır. Yapım eklerinin dağılımsal anlambilim özellikleri değerlendirilmekte, Türkçe türetim morfolojisinin büyük oranda kurallı bir yapı gösterdiği ortaya konmaktadır. Morfemler arasındaki etkileşimden ortaya çıkan karmaşık anlamların temsil edilebilmesi için kategorik gramer çerçevesi kullanılmaktadır. Temel bir gramer oluşturulmakta, keşif denemelerinde kullanılan gözlem listeleri bu gramer üzerinden oluşturulmaktadır. Türkçe türetim morfolojisine odaklanılmakla birlikte, herhangi bir dile özgü varsayımlarda bulunulmamakta, kullanılan yöntemler ve ulaşılan sonuçların parçasal morfoloji içeren diğer diller için de geçerlilik taşıdığı değerlendirilmektedir.

Anahtar Kelimeler: Türetim Morfolojisi, Kategorik Gramer, Bayes Ağları

To Kadife Hanım,

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my advisor, Cem Bozşahin, for his guidance, patience and understanding.

I would like to thank Barbaros Yet, Burcu Can and Umut Özge for their invaluable feedback.

This study was supported by TÜBİTAK under the 2211/A National Graduate Fellowship Program.

TABLE OF CONTENTS

CHAPTERS

LIST OF TABLES

LIST OF FIGURES

LIST OF ALGORITHMS

ALGORITHMS

LIST OF ABBREVIATIONS

A	Adverb
ABL	Ablative (Case)
ACC	Accusative (Case)
AGG	Agglomerative Clustering
BBN	Bayesian Belief Networks
BOR	Bayesian Occam’s Razor
CCG	Combinatory Categorial Grammar
CDS	Child-Directed Speech
CdS	Conventionalized Structure
CG	Categorial Grammar
CnM	Construction Morphology
CS	Child Speech
DAT	Dative (Case)
DdM	Distributed Morphology
DM	Derivational Morphology / Derivational Morpheme(s)
DN	Derivation Node (BBN)
DS	Distributional Semantics
FLH	Full Listing Hypothesis
GEN	Genitive (Case)
GMM	Gaussian Mixture Models (Clustering)
HFr	High-Frequency
HPr	Highly Productive

IA	Item-and-Arrangement
IM	Inflectional Morphology / Inflectional Morpheme(s)
IOC	Initial Observation Count (BBN)
IP	Item-and-Process
J	Adjective
KM	K-Means Algorithm (Clustering)
LF	Logical Form
LFr	Low-Frequency
LIH	Lexical Integrity Hypothesis
LN	Lexical Node (BBN)
LOC	Locative (Case)
LPr	Lowly Productive
LT	Learning Threshold (BBN)
MDL	Minimum Description Length
MED	Maximum Embedding Dissimilarity (BBN)
MN	Meaning Node (BBN)
MWE	Multi-Word Expression
N	Noun
NACC	Noun in the Accusative
NND	Denominal Nominal Derivation
NNI	Nominal Inflection
NP	Noun Phrase
NVD	Deverbal Nominal Derivation
OT	Orkhon Turkic
PCCG	Probabilistic Combinatory Categorial Grammar
POS	Part of Speech

POSS	Possessive Marker
PredP	Predicate Phrase
PStr	Phrase Structure
RC	Relative Clause
SA	Suspended Affixation
SN	Segmentation Node (BBN)
TAM	Tense-Aspect-Modality
TER	Turkish Emphatic Reduplication
V	Verb
VND	Denominal Verbal Derivation
VVD	Deverbal Verbal Derivation
VVI	Verbal Inflection
WP	Word-and-Paradigm
XXD	Derivational Morpheme with Unspecified Base and Lemma Categories

CHAPTER 1

INTRODUCTION

Through decades of research in computational linguistics, derivational morphology (DM) failed to attract the amount of attention syntax and inflectional morphology (IM) enjoyed. This discrepancy is due to DM having properties that complicate computational analyses. There are three main reasons why DM does not lend itself easily to a computational analysis.

First, DM lacks the systematic and regular application rules that can be exploited by computational models. Simple generative rules fail to represent derivational processes due to extensive semantic selection. This is not the case for IM, where rules are typically applicable to all instances of a certain category.

Second, derivational morphemes are often polysemous. The resulting semantic ambiguity cannot be resolved without contextual clues or supervision. Therefore, computational models must either include the context, or some kind of supervision. Right from the start, models of DM must satisfy more complex requirements.

Finally, derived forms often assume non-compositional meanings. While semantic selection and polysemy make it difficult to come up with generative rules, non-compositional meanings make it impossible. Any realistic model of DM has to accept the existence of non-compositional semantics and somehow incorporate them into the theory. This means that, a model that explains DM must also explain why and how lexicalization occurs. Perhaps this is the core problem in DM.

These difficulties meant that a generative approach to DM has been largely unfruitful. To the best of our knowledge, there has yet to be a comprehensive study of DM in major areas of research. Nevertheless, there have been a few lines of research that approach DM from a computational perspective. For instance, unsupervised learning models of morphology, as reviewed by Hammarström and Borin (2011), assume a complete grammar and rely on surface form similarities. We believe that the assumptions in unsupervised learning models are not psychologically plausible. A second group of studies apply connectionist approaches such as Seidenberg and Gonnerman (2000), trying to explain psycholinguistic data. While quite successful, the opaque nature of connectionist models do not lend themselves well to a theoretical investigation. More direct explorations of DM, such as Mayo (1999), fail to exploit linguistic theory, and simply treat the acquisition process as a software engineering problem.

On the other hand, DM has been more actively studied in the fields of distributional morphology Cao and Rei (2016); Musil et al. (2019); Cotterell and Schütze (2019) and psycholinguistics Burani and Caramazza (1987); Laudanna et al. (1992). Perhaps, methods of these fields has so far been better suited to work with the challenges of DM. Distributional methods allow the researchers to draw insights from vast samples from corpus data. At the other end of the spectrum, psycholinguists are able to extract valuable clues from experiments. However, results from such studies are often inconclusive; their hypotheses are open to interpretation in multiple ways. The search for the structure behind derivational morphology continues.

In this thesis, we take a different approach. Acknowledging that DM lacks the rigid, well-defined structures of syntax and IM, we devise a conventionalized structure (CdS) that represents DM in hierarchical levels of complexity and ambiguity. We aim to achieve a simple, mechanistic explanation of DM, that is also plausible psychologically. We build a model based on Bayesian Belief Networks (BBN) that satisfies these requirements and explore the consequences of adopting this approach.

The simplifying assumptions in this model are guided by our analyses of linguistic data and our theoretical decisions. We do not simply wish to present a procedural model of morphology learning and processing, but we aim to build a theory with adequate explanatory power regarding linguistic facts and observations. In order to do that, we first review the literature to explore different approaches and lines of research. This review also helps us position our research question within the wider literature.

The first point of interest concerns the status of morphology itself. There have been several prominent studies siding with the idea that morphology must be considered an autonomous module, such as Aronoff (1994) and Aronoff and Fudeman (2022), while others such as Lieber (1992) claim that any processes attributed to morphology can actually be explained within the theory of syntax; that morphology is simply syntax below the level of X^0. In our view, most derivational processes can be expressed by generative rules. Nevertheless, we do not claim that the need for a distinct module of morphology can be completely eliminated. As Aronoff (1994) explains, some aspects of morphology (the morphomic level) does not seem to be reducible to phonology and syntax.

The second point of interest concerns the meta structure of morphological operations. There are three main approaches with equivalent power but different assumptions: Item-and-Arrangement (IA), Item-and-Process (IP) and Word-and-Paradigm (WP). IA assumes segmental morphology, IP assumes a procedural mechanism, and WP makes neither assumptions. Construction Morphology (CnM) by Booij (2010) and Distributed Morphology (DdM) by Halle and Marantz (1994) are also important theories that shaped the way we think about morphology. CnM schemas are much more flexible than IA. CnM can also be considered a generalization over IP, since consecutive processes can be organized into several layers. While we believe CnM has the adequate expressive power to represent all of DM, we prefer a more restrictive approach in our investigation. Capabilities of IA are adequate, since we restrict our scope to segmental morphology.

Another age-old question in morphology is whether derivational and inflectional operations employ separate mechanisms. At first glance it seems possible to partition morphology into two non-intersecting sets, but quite often classifying a particular affix turns out to be a matter of definition. Individual affixes do not occur on a binary scale; they are distributed on a continuous spectrum. There are several alternative places where a demarcation line could be drawn between inflection and derivation, but all such lines would be arbitrary in some sense.

In a general sense, inflection is closer to syntax, is more productive and contributes well-defined abstract semantics, while derivation is subject to semantic selection and may change the original concept altogether. These are the traditionally accepted differences between the two sides of morphology. A clear-cut distinction between inflection and derivation may or may not be possible eventually, but for our purposes, we simply call morphological processes required by syntax as inflection, and the rest as derivation.

For our analyses and model building, we restrict our focus to Turkish. Being an agglutinating language with a rich and diversified DM, Turkish offers a suitable testing ground for a theory of morphology. Studying Turkish DM, our main source is the excellent grammar book by Göksel and Kerslake (2005). Ergin (2009), Tekin (2016), Erdal (2004) and Erdal (1991) provided us with important insights. Since it offers a very clear and structured demonstration of Turkish DM, we choose the inventory of affixes presented in Bozşahin (2018) as a starting point.

We work on this initial inventory, compare it to the lists of affixes given in several grammars and examine etymological origins of derived forms. Our aim with this investigation is to discover previously unnoticed similarities and differences between affixes and create a more accurate inventory. Results of our investigations guide our theory and allow us to make informed simplifying assumptions. We find that three notions must be consistently taken into consideration: suppletive allomorphy, polysemy and fusion. Many difficulties in studying DM stem from these three notions. We identify instances of these phenomena to the best of our ability. As a result, we construct a new, and arguably more accurate, inventory for Turkish morphology.

To understand the nature of morphology processing better, we look into how complex forms are acquired and processed. Literature on psycholinguistics offer several prominent studies examining how infants and children are able to discover word-internal structure Tyler and Nagy (1989), Nagy et al. (1993), Saffran et al. (1996), Bertram et al. (2000), Duncan et al. (2009), Givon and Slobin (1985), Peters (2013). Evidence accumulated in this line of research suggests that form-meaning relations are first established for larger forms, i.e. whole words and phrases. Word-internal structure becomes available only after segments common across several complex forms can be recognized.

This is a crucial observation with far-reaching implications. If constituents are acquired by decomposition of wholes, the child's lexicon is different from the adult's lexicon not only in terms of its content, but also in terms of its structure. The child's lexicon is unstructured, lacking the hierarchy of application rules present in the adult's lexicon. Over time, analyses on new observations help the child recognize common constituents inside previously acquired lexical items. From these analyses, rules of application emerge. Not only is this view of morphology acquisition reasonable, but also it is backed up by observations in psycholinguistics research. We make our first simplifying assumption based on this view: Constituents are acquired by the decomposition of wholes.

We do not interpret the analysis of complex forms as a purposeful act, aiming to reduce the size of the lexicon; it is simply an automatic process that recognizes patterns. Numerous studies in computational linguistics Goldsmith (2001), Creutz and Lagus (2007) put a great emphasis on parsimony by evaluating their models according to the Minimum Description Length (MDL) principle. From a cognitive perspective, we are not aware of any psychologically motivated arguments why grammar size would put such a critical strain on mental capacity. MDL and equivalent principles may be employed for hypothesis selection, but they cannot serve as a substitute for linguistic structure. Hypothesis selection is only meaningful when hypotheses are generated by an adequate structure.

The more controversial subject concerns how morphologically complex forms are processed. If a speaker cannot recognize the constituents in a derived form, but recognizes the whole, they must process the item as a whole. This is called retrieval. If a speaker can recognize the constituents, but not the whole, they must derive the meaning of the whole by its constituents. This is called decomposition. But what happens when a speaker can recognize both the constituents and the whole? Once the constituents are available, must the speaker dispense with the whole?

There have been three main views regarding this question: Whole word access models Manelis and Tharp (1977), componential access models Taft and Forster (1975) and dual-access models Bybee (1985), Caramazza et al. (1988). In our view, strong versions of both whole word access and componential access are untenable. One reason is that mental processes that carry out syntactic operations and lexical search seem to be automatic. If both the constituents and the whole are recognizable, it is hard to justify how whole word access and componential access can completely block one another. Dual-access models offer a much more flexible platform for modeling the processing of complex forms. Our second simplifying assumption is that morphologically complex forms are processed in a dual-access manner.

This leads us to an investigation of the decisions a hearer has to make in order to interpret an observation. We believe there are three layers of such decisions, each involving some degree of ambiguity. Alternatives at each layer depend on the decisions from the previous layer. By studying these types of ambiguity, we try to devise a structure that accurately represents morphological processing.

The first layer is concerned with the analysis of the observation's form. This is the segmentation layer. Almost always, it is possible to analyze an observation in multiple ways, especially with auditory input. Which way of analysis will be more prominent is dependent on the previous encounters of the hearer. Since this analysis is the first one to be made, and it determines the alternatives downstream, it is the source for primary ambiguity.

The second layer is concerned with lexical selection. Each segment in the analyzed observation potentially matches with multiple items in the lexicon. In order for the hearer to derive a meaning from the parts, he must be able to choose appropriate lexical items. Again, this choice is affected by previous encounters with the relevant lexical items (priming).

The final layer deals with the different ways in which segments can be brought together. Different derivation sequences may lead to different interpretations, or simply multiple derivations may be possible for the same interpretation. A dedicated mechanism must be able to generate these derivations.

At each layer, alternatives are in competition with each other. Their prominence (or salience), which is dependent on previous observations, must be updated with every observation. This constitutes the metric based on which later preferences develop. We believe such a relationship is best modeled probabilistically. The dependence of downstream decisions on upstream ones also suggests that layers cannot be thought of as isolated modules. An adequate model must take into account the interaction between them.

Bayesian Belief Networks (BBN) offers an adequate representation for this structure. A BBN represents a set of probabilistic variables and their conditional dependencies by a directed acyclic graph. In our case, each source of ambiguity acts as a probabilistic variable; different segmentation alternatives are represented in one variable, while valid lexical alternatives for each segment live in a variable dedicated to that segment. Overall, BBN is the framework we choose for building a model to represent the structure we believe is behind morphological processing.

The data structure depends on our theoretical and experimental choices with regards to the grammar and the lexicon. We must use a framework that is linguistically expressive at an adequate level, as well as computationally efficient. It must especially be suitable for segmental morphology, due to our focus being on an agglutinating language. Categorial Grammar (CG) is such a framework. We adopt CG to create a representative grammar of Turkish morphology. This grammar guides the data structure in our modeling efforts.

An alternative path for semantic representation is Distributional Semantics (DS). In essence, DS is a way of quantifying semantics. Word meanings are represented as vectors (called word embeddings) on massive matrices with hundreds of dimensions. We use word embeddings to measure the semantic similarity and dissimilarity between two words, two affixes or a stem and a derived form. These measures are used as reference during affix recognition. We make an attempt at estimating affix embedding in a similar way to Musil et al. (2019) and obtain evidence that Turkish DM is mostly regular.

With these assumptions and decisions in mind, we build a working model of Turkish DM based on a BBN. This model is able to analyze and learn from observations, as well as recognize new affixes. Salience of each lexical item and each segmentation are tracked and guide the analysis of future observations. The gradual development of preference towards certain segmentation and lexical alternatives is exactly what would be expected from humans based on psycholinguistic theory. We demonstrate the model's capabilities and plausibility on several trials.

Ultimately, we believe the proposed structure and model are not just relevant for morphology processing or morphology acquisition. Since they aim to represent the relations between different lexical items and categories, they act as a model of the lexicon. In a simple, mechanistic manner, we put forward a novel way of looking at how an individual builds a lexicon, how they use it, and how the contents of the lexicon evolve over time.

The contributions of this thesis are five-fold.

First, based on results from the psycholinguistics literature, we argue that the traditional view of morphological structure is not sufficient. Previous studies model only lexical selection and sometimes derivation in order to explain asymmetries in the data, but segmentation must be incorporated as the first step of morphological processing. We present a more adequate structure to represent DM, called the Conventionalized Structure.

Second, we argue that the existing classification of Turkish morphology is not enough for a computational investigation. We examine contemporary and Old Turkic grammars to better understand the overlaps and differences between affix groups. We present a new classification of Turkish morphology. This new classification constitutes the basis for the baseline grammar.

Third, we recognize the drawbacks of a purely theoretical investigation of lexical items. While distributional semantics of words are well-studied, less effort has been paid on finding the distributional semantics of affixes. We show that affix embeddings can be estimated to a reasonable degree by simple vector arithmetic. Based on this investigation, we show that most Turkish DM is regular. Also, we use the dissimilarity between stem and lemma embeddings in our morphology processing model, as a factor that potentially prevents discovery of word-internal structure.

Fourth, we develop a set of grammatical representation rules that is suitable for both morphological and syntactic operations across all syntactic categories. We establish rules for consistency in representation and derivation. We demonstrate that there is structural asymmetry between morphological and syntactic operations in the way they process bound variables from stem logical form (LF). These findings and the baseline grammar are used while conducting trials on the processing model.

Fifth, we show that BBN is an adequate tool to represent the CdS. Network structure of the BBN represents in a compact way the statistical dependence relations between different layers of ambiguity. Based on this structure, our algorithm generates alternative interpretations (hypotheses) for each observation. Hypothesis selection is carried out by the Bayesian Occam's Razor (BOR). We present the full implementation of the algorithm, along with several custom libraries. We conduct trials based on this model. We observe that model predictions regarding the preference between retrieval and decomposition are in line with observations in the psycholinguistics literature.

The rest of the thesis is structured as follows. In Chapter 2, we review the literature on linguistics and psycholinguistics to clearly define the aims, objectives and limitations of this thesis. Based on this review, we propose a novel structure for representing the components of morphological processing. We also go over many different lines of research in search for an appropriate modeling framework.

Chapter 3 clarifies what DM is, with a focus on Turkish. The discussion in that chapter illustrates how different components of DM come together, and how the complex nature of DM defies simple explanations. That complexity requires us to consider the CdS.

In Chapter 4, we start the computational investigation of DM. We first delve into the distributional semantics of derivational affixes. Second, we devise their CG representations in order to represent both the semantic and syntactic properties of derivational affixes.

In Chapter 5, we finally create a Bayesian model of Turkish DM. This model represents the conventionalized structure proposed in Chapter 2, is fed the material collected in Chapter 3 and is built on the data structure developed in Chapter 4. We carry out several trials to examine how the model reacts to input, and how its behavior fits the findings in psycholinguistics research.

The final chapter is reserved for general discussion and future work.

CHAPTER 2

UNDERSTANDING DERIVATIONAL MORPHOLOGY

We start with a descriptive survey of derivational morphology (DM). In this chapter, we develop a novel perspective towards DM, reviewing studies on acquisition, processing and morphological structure.

2.1 Definitions and Dichotomies

Morphology in general, and DM in particular, are not clearly demarcated modules of language. They are regions on a continuum of a variety of linguistic phenomena, rather like colors on a color spectrum. Since there is no consensus on their definitions, the literature is full of different takes on what constitutes morphology and DM. In this section, we present our point of view regarding this debate.

2.1.1 Syntax and Morphology

Possibly one of the most central and controversial dichotomies in the literature is the one concerning syntax and morphology. There have been important works siding with the idea that morphology must be considered a distinct module, such as Aronoff (1994) and Aronoff and Fudeman (2022); while others such as Lieber (1992) claim the exact opposite. On one hand, it really seems morphology distributes itself to different linguistic modules and operations. On the other hand, we feel this does not eliminate the necessity for us to come up with a theory of morphology.

Lieber (1992) claims that any processes attributed to morphology can actually be explained within the theory of syntax; succinctly, in her theory, morphology is what we call syntax below the level of X^0. She argues that compounds are syntactically formed and inflectional processes are basically a part of syntax, but concedes that no one has yet succeeded in deriving the properties of words and sentences from the same basic principles of grammar. Indeed, certain parts of morphology are easier to explain in terms of syntax. The situation also differs across languages; what is relegated to grammar in one language may be expressed in the lexicon of another, and vice versa. Ultimately, we aim to represent the entire set of possibilities from syntax and morphology in the simplest possible framework.

Aronoff (1994) argues for the autonomy of a morphological module. He starts with the Separation Hypothesis:

> Separation Hypothesis: Morphological operations should be separated from accompanying phonological operations.

Aronoff (1994) demonstrates the existence of a morphomic level, which he claims is an autonomous morphological level. He distinguishes between inflectional classes and gender, as two autonomous levels and argue that inflectional classes are purely morphological. He admits that morphology is not necessary for a language, but many languages do have elaborate morphology which cannot be adequately explained without a dedicated theory of morphology.

Our position regarding the existence of a separate morphological module is closer to Lieber (1992), but not entirely in agreement with her thesis that morphology is simply "syntax below the level of X^0". There are three reasons for this: First, syntactic rules are highly regular and (mostly) blind to the semantics of constituents, while morphological processes are often quite sensitive to semantics. Second, we still have plenty of processes that could not be attributed to syntactic operations, despite repetitive attempts, such as the morphomic level in Aronoff (1994).

Grimshaw (1990) puts forward a convincing account of argument structure (a-structure), demonstrating how and why it exists as a distinct layer. She argues that a-structure exists as a layer of linguistic structure distinct from the well-accepted thematic structure (θ-structure). According to her, the lexical conceptual structure (lc-structure) projected directly by the lexical entry generates the θ-structure as well as the event structure (e-structure). The latter contributes an aspectual dimension on the θ-structure, licensing the arguments of the item. The combination of the θ-structure and the e-structure produces the a-structure. While the θ-structure organizes thematic relations into a hierarchy, the a-structure does not include any θ-marks, but organizes and licenses the arguments projected from the lexical entry.

Grimshaw (1990) demonstrates the significance and validity of her arguments on examples selected from psychological verbs and complex event nominalization. These classes of items are special in that their θ-structures and a-structures do not have to coincide, unlike most other constructions.

Since the hierarchies from the θ-structure and from the e-structure coincide most of the time, the existence of a separate a-structure has not always been obvious. What Grimshaw (1990) does is to put forward a theory that is able to explain the outcomes produced in the presence of conflicting structures. Indeed, analyses of several linguistic phenomena involving psychological verbs and complex event nominalization (as well as many others) show how the two structures do not have to coincide. With clear examples, Grimshaw (1990) demonstrates how her theory explains these phenomena quite elegantly. Hale and Keyser (2002) provides a simpler, beginner-friendly discussion of a theory of argument structure.

Assuming every lexical entry is associated with a thematic structure, and citing several sources for the possibility of an a-structure distinct from the θ-structure, Sezer (1991) claims that syntax operates on the thematic structure, while morphology operates on the argument structure. He points out how most derivations involve the addition or suppression of an argument, giving as a typical example passive by-arguments that act both like adjuncts and arguments. This is a powerful claim that touches the heart of the issues we are attempting to investigate. The deepest difference between syntactic and morphological processes may be rooted in their being responsive to different linguistic structures. We have found a similar asymmetry between categorial representations of syntax and morphology, which is discussed in Section 4.5.

Sezer (1991) also emphasizes the distinction between semantic selection (s-selection) and category selection (c-selection). This distinction plays an important role in our investigation of the behavior of

DM. One could say s-selection is a more general kind of restriction and c-selection is a subset of it. While scanning the semantic content of a candidate, s-selection implicitly takes a position regarding its grammatical category. On the other hand, c-selection only takes into account the grammatical categories of candidates. Perhaps inflectional processes are constrained by the principles of c-selection only, while derivational processes are subject to a wider set of constraints resulting from s-selection.

Dowty (1979) offers a different, perhaps the most plausible perspective regarding the delineation of syntactic and morphological processes. Dowty (1979) separates the notions of rule and operation, to partition morphological processes in two dimensions instead of just one.

Table 1: Typology of rules and operations in Dowty (1979)

Operation / Rule	Syntactic Rules	Morphological Rules
Syntactic Operations	A. traditional syntactic rules (PStr-like and transformation-like)	B. rules forming lexical units of more than one word, e.g., Eng. V-Prt combinations
Morphological Operations	C. IM, 'derivational' morphology when unrestricted and semantically regular (polysynthetic lang.)	D. rules introducing deriv. morphology, zero-derivation and compounding, partially productive and less than predictable semantically

With this partition of linguistic processes, Dowty (1979) rejects the syntax-morphology dichotomy and introduces a four-way split. In Table 1, Cell A corresponds to what is traditionally accepted as syntax and D corresponds to word formation. Cell B covers multi-word expressions (MWE) and Cell C's domain is largely inflectional morphology (IM). Sugioka (1986) comments on this classification and emphasizes that the demarcation between Cell A and Cell C presents problems for strictly lexicalist theories. This typology is not immune to problematic examples, as Sugioka (1986) demonstrates. In the next section, we try to find the boundary between inflection and derivation, or in Dowty (1979)'s typology, syntactic and morphological rules.

2.1.2 Inflection and Derivation

At first glance, it seems possible to partition morphology into two non-intersecting sets, inflection and derivation. However, individual affixes do not occur on a binary scale; they are distributed on a continuous spectrum. There are several alternative places where a demarcation line could be drawn between inflection and derivation, but all such lines turn out to depend on arbitrary decisions in some sense. In this section, we compare some properties of inflection and derivation.

Aronoff and Fudeman (2022) provide interesting cases and detailed explanations on both inflectional and derivational morphology, covering a wide range of examples from many languages of the world. As in Aronoff (1994), they argue for the existence of a distinct linguistic component called morphology, claiming some aspects of morphology cannot be attributed to anything else. They define inflection as the formation of grammatical forms of a single lexeme, uses of which are determined by syntax, and derivation as the creation of one lexeme from another, including compounding. Of course these definitions only establish familiar prototypes, and do not clearly distinguish the two kinds of morphology.

In a general sense, inflection is closer to syntax, is more (often fully) productive and contributes a well-defined abstract meaning. On the other hand, derivation is subject to semantic constraints and may change the stem meaning altogether. These are the traditionally accepted differences between the two sides of morphology. Numerous authors studied this question, often suggesting their own criteria for identifying a derivational process. These criteria are usually too vague or too subjective to reliably guide annotation. Most of them have been shown to fail outside the language or language family they were originally tested in. We have reviewed five primary sources for this investigation:

(1) Studies on the distinction between inflection and derivation

a. Scalise (1988) presents the first extensive list of diagnostic criteria.

b. Dressler (1989) builds a larger list, based on an extensive literature review.

c. Stump (2017) offers some insight on the prototypical nature of morphological processes.

d. Booij (2000) presents a systematic investigation.

e. Ten Hacken (2014) offers a relatively recent overview of the academic positions regarding the issue.

Scalise (1988) mentions four main points of view:

(2) a. Inflection and derivation are bound by the same set of rules. They do not differ in a meaningful way.

b. Inflection and derivation occur on a continuous spectrum. Therefore, they only differ in degree, not in kind.

c. Inflection and derivation are different in the way they interact with syntax. They should be considered in different subcomponents of the grammar.

d. Inflection and derivation are both in the lexicon, but they are handled by rules with different formal properties.

All these positions, except the final one, still enjoy significant support among researchers. Demonstrating that the debate still continues decades after Scalise (1988), Ten Hacken (2014) draws a similar picture:

(3) a. Categorizing tradition: Inflection and derivation are distinct categories with a clear boundary between them.

b. Skeptical tradition 1: Drawing a clear boundary between inflection and derivation is not possible. Therefore, our theories should not require it.

c. Skeptical tradition 2: The best theory would not require a clear boundary between inflection and derivation. Therefore, we do not need to find it.

Ten Hacken (2014) points out the reason why this debate has continued for so long. He perfectly captures why efforts for converting prototypes into precise concepts are futile:

> The strength of the prototype is the result of converging criteria, but when these criteria are used in a definition, the differences between the sets of phenomena they identify are highlighted.

According to this point of view, concepts of inflection and derivation are prototypes. A contradiction arises when we use the criteria in the definition of these concepts, because individual affixes may not exactly fit the definition. For the definitions to be meaningful, they must be vague, but if they are vague, it is not possible to make a clear demarcation between inflection and derivation.

We believe that morphological processes occur on a continuous spectrum with inflection on one end and derivation on the other. Any theory that depends on a clear distinction will fail to explain at least a portion of the data. Nevertheless, one can bring clarity to the problem by systematically analyzing the behavior of individual affixes. Here, we present our own criteria with respect to certain linguistic operations. We revisit this question in Section 3.1 with a focus on Turkish DM.

Table 2 shows the criteria we found meaningful in this investigation.

Regarding Criterion A, it has been shown that only derivational affixes may occur more than once around the same stem. This is only logical, as inflectional affixes tend to occur in dedicated slots. the recycler *-ki* provides a way of avoiding this restriction, but it creates a whole new stem in the process. Therefore, we disregard the counterexamples constructed with *-ki*. Although voice markers are generally considered to be part of inflection, Turkish causative markers undeniably display recursivity.

Criterion B concerns polysemy. For this criterion, we do not consider homophonous affixes such as the voice marker *-Iş* and the deverbal nominal derivational affix (NVD) *-Iş*. We consider affixes like *-lIK*, that exhibit a range of polysemous uses; provided that individual uses are relatively productive in their own right.

Many derivational affixes exhibit several forms that are apparently slight phonological variations of each other. These forms cannot be considered ordinary allomorphy, as they often compete with each other to fulfill comparable / slightly different functions. Criterion C tracks whether such variations are present for an affix.

Most authors use the notion of productiveness, pointing out that inflectional affixes tend to be more productive. Perhaps a more accurate formulation of this criterion would refer to the type of selection by the affix. Generally, inflectional affixes only employ syntactic selection (category selection, or c-selection), applying on all instances of a syntactic category. On the other hand, derivational affixes have semantic selection criteria (s-selection); therefore, they apply on a restricted set of stems. Criterion D tracks the existence of semantic selection.

Only Dressler (1989) suggested that an affix appearing on a dictionary entry is probably a derivational affix. Although inflected forms may also lexicalize and be listed in the dictionary, closer inspection reveals that they have gained a totally different role during the lexicalization process. According to Criterion E, an affix is likely to be a derivational affix, if it appears in dictionary entries.

Table 2: Criteria for distinguishing between derivation and inflection

Code	Criterion	Possible Values	References	Verdict	Interesting Cases
A	Recursivity (without -ki)	No, Yes	Scalise (1988), Dressler (1989), Booij (2000)	Decisively suggests DM.	Causative
B	Irregularity in Semantics / Polysemy	None, 2, 3, 4, ...	Dressler (1989), Stump (2017), Booij (2000)	Strongly suggests DM.	
C	Rule Variation / Competition / Deriv. Allomorphy	None, 2, 3, 4, ...	Scalise (1988), Dressler (1989)	Strongly suggests DM.	-(y)An / -AGAn / -GAn / -GIn, -(G)AÇ / -GIÇ
D	Semantic Selection	None, Exceptional, Extensive	Scalise (1988), Dressler (1989), Stump (2017), Booij (2000), Ten Hacken (2014)	Strongly suggests DM.	
E	Dictionary entries	None, 1, 2, 3, . . .	Dressler (1989)	Weakly suggests DM.	
F	Change in Stem Argument Structure	N/A, No, Yes		Weakly suggests DM.	Voice markers, POSS Öztürk and Taylan (2016)
G	Change in Stem POS	No, Yes	Scalise (1988), Dressler (1989), Stump (2017), Booij (2000), Ten Hacken (2014)	Weakly suggests DM.	Participles
*	Order of Application	V-1, V-2, . . . , N-1, N-2, . . .	Scalise (1988), Dressler (1989), Stump (2017), Booij (2000), Ten Hacken (2014)	Being further from the stem weakly suggests IM.	NVD affixes sharing the V-4 slot with TAM markers
H	Phrasal Scope	No, Yes		Weakly suggests IM.	ters tatrü törö-çi Erdal (2004)
I	Member of a Paradigm	No, Yes	Dressler (1989), Booij (2000)	Weakly suggests IM.	
J	Invariability of Order of Application	No, Yes	Scalise (1988)	Strongly suggests IM.	
K	Suspended Affixation	No, Yes		Decisively suggests IM.	tuz ve limonluk Bozşahin (2007); limon ve tuzluk Kornfilt (2012)
L	Required by Syntax	No, Yes	Scalise (1988), Dressler (1989), Stump (2017), Booij (2000)	Decisively suggests IM.	
M	Morpheme Class	Prefix, Suffix, Clitic		A DM can never be a clitic.	

Criterion F and Criterion G complement each other. Any change in either the argument structure or stem part-of-speech (POS) suggests a derivational process. Inflectional processes are only expected to insert information in a dedicated slot, not change the argument structure. An affix might change the argument structure of its stem, without changing its POS. Criterion F is relevant in those cases.

We are aware that, conventionally, voice markers are considered to be inflectional affixes fulfilling a grammatical feature. Nevertheless, we believe their unique position must be recognized. Observing the behavior of voice markers from a categorial point of view, it can be said that they change the stem in a more fundamental way than any derivational affix. Manipulation of the stem argument structure of a verb substantially changes its logical form as well as its syntactic category. Voice markers' ability of changing the argument structure without changing the stem POS does not put them in either classification, but demonstrate the fluidity of the boundary between inflection and derivation.

Again regarding argument structure, we also value the claim by Öztürk and Taylan (2016) that possessives (POSS) are derivational markers due to their licensing an NP argument for the stem, which is also NP. They also demonstrate that POSS interacts with other DM (such as *-CI*) in an interesting way. While the conclusions drawn in Öztürk and Taylan (2016) are less than certain, the controversial status of POSS must be recognized.

Inflectional affixes are not expected to change the stem POS, but derivational affixes often do. This is a weak indicator, though, because affixes such as participles also change the stem POS, but are still generally considered inflectional affixes. Dressler (1989) adds that infinitives act like nouns, despite being an inflected form.

A popular criterion states that DM occur closer to the stem, while IM occur further away. This observation is generally correct; for instance, if we had a definitive derivational affix and another affix closer to the stem, the latter would probably also be a derivational affix. Similarly, an affix that is further from the stem than a definitive inflectional affix would probably be an inflectional affix.

Uniformity of Theta Assignment Hypothesis (UTAH) presented in Baker (1988) (and expanded in Hale and Keyser (2002)) makes the same prediction. As a consequence of UTAH, the distance of an affix to the stem is determined by its place in the thematic structure and derivational affixes are closer to the stem than inflectional affixes.

This cannot be considered a genuine criterion, though, because it operates on relative terms. For instance, it cannot be used on an affix with no neighboring affixes. It is not without exceptions either, NVD affixes like *-GAn* and *-GAÇ* often share the same slot with TAM markers, succeeding voice and negation. Should this mean voice and negation are derivational affixes or *-GAn* and *-GAÇ* are inflectional affixes? Both could be true, but the relative positions of affixes cannot answer this question.

Most inflectional affixes and a few derivational affixes can be shown to take phrasal scope. Criterion H focuses on this difference. Phrasal scope is a weak indicator at best. Among the examples of derivation on a phrasal scope is the Orkhon Turkic *ters tätrü törö-çi* 'followers of wrong teachings' from Erdal (2004). We further investigate this property in Section 3.1 using many examples.

Level-ordering hypothesis in Siegel (1974) is one of the many attempts at describing the organization of morphological operations. However, Sugioka (1986) finds many examples violating the proposed rule. Phrasal scope does not seem to be exclusive to inflectional affixes. Agglutinating languages such as Japanese (and Turkish) offer plenty of such examples.

Membership in a paradigm (Criterion I) correlates with being required by syntax and with having an invariable position with respect to other affixes. All of these, in their own right, are criteria suggesting IM. However, it is not a trivial task to identify paradigms beyond the generally accepted morphosyntactic features such as case, tense and person. Moreover, Stump (1991) and Štekauer (2014) give many examples of derivational paradigms, demonstrating that membership in a paradigm is not exclusive to inflectional affixes. We consider membership in a paradigm as a weak indicator of IM.

As Scalise (1988) points out, there is usually a strict order among inflectional affixes and between inflectional affixes and derivational affixes (Criterion J). However, no such order exists among derivational affixes. Therefore, two derivational affixes X and Y may appear in the XY order on one stem, and in the YX order on another. This is not possible for inflectional affixes, as their position is dictated by syntactic rules. Turkish TAM markers and copular markers are no counterexamples to this generalization, as a copular marker appends on a different stem than the neighboring TAM marker.

As we explain in Section 3.2.4, suspended affixation (SA) is a largely overlooked linguistic phenomenon with significant consequences concerning lexicalism and related concepts. Despite our extensive literature review, we have not found any convincing examples demonstrating SA of Turkish derivational affixes. We do not deny the possibility that a case in Turkish can be found in the future, but until then Criterion K decisively suggests IM for affixes displaying SA.

Several morphosyntactic features are expressed by affixes in dedicated slots around the stem. If a sentence cannot be grammatical when such a slot remains vacant, then the morphemes belonging to that slot are said to be required by syntax. Case and TAM markers are good examples of this. On the other hand, NVD affixes such as *-GAn* and *-GAÇ* cannot be considered as required by syntax. They occupy the same slot as TAM markers, but they do not realize a morphological feature. Instead, they make deep and irregular changes to the semantics of the stem. Criterion L states that affixes required by syntax are part of IM.

Criterion M does not play a crucial role in the diagnostics, because it applies to few morphemes. It still deserves a place among our criteria because it is definitive. A DM can never be a clitic for obvious reasons. Although there have been many attempts at drawing a boundary around clitics, the set of clitics in Turkish is still not entirely agreed upon. As Erdal (2000) so masterfully points out, there are many clitics that are generally accepted as affixes.

2.1.3 Subtypes and Mechanisms

Regarding the mechanism of morphological operations, there are mainly three approaches: item-and-arrangement (IA), item-and-process (IP) Hockett (1954), Aronoff and Fudeman (2022) and the classical word-and-paradigm (WP) Blevins (2016). The reader is referred to Roark and Sproat (2007) for a comprehensive review of these lines of research. While these approaches are ultimately equivalent in power (Aronoff and Fudeman, 2022), they adopt different mechanisms and make different assumptions. IA assumes segmental morphology; it must find ways to represent apparently non-segmental morphology in a segmental manner. In contrast to IA, IP assumes a procedural mechanism, and represents even obviously segmental morphology in terms of processes. WP makes neither assumptions and focuses on paradigms.

Booij (2010) describes the theory of construction morphology (CnM). Constructional schemas are much more flexible than IA. They can also be considered a generalization over IP, since consecutive processes can be organized into layers of construction. This framework is quite interesting but it does not seem practical for a computational study, because generating and organizing construction schemas for a large dataset would be very difficult. Also, it would be more desirable to aim for a theory of morphology where every process operates autonomously, without regard to lower level and higher level categorial / semantic operations. Distributed Morphology (DdM) by Halle and Marantz (1994) was also an important theory that shaped the way we think about morphology.

Several of these mechanisms may coexist during language processing. Segmental morphological operations in Turkic, Japonic and Finno-Ugric languages could perhaps be represented by IA, while templatic morphology of Semitic languages could be represented by IP. It may be possible to apply WP to some aspects of Latin inflection and CnM to others. It is even possible that some morphological operations, such as Turkish Tense-Aspect-Modality (TAM), are processed with both IA and WP concurrently. We simply do not need to make a choice between these mechanisms. Rather, we must apply each one on adequate processes. As we will see in later chapters, Turkish DM is overwhelmingly segmental; lending itself to an adequate representation by IA. In this thesis, we adopt the IA approach.

Nikolaeva (2014) lists the types of Altaic derivational processes as suffixation, prefixation, conversion, compounding and reduplication. The status of conversion is controversial, due to widespread use of words in multiple functions, such as noun-verb (i.e. *boya* 'paint', *boya-*, 'to paint'), adjective-verb (i.e. *kuru* 'dry', *kuru-* 'to dry') and noun-adjective (i.e. *kırmızı* 'red' / 'the red one', *eski* 'old' / 'the old one'). Especially the hypothetical noun-adjective conversion is too productive. Working with a wider category of Turkish substantives (Chomsky, 1993) instead of nouns and adjectives is perhaps more adequate. Noun-verb and adjective-verb conversions are quite rare and could be fully relegated to the lexicon.

Göksel and Kerslake (2005) lists two types of noun compounds, bare compounds and *-(s)I* compounds. These are not based on standard morphemes as in affixation, but they are semantically-driven constructions. As Göksel and Kerslake (2005) emphasizes, the possessive marker *-(s)I* does not signify possession in compounds. Rather it is the compound marker, serving several distinct functions depending on context. Other types of compounds such as auxiliary verbs and incorporation are also semantically-driven and should be relegated to the lexicon. Following Scalise (2011), we believe it is possible that the origin of compounds is sentential. We separate compounding from DM following Olsen (2014) and leave compounding entirely outside the scope of this thesis.

The IA approach adequately represents suffixation, prefixation and, to some extent, reduplication. Among these, prefixation in Turkish is rare and limited to foreign stems. Reduplication has only three kinds listed in Nikolaeva (2014) and in Göksel and Kerslake (2005) as emphatic left-reduplication, generic plural (m-reduplication) and adverbial doubling. In later chapters, we focus on suffixation for the sake of simplicity, without sacrificing much content.

Sugioka (1986) studies the interaction of DM and syntax in Japanese and English. The author examines the plausibility of a universal lexicalist hypothesis based on a wide range of word formation processes in Japanese and English. Japanese being an agglutinating language and English being closer to the analytic side, she believes the examples drawn from these two constitute a good testing ground for her investigation. Observations made in this book are relevant to our investigation.

Sugioka (1986) reviews three main proposals regarding the nature of word formation: Lexical transformation, phrase / word structure rules and lexical insertion. Like Sugioka (1986), we find it hard to believe that the transformationalist hypothesis (Scalise, 2011) could be an adequate explanation for word formation. The proposal that the main mechanism for word formation is lexical insertion is not plausible either. It is clear that there is some regularity in word formation. New words do not just emerge randomly; their semantics are usually compositional; they admit the categories of their internal heads and there are a limited number of bound morphemes used in derivation. Arguing for lexical insertion as a linguistic mechanism would not contribute much to our understanding of Language, anyway.

Sugioka (1986) presents and analyzes a wealth of ideas and examples throughout the book. She puts special emphasis on the analysis of phrasal suffixes, since phrasal scope is probably the clearest clue for positioning a process within the rule typology. The following list of examples is borrowed from Sugioka (1986).

(4) Phrasal suffix examples from Sugioka (1986)

a. story teller

b. pencil pointed

c. two legged

d. cold blooded

e. hydro electricity

f. Godel numbering

g. atomic scientist

h. lexically relatedness

i. reairconditioning

j. set theoretic

k. bathroomless

Argument structures are normally expected to be licensed by a verb, but some noun phrases clearly involve an argument structure without a verbal constituent. It is easy to notice that most such noun phrases include a deverbal nominal, apparently playing the role of the missing verbal constituent. This may either indicate that the argument structure persists despite the change in the host's category or that the suffix takes phrasal scope and applies only after the argument to the verb is fulfilled.

The former alternative means we have to accept that nominals are able to license arguments. The latter alternative means we have to reject that derivational affixes apply before inflectional ones. Correct bracketing of the following examples is crucial to understanding morphology.

(5) Examples on persistent argument structure

a. story teller

b. military historian

c. rocket scientist

d. *metroya yakın ev* ‘a house close to the metro’

e. *benden uzak Allaha yakın* ‘distant from me close to the God’

f. *tüm katılımcılara uygun toplantı saati* ‘a meeting schedule suitable to all participants’

g. *öğle namazını müteakip* ‘following the noon prayer’

Leaving the details to later sections, we argue that it is significant for an affix to take phrasal scope. This is evidence for an underlying syntactic rule in the sense of Dowty (1979), as opposed to what would be expected from conventionally accepted derivational processes.

Finally, let us comment on morphological processes with respect to linguistic typology. With our focus on Turkish, we develop methods particularly suitable for agglutinating languages. These methods are unlikely to be applicable on isolating languages or suitable for analytic languages. On the other hand, to the extent that words can be meaningfully decomposed into morphemes, we aim to generate insights relevant to many different languages. More on this in Section 2.3.4.

2.2 Facts and Data

From the perspective of cognitive science, studying DM is interesting because we assume the speakers acquire and make use of morphological knowledge. This section will investigate the extent to which this knowledge is active, drawing from the literature on psycholinguistics research, including morphology acquisition and processing.

2.2.1 Awareness of DM and Categories

First, we review evidence on speakers’ awareness of DM, starting with studies on youngest speakers. Unfortunately, data on Turkish is not extensive; therefore for some age groups, we refer to studies on other languages.

Longitudinal studies constitute the most direct and complete way of collecting data on acquisition. However, such studies are few, because monitoring a large group of children for a long time is a formidable task. Aksu-Koç and Slobin (2017) review several datasets (including Slobin (1982)) up to the time of their writing and make important observations. Their data and observations have been an essential resource for the arguments in this section.

The better half of later studies focus on L2 acquisition and fall outside the scope of our investigation. Only a small portion of the remaining studies are concerned with morphology acquisition, and much fewer studies explicitly deal with DM. Among these studies, very few are supplemented with longi-

Table 3: Analysis of Burcu's CS over 5 sessions

Age	Words		Inflectional Affixes		Derivational Affixes	
	Tokens	Types	Tokens	Types	Tokens	Types
2;0	63	32	13	5	0	0
2;2	105	48	51	12	0	0
2;5	486	181	354	26	1	1
2;8	770	281	478	27	10	6
3;0	736	325	639	32	19	4

tudinal data collection. In this context, CHILDES is an important resource which contains Turkish datasets.

Other datasets include Ketrez (1999) who defines derivational suffixes as the ones that create a valency change in verbs. She conducts a longitudinal study of four Turkish children between the ages 1;1 and 3;3, but makes no remarks on true derivational affixes. Sofu (2005) studies acquisition of Turkish reduplication. Aksu-Koç et al. (2007) collects data from a single Turkish child between the ages 1;3 and 2;0, but this does not yield many examples or insights on DM. Ketrez and Aksu-Koç (2007) looks specifically into the acquisition of diminutives, based on the same data as Aksu-Koç et al. (2007).

Avcu (2014) relies on a longitudinal corpus of three Turkish infants. The data is collected for a longer duration. We acquired and studied this dataset, in order to test the plausibility of analyzing the CS (child speech) data from younger children with a focus on DM. It includes transcripts of CS and CDS (child-directed speech) with three children:

(6) a. Burcu: 44 sessions between 0;8 and 3;0

b. Can: 26 sessions between 0;8 and 2;10

c. Ekin: 28 sessions between 0;8 and 1;10

We worked with Burcu's child-speech (CS) data due to its longer span. We annotated the inflectional and derivational affixes in CS and analyzed the data from 5 sessions. For all sessions, TAM markers constituted more than half of all inflectional affixes. Due to the rarity of inflected and derived forms, we calculated type and token counts for individual affixes.

In the first two sessions, none of the 168 words in CS were derived forms. Furthermore, in the third session, which involved 486 words, only one of them was a derived form: *almaya* 'taking-DAT'. The number and variety of derivational affixes used by Burcu increases in the following 12 months, but only slightly. None of the derived forms show an uncommon (possibly creative) use. It is likely that Burcu is simply uttering derived forms from memory.

The obvious finding is that derivational affixes in CS are quite few and far between, especially before the age of 2;5. The variety is also quite low. Finally, the derived forms that appear in CS are completely ordinary without any hint of productive use: *başla-* 'begin', *bağla-* 'tie', *fırçala-* 'brush', *üfle-* 'blow' and *kınalı* 'hennaed'. Perhaps better evidence could be collected with older children. The sparseness of the data increases the value of insights from large datasets and even anecdotal evidence.

Aksu-Koç and Slobin (2017) present such crucial insights. They report full mastery of nominal inflection and most verbal inflection before the age of 24 months, with remarkably rare errors. They attribute these to the paradigms' being almost entirely regular and without exception, with almost to homonymy. Furthermore, the mapping of functions onto form is generally non-synthetic. Interestingly, Aksu-Koç and Slobin (2017) witness no errors in the order of application, despite the range and complexity in both nominal and verbal inflectional systems. Thus, position classes are also learned in the same time-frame.

Their discussion on the typical morphological errors by Turkish children is especially relevant to our study, as a large part of these errors concern DM.

(7) Typical DM errors in deverbal derivation presented in Aksu-Koç and Slobin (2017):

a. *kes-il-me elma

b. *ısır-ıl-ma elma

c. *ısır-an elma

d. *yi-yen elma

The correct forms of the first pair of examples are *kesilmiş elma* 'cut apple' and *ısırılmış elma* 'bitten apple'. The mistake here is the use of the nominal *-mA* instead of the participle *-mIş*. The second pair of examples can be corrected by the application of passive voice: *ısırılan elma* 'bitten apple' and *yenilen elma* 'eaten apple'.

(8) Typical DM errors in denominal derivation presented in Aksu-Koç and Slobin (2017):

a. *ye-me-li elma

This example is another child's description of a bitten apple. The expression *ye-me*, is the child's attempt at describing a bite. If this expression were acceptable, the possession affix *-lI* would also be appropriate both categorially and semantically. Therefore, the example not only demonstrates the child's invention of a new word *yeme*, but it also demonstrates the productive use of *-lI*, a derivational affix. While the end result may be judged incorrect by an adult speaker, affix application in *yemeli* is categorially correct. *-mA* derives deverbal noun and *-lI* derives denominal adjectives, correctly producing an adjective to modify *elma* 'apple'.

The fact that children may make mistakes in the selection of appropriate affixes but not in the affix sequence, can also be considered evidence that they are aware of categories. In that case, bare verbs must be represented with one category (verb, V), while TAM markers convert them to another (predicate phrase, PredP) and person markers to yet another (perhaps sentence, S). This task only requires the child to recognize the base as a verb and apply affixes following their correct subcategorization.

Another, perhaps more mentally demanding task is to handle the argument structure. This task requires tracking the valency of the verb (in addition to its category) and computing the new valency if any voice markers are/should be used. The additional steps lead to a higher level of difficulty. Indeed, children have been observed to make mistakes in the selection of voice.

Based on her observations and recordings of Turkish children between the ages 3;0 and 6;0, Ekmekçi (1987) reports many interesting examples of creative application of DM:

(9) Creative application of DM in Ekmekçi (1987)

a. *bakkal-cı

b. *berber-ci

c. *para-cı

d. ?okul-culuk

e. ?para-cılık

f. ?öğretmen-cilik

g. ?ev-cilik

h. *küs-tür-ücü

i. *yan-ıt-ıcı

j. *saldır-ıcı

k. *emek-len-di-niz

l. *gül-dür-mece

m. *yırt-maca

n. *buz-da kay-dır-maca

-CI is a denominal nominal that derives agent names. It can be used with stems such as *banka* 'bank' deriving *bankacı* 'banker', indicating the profession of a person based on their workplace. Therefore, derivations such as **bakkalcı* 'grocery store agent' and **berberci* 'barbershop agent' are hardly unacceptable. Still, adult language rejects these two forms, because their meaning is already covered by *bakkal* 'grocery store owner' and *berber* 'barber'. The fact that the same words are used to indicate both the worker and the workplace must be confusing for children.

We believe the meaning that is learnt first determines when decomposition is possible. For instance, if the child learns the *bakkal* 'grocery shop' meaning first, the **bakkalcı* derivation becomes available when the need arises. An accurate model of DM processing should allow this derivation to take place, but phase it out of the mental lexicon, as new evidence suggests that the correct word is *bakkal*. On the other hand, if the *bakkal* 'grocery shop owner' meaning is learnt first, **bakkalcı* is never derived. We refer to this phenomenon as "the order of exposure" (not to be confused with the order of exposure in the language teaching literature). Blocking ensures that no two forms have exactly the same semantics and distribution in adult language. The numerous pairs of rival words can only continue to exist by sharing either of these dimensions.

-CIlIK also seems to be a productive affix for small children. While it is obviously made up of two affixes *-CI* and *-lIK*, the combined form assumes several functions that is different than the sum of its constituents. One such function is deriving names of role-play games such as *doktorculuk* 'pretending to be a doctor'. Some examples in Ekmekçi (1987) derived by *-CIlIK* may or may not be assessed acceptable, but they definitely indicate the productive use of *-CIlIK*, as these words are unlikely to be part of CDS. Other examples derived by *-CIlIK* include *muayenecilik* 'pretending to go through a physical examination' and *penguencilik* 'pretending to be a penguin' (Yet, 2021).

Other examples can be analyzed in a similar fashion. *küstürücü* 'causer of an offense' is unacceptable in adult language, but is grammatical. *Yanıtıcı* is an attempted substitute for *yakıcı* 'burner'. If *emekli olmak* 'to retire' was not well-established, *emeklenmek* 'to retire' would indeed be a more practical way of expression.

These examples demonstrate two important points. First, children make productive use of morphology. Second, they may not use the correct word, affix or allomorph, but they are consistent in their choices of category. This indicates that even when the lexical knowledge is limited and the command of semantic selection rules is not complete, children have already mastered categories.

Using the same data, Küntay and Slobin (1999) mentions these examples:

(10) Creative application of DM in Küntay and Slobin (1999)

a. *güzel-t

b. *gerçek-leş-miş-im

c. *dil-li-yor-um

d. *öpücük-le-y-ebil-ir mi-yim

In these examples (and the ones we did not include here), we observe mistakes with respect to the argument structure (failing to use passive voice) or failure to select the correct affix. However, we never observe application of an affix of the subcategorization. Aksu-Koç and Slobin (2017) also points out that consistent application of categorial rules reflect an underlying knowledge of affix categories.

Aksu-Koç and Slobin (2017) wonders the reason why children insist on using *-mA* in every subordinate clause. This preference is common in both nominalizations and relative clauses. After all, there is no apparent reason for choosing *-mA* over *-DIK* or *-mIş*. We believe the actual reason might be children's awareness of the distinction between main and subordinate clauses. While these two kinds of clauses have similar underlying structures, cliticization of person markers and the use of GEN in subordinate clauses conceal their similarity. As a result, children might be assuming that TAM markers are reserved for main clauses. Eventually, they accept that a similar set of TAM markers (*-DIK* instead of *-DI*) plus *-(y)An* and *-mA* can be used in subordinate clauses. The similarity between main and subordinate clauses is not transparent in Modern Turkish (MT), but it was quite obvious in Orkhon Turkic (OT). We come back to this point in Section 3.1. Perhaps Göktürk children in the 8th century did not make the mistake studied by Aksu-Koç and Slobin (2017).

Anglin et al. (1993) works with 1st, 3rd and 5th grade English elementary school children. He finds that comprehension of derived forms in grade 5 far surpasses that in grade 1. Similarly, multimorphemic words are much better understood by grade 5, compared to grade 1. These are considered as evidence towards the idea that increasing morphological complexity define the character of lexical development.

Anglin et al. (1993) makes a crucial remark which constitutes the core of this thesis: Evidence of morphological analysis increases with age. In other words, older children are more likely to be able to analyze complex forms. In order to qualify that observation, he recognizes the importance of the distinction between retrieval and decomposition, and the difficulty of assessing which access path is used by the child. We believe this distinction is key to understanding DM. We dedicate the next section to a discussion of retrieval and decomposition.

Bertram et al. (2000) works with 3rd and 6th grade Finnish elementary school children. Based on a statistical analysis of experimental results, they report that subjects make use of morphology while interpreting word meanings. Three observations are crucial for our thesis:

(11) a. Both grades perform better on derivations with highly-productive (HPr) suffixes, compared to monomorphemic words. This is true for both high-frequency (HFr) and low-frequency (LFr) words. The difference is larger with LFr words.

b. Both grades perform better on HFr monomorphemic words, compared to HFr derivations with lowly-productive (LPr) suffixes. The opposite is true between LFr monomorphemic words and LFr derivations with LPr affixes.

c. Performance across all categories improve significantly from the 3rd grade to the 6th grade.

These observations strongly suggest that internal structure aids comprehension. To arrive at a correct interpretation, the subject must either know the meaning of the whole word, or the meanings of all constituents. When the word is HFr, the probability of its being known is higher. This is why the performance on HFr monomorphemic words are high. When the whole word is a derived one, and the constituent suffix is also HPr; the subject may not know the whole form, but may be able to derive its meaning from the constituents. Therefore, better performance can be achieved with derived forms compared to monomorphemic words.

The opposite effect can be observed between HFr monomorphemic words and HFr derived forms with LPr suffixes. The LPr suffix is often unknown, hindering the subject's ability to find the correct interpretation. If knowledge of DM was not in effect, we would expect HFr words to produce similar performance across all categories. The fact that HPr suffixes improve and LPr suffixes deteriorate performance demonstrate that DM is an integral part of processing.

Further evidence can be found in LFr words. This time, monomorphemic words produce the lowest performance. With monomorphemic words, the subject either knows or does not know the word; there is no contingency. However, derived forms allow an alternative path with decomposition. Indeed, both HPr and LPr suffixes aid comprehension to such great extent that subjects perform much better with LFr derived forms, compared to LFr monomorphemic words.

According to Bertram et al. (2000), these results add credibility to the claim that different paths of processing operate in parallel, as in Caramazza et al. (1988) and Frauenfelder and Schreuder (1992). They point out that for HFr words, retrieval is more dominant; while for LFr words, the hearer leans more on decomposition. This is a crucial finding for our understanding of DM and morphological processing. It is one of the central objectives of this thesis to describe and model the mechanism behind this asymmetry.

Creative application of DM is not limited to children. Adults also invent derived forms. For instance, there are interesting cases where speakers expand the use of a foreign affix that is previously used in a few borrowed words. Rarely, speakers also adopt completely foreign affixes. The following examples are taken from Nişanyan (2021):

(12) Adult invention of new derived forms using foreign affixes

a. *olabilite* 'possibility'

b. *atmasyon* 'made up story'

c. *sallamasyon* 'made up story'

d. *sınavzede* ‘exam victim’

e. *depremzede* ‘earthquake victim’

f. *bankerzede* ‘financier victim’

g. *afetzede* ‘disaster victim’

h. *kazazede* ‘accident victim’

The first three examples are somewhat-inventive combinations of a Turkish base with a foreign affix. The *-ability* affix from English according to Nişanyan (2021) began around 1990s to be applied on Turkish stems. Its adoption is probably facilitated by the similarity between the *-abil-ity* suffix set in English and *-abil-ir-lik* suffix set in Turkish, with strangely similar pronunciation and meaning.

The second and third examples demonstrate the application of the NVD class affix *-(t)ion*, on a deverbal nominal. Nişanyan (2021) argues that since there is no rule preventing the adoption of foreign affixes, these words should be considered grammatical. It is true that foreign affixes routinely make their way into a language and it is also true that affixes of Turkish origin are also occasionally used on non-standard base categories.

The rest of the examples are more familiar for the general public. While *-zede* is a Persian affix, it can be applied on both Turkish and borrowed stems without problems.

These examples show that the knowledge of DM is real and active in the minds of speakers. New uses of affixes and derived forms are frequently invented. Second, speakers are not restricted by the etymological origins of affixes or stems. (They could be restricted culturally, but not linguistically.) All that matters is their semantic content and an appropriate subcategorization.

So far, we have seen cases of spontaneous derivation. Derivational affixes are also used deliberately to create new words. Many words coined by TDK make it to common use, while others don’t. There are also ones that are misanalyzed by the public and used in a different way than intended. The following examples are taken from Nişanyan (2021) and Ergin (2009).

(13) Similarity affixes

a. *özümle-* > *özümse-* ‘absorb’

b. *gülümsin-* > *gülümse-* ‘smile’

c. *acısı* > *acımsı* ‘bitterish’

d. *ekşitırak* > *ekşimtrak* ‘sourish’

e. *azsa-* > *azımsa-* ‘underestimate’

In these cases, either a more productive affix is imposed in the place of another phonologically similar affix (as in the first two examples) or a phoneme from a base frequently used with the affix gets inserted in other derivations as if part of the affix (as in the last three examples). Whatever the underlying processes are, morphological operations are so blended with phonological processes, it is often impossible to identify purely morphological rules.

2.2.2 Retrieval and Decomposition

The previous section laid out the psycholinguistic evidence on the acquisition of DM. It is complemented with examples from inventions of new derived forms. Now, we turn to the psycholinguistic evidence regarding the processing of DM.

The last chapter of Anglin et al. (1993) presents an insightful discussion of the distinction between learning and constructing word meanings. Due to the emphasis on comprehension in this thesis, we refer to this distinction as retrieval and decomposition. Anglin et al. (1993) discusses how the definition of "word" determines our estimations of a person's vocabulary development, and argues that a consistent definition may not even be possible. (Haspelmath (2011) makes a thorough investigation of whether such a definition is possible. His conclusion is in the negative.)

According to Anglin et al. (1993), even if "word" can be defined in a consistent manner, there are further issues. Homography and polysemy play an important role in the development of vocabulary. He cites previous research that failed to distinguish different semantics due to homography. Anglin et al. (1993) believes that homographs, just as they are listed as separate entries in the dictionary, should be taken into account in a vocabulary development study; but it would be difficult to achieve. Concerning polysemy, he points out another layer of uncertainty. Similar to how one cannot directly determine whether the listener simply retrieves or decomposes the word, one cannot directly determine whether the listener knows the exact meaning of the word, or guesses its meaning based on known polysemous uses. We consider homography (or homophony in spoken language) and polysemy crucial in an investigation of DM, and we return to these issues in later sections.

According to Seidenberg and Gonnerman (2000), there are three theoretical approaches to studying morphological processing: hybrid models (accounting for the retrieval and decomposition routes for the derived form), interactive activation models (following the decomposition of the concept into its orthographic and phonological constituents) and distributed connectionist models (representing the association between context, semantics, orthography and phonology).

Distributed connectionist models such as Seidenberg and Gonnerman (2000) do not represent morphemes as distinct entities. Rather, morphology emerges in the correlations between the sound-meaning mappings. Perhaps this is an adequate approach to model human processing, given the distributed representation in the human mind. Even if this is true, it is hard to imagine how such a model would contribute to our understanding of DM. Granted, an adequate representation of morphology could be possible without morphemes, but would it be transparent to human examination? In this thesis, we aim to improve the theoretical understanding of DM, thus prefer not to completely neglect symbolic representation. Nevertheless, Gonnerman (1999)'s remarkable discovery that morphological complexity occurs on a graded scale, hints at the need for a model that accommodates blurred boundaries.

Interactive activation models such as Taft and Zhu (1995) do not suffer from this defect. However, they also present several important problems. The most central issue is adequately determining the levels of representation along the chain of decomposition. Concept and morpheme levels are indispensable, but the necessity of word and sub-morpheme levels is open to debate. First, for the word level to be meaningful, we need a rigorous method to define word, which may not be possible according to Anglin et al. (1993) and Haspelmath (2011). Second, representing words and (bound) morphemes in different layers is hard to explain without credible evidence to suggest such an asymmetry being in effect during processing. Third, the sub-morpheme layer cannot be justified by semantics; at least its effect could be negligible compared to semantics. It is certainly possible that phonological coincidence may have an effect in morphological processing, but including such a layer adds (possibly) unnecessary degrees of freedom. When the only layers are the concept and morphemes, the interactive activation model is not that different from hybrid models.

Seidenberg and Gonnerman (2000) list the types of hybrid models as the following:

(14) a. Models where all words are decomposed

b. Models where all words are retrieved

c. Models where semantically transparent forms are decomposed, opaque forms are not

d. Models where suffixed forms are decomposed, prefixed forms are not

e. Models where inflected forms are decomposed, derived forms are not

f. Models investigating the existence of possible decomposition on lexical access

g. Morphological race models where variable processing speeds are assumed for retrieval and decomposition

Each type of hybrid model must be evaluated based on its underlying assumptions. Early models where all words are decomposed ignore the fact that some forms are non-compositional. Even if the hearer attempts to decompose, an alternative path of retrieval must be available. On the other end of the spectrum, the claim that all words are retrieved simply contradict with the fact that speakers are aware of DM, as detailed in the previous section.

Decomposition of transparent forms is to be expected, but it is controversial whether opaque forms are decomposed. The hearer cannot be expected to know if a derived form is opaque or transparent, without first decomposing it. The access route taken by the hearer should not depend on whether the derived form is transparent.

Regarding the fourth type, it is well-known that prefixes are acquired later than suffixes (Clark, 2017), but prefixes are still acquired. Awareness of prefixation means that models should account for decomposition of prefixed forms. Again, based on evidence presented in the previous section, decomposition of derived forms must be available; making the fifth type also inadequate.

The last two types are much more reasonable. Andrews (1986) make two important discoveries. First, morphological effects are not due to obligatory decomposition taking place before retrieval. Second, it is unlikely that words are retrieved based on their stems. This means that constituent morphemes are properly decomposed; each morpheme can be individually retrieved without reference to the others.

The Augmented Addressed Morphology (AAM) model of Caramazza et al. (1988), the Morphological Race Model (MRM) of Frauenfelder and Schreuder (1992) and the meta-model of Schreuder and

Baayen (1995) argue that both access routes are open for complex words. While the first two can be considered hybrid models, where alternative access routes operate in parallel; Schreuder and Baayen (1995) is an interactive activation model.

Caramazza et al. (1988) posit that in addition to the retrieval and decomposition processes, orthographically similar forms are processed in parallel. This is essentially a race model, but the retrieval route is always faster. When it is possible to both retrieve and decompose a complex word, the retrieval route will always win. Therefore, the decomposition route is just a contingency for cases where retrieval will be attempted without success. We do not believe this assumption is well-justified. Frauenfelder and Schreuder (1992) point out several other issues with their approach.

Principles such as economy of storage and economy of processing cited in many studies on psycholinguistics (and reviewed by Frauenfelder and Schreuder (1992)) are not justified by data, nor the principles of elegance and parsimony cited in their support are grounded in real-life observations. We see little reason to mold the theory in accordance with such principles. Frauenfelder and Schreuder (1992) also recognize this, and that race models are in conflict with both principles. Nevertheless, they argue that race models can be justified in terms of efficiency.

Frauenfelder and Schreuder (1992) predict that the decomposition route will win the race most probably for low-frequency transparent complex words. This prediction leads to another prediction that inflected forms (IM being more productive and transparent) must be easier to access by decomposition, compared to derived forms. These predictions are in agreement with Schreuder and Baayen (1995)'s list of empirical facts:

(15) a. Cumulative root frequency effects

b. Stronger effects for IM, compared to DM

c. Stronger effects for semantically transparent, compared to opaque complex words

d. Productivity (Processing of unfamiliar complex words)

e. Affixal homophony

f. Pseudo-prefixation

g. Structural differences between languages

The first three facts are similar to the ones stated in Frauenfelder and Schreuder (1992). The fourth one, productivity, does not require further justification. Affixal homophony and pseudo-prefixation are not trivial facts. Affixal homophony (combined with polysemy in our case) creates a layer of ambiguity that must be explicitly represented in the model. Pseudo-prefixation, on the other hand, must be accounted for by a segmentation stage and later eliminated by subcategorization rules (licensing in Anglin et al. (1993)). Both these stages are discussed in depth in the next section.

Perhaps an additional fact is "blocking", as described by Frauenfelder and Schreuder (1992). Competing forms that have been learnt earlier, preserve and increase their advantage over alternatives.

An important issue concerns all past and future models of this kind. It is extremely difficult to validate the predictions of such models with direct evidence. By direct evidence, we mean data on acquisition, real-life conversation and similar linguistic corpora. Such data is sparse and occasional experiments have small sample sizes. Also, there are too many moving parts for a comprehensive controlled experiment to be conducted.

Based on the above discussion, we believe the two access routes, retrieval and decomposition, must be available simultaneously. We do not comment on whether they are interacting or not.

2.3 Beyond Morphological Structure

When investigating DM, the literature mainly focuses on morphological structure. While this kind of investigation may be sufficient for IM, which is much more regular, it is definitely inadequate for an investigation of DM. There are two reasons why a study of DM must be complemented with a new analysis.

First, DM exhibits allomorphy and polysemy to a much greater extent than IM. While inflectional affixes usually have a one-to-one correspondence with morphosyntactic properties (at least in Turkish), derivational affixes have a many-to-many correspondence with semantic content. The structure of DM incorporates the larger flexibility in lexical choice and semantics.

This is not to say that IM is completely devoid of ambiguity. To give a few examples from Turkish, possessive markers (thoroughly analyzed in Öztürk and Taylan (2016)), the overlap between person and possessive markers and the overlap between TAM and copula complicate models considerably. There are also cases where inflectional affixes appear in a derivational capacity: *alın-dı* 'receipt', *yemek* 'food', *alacak* 'receivable' etc.

Nevertheless, a difference arises between IM and DM, because in IM such exceptional cases are both fewer and easier to disambiguate. While similar in form, person markers and possessive markers have an unmistakable difference: They append on stems of different POS. The case with *-DI*, *-mIş* and *-sA* is similar in their double function as TAM and copula. While both TAM and copula append on verbs, they occupy different position classes. It is impossible to mix the two, because copula always comes after TAM. On the other hand, position classes cannot be used as a clue for disambiguating derivational affixes, because DM order of application is unrestricted.

If we turn to cases where inflectional affixes assume a derivational role, we observe that such uses are not really productive. These forms are syntactic constructions that happen to be lexicalized with a non-compositional meaning. Despite its full productiveness as a TAM marker, there are only 22 forms derived by *-AcAK* in the dictionary. The situation is similar with *-mIş*, *-DI* and the infinitive *-mAK*.

Participles introduce another source of ambiguity. Turkish participles except *-(y)An*, share their form with TAM markers: *-(y)AcAK*, *-mIş*, *-(A)r* / *-(I)r*. *-DIK* is not identical in form to its TAM counterpart *-DI*, but our analyses in Chapter 3.2.3 demonstrate that the structure of subordinate clauses are almost the same as main clauses, except the GEN marked topic. As a result, it is possible to interpret the verbal predicates in main clauses as participles, making TAM markers participles. *-DIK* is used inside subordinate clauses, while *-DI* derives main clause participles. These results are supported by Kuznetsov (1997)'s claim that Turkish predicates are always based on nominals.

DM exhibits ambiguity in both syntactic and semantic aspects. As a result, it is crucial for an analysis of DM to take into account the alternative ways an affix can be interpreted (during comprehension) or the alternative affixes that can be selected (during production). More specifically, the proposed structure of DM must be able to develop preferences towards certain alternatives in certain contexts.

Another difference relates to the knowledge of DM being dependent on linguistic exposure. Since inflectional affixes often realize morphosyntactic features, they are much more frequently observed and have a clearer function. There is evidence that inflectional paradigms are established much earlier in children than derivational processes Aksu-Koç and Slobin (2017). This ensures that knowledge of IM is distributed more completely among speakers.

This is not true for DM. A speaker's repertoire of derivational processes is determined by the derived forms they are exposed to. Rarely used derivational affixes may remain unrecognized by certain individuals, even in adulthood. Therefore, the list of recognized derivational affixes is different from individual to individual. Indeed, Avcu (2014) finds that the frequency of affixes in CS is proportional to the frequency of affixes in CDS. At a larger scale, a difference in productivity may cause an affix to be forgotten in one community, while it continues to be used in another. The same is true in the time dimension. While the use of IM is relatively stable due to their full productivity, the fate of a derivational affix is completely tied to the preference of the community.

In a nutshell, IM (and syntax) consists of a set of rules to be learned; in contrast, DM is a set of conventions that is adopted. While morphological structure suffices to explain the relatively rigid and regular inflectional processes, a new layer of analysis must be added to explain derivational processes.

Due to its dependence on exposure, a study of DM is in large part a study of the lexicon. Ideally, we need to be able to explain how a derivational affix enters the lexicon, how it is represented as a lexical item, and how it is employed during the processing of a derived form. Since this whole ordeal begins with exposure, we start with the analysis of linguistic observations.

This analysis takes place in three layers: segmentation, lexical selection (borrowing the term from language production literature) and derivation. Each of these layers are affected by different sources of ambiguity: segmentation ambiguity, lexical / semantic ambiguity and structural ambiguity. Now, we look into these layers in more detail.

2.3.1 Primary Layer: Segmentation

The obvious first step in interpreting a linguistic observation is analyzing its form. Throughout the thesis, we focus on segmental morphology; therefore, we do not go into an analysis of non-segmental morphology. We aim to explore morphology processing specifically in agglutinating languages.

The possibilities and limitations regarding segmentation have been considered in many studies of morphology. Hammarström and Borin (2011) and Ruokolainen et al. (2016) give extensive reviews of studies on unsupervised and minimally supervised learning of morphology. Most of these studies learn segmentation based on a large dataset. Gaussier (1999), Goldsmith (2001) and Creutz and Lagus (2007) are other notable studies that we consulted for insights into the segmentation problem. Three more recent studies Üstün and Can (2016), Can and Manandhar (2018) and Can et al. (2022)

suggest sophisticated word-embeddings-based, dirichlet-process-based and LSTM-based architectures for learning segmentation.

These studies focus on computationally analyzing a large dataset and identifying morphemes with the highest accuracy. They are quite successful in attaining their objectives and new models with higher accuracy are introduced every year. However, they do not claim psychological plausibility. The human mechanism for segmentation must be incremental and based on a limited dataset, especially for children. Perhaps more importantly, the acquisition of morphology is not unsupervised. CDS and the context constitutes enough supervision for learning. The child is not expected to learn from form-based correspondences between the contents of a large lexicon; he is expected to learn from semantically-motivated correspondences between a limited set of lexical items.

There are also studies that focus on the psycholinguistic aspects of the problem. Cutler and Norris (1988) points out the phonological clues for segmentation. Schreuder and Baayen (1995) proposes an integrated model that incorporates segmentation and lexical selection.

We do not focus too much on computational efficiency, nor on the subtly phonological clues that aid segmentation. Our aim is simply to gain an understanding of the place of segmentation in the larger problem of morphology processing. We leave the concerns of speed and accuracy to other authors. We take inspiration from proposals in the literature and pursue a supervised, incremental and psychologically plausible mechanism.

We start by evaluating two extreme possibilities regarding compositionality: Either all derived forms are to some extent non-compositional, or they are all completely compositional (with some homonymy, polysemy and synonymy). If the former were the case, we would not expect people to benefit from the knowledge of DM. In fact, a knowledge of DM would not be possible, at all. This is not the case. In Section 2.2.1, we reviewed the large collection of evidence supporting the idea that speakers, including children, are aware of DM and actively use DM knowledge to make sense of their observations. Another related observation is that speakers, again including children, are aware of categories of not only stems, but also morphemes.

On the other end of the spectrum, DM's being completely compositional is not a tenable position, either. It is easy to find derived forms that assume different meanings than the ones recoverable by the decomposition of their constituents. *dolmuş* 'shared taxi' (literally 'full') is a frozen form that is widely used by Turkish speakers. Its being frozen does not completely prevent one from attempting from analyzing it, though. *dolmuş* 'get full-NARR' still exhibits transparent morphology and its current meaning is still clearly tied to its original semantics. Therefore, while its meaning is clearly non-compositional, it is still a derived form. There are thousands of similar examples, some of which are presented in Section 3.1.

Another piece of evidence comes from distributional semantics. If derived forms were mostly compositional, we would expect there to be some way to consistently estimate affix embeddings from base and lemma embeddings. Kunter et al. (2020) estimate affix embeddings from thousands of base-lemma pairs and demonstrate that these estimations cluster according to their respective semantics. Section 4.1 expands on these results.

These observations do not provide enough evidence to suggest that DM is simply word-internal syntax. Nevertheless, they justify modeling a robust decomposition path for interpreting derived forms.

In Section 2.2.2, we reviewed the theoretical landscape concerning morphological processing. Focusing on the dichotomy between retrieval and decomposition, we observed that more recent studies are more likely to posit a dual mechanism to process morphologically complex forms. Alternative paths of retrieval and decomposition are thought to work sequentially or simultaneously. Our understanding of morphology is in line with this point of view.

For decomposition to take place, the hearer must be able to identify the constituents in an expression. We call this process segmentation. Multiple alternative segmentations might be possible for an expression, based on the contents of the hearer's lexicon. It is impossible to know which interpretation is correct (if the correct interpretation is attested at all) before the interpretation is fully derived and checked against the context. Therefore, the hearer does not have reason to disregard any segmentation alternative at the beginning of processing.

Caramazza et al. (1988) present evidence that retrieval and decomposition are in competition to generate the correct interpretation. This competition may only take place after the contents of the lexicon make both paths available. If an expression is observed for the first time, it is impossible to use the retrieval path. If the constituents of an expression cannot be recognized, it is impossible to use the decomposition path. Alternative paths are simply the result of different decisions regarding segmentation.

The simplest way to carry out segmentation is recursively dividing the expression. For instance, *kitaplık* 'bookshelf' can be segmented in the following way, if the lexicon only contains *kitaplık* 'bookshelf', *kitap* 'book' and *-lIK* 'container'.

(16) a. Kitaplık: Attested. All segments present in the lexicon.

b. Kitaplı-k: Unattested.

c. Kitapl-ık: Unattested.

d. Kitap-lık: Attested. All segments present in the lexicon.

e. Kita-plık: Unattested

f. ...

g. Kitapl-ı-k: Unattested.

h. ...

i. K-i-t-a-p-l-ı-k: Unattested

The valid alternatives are *kitaplık* and *kitap-lık*. If we worked with the full lexicon (including morphemes such as *-A*, *-I*, *-l* and *-k*) and did not apply categorial selection, the number of alternatives would be close to 2^7. When categorial selection rules are applied, the number of alternatives does not grow exponentially.

If done in an unrestricted fashion, solely based on forms, the segmentation problem would be intractable. Given many one-character morphemes in the Turkish lexicon, it would be possible to generate millions of alternative segmentations for an especially complex derived form. However, the problem can be kept tractable, because most alternatives are eliminated by the hearer's awareness of morpheme categories (categorial selection) and application rules (semantic selection). In other words, categorial selection keeps the segmentation tree in check. Most affixes only apply on a certain category, greatly reducing the number of possible segmentations. When semantic selection is added to the restrictions, the segmentation problem can be kept tractable even for large sentences.

Given that children are aware of DM, the segmentation problem must be tractable from their perspective, too. Children can be thought to operate with a smaller lexicon, generating fewer alternatives at each step of segmentation. The segmentation tree grows, as their lexicon grows. At the same time, the size of the segmentation tree is restricted by newly learned categorial and semantic selection rules.

A larger example is *gözlükçülük* 'profession of an optician'. For this example, the following lexicon is assumed (semantic selection rules are given in parantheses):

(17) a. *gözlükçülük* 'profession of an optician'

b. *gözlükçü* 'optician'

c. *gözlük* 'glasses'

d. *göz* 'eye'

e. *-lIK*: affix deriving names of apparels (only applies on names of body parts)

f. *-lIK*: affix deriving names of professions (only applies on names of professions)

g. *-CI*: affix deriving names of salesmen (only applies on names of sellable items)

h. *-CIlIK*: affix deriving a game where children pretend to practice a profession (only applies on names of professions)

Figure 1 presents the segmentation tree for *gözlükçülük*. Attested alternatives are colored black. Alternatives that are not attested due to a lack of form-matching morphemes are colored white. One alternative, where each segment has a matching morpheme, is still not attested due to semantic selection. That alternative, *gözlük-çülük*, is colored gray. Overall, only four segmentation alternatives are attested.

Most Turkish DM are suffixes. Therefore, it is sufficient to carry out the recursive operation on the left constituent. This is an algorithmic choice and can be modified to account for prefixation with little effort. Segmentation from one side is lighter on computational complexity and is probably plausible psychologically. Perhaps such an asymmetry could explain why prefixes are learned later than suffixes, as observed in Clark (2017). Also, segmentation on the left emphasizes the learning of affixes, instead of the learning of bases.

Section 5.2.2 presents the application of a segmentation algorithm based on this approach. Once segmentation alternatives are obtained, the hearer moves on to the second step: lexical selection.

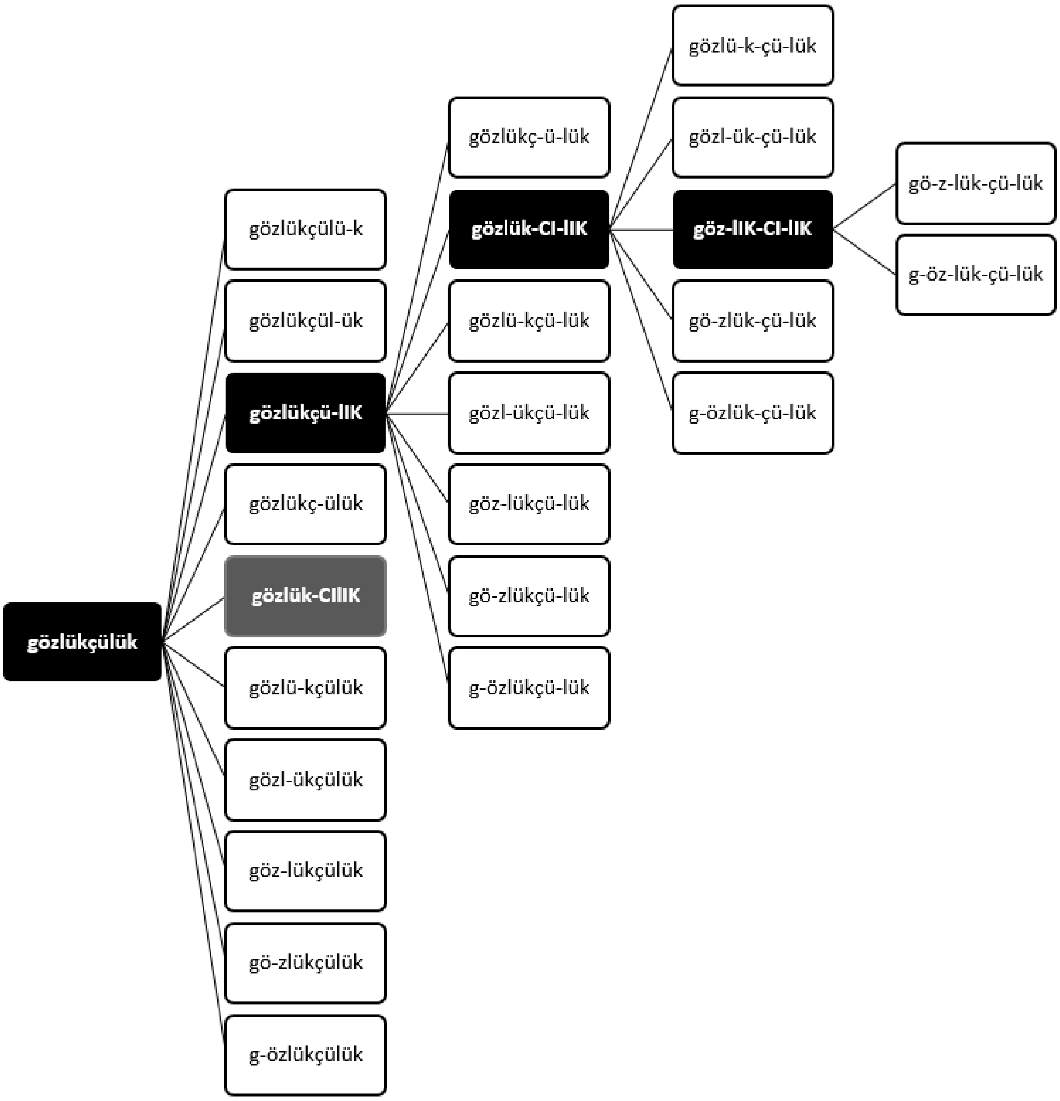
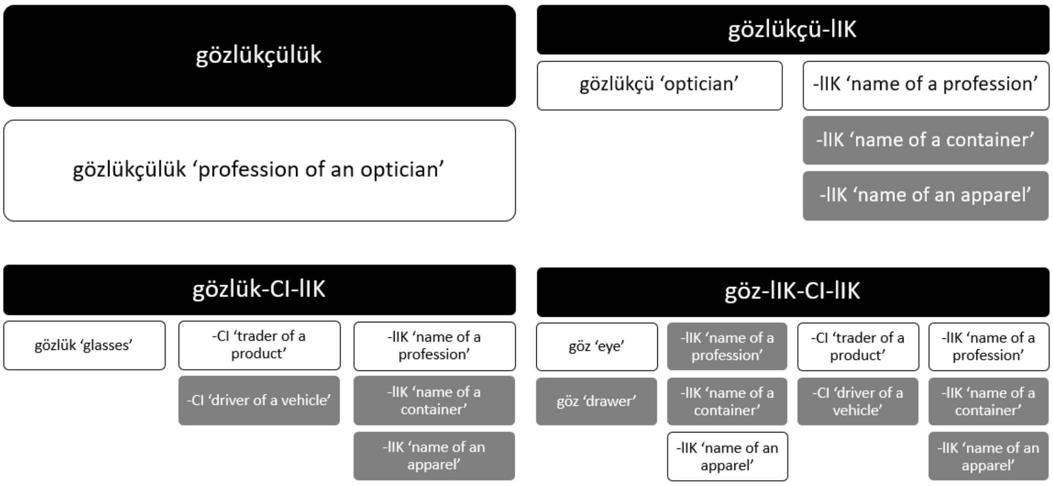


Figure 1: Segmentation alternatives for *gözlükçülük*



gözlükçülük
gözlükçülük 'profession of an optician'
gözlükçü-lIK
gözlükçü 'optician'
-lIK 'name of a profession'
-lIK 'name of a container'
-lIK 'name of an apparel'
gözlük-CI-lIK
gözlük 'glasses'
-CI 'trader of a product'
-lIK 'name of a profession'
-CI 'driver of a vehicle'
-lIK 'name of a container'
-lIK 'name of an apparel'
göz-lIK-CI-lIK
göz 'eye'
-lIK 'name of a profession'
-CI 'trader of a product'
-lIK 'name of a profession'
göz 'drawer'
-lIK 'name of a container'
-CI 'driver of a vehicle'
-lIK 'name of a container'
-lIK 'name of an apparel'
-lIK 'name of an apparel'

Figure 2: Lexical selection for *gözlükçülük*

2.3.2 Secondary Layer: Lexical Selection

The segmentation stage only divides the form of the observation into segments, but it does not assign meaning to individual segments. Meaning is assigned in the second stage. Since a segment's form may match multiple lexical items, a second layer of ambiguity emerges.

Lexical selection (or lexical choice) is a term borrowed from the language production literature. It refers to the selection of a lexical item that is appropriate for the context and the intended meaning. In production, lexical selection is the selection of a form, given a meaning. The term is somewhat appropriate for the current setting, because the hearer still needs to choose a lexical item. However, this time, lexical selection is the selection of a meaning given a form. Licensing is a related term by Baayen and Schreuder (2000). It refers to the "checking of subcategorization compatibilities".

Lexical selection in this sense is complicated by homonymy and polysemy. We study these lexical relations in Turkish DM in Chapter 3. For the moment, let it suffice that the nature of DM tends to produce plenty of polysemy. Most derivational affixes in Turkish (65%) have multiple meanings. An average derivational affix serves in 3 different functions. Compared to derivational affixes, the portion of Turkish inflectional affixes that have multiple meanings is much smaller (5%) .

To continue the example from 2.3.1, Figure 2 demonstrates the lexical alternatives. We assume the lexicon includes 1 lexical item matching *gözlükçülük*, *gözlükçü* and *gözlük* each; 2 items matching *-CI*; and 3 items matching *-lIK*. As a result, the segmentation *gözlükçülük* produces only 1 alternative interpretation, while *gözlükçü-lük* produces 3, *gözlük-çü-lük* produces 6; and *göz-lük-çü-lük* produces 36.

A model of DM must be able to generate and derive such a wide variety of possible interpretations due to lexical selection. Ideally, the model should also be able to evaluate the plausibility of each interpretation, as well as of each alternative lexical item.

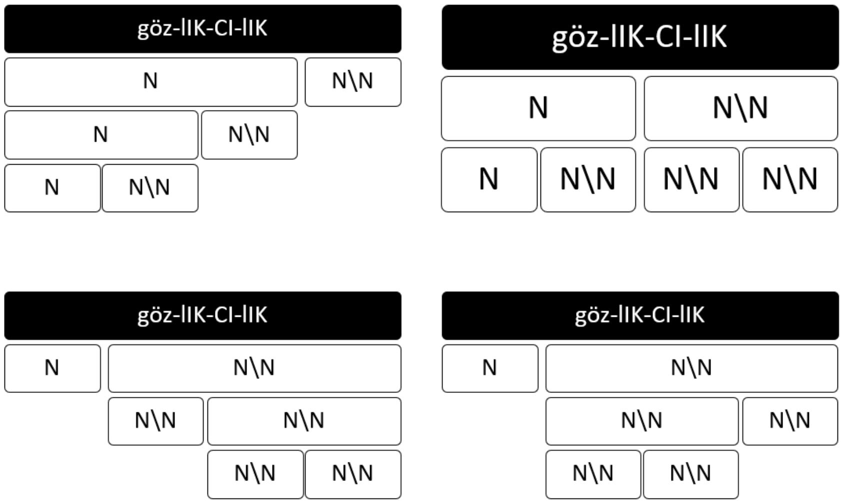


Figure 3: Derivation paths for *gözlükçülük*

2.3.3 Tertiary Layer: Derivation

After appropriate lexical items are identified for each segment, the hearer needs to bring them together to derive the alternative interpretations of the expression. Even if each segment can only be matched with one lexical item, there may be multiple alternative paths for derivation. This is called structural ambiguity.

In the example of *gözlükçülük*, if we focus on the segmentation alternative *göz-lük-çü-lük*, 4 derivation paths are possible. All lexical items matching *göz* are of category N, while *-CI* and *-lIK* are of category N N (subcategorizing for N, producing N). Using the application modes in Steedman and Baldridge (2011); Figure 3 summarizes the possibilities.

For simplicity, we work with only forward and backward application between adjacent constituents (for details, see Section 4.4), and assume these application modes are available to all speakers. Observations of multiple authors reviewed in Section 2.2.1 must suffice to convince the reader that categories are also available to speakers, at least by the time they start making use of DM.

CCG is an appropriate tool for this analysis, because it represents linguistic processes with semantics and syntax in lockstep. Section 4.4 provides a brief introduction to CCG and demonstrates its capabilities toward present purposes.

2.3.4 Conventionalized Structure

Due to the sources of ambiguity described in previous sections, derived forms often evoke multiple interpretations. Yet, when a hearer observes a derived form, they often effortlessly choose one interpretation over others. As reviewed in Section 2.2.2, previous work on morphological processing point out some possible causal mechanisms for the observed asymmetries, but cannot quite propose a computational model with adequate explanatory power.

The facts regarding morphological processing are clear. (In this thesis, we only approach the problem from the comprehension side, but the arguments can be generalized to cover production.) First, hearers are aware of morphological knowledge. Second, if the lexicon permits, it is possible to interpret morphologically complex words following two paths: retrieval and decomposition. Third, given that decomposition is possible, three stages must take place for the whole to be interpreted: segmentation, lexical selection and derivation. Each of these stages create a layer of ambiguity. Additional facts such as frequency effects are also presented in the literature, depending on the data and focus adopted by authors.

In order to explain these facts, linguists often focus on morphological structure, based on form, category and semantics. Their methods adopt the latest knowledge on linguistic structure, but often lack the real-life data that would justify or contradict their conclusions. Psycholinguists collect data from relatively small samples of children and adults. They analyze these data often with a focus on frequency and reaction time. While frequency obviously plays a role; when it is not complemented with theoretical knowledge, it cannot adequately explain the data. Computational linguists invent quite complicated algorithms to achieve the highest possible rate of correct interpretation. Their methods are data-driven, but do not try to discover an underlying structure that explains why an algorithm succeeds, where another fails.

Ideally, one would work with a single integrated structure, which encompasses all the stages of processing and which accommodates all the parameters observed in the data. Such a structure would give rise to a model that seamlessly processes observations, maintains a lexicon and develops preferences regarding alternative interpretations. In this thesis, we propose such a structure and develop such a model. We call the former the Conventionalized Structure (CdS), giving rise to a Bayesian Belief model laid out in Chapter 5.

CdS is composed of three layers: segmentation, lexical selection and derivation. These layers are configured in such a way that the output of each layer constitutes the input for the next. Ambiguity is not resolved by the selection of a single alternative, rather, all alternatives are processed in parallel. After all alternatives are processed, their degrees of plausibility are ranked.

Whatever the contents of the lexicon, the hearer cannot be sure of having obtained the correct interpretation, solely on the basis of the output of CdS. The correctness of an interpretation depends on the real-life context. If the internal structure produces a single interpretation, it may or may not be contextually appropriate. If the internal structure produces dozens of interpretations, multiple alternatives may be correct, or all may be incorrect.

Given that preferences emerge over time for certain meanings of certain forms, CdS must accommodate the development of such asymmetries. For instance, *göz* ‘eye’ and *göz* ‘drawer’ are homonyms, but the ‘eye’ meaning is the one that comes to mind when one encounters the form *göz*. This is because in the vast majority of cases, *göz* is used in the ‘eye’ meaning. The ‘drawer’ meaning is not forgotten, though. It is still invoked every time *göz* is encountered, but is much less salient than the ‘eye’ meaning.

This asymmetry can be modeled with the help of a probability distribution for each unique form. Lexical items with the same form may be ranked according to their probability values. We can achieve this by distributing a limited total probability (100%) among alternatives. This way CdS sets off a competition between alternative lexical items. Feedback from the real-life context serves to highlight the context-appropriate alternatives, while suppressing others.

The competition between lexical items is not the only one. Alternative segmentations of the same observation are also in competition. Similar to lexical alternatives, alternative segmentations of an observation are represented on a probability distribution. Again, this probability distribution evolves with feedback from the real-life context.

All contents of the network are incrementally learned from observations. No other input is needed. Observations are assumed to have three components: form, context and category. The first component is uncontroversial. After all, the word "observation" implies that something is observed; that something is the form. Whether verbal or written, a form is a necessary component of any observation.

The second component is not so straightforward. Any linguistic observation takes place in the presence of a countless number of related and unrelated pieces of information in the environment. Context is the set of pieces of information that aids in the interpretation of the observation. Unfortunately, context is complex and chaotic, impossible to control and experiment with. Therefore, a comprehensive model of context is unattainable.

Instead of modeling context, we approximate it. Given that children learn their first words without any command of the language, we assume that in many cases, the context is strong enough to permit a direct match between form and meaning. The contribution of context occurs on a scale: In many cases, the context provides clues so strong that the form is made unnecessary. In other cases, the context is so deficient that the hearer struggles to choose the correct interpretation from the alternatives generated by CdS.

Considering a child during DM acquisition, we can simply discard the observations of the second sort, since they do not contribute to the evolution of the structure. Instead, we work with a set of observations with strong context. At minimum, we need the context to be strong enough for the hearer to be able to discern the correct interpretation. This way, the hearer can develop preference between alternatives. At other times, we need the context to directly let on the meaning. This way, the hearer can learn new phrases, words and morphemes, even if both retrieval and decomposition paths are unavailable. Following the semantic bootstrapping hypothesis of Abend et al. (2017), we assume the existence of such strong contexts during acquisition. We represent the context of an observation by the logical form of the observation itself.

Based on this discussion on context, the third component of an observation is easier to justify. We reviewed the evidence suggesting hearers' awareness of categories in Section 2.2.1. Even very weak contexts should suffice for the hearer to understand whether the observation is a noun or a sentence. As we will see in Section 4.4, only distinguishing between these two categories is enough to infer other categories and semantics.

CdS is a multi-level structure, because it is composed of three distinct, hierarchically organized layers. It is dynamic, because salience of segmentation and lexical alternatives are constantly changing based on new observations.

Adequacy of CdS for representing morphology in various languages can be evaluated on two aspects: whether segmentation or an equivalent operation can be carried out and whether the competition between segmentation alternatives provides valuable insights. Regarding the former aspect, templatic morphology presents a challenging case. Templatic morphology resists analysis at the segmentation stage, because it is not segmental. Nevertheless, templates can be recognized by an algorithm and encoded as separate linguistic objects. One can apply CdS on templatic morphology, by replacing the segmentation stage by a template recognition stage. We do not attempt this.

Regarding the latter aspect, we admit that segmentation alternatives may not always point to interesting trade-offs. For instance in analytic languages, isolating languages especially, modeling the competition between segmentation alternatives below the word level makes less sense. However, even in those languages, CdS might be useful for modeling multi-word expressions (MWE) and phrases.

CdS is much more readily applicable on synthetic languages. These languages display a rich word-internal structure that can be exploited via the consecutive layers of analysis proposed in this section. The approach is especially suitable to agglutinating languages, as the exploration in Chapter 5 demonstrates for Turkish. Fusional languages can be studied in the same way, as form-meaning relationships being one-to-many does not bring any additional challenge for CdS. Working on languages with rich segmental morphology, the main issue would be handling phonological variation.

Throughout this thesis, a large majority of the examples we study are suffixes and prefixes. In order to represent infixes and circumfixes, we would need a more sophisticated segmentation algorithm. Ultimately, all types and mechanisms of morphology can be formalized in terms of IA, IP or WP approaches and the resulting formalization can be represented on CdS. Still in its simplest form, the structure we propose is most readily suited to agglutinating languages.

2.4 Methodology

So far, we cited studies from several different areas of research: computational linguistics, linguistics, psycholinguistics and neuroscience. Each area of research adopts a different focus and a different framework for approaching the same problem. In turn, each framework has its strengths and weaknesses. Therefore, the variety of points of view presents an abundance of ideas with which the individual topics of research can be attacked. The collection of these ideas is a great asset for the research community.

Nevertheless, after decades of research, there are still great gaps in our knowledge of certain subjects. Existing methods somehow fail or aid progress too slowly in some directions. This is especially true in subjects where one has to work with prototypes rather than definitions, with little data, or in subjects where controlled experiments are impossible. Without clear definitions, we run the risk of not being able to communicate our ideas. Without large, diversified and standardized data, we run the risk of speculating over anecdotes. Without controlled experiments, we run the risk of never really establishing cause-effect relations.

Linguistics lives under the burden of all three deficiencies. Haspelmath (2011) explains why a definition of word, consistent across languages, may never be possible. Section 2.1 reviews the great morphologists who still debate on the existence and limits of DM. Section 2.2 reviews existing data on the acquisition of DM and their insufficiency. It has been established decades ago that behaviorist controlled experiments on language learning and processing would not be as fruitful as once imagined. Perhaps the best researchers can do is to establish statistical correlations between performance and linguistic observations, provided these are based on meaningful independent variables. Bertram et al. (2000) is a good example of such a study. However, a computational model of adequate explanatory power is still elusive.

Perfectly aware of these difficulties, we focus on a neglected research problem. DM is an underexplored area of research within linguistics. By proposing a novel structure we deprive ourselves from established research paradigms. In order to fully develop CdS, we need to find an appropriate modeling framework and test our proposal against real-life scenarios.

In this section, we present a short review of existing methods in related areas of research. There is an obvious dichotomy between two major research paradigms. On one side, we have theoretical methods, which explore the structure behind linguistic phenomena. On the opposite side, we have data-driven methods, which exploit large datasets to create AI models. Psycholinguists' studies based on survey and reaction time data may constitute a third paradigm, powered by statistical correlations. We do not go into the details of studies previously discussed; this time, we only point out their limitations and their appropriateness for the current study.

Purely theoretical approaches use symbolic representations of linguistic objects and phenomena. Such approaches are generally put into use in syntax and phonology, because mechanical rules have been proposed for many processes within these domains. Validity of a theoretical approach depends on the possibility for a process to be described in mechanical rules, and in turn, theoretical approaches look for mechanical rules to describe phenomena.

With data-driven approaches, such as distributional semantics, the emphasis is on semantics rather than syntax. They do not represent semantics by logical forms and equivalent systems of symbolic representation, but they extract it from a corpus in the form of vectors. On these vectors, all properties of a lexical item are represented in a distributed and superpositioned fashion, virtually opaque to an analytical investigation.

DM reaches across multiple domains. Its syntactic and phonological components can be reduced in a rule-based fashion; albeit with more effort due to derivational allomorphy. The fundamental issue lies with the chaotic nature of lexicalization. It is too integral a part of DM to be eliminated by simplifying assumptions. At the same time, we cannot lose sight of the syntactic component, if we are to remain in the realm of morphology. To sum up, we need a framework to study both syntax and the lexicon.

In order to escape the limitations of existing paradigms, we require a novel method. Ideally, this method should be based on an adequate linguistic structure, be able to learn from data and operate in a manner that is congruent with linguistic facts. In the words of Herbert Simon, we are looking for a theory-driven method.

2.4.1 Theoretical Approaches

In this category, studies of morphology from theoretical linguistics naturally occupy a large place. Ideas based on various layers of linguistic structure can be found in the literature. Some of these studies rely on examples from a variety of languages in order to deduce general rules. Others try to model selected processes, often with the help of grammar frameworks (HPSG, LFG, CCG etc.).

Regarding morphology and its status as a distinct linguistic module, prominent authors have presented their ideas: Lieber (1992), Aronoff (1994), Aronoff and Fudeman (2022) and Carstairs-McCarthy (2010). These were reviewed in Section 2.1. Methods of such studies are way too high-level for our purposes.

Another long debate has been ongoing over the Lexical Integrity Hypothesis (LIH). Lieber and Scalise (2006) give a comprehensive an insightful review of various ideas and turning points in this debate. We simply comment that our investigations regarding Turkish DM in Chapter 3 do not find definitive evidence for or against LIH; in turn, the extensive debate on LIH does not provide definitive evidence for or against the approach proposed in this thesis.

Halle et al. (1993), Halle and Marantz (1994) and Marantz (1997) argue for Distributed Morphology (DdM). DdM states that morphemes are the basic construction blocks of phrases, not words. Word formation follows the rules of word-internal syntax, while inflection still operates on the ordinary syntactic layer. Marantz (1997) even makes the claim that DdM holds the key to getting rid of lexicalism "once and for all". Williams (2007) concisely explains the difference between DdM and LIH. He states that DdM proposes that sentences are made of morphemes; while LIH proposes that sentences are made of words, which are made of morphemes. We lean towards adopting Marantz (1997) and Marantz (2013)'s ideas that morphology cannot be studied in isolation from syntax or morphemes. The best theory would need to recognize the different paths morphology may take.

Positioning himself with respect to LIH, Booij (2009) and Booij (2010) propose Construction Morphology (CnM). Booij (2009) claims that LIH has two components: non-interruptibility and non-accessibility of word-internal structure. He accepts the former, while rejecting the latter. This line of thinking is in line with Lieber and Scalise (2006) and our own. Chapter 3 goes into great detail looking for evidence for or against non-interruptibility. Section 3.2.4 investigates suspended affixation, finding that DM cannot be suspended, while IM often can. Perhaps the boundary of a word must be drawn before IM. Even this line of demarcation creates problems, though; for instance with the Turkish recycler *-ki*.

One extreme idea concerning the processing of morphology, that we cannot leave unmentioned, was the Full Listing Hypothesis (FLH), which suggests that all forms were listed in the mental lexicon. Hankamer (1989) convincingly argues based on observations from agglutinative languages that FLH, and its variants, cannot be true. Unsurprisingly, FLH has not been popular in recent literature. We believe the evidence in Section 2.2 is enough to disprove this claim.

Theoretical approaches such as the ones reviewed above are focused on discovering a structure behind observations that the authors use as reference. While such approaches has yielded quite an impressive literature and advancement on our knowledge of Language, they present a couple of common drawbacks. The building blocks on which the proposed structures are developed are not really definitions, they are prototypes. Therefore, strong structural claims on even very narrow domains fail when presented with counterexamples. We will review several such cases from Turkish in Chapter 3. The solution cannot be to keep making strong claims over narrower domains, because such claims would not be generalizable or meaningful. Perhaps, it would help to make weaker but more defensible claims over domains large enough to be meaningful. This is why weaker versions of LIH, DdM and CnM are still being discussed. These are relaxed versions of past approaches, updated according to the latest theoretical setting. One particularly effective method of relaxing structural theories is minimizing the structure to cover the basic principles and transferring the rest to the lexicon as idiosyncratic learning blocks. This is what makes CnM so convincing.

Several grammatical frameworks have gained and lost popularity over the years. Work on these frameworks can be considered a theoretical endeavour, since they essentially act as testing grounds for theoretical claims. HPSG, LFG and and Categorial grammars (CG) are among these frameworks. Combinatory Categorial Grammar (CCG) Steedman and Baldridge (2011) is an exceptionally successful framework that handles syntactic and semantic operations in lock-step, achieving seamless integration between the two modules. The syntactic aspects of the present issues may be represented very well in CG. In Section 4.4 we go into more detail about the capabilities of CG and how we use it to represent meaning.

Although CG offers many advantages, DM does not lend itself easily to a categorial representation. First, derivational processes are not fully productive; therefore, extensive semantic selection must be taken into account in order to avoid generating a large number of invalid constructions. Second, widespread polysemy in DM forces us to multiply lexical items with the same form. Distinct meanings can be represented in this way, at the expense of complicating the model.

For these reasons, the literature on CG does not provide a large number of studies focused on DM. We believe that CG alone would not be an adequate framework for studying DM. However, within the CdS framework and in a probabilistic setting, CG would be quite adequate for representing meaning and deriving constructions.

2.4.2 Statistical Approaches

Statistical approaches largely fall into the psycholinguistics paradigm. Two most prominent sources of data for a statistical analysis are corpora and experimental data. Experimental settings further divide into two major categories: evaluation of responses and evaluation of reaction times.

Seidenberg and Gonnerman (2000) give a comprehensive review of literature on morphological processing. In order to avoid repetition, we refrain from extensively discussing the studies already reviewed in Section 2.2.2.

Whole word access models Manelis and Tharp (1977) and componential access models Taft and Forster (1975) predate hybrid models. At that stage, the theoretical space was divided between proponents of the decomposition route and proponents of the retrieval route. First true hybrid models were dual-access models Bybee (1985). Some studies Marslen-Wilson et al. (1994) distinguished between regular forms that allowed decomposition and irregular forms that required retrieval. Similarly, some studies distinguished between derivation (which is often wrongly associated with irregularity) and inflection (which is correctly associated with regularity). Yet some others distinguished between suffixed and prefixed forms. The reader can find the rest of the references in Seidenberg and Gonnerman (2000). One important question is whether these claims originate from a theoretical understanding of Language, or they are motivated by data. More often than not, the latter is true. The experimental data is often based on reaction times, which constitutes only indirect evidence towards rapid or slow recognition, in the presence of confounding factors. Furthermore, the data is collected on a single language (usually English), on a sample size of 20-40 subjects (sample sizes rarely go above 60). Thus, even if the results are completely valid, their generality across languages is open to debate.

More recent versions of hybrid models are race models as in Caramazza et al. (1988) and Frauenfelder and Schreuder (1992). Race models typically assume that both retrieval and decomposition routes remain open for the processing of a complex form. Therefore, processing becomes a race between these two routes. Like in any race, the winner is the fastest alternative. Since the speed of processing is known to change due to priming, analysis of priming effects as in Laudanna et al. (1992) became an important part of such studies. The recognition of dual-route processing in these studies is noteworthy. We also adopt this position. However, the experimental evidence regarding processing speed and priming are quite indirect. The high cost of collecting experimental data and the large number of confounding variables in these experiments make it virtually impossible to validate claims.

Even the hybrid models reviewed above talk of activation of words and morphemes by letter strings (Caramazza et al., 1988). Interactive activation models can therefore be considered a natural continuation of this line of research. One important example for such studies is Taft and Zhu (1995). As discussed in Section 2.2.2, these models require the researcher to decide on the levels of representation.

If one can make justified decisions concerning representation levels, interactive-activation models are quite promising for several reasons. First, they make it possible to setup a hierarchical structure, where different levels of linguistic processes can take place in relative isolation. This is crucial, because the combination of all such processes create a very large and complicated picture resisting analysis. An appropriate level of modularity facilitates the modeling effort. Second, interactive activation models are adaptable. Taft and Zhu (1995) proposes several architectures such as concept & sub-morpheme, concept & morpheme & sub-morpheme, concept & word & bound-morpheme & sub-morpheme and concept & character & radical. The architecture of an interactive activation model can be adapted to the structure of the language, as can be seen in the character and radical levels designed for Chinese. For an agglutinating language such as Turkish, perhaps another architecture would be more fitting. Third, a hierarchical representation is naturally suited for information compression. Processing large sentences would become intractable even with a light unstructured representation. The number of possibilities for segmentation, lexical selection and derivation are simply enormous. Only a hierarchical model can organize the necessary information in a compact and digestible format. In this thesis, we do not directly adopt the interactive activation approach, but we propose a novel hierarchical model that benefits from the advantages listed above.

Mayo (1999) is an interesting study that attempts to build what is essentially an interactive activation model. Adopting a software engineering perspective, the author gives detailed descriptions of the modules and algorithms that make up the model. The requirements for the software are supplied by the facts and observations based on psycholinguistics literature. Crucially, function words and DM are among the initial input and cannot be learned. Program traces for just three words are provided as experimental results. Overall, the author does not make an attempt at attaining theoretical insights.

Many other studies look at the statistical correlations between words and characters in a corpus, or in an experimental design. Saffran et al. (1996) ties segmentation to statistical analysis, based on an experiment with 8-month-olds. In another research program, Schreuder and Baayen (1995), Baayen and Schreuder (2000), Baayen (2001) and Baayen et al. (2011) explore the effect of frequency on morphological processing. Starting with an interactive activation model in Schreuder and Baayen (1995), they move closer to a data-driven computational model in Baayen and Schreuder (2000), present a guide for word frequency research in Baayen (2001) and propose an “amorphous model for morphological processing” in Baayen et al. (2011). Statistical information on morphemes and words can indeed be useful in expanding our understanding of language. However, statistical relations are unlikely to capture the other, more potent components of processing. Semantics, for one, may be playing a much more important role in learning and processing. Morphemes, which are the building blocks of morphology, have categorial relations and semantic content in interaction with their context. As Marantz (2013) posits right in the title of his study, we do not think morphemes can be ignored in an investigation of morphology.

Seidenberg and Gonnerman (2000) point out the drawbacks of studies on frequency effects and propose a distributed connectionist model. Gonnerman (1999) explores morphological priming effects due to semantics and phonology. Such models eliminate the multi-layered structure that is the essence of interactive activation models. Seidenberg and Gonnerman (2000) consider this a positive, because a more straightforward model, with fewer moving parts is desirable. Connectionist models also do not explicitly classify words as morphologically simple or complex, which is considered another advantage. If the goal is to account for the frequency effect with the simplest model, they are correct. However, in order to understand human processing, we need more transparent models. Connectionist models are opaque and hard to interpret.

Finally, we have studies on acquisition, which are based on similar experimental settings as hybrid models, but focus more on acquisition than processing. Most of these are reviewed in detail in Section 2.2, therefore a quick glance will suffice. Tyler and Nagy (1989) and Nagy et al. (1993) explore the acquisition of English DM. Anglin et al. (1993) studies the effect of morphological analysis on vocabulary development. Bertram et al. (2000), an excellent source in this line of research, work with Finnish school children to establish the extent to which the awareness of DM aids vocabulary acquisition. Due to the agglutinating nature of Finnish, this study has been especially important in our investigation of Turkish DM. Last, but not least, Clark (2014) looks at the acquisition of DM across multiple languages, trying to find common tendencies. The studies in this final group do not model processing, nor are they really statistically motivated. However, they deserve mentioning a second time, because they are reliable in their weaker claims; and their findings are indispensable for validating model behavior.

2.4.3 Computational Approaches

Some studies reviewed in previous sections could also be characterized as computational. However, they were successors to theoretical paradigms, closer to their predecessors than the studies we review in this section. By computational approaches, we mean algorithms run on large corpora for learning Part-of-Speech (POS), morphology, word embeddings or other related information. Learning is usually unsupervised or semi-supervised.

Recent years have seen an explosion in computational studies on linguistics. Numerous different topics and paradigms have been gaining popularity. Therefore, it is impossible to give a comprehensive overview of the field. We only go over a few lines of research that are especially relevant for present purposes. For a more comprehensive review of computational linguistics, the reader is referred to Roark and Sproat (2007) and Mitkov (2022).

One prominent area of research is unsupervised learning of morphology. Starting with Gaussier (1999) and Goldsmith (2001), this has been a popular topic among computational linguists. Hammarström and Borin (2011) provides a review of literature. The central idea is to devise an algorithm to recognize and exploit the consistencies across different but related words. According to Hammarström and

Borin (2011), there are mainly four approaches for attacking this problem. Substrings' frequent cooccurrence with others, string edit distances, distributional similarities, n-grams, phonological categories and many other features are considered as evidence in these studies. Their objective is typically to learn segmentation (usually), paradigms, affixes or rewrite rules.

Ruokolainen et al. (2016) reviews the literature on the related problem of semi-supervised learning of morphological segmentation. In this line of research, the objective is exclusively to learn segmentation boundaries. Creutz and Lagus (2007) describes the well-known Morfessor model family and continues to influence the literature to date. Üstün and Can (2016) and Can and Manandhar (2018) are more recent studies on morphological segmentation. Üstün and Can (2016) exploits distributional semantics to guess segmentation boundaries. Can and Manandhar (2018) proposes a hierarchical clustering model based on a hierarchical Dirichlet process. They obtain the best F-measure among comparable algorithms for MorphoChallenge 2010-Turkish.

Many other tasks are carried out by computational methods. For instance, Goldwater and Griffiths (2007) and Ravi and Knight (2009) make unsupervised POS-tagging. Integrated models have also been constructed by Bowman et al. (2016) for parsing and understanding sentences, and by Can et al. (2022) for segmentation, morpheme-tagging, POS-tagging and dependency parsing. Both models are based on the LSTM architecture.

One key disadvantage of connectionist models, as discussed in Section 2.4.2, is their opacity to interpretation. This is true for many computational methods, too. Smolensky (1999) argues that connectionism and generative grammar are not incompatible. He makes several proposals regarding how a bridge can be built between the two paradigms. Perhaps such a bridge is possible.

Distributional semantics (DS) is another important area in the computational linguistics research. Especially in the recent years, the number and complexity of DS research has been rapidly increasing. Here, we only review a small number of studies relevant to an investigation of DM. We keep this part brief, saving the details for Section 4.1.

Embeddings do not have to be constructed for words. Character-based Bojanowski et al. (2017), morpheme-based Cotterell and Schütze (2019), and syllable-based Choi et al. (2017), Üstün et al. (2018), Şahin and Steedman (2018) embeddings can be computed. However, some of these models require extensive pre-processing. Word embeddings can be indirectly used to infer morphological relations as in Zargayouna et al. (2017)'s search for analogy pairs and Gladkova et al. (2016)'s search for morphological and lexical relations. Similarly, Musil et al. (2019) test whether differences between lemma and base embeddings have a strong correlation with derivational relations. Botha and Blunsom (2014) follows the same line of thinking by assuming that compositional derivational relations can be modeled with simple addition over embeddings. Studies that improve word embedding quality with the help of supplementary features include Cui et al. (2015) and Jurdzinski (2017).

DS is still a relatively novel way of understanding semantics. It offers a way to represent meaning in computational methods. However, by themselves, word embeddings (or morpheme embeddings for that matter) do not give us much insight on linguistic processes. Their usefulness ultimately depends on the quality of the structure in which they are positioned. We believe that DS can be a useful component within CdS.

2.4.4 Theory-Driven Approaches

A distinct category must be reserved for the so-called "theory-driven" studies. These studies are not purely theoretical or computational; they integrate a theoretical framework into a model that can process large amounts of data. This is exactly what this thesis aims to do.

We focus mainly on two lines of research: probabilistic categorial grammars and Bayesian belief networks. Probabilistic grammar is hardly a new paradigm, but it is still developing. Clark and Curran (2003) and Clark and Curran (2007) attempt wide-coverage parsing based on the Combinatory Categorial Grammar framework (CCG). They use log-linear models to estimate parameters for lexical items, and indirectly track probability ranking of alternative parses. Zettlemoyer and Collins (2007), Zettlemoyer (2009) and Zettlemoyer and Collins (2012) work on semantic parsing based on probabilistic CCG (PCCG), introducing a new set of combinators. Later, Kwiatkowksi et al. (2010), Kwiatkowski et al. (2011) and Wang et al. (2014) attempt to generalize the algorithm in several ways. A more recent study is Abend et al. (2017) that models word learning and syntax learning concurrently. It is also an incremental learning model, making it more psychologically plausible. In fact, the authors simulate learning on the CHILDES corpus.

PCCG is a nice way of introducing probability ranking into the lexicon, but it is not perfect. It only tracks the prominence relations between lexical alternatives, but not on segmentation alternatives, at least not yet. In other words, the segmentation layer described in Section 2.3.1 is not represented on PCCG. Granted, alternative derivations produced by a PCCG model indirectly lay out the segmentation alternatives, but this does not give the whole picture. Most studies depend on word boundaries for distinguishing lexical items from each other. When word-internal syntax is in question, this approach falls short. More recent studies such as Wang et al. (2014) solve this problem by obtaining morphological information in pre-processing.

Another drawback of the PCCG approach, is its indirect representation of salience asymmetries between segmentation alternatives. The salience of a segmentation alternative can be computed from the proportion of correct derivations it appears in. In PCCG, segmentation alternatives are not tracked with dedicated parameters like lexical items. If speakers develop tendencies towards retrieval or decomposition, this must be an independent parameter to track; not just an indirect consequence of parameters on some lexical items. The same lexical item may take part in many constructions, but one construction may be more likely to be retrieved, while another is more likely to be decomposed. The two layers must be kept separate.

Of course, this shortcoming may be resolved by modifying the PCCG architecture. However, another framework offers a more natural fit to the structure we propose. Bayesian Belief Networks (BBN) are directed acyclic graphs that model probabilistic variables with conditional dependence relations. They have seen application in many different areas of research. They can be custom built to represent a particular theoretical structure, they can learn from data and hold information in a hierarchical structure in greatly compressed fashion.

There is also great effort towards applying Bayesian principles to cognitive processes. Gopnik et al. (2004), Gopnik and Schulz (2004) and Chater et al. (2006) explore this possibility in several fields. Takahashi and Ichisugi (2017) and Ichisugi and Takahashi (2018) study the plausibility of this from a neuroscience perspective.

Several authors were convinced by these qualities to model human language learning with the help of BBNs. Tenenbaum (1999) sets the stage with a framework for concept learning. Xu and Tenenbaum (2007), Piantadosi et al. (2008), Tenenbaum et al. (2011), Perfors et al. (2011), Piantadosi (2011) and Lake et al. (2015) study word learning, concept learning and discovery of syntax using the Bayesian framework. Especially Piantadosi et al. (2008) is relevant for our purposes, as they model the learning of compositional semantics.

Two drawbacks can be observed in these studies. First, they do not simulate incremental or naturalistic learning. Forms and meanings are input to the algorithm without being matched. Over time, the algorithm is expected to establish the correct matches between the two lists. Second, the authors focus on the statistical learning capabilities of BBN, rather than modeling human processing. Therefore, little linguistic theory has been integrated into the models, so far.

On the other hand, BBN have two important advantages over PCCG. First, they can represent hierarchical structures more explicitly. This is crucial, especially considering the segmentation layer. Second, they use and store information more efficiently. This is partly due to their hierarchical structure. Another reason is that independence assumptions can be used to constrain the model. We believe BBN offers a convenient representation scheme for CdS in Section 2.3.4.

2.4.5 A Simple Processing Model

In this section, we do not propose a computational model that correctly converts input into output. Data-driven models cited above are perhaps more suited to that purpose. We are after a model that is transparent like a theoretical model, responsive to data like a data-driven model and able to explain the statistical findings from psycholinguistics. The model must serve as an implementation of CdS described in Section 2.3, not stand alone as a black box representing input-output relations.

Bayesian Belief Networks (BBN) is a quite suitable modeling framework for this purpose. The hierarchical structure of BBN lends itself well to the layers of CdS. We can use discrete nodes in a BBN to represent segmentation and lexical alternatives, making the model transparent to interpretation. Probabilistic nature of BBN makes it possible to explain findings in relation to statistical analyses elsewhere.

In this section, we present minimal examples to describe a Bayesian processing model that represents CdS. We focus on the comprehension process. We also run the process assuming an early stage of acquisition, namely the first few attempts at comprehending a given observation. This is to ensure that the reader may easily observe the competition and trade-offs embedded in CdS. With adult-level comprehension, new observations would still have an effect on the comparative salience of segmentation and lexical alternatives, but the effect would be much smaller.

Let us first demonstrate the basic principles on nouns, instead of full sentences. Simple expressions initially dominate Child-Directed Speech (CDS), coupled with gestures and motions to emphasize the contextual clues:

(18) a. CDS: *kitap* ‘book’ (Pointing at a book...)

b. CDS: *kalem* ‘pencil’ (Pointing at a pencil...)

c. CDS: *su* 'water' (Pointing at water...)

These observations are much simpler than an average sentence of adult language. They are morphologically simple, too. The following examples are morphologically complex, but the child has no way of knowing this:

(19) a. CDS: *kitaplık* 'bookshelf'

b. CDS: *kalemlik* 'penholder'

c. CDS: *suluk* 'water bottle'

The Turkish suffix *-lIK* has many functions, one of them being the derivation of container names. Recognizing the affix is straightforward for an adult; but it has to be discovered by the child. Since children are able to acquire derivational affixes before any formal training, we can assumed that affixes can be discovered without explicit instruction. There must be a consistent mechanism behind this.

For the first set of words, we assume that they are simply matched with the contextual clues and absorbed into the lexicon. They are monomorphemic, so there is no meaningful way of discovering internal structure. The child cannot analyze them, either.

On the other hand, alternative approaches are available for the second set. From the hearer's point of view, each observation is treated as a simple form until an internal structure can be discovered. Therefore, lexical items are generated for whole forms, regardless of their being simple or complex. Internal structure can be discovered based on common constituents (common form and common semantics) across multiple observations. When an observation is encountered a second time, the hearer is expected to retrieve its meaning from the lexicon. Once the lexicon contains the constituents, decomposition also becomes available.

Content of the lexicon depends only on the hearer's past observations. The child starts with an empty lexicon, collects new lexical items and recognizes affixes. The following is a short list of CDS observations to simulate this process. We shorten the simulation by including only the observations where the context is strong enough for the hearer to understand the meaning. Other observations are simply discarded. This is a minimal example, because each observation is included only once. Finally, we use simplified logical forms to track the semantics of each observation and extract constituent semantics. As always, we start with an empty lexicon.

(20) a. CDS: *su* (New lexical item: *su* 'water')

b. CDS: *kalem* (New lexical item: *kalem* 'pencil')

c. CDS: *suluk* (New lexical item: *suluk* 'water container')

d. CDS: *kalemlik* (New lexical item: *kalemlik* 'pencil container')

e. Mental process (New lexical item: *-lIK* 'x container')

The first four observations are impossible for the hearer to analyze. They must be deduced from the context. For now, we assume affix recognition is possible with only two sets of observations with common constituents. Following these observations, a mental mechanism recognizes the affix *-lIK*. Now, the lexicon contains five entries.

(21) a. CDS: *Kitap* (New lexical item: *kitap* 'book')

b. CDS: *Kitaplık* (Decomposition)

The next problem is how the hearer handles a new morphologically complex word. The root *kitap* is encountered first and *-lIK* was already in the lexicon; therefore, decomposition of *kitaplık* is possible. In the next observation, retrieval of *kitaplık* is not possible, because there is no entry for *kitaplık* in the lexicon. However, decomposition is possible this time. Segmentation is the first step.

(22) a. kitaplık: Unattested.

b. kitaplı-k: Unattested.

c. kitapl-ık: Unattested.

d. kitap-lık: Attested.

e. ...

f. kitapl-ı-k: Unattested.

g. ...

The only attested segmentation is *kitap-lık*. One lexical item matches each segment. The only interpretation resulting from decomposition is 'book container', which is correct. *kitaplık*, as a whole, is also added to the lexicon. Next time it is encountered, both retrieval and decomposition will be possible. Competition on the segmentation layer does not take place between retrieval and decomposition, it is rather between the different segmentation alternatives of an observation. Retrieval is simply what we call segmentations that consist of a single segment.

Segments may have multiple lexical matches. Adding the following observations to the lexicon adds some complexity to the process.

(23) a. CDS: *göz* (New lexical item: *göz* 'eye')

b. CDS: *gözlük* (New lexical item: *gözlük* 'eye apparel')

c. CDS: *baş* (New lexical item: *baş* 'head')

d. CDS: *başlık* (New lexical item: *başlık* 'head apparel')

e. Mental process (New lexical item: *-lIK* 'x apparel')

Now, processing *kitaplık* will be slightly more demanding, because there is one more layer of ambiguity. Multiple segmentation alternatives and multiple lexical alternatives are attested.

(24) a. kitaplık ‘book container’

b. kitap-lık ‘book container’

c. kitap-lık ‘book apparel’

Another short example suffices to demonstrate the application of the same approach on the syntactic level. Again assuming an empty lexicon, the following observations are given to a hearer, leading to the recognition of the plural marker:

(25) a. CDS: *kitap* (New lexical item: *kitap* ‘book’)

b. CDS: *kalem* (New lexical item: *kalem* ‘pen’)

c. CDS: *kitaplar* (New lexical item: *kitaplar* ‘book multiple’)

d. CDS: *kalemler* (New lexical item: *kalemler* ‘pen multiple’)

e. Mental process (New lexical item: *-lAr* ‘x multiple’)

The same process can apply on the phrase level, as well.

(26) a. CDS: *kitap* (New lexical item: *kitap* ‘book’)

b. CDS: *kalem* (New lexical item: *kalem* ‘pen’)

c. CDS: *kitap geldi* (New lexical item: *kitap geldi* ‘book came’)

d. CDS: *kalem geldi* (New lexical item: *kalem geldi* ‘pen came’)

e. Mental process (New lexical item: *geldi* ‘x came’)

These three examples are designed to demonstrate how the model works from the comprehension perspective. A look from the production perspective could shed light on the lexicalization process. The Nootka example from Swadesh (1938) (the evolution from “large thing” to “whale”) inspires our fourth example. Imagine there is no word for whale in a language. The concept could be conveyed with a construction such as “large thing”. Starting with an empty lexicon, assume the following observations are encountered.

(27) a. CDS: large (New lexical item: large ‘large x’)

b. CDS: thing (New lexical item: thing ‘thing’)

c. CDS: large thing (New lexical item: large thing ‘large thing’)

d. CDS: large thing (New lexical item: large thing ‘whale’)

“large” and “thing” are quite frequent words, used in countless contexts. Their meanings are probably understood quite early. The phrase “large thing” is easy to derive once the constituents are known.

Even if the concept of a whale is indicated by “large thing” only once, it becomes the sole alternative for this role. “large” and “thing” do not change their meaning at all; but the phrase “large thing” assumes a narrower, non-compositional meaning, in addition to its ordinary meaning. Depending on the frequency of the whale concept in daily encounters, the narrower meaning may even dominate the probability distribution for the semantics of “large thing”. This is the process of semantic fixation (Swadesh, 1938). The process would be the same if one of the constituents were a bound form, as in the Nootka word for whale.

Contrary to common perception, lexicon does not store linguistic irregularities. Lexicalization precedes irregularity (or non-compositionality, in the context of DM). CdS and its inherent trade-offs ensure that form-meaning pairs gain or lose salience according to the habits of speakers. A slight divergence from the accepted meanings first establishes a novel branch of polysemy, but starts to evolve separately from the original semantics. Only when branches diverge enough, we can speak of irregularity.

Choices in one layer determine valid alternatives in the next. Therefore, CdS displays a hierarchy of preferences. The forward stage of processing takes place from segmentation, through lexical selection to derivation. After all alternative interpretations are obtained, they are checked against the context, and labeled as correct or incorrect. This starts the backward stage of processing. This time, the network is traversed in the opposite direction towards segmentation alternatives. Segmentation and lexical alternatives that contribute to correct interpretations are rewarded with higher probability in future observations.

Bayesian Belief Networks (BBN) offer a suitable framework for the structure and processes we described above. A BBN graphically represents a set of probabilistic variables, as well as the conditional dependency relations between them. It is capable of representing the hierarchical relations between successive stages of processing, as well as the probability distributions that represent the competition between segmentation and lexical alternatives. Chapter 5 describes this framework in further detail.

Such a model promises to alleviate the drawbacks of classical methods used in research on DM. It incorporates a data module to drive incremental learning, but it does not converge to the massively data-driven unsupervised learning models of computational linguists. It exploits a theoretical understanding of linguistic categories and morphological structure, but does not stop at proposing a structure. The steady-state of this model is also able to produce the frequency effects reviewed in the psycholinguistics literature.

2.5 Claims and Objectives

The primary issue to consider in our study is DM’s unique nature. Derivational processes occur on a gradient from lexicon-like to grammar-like. Being lexicon-like implies irregularity, non-compositionality, non-productivity and retrieval. Being grammar-like implies regularity, compositionality, productivity and decomposition. It is established that syntax and IM are grammar-like processes. The dichotomy between DM and IM has for long led researchers to consider DM a purely lexicon-like process. Reviewed evidence contradicts this. Also, the possibility for a computational study of DM depends on DM displaying regularity to some extent. In Section 2.2, we observed that DM occurs on a spectrum from non-compositional to compositional forms, with a large portion fully compositional. This constitutes the starting for this thesis. The corresponding aim is to provide further evidence for this regularity, based on an investigation of DS.

On the opposite side of the spectrum, we have to account for the many derived forms that have lexicalized with non-compositional semantics. In this process, derived forms lose their internal structure in the eyes of the speakers. If no affix can be recognized in a derived form, it should no longer be considered a derived form. Numerous examples of this process can be found in Turkish. As a principle, we accept that the existence of a DM is tied to its learnability. We do not speak of a DM, if it cannot be discovered by speakers. Admittedly, the linguistic exposure of each speaker is unique; therefore, their linguistic knowledge must also be unique. We respect this fact and point out possible effects of individual exposure on the acquisition of DM.

Another point concerns the appropriateness of basing a study of DM solely on morphological structure. Morphology is usually analyzed in terms of morphological structure, morphemes and morphomes. This method works well with IM, since IM regularly follow the rules of phonology and syntax. Its semantics is also regular. As a result, IM can be represented by clear-cut rules and processes. This is not the case for DM. DM occurs on an unrestricted semantic space, in the presence of complicating phenomena such as suppletive allomorphy and extensive polysemy. There is many-to-many correspondence between semantic content and derivational affixes. This creates ambiguity and competition. The second principle we adopt is that an analysis of morphological structure is not sufficient to explain DM. The corresponding aim is to propose an alternative structure that accommodates the distinctive characteristics of DM. With CdS of DM we propose in Section 2.3.4, ambiguity and competition can be clearly represented.

According to this structure, layers of ambiguity in morphology processing manifest themselves as layers of probability distributions. For the segmentation layer, the probability distribution represents the relative salience of alternative segmentations. Similarly for the lexical selection layer, the probability distribution represents the relative salience of available lexical items. The alternatives generated by the derivation layer do not need to be represented in the same way, as they are aggregated according to whether they lead to a correct interpretation or not. When these layers are considered together, we obtain a multi-level probabilistic structure, with some independence assumptions.

Proposing a structure for a linguistic process is one thing. Demonstrating its operation with the help of a model or an algorithm is crucial for communicating its principles and assessing its applicability in real world. The literature on DM does not really offer an appropriate modeling framework for this structure. Studies on the theoretical side, the statistical side and the computational side each present unavoidable drawbacks on the account of responsiveness to data, transparency to analysis and psychological plausibility. On the other hand, theory-driven approaches such as PCCG, DS and BBN can be considered more promising for gaining insights into how the mind works. We believe a model based on BBN constitutes a good approximation of CdS of DM. The corresponding aim is to create such a model and an accompanying algorithm.

Finally, based on evidence from psycholinguistics, we recognize that two processing routes are available for complex forms: retrieval and decomposition. These routes operate in parallel. The segmentation layer directly represents the competition between these two routes, although the latter may give rise to multiple alternatives. Again, based on evidence, we recognize that it is possible for one route to gain prominence over others.

The nature of DM knowledge makes it very difficult to gather data. This is especially true without an adequate structure to direct data collection and interpretation. In this thesis, we provide the proof-of-concept for such a structure. Based on this structure, we also propose a model and develop an algorithm. However, experimental data for validating our proposals remain as the bottleneck. Gathering such data is a formidable task that requires several years of dedicated work. We hope future work can plug this gap.

(28) Principles for a study of DM

a. DM occurs on a spectrum from non-compositional to compositional forms. A large portion of derived forms are compositional. A computational study is meaningful only if this is true.

b. Contents of DM is based on the speakers' ability to discover it. Knowledge of DM varies from individual to individual, based on previous linguistic exposure.

c. DM has lexicon-like and grammar-like components. There is a unifying structure behind these components.

d. The morphological structure may be sufficient to explain IM. DM must be studied with a CdS laid over the morphological structure.

e. CdS of DM consists of probabilistic layers of competing segmentation alternatives and lexical alternatives.

f. Theory-driven approaches, especially BBN, are more suitable for building a model that fits CdS.

g. Based on available segmentation alternatives, morphological processing takes two paths concurrently: retrieval and decomposition.

h. Gathering and assessing experimental data without an adequate structure and model of DM is hardly fruitful. Our proposals may aid data gathering efforts by suggesting where to look and how to interpret data.

CHAPTER 3

A NEW CLASSIFICATION OF TURKISH DERIVATIONAL MORPHOLOGY

The necessity for and the plausibility of CdS proposed in the previous chapter depends to some extent on the properties of the target language. For instance, decomposing words in a purely isolating language would not be possible. The segmentation layer would lose its meaning in that case, because the only segmentation alternative would be the single-segment one. In a language where there is very little homonymy or polysemy, lexical selection would lose its significance. For each form, there would usually be a single interpretation. Without any competition (or ambiguity) due to segmentation and lexical alternatives, we would not need CdS.

On the other hand, in a language where words license many segmentation alternatives and where segments license many lexical alternatives, the advantages of CdS would be significant. We believe agglutinating languages are especially suited to an investigation based on CdS. In more general terms, the proposed structure is appropriate for languages that lend themselves well to an analysis based on Item and Arrangement. In this chapter, we take Turkish as an example. Based on an analysis of Turkish morphology, we map the space of morphological possibilities both in terms of form and meaning.

This chapter is not an attempt to "fix" Turkish grammar. There are already excellent resources on the capabilities of Turkish DM such as Göksel and Kerslake (2005). Existing grammars serve their purpose well. This chapter is an attempt to organize Turkish DM as preparation for computational processing. The coming sections are full of examples demonstrating the messy, complicated nature of DM. Our aim is to untangle the many different pieces of the puzzle, to the best of our ability.

This chapter contributes to our understanding of Turkish DM in several ways. First, it helps dispel the myth that DM is highly irregular or non-compositional. This cannot be further from the truth. Most derived forms are completely regular. The confusion is partly due to the polysemy relations common in DM. This regularity, even if it is only partially true, is a necessary condition for the applicability of CdS.

Second, we classify morphemes under a different light than previous approaches. When we find it difficult to classify a morpheme, we turn to Orkhon Turkic (OT) for insights. Semantic relations in Modern Turkish (MT) are at times very complicated or very specific. As expected from millenia of evolution, morphological processes operate as an intricate web of semantic connections. Some morphemes gained very specific meanings and selection criteria, while others lost productivity. As a result, it is very hard to discover the underlying network that binds together different morphemes. With the relative simplicity of OT, this is much easier. Of course, insights from OT can only be used when they are also discoverable naturally by speakers of MT. Therefore, OT only helps us discover the original connections and we assess whether those connections are still valid.

Third, we build two dimensions against which a new classification may take place. These dimensions help us organize different aspects of the forms and meanings of morphemes. In the form dimension, grouping morphemes according to suppletive allomorphy is crucial. This is often overlooked. Simple allomorphy does not suffice when DM is considered. In the meaning dimension, we group morphemes according to polysemy. For denominal verbs and deverbal nominals, we propose a more universal scale for the meaning dimension: thematic relations.

Fourth and most importantly, we propose a novel classification of Turkish DM, based on the dimensions described above. The thinking that lead to this classification is explained in Section 3.4.

The findings of this chapter brought us to the conclusion that a multi-layered probabilistic structure such as CdS is suitable, even required, for representing Turkish DM. The fact that Turkish derivational morphemes can be organized in this way, both with respect to form and meaning, is what encouraged us to develop CdS.

We first present an overview of Turkish morphology, point out issues in the current understanding of Turkish DM, discuss our observations and propose a new classification of Turkish DM. The new classification suggested here serves as a basis for computational analyses in later chapters.

3.1 Overview of Turkish Morphology

In this section, we present a review of Turkish morphology based on several sources. This review covers both synchronic and diachronic perspectives, incorporating insights from our study of OT. The differences between synchronic and diachronic analyses highlight the issues present in the contemporary understanding of Turkish morphology.

3.1.1 Main Sources for Modern Turkish

Oflazer et al. (1995) gives "an outline of Turkish morphology", complete with a list of morphophonemic processes, an affix inventory and finite-state machines for nominal and verbal morphotactics. Bozşahin (2018) builds on the affix inventory presented in this work. He creates a list of primitive binary concepts and analyzes each affix in the inventory using these concepts. The idea is to come up with a conceptual analysis of these processes to provide a reliable basis for further research on Turkish morphology, syntax and semantics. A few of these binary concepts (out of a total 24) are CAUS (causative), FACIL (facilitative), GRAD (gradable), PRED (predicative) and STAT (stative). This inventory is a good starting point for studying Turkish DM.

Many affixes are omitted from this list, despite their productive use, due to their non-Turkish origins. Affixes like *na-* from Arabic (i.e. *natamam* 'incomplete') and *-syon* from French (i.e. *fermantasyon* 'fermentation') are quite productive and they are consciously deployed (and even sometimes removed) by Turkish speakers. Nevertheless, including these affixes in an inventory of Turkish affixes is neither plausible nor would provide any additional insight. Some other affixes, like *mü-* from Arabic (i.e. *müdahale* 'intervention') and *-aj* from French (i.e. *arbitraj* 'arbitrage'), cannot be considered productive at all. Words derived by these affixes cannot be analyzed by (most) Turkish speakers, so they must be kept in the lexicon in their final form. Therefore, we do not consider these as valid derivational processes in Turkish.

A more conventional source is Göksel and Kerslake (2005) which provides a truly comprehensive grammar of Turkish. This extremely detailed and perfectly organized book is an invaluable resource for Turkish grammar. Erguvanlı-Taylan (2001) provides an even deeper analysis of Turkish verbs. Looking at various aspects of verbs, Erguvanlı-Taylan (2001) is full of insights. Ergin (2009) is another grammar book for Turkish that we reviewed in search for insights.

Aslan et al. (2018) develop a large database of Turkish words, each annotated with a selection of morphological features such as POS (both base and lemma), reciprocal, passive, phonetic transformation (in the last morpheme of the lemma), attachment (phonetic events like consonant epenthesis), deletion (haplology) and source language. Most of the entries come from the TDK (Turkish Language Association) dictionary. They also provide statistics on the morphological complexity of the items: Only 37.8% of the items are roots, 56.7% are derived and 2.7% are compounds. We have found many interesting examples in this database.

3.1.2 The Coding Scheme

Bozşahin (2018) uses a coding scheme for classifying affixes. We modify Bozşahin (2018) to cover all the examples we came across in our annotation of the derivational relations in Aslan et al. (2018). The code consists of 3 letters and a description. The first letter represents the lemma category (N, V, J, A), the second letter represents the stem category (N, V, J, A) and the third letter represents the type of operation (I, D). For derivational affixes the description is a representative allomorph, and for inflectional affixes it is the description of their function. Each affix has a unique code.

(29) Modified coding scheme

a. NNI_PLU: Inflection on a noun, producing a noun, making it plural

b. VVI_REFX: Inflection on a verb, producing a verb, making it reflexive

c. VJD_AL: Derivation on an adjective, producing a verb, using a suffix of the form AL

d. AAD_CAK: Derivation on an adverb, producing an adverb, using a suffix of the form CEK

e. XXD_LIK: Derivation on a stem of unspecified category, producing a lemma of unspecified category, using a suffix of the form LIK

Homophony is common among Turkish affixes. Emphasizing stem and lemma categories makes it easier to distinguish such affixes. An important example is the large family of affixes of the form *-CA*. Most of these affixes have their origins in the OT equative case *-CA* and they are still phonologically indistinguishable from each other. Most sources represent them as one affix, but they are categorially and semantically different. As a consequence, they must be represented by different entries in the lexicon. When we have to distinguish between affixes of the same form and category but different semantics, we distinguish them with a number after the description. If an affix was not listed in

Bozşahin (2018), we indicate this with a plus sign at the end of the code. (30) shows the wide variety of functions in which *-CA* is used. Most of these affixes (except NVD_CA+ and JND_CA1+) are productive.

(30) The eight types of *-CA*

a. NND_CA: *arapça* ‘Arabic’, *çince* ‘Chinese’, *katalanca* ‘Catalan’

b. NJD_CA: *kaplıca* ‘thermal spring’, *kokarca* ‘skunk’

c. NVD_CA+: *düşünce* ‘thought’, *güvence* ‘guarantee’, *izlence* ‘show’

d. JND_CA1+: *kesmece* ‘by cut’, *seçmece* ‘by choice’

e. JND_CA2+: *binlerce* ‘thousands’, *yüzlerce* ‘hundreds’

f. JJD_CA: *irice* ‘largish’, *serince* ‘coolish’

g. AND_CA1: *ailece* ‘as a family’, *arkadaşça* ‘as a friend’

h. AND_CA2: *boyca* ‘in terms of height’

i. AJD_CA: *acımasızca* ‘cruelly’, *hesapsızca* ‘rashly’, *kolayca* ‘easily’

j. AAD_CA+: *beraberce* ‘together’, *böylece* ‘in this way’, *evvelce* ‘previously’

Indicating categories also serves to emphasize the inflectional / derivational status of an affix. For instance, *birazdan* ‘soon’ was derived by the locative case *-DAn*, but assumed a completely different meaning through lexicalization. Annotating this derivational relation as NNI_LOC would both distort our data on NNI_LOC and lose an opportunity for examining cases where *-DAn* operates as a derivational affix. We mark such cases as AND_DAN+, provided that a minimum level of productiveness can be observed for this use.

Tables 4 and 5 show the distribution of affixes with respect to their function. Each person affix and each possessive affix is counted separately. Bozşahin (2018) does not list any zero affixes, but we include them for the sake of completeness.

Table 4: Syntactic categories of inflectional affixes

Source / Result	Noun	Verb
Noun	15	0
Verb	0	41

Nominal inflection only applies on nouns and noun phrases; if they appear on an adjective, we assume it to be first converted into a noun phrase. Adverbs cannot take inflectional affixes. Verbal inflection consists of voice markers, negation marker, tense-aspect-modality (TAM) markers, person markers and copula (*-DIr*).

Table 5: Syntactic categories of derivational affixes

Source / Result	Noun	Verb	Adjective	Adverb
Noun	9	6	10	6
Verb	27	6	20	2
Adjective	2	6	6	3
Adverb	0	0	0	1

As Bozşahin (2018) and Nikolaeva (2014) point out, some Turkish words appear in multiple categories with similar semantics. Whether these are cases of homonymy or conversion is open to debate. This is especially true for noun-adjective pairs. If one is to talk about adjective-to-noun conversion in Turkish, this may well be a fully productive process. An alternative view is to treat Turkish nouns and adjectives as substantives in the sense used by Chomsky (1993). As far as DM is concerned, it is usually possible to distinguish stem / lemma categories as noun or adjective. We classify affixes according to these categories whenever possible, and point out exceptions.

We generally follow the categories suggested in TDK (2019), which are mostly consistent. For instance, words indicating a person's hometown (*İstanbullu* 'Istanbulite') are always marked as nouns. Occasionally, there are inconsistent category assignments. For instance, words describing the followers of a religion or sect are sometimes marked as nouns, and sometimes as adjectives. It seems to us that words with a more widespread usage tended to be marked as nouns, while rarer words are marked as adjectives.

In a categorial grammar (which we build in Section 4.4), it is possible to represent this fluidity between the two categories with zero derivation / conversion. A similar state of affairs is in effect between English verbs and nouns. Nevertheless, the ambiguity disappears when the word is observed within a sentence; so, speakers should be able to assign the appropriate category to the accompanying affix. We chose the affix category according to the most prominent usage.

3.1.3 Orkhon Turkic

Lyons (1981) explains that in order to ensure cohesion among streams of research, modern linguists restricted themselves to theoretical synchronic microlinguistics. This is despite the fact that there are countless related research fields; or perhaps due to the very fact that there are so many related fields. There are convincing arguments for such a restriction to be in place. It was logical to be wary of overextension without a strong theoretical foundation concerning the core issues of linguistics. Failing to recognize this danger would only contribute to the innumerable complexities and confounding factors within linguistics.

From the viewpoint of cognitive science, additional arguments might be raised for the necessity of keeping one's focus on a synchronic investigation. Children acquire language based on a sample of contemporary data. The grammaticality of adult speech is judged based on the contemporary grammar. Therefore, one might say, it would be best for a cognitive scientist to solely focus on contemporary data; there is no benefit in looking into etymology, since the results cannot be applied in the framework of language acquisition or processing.

Researchers of syntax could be comfortable with these arguments. Syntactic rules remain largely unchanged throughout centuries of linguistic evolution, so there would be little to gain from studying cognate languages. They could exclusively focus on contemporary data; because syntactic rules operate above the word level and they are relatively easy to observe. No matter how wildly the form and semantic content of lexical items might change, the validity of the syntactic rules “autonomously” operating on them would not be harmed. (By the way, the idea of “autonomy of syntax” attributed to Chomsky (1977) is not uncontroversial at all. Ney (1982) and Anderson (2006) are among the studies that dispute it.)

This is not true for researchers of morphology. The rules governing morphology, which are often considered part of syntax, are also largely constant. However, morphology does not operate at the level of autonomy enjoyed by syntax. Morphological structures occur below word-level, so they are much more sensitive to phonological changes. As a result, morphological forms that were once clear syntactic constructions with clear categorial and semantic structure, may very quickly become unrecognizable.

This makes it more and more difficult for both speakers and researchers to disentangle the interactions between surface elements and identify the structures that produce them. Researchers that work exclusively on contemporary data, at the very least, risk having to spend a lot of time seeing through the many layers of phonological interactions. The mental structures assumed to be reflected on linguistic processing should be sought at the levels of morphology, syntax and semantics, not phonology.

Research on a parent language might help remove the layers of phonological changes that hide a more orderly, clearer picture of linguistic structure, facilitating research. In this respect, analysis of a parent language would not only be more fruitful, it would provide clues on the correct analyses of complex structures of its child languages. As grammatical rules evolve quite slowly, we believe these clues would also be applicable in the analysis of a contemporary language.

Our study of Turkish morpho-etymology started with these arguments in mind. We are certainly after a picture of Turkish morphology that is learnable by children, thus a “correct” analysis must be possible to acquire based on data from contemporary Turkish. However, contemporary data offers support to a large variety of hypotheses, most of which remain unexplored. We look into the grammar of Orkhon Turkic (OT) to gather insights and form new ideas on how morphological structures should be interpreted. We then apply these ideas to the analysis of contemporary data. New evidence both provide valuable support for some hypotheses, and often guide us into uncharted territory. Both outcomes contribute to our understanding of morphology and Turkish.

The regularity in OT grammar is surprising. Simpler and more regular alternative mechanisms to explain the structures of MT can be devised with this inspiration. We believe such mechanisms may be better candidates to match the deep interactions between mental concepts. They also seem to require fewer rules and exceptions for a computational analysis. We reiterate that our methodology respects the priority of synchronic description emphasized so vehemently in the field, but priority should not be confused with exclusivity. It is no reason for disregarding the benefits of an investigation of OT.

3.1.4 Main Sources on Orkhon Turkic

We have studied three main sources on OT: Tekin (1968), Erdal (2004) and Tekin (2016). Tekin (1968) is the first comprehensive grammar of OT and remained the only one for decades. The first edition of its translation into Turkish was published in 2000 by TDK; we have been using the fourth edition, Tekin (2016). The author scans inscriptions from Orkhon and Yenisey as well as some later writings such as Irk Bitig to construct a full grammar of Orkhon Turkic, listing rules and examples on phonology, morphology and syntax. Many of the source texts are given in Latin letters at the end of the book. Also, a word index and a dictionary are provided.

Erdal (2004) surveys more sources than Tekin (1968), including many Manicheist texts from the Uyghur period. The focus is not just on Orkhon Turkic, but on Old Turkic. The author never struggles to find perfect examples to demonstrate the wide variety of structures in the grammar, thanks to the richness of his sources. Erdal (2004) often puts forward interesting and compelling arguments regarding the nature of some morphological and syntactic rules. The book contains an index of Orkhon Turkic affixes and an index of grammatical terms.

A supplementary source is Aydın (2015), reporting extensive analyses of Yenisey inscriptions. Most Yenisey inscriptions are only a few lines long, and do not present a large vocabulary; nevertheless, there are several inscriptions offering a number of sentences of interest to us.

Literature on diachronic studies of Turkic languages provide plenty of ideas concerning the evolution of Turkish morphology. Some of these ideas are especially insightful such as Şçerbak (1989) and Kuznetsov (1997).

Şçerbak (1989) claims that, in Turkish, there are only two ways for morphological constructions to emerge: Either a noun joins into another noun, first as a particle, then as an affix; or a verb joins into another verb, first as an auxiliary verb, then as a morphological element. This is quite a strong claim. Even if it is to be rejected, discovering a small number of possible evolutionary paths to classify Turkish affixes could solve important puzzles. Indeed, from what we observed in the grammar of Orkhon Turkic, these two paths seem to constitute a suitable explanation for at least a large portion of Turkish morpho-etymology.

One convincing example in Şçerbak (1989) for this claim is the one linking the modern affix *-sI*, to the old Turkic verb *sı-*. The modern affix *-sI* forms denominal verbs that indicate an act of becoming similar to the root noun. To demonstrate the strength of the claim, we provide his comparative examination involving these morphemes across several Turkic languages.

(31) -sI vs. sı-

a. Orkhon Turkic: *sı-* ‘to resemble something’

b. Orkhon Turkic: *yagsı-* ‘to taste like oil’

c. Orkhon Turkic: *begsig* ‘like a beg’

d. Kazakh: *siyakti*, *sikildi* ‘like, similar to’

e. Kyrgyz: *algansı-* ‘to pose as somebody taking something’

f. Kyrgyz: *kün sımak* ‘like the sun’

g. Karachay-Balkar: *kızılsıman* ‘reddish’

h. Nogai: *avansı-* ‘to pose as ingenuous’

i. Tuvan: *kızılzımar* ‘reddish’

Another important idea expressed by Şçerbak (1989) is morphological constructions’ being more resistant to the influence of foreign languages than individual lexical items. The borrowing of morphological constructions can only be the result of persistent and strong influence by the source language. He gives the feminine noun suffix *-ka* as an example in Gagauz due to Slavic influence. Turkish has also influenced other languages. Masliyah (1996) lists four suffixes in Iraqi Arabic that originated in Turkish: *-lI*, *-lIk*, *-sIz* and *-çI*. The professional name forming suffix *-çI*, is exemplified in *kahrbijiun* ‘electrician’.

Kuznetsov (1997) is another study of Turkish morpho-etymology. He argues that most affixes in agglutinative languages resemble in form to single-syllable words. If languages evolve from concrete to abstract in terms of semantics and from simple to complex in terms of syntax, and if affixes are more abstract in meaning than roots, ancient languages must have been analytic. Taking this approach, he cites a long list of important works that seek the roots of many contemporary affixes in free morphemes, most of which are still in use themselves.

Kuznetsov (1997) claims that there is consensus among Western linguists that Turkish predicates are always based on nominals. In other words, finite verbs are replaced by gerunds and participles in Turkish. He claims that Turkish participles separated from the tense, aspect, modality base and went on developing separately. At this point, Kuznetsov (1997) presents a small survey on the morpho-etymology of *-DIK* and *-DI*, as a major area of the debate concerning participles. The proposal that *-DIK* is the original affix prevails, making *-DI* a contracted variation of it.

Alibekiroğlu (2019) also points out that some affixes in Turkish are the result of free morphemes gradually fusing into the stem. He claims that usually the original morpheme is lost during the evolution of such affixes, but several such morphemes and their affix versions continue to live on in contemporary Turkish. The examples and ideas given in this work proved useful in determining the correct semantics for quite a few affixes. İlhan and Öz (2019) is a similar work, providing a rich list of affixes and possible evolutionary roots.

We frequently needed to check the morphology of OT words. To this end, the dictionaries at the end of Tekin (1968) and Tekin (2016) were helpful, since the entries they listed were morphologically analyzed. Other times, we consulted with Nişanyan (2021) or Eyüboğlu (2017) to at least find an OT origin for a word.

We follow Tekin (2016) in the presentation of OT affix inventory but we add comments based on data from other sources, mainly Erdal (2004). Erdal (2004) is sometimes more broad in coverage, since it covers a longer time frame. In Tekin (2016), affixes are first divided into two, inflectional and derivational affixes. Inflectional affixes are divided into nominal and verbal inflectional affixes, while derivational affixes are divided into denominal nominals, denominal verbs, deverbal nominals and deverbal verbs. Tekin (2016) distinguishes between nouns, adjectives and adverbs, but only as comments on the entries of individual affixes. Participles and gerunds are separately listed, as they can be considered neither inflection nor derivation.

3.1.5 Inflection

Nominal inflectional affixes in MT can be grouped into four: plurality, possession, case and the relative marker. There are three kinds of nominal inflection in OT: Plurality, possession and case.

All nominal inflection takes phrasal scope. This is uncontroversial, as inflectional operations take place at the syntactic level. However, they are subject to the phonological rules operating below-word-level. In this way, all inflectional affixes we have reviewed fall into Type C (syntactic rules and morphological operations) in the Typology of Dowty (1979).

There is only one productive plural marker in Turkish: *-lAr*. It clearly takes phrasal scope over constructions. Suspended affixation (SA) is also possible.

Nizam (2017) is a survey of Turkic affixes marking plurals and collectives. Plural markers during OT period are listed as *-lAr*, *-t* and *-s* (not listed in Tekin (2016)). From the Uyghur period onwards, only *-lAr* remains as the productive plural marker. Nizam (2017) lists *-An*, *-AgU* and *-AgUn* as collective markers, but remarks that collective markers gradually fell out of use in favor of plural markers as speakers abandoned nomadic life. Nizam (2017) also cites a dozen studies considering *-z* an ancient plural / dual marker, possibly exemplified in *biz* 'we', *siz* 'you', *ikiz* 'twins', *göz* 'eye', *diz* 'knee', *Oğuz* 'Oghuz', *Gagavuz* 'Gagauz', *Kırgız* 'Kyrgyz' among others. Plurals were usually not marked until the Uyghur period, and there is still no perfect regularity in the use of the plural marker. Only one of these affixes, *-lAr*, remain in Modern Turkish.

The second class of nominal inflectional affixes is possessive markers. Possessive markers also have phrasal scope and SA, demonstrated in the following examples. Possessive markers in OT are extensively studied in Kürüm (2015). Differences with MT are only phonological.

(32) Possessive Markers in Orkhon Turkic

a. 1sg: *-(X)m*

b. 1pl: *-(X)mXz*

c. 2sg: *-(X)ŋ / -(X)g*

d. 2pl: *-(X)ŋXz*

e. 3sg/pl: *-(s)I*

Most grammars of MT list 5 cases: nominative, accusative, dative, locative, ablative. Bozşahin (2018) recognizes two additional case markers: genitive and instrumental. He also includes in the list, a second variation for each one of four cases (accusative, dative, locative, ablative). These variations are used when case is preceded by the 3rd person singular possessive marker. They bear the so-called pronominal N (Kürüm, 2015), the residual from an ancient personal pronoun for the third person. It is easy to demonstrate how case markers take phrasal scope.

There are 9 case markers in Orkhon Turkic.

(33) Case Markers in Orkhon Turkic

a. Nominative: *-∅*

b. Genitive: *-(n)Iŋ* / *-Ig* / *-In*

c. Accusative: *-(X)g* / *-(I)n* / *-nI*

d. Dative-Locative: *-kA* / *-ŋA* / *-A*

e. Locative-Ablative: *-dA* / *-tA* / *-tAn*

f. Directive: *-gArU* / *-ŋArU* / *-ArU* / *-rA*

g. Equative: *-çA*

h. Instrumental: *-(X)n*

i. Comitative: *-lIgU*

Sagidolda (2016) reviews some arguments and data concerning the declension system of Turkic languages. This study includes the coordination marker *-lI...-lI* in the list of case markers, but excludes *-ça* and *-lIgU*. Tekin (2013) explains his hypothesis on the etymology of the genitive marker in Turkish, citing several other hypotheses. Erickson (2002) does the same for the directive suffix, attributing its emergence to the combination of a dative-locative marker, the verb stem *är-* and gerund forming suffix *-U*.

The relative marker *-ki* is a controversial one. It behaves like an inflectional suffix, but it also behaves like a derivational suffix, changing the category of the base, usually into an adjective. It often comes after locative and genitive cases (Göksel and Kerslake, 2005).

Bozşahin (2018) lists 41 affixes employed in verbal inflection. Among these are voice markers, TAM markers, person markers and copular (*-DIr*). We leave person markers to Section 3.2.1. Göksel and Kerslake (2005) make a similar list, but also classify the affixes according to their position within the lemma. All affixes seem to have a fixed position relative to others. If we include voice, a total of 6 positions can be identified. Person markers follow all other affixes except in a few cases (Göksel and Kerslake, 2005).

Voice markers are, as expected, closest to the root. They are a unique class of affixes, because their very purpose is to alter the argument structure; therefore, they must apply before any arguments are fulfilled. This is why they cannot take phrasal scope. Although they are often considered inflectional affixes, perhaps voice markers should be considered among the true derivational affixes.

The possibility and negative markers are located on Position 1. Position 2 contains the compound verbs. Notice the vowels (/A/ and /I/) preceding the root of compound verbs. These vowels are not optional, because they are the deverbal adverb forming suffixes of the Orkhon Turkic, now squeezed between the two verbs. Position 3 holds all TAM markers. This position must always be filled in finite verbs. Only nominal sentences in the present tense may skip this position.

Table 6: Positions of verbal inflectional affixes given in Göksel and Kerslake (2005)

Position 0	Position 1	Position 2	Position 3	Position 4	Position 5
-DIr/...	*-(y)A*	*-(y)Abil*	*-DI*	*-(y)DI*	*-DIr*
-Il/-n/-In	*-mA*	*-(y)Iver*	*-mIş*	*-(y)mIş*	
-(I)n		*-(y)Agel*	*-sA*	*-(y)sA*	
-(I)ş		*-(y)Ayaz*	*-(A/I)r/-z*		
		-(y)Akal	*-(y)AcAK*		
		-(y)Adur	*-(I)yor*		
			-mAlI		
			-(y)A		

Copula marking past, evidentiality and conditionality take their place at position 4. The /y/ at the beginning of these markers is residual from *i-*, which is the contracted form of the Orkhon Turkic verb *är-*. In Modern Turkish, the auxiliary verb *i-* is also used occasionally. We explain in Section 3.2.3 the reasons why we are convinced that TAM markers change the category of their host. In sentences involving *i-*, there is no issue; *i-* carries out its task of reconverting its host into a V. However, when *i-* is missing, parsing the sentences becomes problematic. Perhaps using a zero-morpheme here is justifiable, because it reserves the position of an actual morpheme that is removed phonologically. Other uses of zero-morphemes are often based on paradigms that lack concrete evidence.

Finally, the generalized modality marker *-DIr* is the only possible occupier of Position 5.

Tekin (2016) lists two kinds of finite verbal forms in OT: primary and secondary. Primary forms can only function as predicates, while secondary forms are constructed from gerunds and participles. Moreover, there are only three verb forms that make primary forms: imperative, voluntative and perfective. While Tekin (2016) does list voluntative and imperative modalities separately, Tekin (1968) does not. These two modalities logically and perfectly complete each others' paradigms. We prefer the presentation in Tekin (1968) given below. Affixes in this group are quite recognizable, despite phonological changes.

(34) The Voluntative-Imperative

a. 1sg: *-(A)yIn* (Voluntative)

b. 1pl: *-(A)lIm* (Voluntative)

c. 2sg: -∅ (Imperative)

d. 2sg: *-gIl* (Imperative with emphasis)

e. 2pl: *-(X)ŋ* (Imperative with emphasis)

f. 2pl: *-(X)ŋlAr* (Imperative with emphasis)

g. 3sg/pl: *-zU(n)* (Imperative with emphasis)

Tekin (2016) also lists the paradigm of the perfective among the primary finite verbal forms, incorporating the person affixes. This choice may be due to the fact that these affixes are only used with the perfective, the verbal noun *-DOk* and the verbal noun *-sIk*, so they may not constitute a paradigm by themselves. We find it unnecessary to construct a whole paradigm for this modality, as the semantics of the perfective is obviously contained in a separate affix *-D* or *-DX*. Erdal (2004) also takes this position and describes the paradigm as *-d* followed by possessive suffixes, without mentioning why there is no 3rd person marker on finite verb forms, while a 3rd person possessive marker exists.

Tekin (2016) sees two patterns in secondary finite forms: verbal noun + possessive suffix, participle + personal pronoun. These are the known past tense (*-DOk*) and the future modality (also conveying necessity) (*-sIk*).

There are three points to notice: The perfective *-D* / *-DX* and the past tense *-DOk* are semantically very similar. They are also phonologically quite similar. Unlike most other finite forms (only other exception is *-sIk*) they both take affixes to determine person. There are quite a few studies proposing a genetic connection, including Tekin (1997), Nalbant (2002) and Koç (2012).

The affirmative present tense forms are all based on /r/. Also, considering the /r/-/z/ alternation, we find it unnecessary to list *-mAz* as a separate finite verb form and prefer it be considered as a combination of the negation marker and the present tense marker.

The voluntative-imperative is the only verb form that could not logically be made into a participle or gerund, since it expresses more of a speech-act than a modality. All other verb forms accept the separation of the morpheme determining tense-aspect-modality (TAM) and the morpheme determining person. The perfective (*-d* / *-dX*), the past tense (*-dOk*) and the future modality (*-sIk*) falls into the second group on account of their being accompanied by person affixes rather than personal pronouns. The rest falls into the third group, accompanied by personal pronouns.

While the future modalities of OT disappeared and MT adopted several new modalities, the connections between the two inventories are clear. The entirety of Position 2 in MT is residue of verb-verb compounding. Conditional and future markers in MT fall into the second and third groups, respectively.

3.1.6 Denominal Nominals

An important distinction is whether an affix falls into the Type C or Type D categories in Dowty (1979). To reiterate, Type C includes IM along with unrestricted and semantically regular DM; while Type D includes partially productive, semantically unpredictable DM, such as zero-derivation and compounding.

To understand whether an affix works on the syntax layer, we observe if it takes phrasal scope. The affixes that can, usually prove to be more productive, more regular and compositional in their meaning contribution.

Tekin (2016) lists 31 suffixes forming denominal nominals. We divide these suffixes into 8 groups considering their origins, functions and level of productivity, in order to emphasize the similarities between them. It will be easier for us to refer to these groups while examining the suffixes in Modern Turkish. Erdal (2004) claims that denominal nouns are derived with 8 purposes: fulfilling syntactic requirements, forming diminutives or denoting similarity, class-membership, collectivity, related functions and presence / absence.

(35) Lexicalized Case Markers

a. *-A*: *üzä*, *kiçä* ‘past’

b. *-gArU*: *yüggärü* > *yukarı* ‘upwards’, *ilgärü* > *ileri* ‘forwards’

c. *-kA*: *arka* ‘behind’

We believe listing case markers that are still in use as derivational affixes is a controversial move, but it serves to demonstrate the blurred boundaries between inflection and derivation. Lexicalization is a significant force that complicates our classification efforts time and again.

Four distinct suffixes were used to form diminutives. *-(X)ç* is somewhat uncontroversial, as it can still be found inside *-CIk*, its modern-day version. The evidence is too little for us to be certain about others. Erdal (2004) lists only two productive diminutive suffixes *-(X)ç* and *-kIñA* / *-kIyA*. He includes *-(I)çAk* in this group, but at a different status due to a lower level of productivity. Diminutives routinely take phrasal scope.

(36) Diminutives

a. *-(X)ç*: *ataç* ‘dad’

b. *-gAç*: *ıgaç* ‘tree’

c. *-kIñA*: *azkıñA* ‘a little bit’

d. *-mAn*: *ataman* ‘leader’

Erdal (2004) gives three suffixes forming color terms: *-gXl*, *-sIl* and *-Xş*, but refers us to Erdal (1991) for more details. These affixes do not present enough examples for any judgment on phrasal scope.

(37) Affixes Forming Color Terms

a. *-An*: *yägrän* ‘chestnut color’

b. *-Il*: *yaşıl* ‘green’, *kızıl* ‘red’

The following suffixes can be considered productive to some extent, with clear and constant meanings. Half of these form adverbs.

(38) Semi-Productive

a. *-dI* / *-tI*: *ädgüti* ‘properly’, *katıgtı* ‘properly’: Adverbs

b. *-dXn*: *birdin*, *yırdınta*: Adverbs of place and direction

c. *-rA*: *asra* ‘below’, *içrä* ‘inside’, *taşra* ‘outside’: Adverbs of place

d. *-kAn*: *tarkan* ‘title for a soldier’, *kadırkan* ‘Kadırgan mountains’: Only on names and titles

e. *-(X)nç*: *törtünç* ‘fourth’, *bişinç* ‘fifth’: Ordinal numbers

-dXn could be the original form of the ablative suffix *-dAn* emerging around the Uyghur period. *-rA* could be a form of the directive case, frozen on some items. Semi-productive affixes forming adverbs of place operate like case markers, and take phrasal scope.

There are several more suffixes that are among the most productive in OT. All of these are still in use in Modern Turkish without much change in form.

(39) Productive

a. *-çI*: *bädizçi* ‘artist’, *äbçi*: Profession

b. *-dAş*: *kadaş* ‘sibling’: Persons who are related through some entity

c. *-gI / -kI*: *içräki* ‘palace related’, *biryäki*: Possession, affiliation

d. *-gU*: *ädgü* ‘good’, *nägüdä*: Nouns designating qualities

e. *-lIg*: *atlıg* ‘horseman’, *ärklig* ‘powerful’, *kullug* ‘slave owner’: Ownership

f. *-lXk*: *bäglik* ‘worthy of princedom’, *özlük* ‘private’: Various meanings and functions

g. *-sIg*: *yılsıg* ‘prosperous’: Similarity

h. *-sIz / -sUz*: *aşsız* ‘hungry’, *buŋsız* ‘untroubled’, *kalısız* ‘complete’: Privative

Modern-day version of *-lIg* and *-sIg* have lost the /g/ in the end. Also, the vowel in *-sIg* has changed and it now forms denominal verbs, as in *yakınsa-* ‘converge’ and *ıraksa-* ‘diverge’. In Erdal (2004), *-sIg* is the only suffix expressing similarity. It is said to be productive. Erdal (2004) presents a few examples where some of these affixes take phrasal scope.

(40) Derivational affixes taking phrasal scope

a. *körümçi ulatı [tärs tätrü törö]çi* ‘diviners and other followers of wrong teachings’

b. *[bir iş]däş* ‘having a common cause’

c. *[bir yin]täm* ‘exclusively’

d. *[bir yaŋ]lıg* ‘uniform’

e. *[akar suv]luk* ‘a place with flowing water’

f. *[öŋi yer]sig ak-* ‘flowing as if at different places of a river’

g. *[tümän mıŋ tü]sig* ‘as if in 1000s of myriads of shapes’

h. *nomlarnıŋ [çın kertü töz]süzin ... bilirlär* ‘they know that dharmas are without any root’

i. *[bir ägsük]süz* ‘not one missing, completely’

Majority of the most productive denominal noun forming affixes, which are also still in use, are able to take phrasal scope. This is an important finding, providing evidence to the idea that derivational affixes have syntactic origins. Unfortunately, we cannot go back far enough in the evolution of Turkic languages to discover the free form origins claimed in Şçerbak (1989).

Erdal (2004) adds to the list of derivational affixes *-AgUt* forming status designations and *-(I)dUrXk* forming names of apparels worn on specific body parts such as *boyunduruk* ‘headlock, shackle’, *beldürük* ‘belt’ and *sakalduruk* ‘tie beneath the beard’. Since the deverbal nominal forming suffix *-uk* and the verb *dur* ‘stop’ exists, it is quite possible that the phrasal origins of X-(I)dUrXk is of the form ‘the apparel sitting on X’.

Erdal (2004) also examines intensification of adjectives and adverbs, giving examples to cliticized particle *(O)k* and *-rAk* as well as reduplication and superlatives. These affixes are rarely recognized in MT.

The class of MT affixes forming denominal nominals contains 11 items. It is represented in 2 subclasses in Bozşahin (2018): 9 denominal nominals and 2 deadjectival nominals.

Affixes forming denominal nouns

- NJD_CA
- NND_CA
- NND_CAGZ
- NND_CAK
- NND_CI
- NND_CIK
- NND_CIL
- NND_DAS
- NND_GEN
- NJD_LIK
- NND_LIK

NJD_CA and NND_CA originate from the OT equative suffix *-çA*. Since case markers operate on the syntactic level, they take phrasal scope. Erdal (2004) present many examples where *-çI* (NND_CI), *-dAş* (NND_DAS), *-lXk* (NJD_LIK, NND_LIK) taking phrasal scope. It is only natural for these affixes to also continue taking phrasal scope.

(41) Interesting examples of denominal nouns

a. *kıdemli uzmanca* ‘like a senior specialist’

b. *askeri tarihçi* ‘military historian’

c. *yapısal analizci* ‘structural analyst’

d. *kapkaççı* ‘snatcher’

e. *sıhhi tesisatçı* ‘plumber’

f. *tek adamcı* ‘supporter of a one-man management’

g. *ben bilmem beyim bilircilik* ‘following the motto “I do not know, but my husband does”’

h. *sen anlamazsıncılık* ‘following the motto “you wouldn’t understand”’

i. *ben bunu bilmemcilik* ‘following the motto “i don’t know about this”’

j. *adamsendeci* ‘being easygoing’

k. *burnu büyüklük* ‘think one is the bee’s knees’

l. *tanrı tanımazlık* ‘atheism’

m. *sonradan görmelik* ‘la-de-da’

n. *üç kelimelik* ‘three-word’

Some of these examples can be completely new for some speakers, but they are easy to understand with little ambiguity.

There are 16 derivational affixes that form adjectives: 10 denominal adjectives and 6 deadjectival adjectives.

Affixes forming denominal adjectives

- JJD_CA
- JND_CI
- JND_CIK
- JJD_CIL
- JJD_IMSI
- JND_INCI
- JND_IZ
- JND_LI
- JND_LIK
- JJD_MAN
- JJD_MSAR
- JND_MSI
- JJD_MTRAK
- JND_SAL
- JND_SER
- JND_SIZ

JJD_CA also seems to originate from the OT equative suffix *-çA*; so it would be expected to be able to take phrasal scope. All JJD affixes except one modify the intensity of the adjective, applying only on gradable adjectives. The exception is JJD_MSAR which derives *iyimser* ‘optimist’, *karamsar* ‘pessimist’ and the like.

(42) Interesting examples of denominal adjectives

a. *koyu yeşilimsi* ‘dark greenish’

b. *bilmem kaçıncı* ‘I don’t know which’

c. *Ankaragüçlü* ‘supporter of Ankaragücü’

d. *servi boylu* ‘tall like a cypress’

e. *yabancı kökenli* ‘of foreign origin’

f. *maymun iştahlı* ‘whimsical’

g. *Mart 95’li* ‘born in March 1995’

h. *turşu yapmalık* ‘suitable for pickling’

i. *sahibinden satılık* ‘for sale by owner’

j. *modern sanatımsı* ‘contemporary artish’

k. *sonunu düşünmeksizin* ‘without thinking to the end’

l. *çoluksuz çocuksuz* ‘without children’

There are 10 derivational affixes that form denominal adverbs: 6 denominal adverbs, 3 deadjectival adverbs and 1 deadverbial adverb.

Affixes forming denominal adverbs

- AJD_CA
- AND_CA
- AND_CAK
- AAD_CEK
- AJD_IN
- AND_IN
- AND_LA
- AND_LAYIN
- AJD_YA
- AND_YA

AJD_CA and AND_CA also originate from the OT equative case marker *-çA*. Gökdayı and Sebzecioğlu (2011) review the numerous types of *-CA* morphemes in Modern Turkish. AJD_IN and AND_IN probably originated from the OT instrumental case marker *-In*. AND_LA clearly originates in the modern instrumental case marker *ile*. AJD_YA and AND_YA may have originated in the modern dative case marker *-A*. Since case markers operate on the syntactic level, these would be expected to be able to take phrasal scope.

(43) Interesting examples of denominal adverbs

a. *iç güveysinden hallice* ‘mustn’t grumble’

b. *kendini adamışçasına* ‘as a devotee’

c. *geri dönmemecesine* ‘never to turn back’

There is a single affix that applies on adverbs, and it is not productive. Derivation from adverbs seems to be almost impossible in Turkish. Possibly, deriving from Turkish adverbs requires a too complex argument structure. Since most complex adverbs are derived by case markers, and case markers are the final affix that may append on a word, normally we would not expect them to be derived further.

3.1.7 Denominal Verbs

Tekin (2016) lists 13 denominal verb forming suffixes. We group these according to the regularity of their meaning, separating the unproductive ones from the main group.

Six denominal verb forming suffixes contribute a quite regular meaning to their base. *-(A)d* is quite similar to a bound version of ‘to be / to become’. *-(I)k* indicates a movement towards the thing denoted by the base. Verbs formed with the help of *-(I)k* are always intransitive. *-kA* forms intransitive verbs. *-rA* forms onomatopoeia based verbs.

(44) Regular Meaning

a. *-(A)d*: *başad-* ‘to lead’, *buŋad-* ‘to be troubled’, *kulad-* ‘to become a slave’

b. *-gAr*: *içgär-* ‘to subjugate’

c. *-(I)k*: *içik-* ‘to obey’, *taşık-* ‘to revolt’, *tagık-* ‘to climb a mountain’

d. *-kA*: *isirkä-*, *yarlıka-* ‘to grant’

e. *-rA*: *möŋrä-*, *yaŋra-*

f. *-sIrA*: *elsirä-* ‘to be without a country’, *kagansıra-* ‘to be without a king’

-sIrA contributes a strangely specific meaning; ‘to be / to become without’ the thing denoted by the base. It is easy to imagine *-sIrA* is related to *-sIz* ‘privative’ and some denominal verb forming affix, considering the /r/-/z/ alternation in OT.

-gAr complements *-(I)k*; while the latter forms intransitive verbs, the former forms transitive ones. It indicates a movement towards a direction denoted by the base, that is caused by the subject (agent) but carried out by the object (patient). Perhaps *-gAr* is the combination of *-(I)k* and some causative marker, such as *-Ur* or *-gUr*. Possibly the directive case marker *-gArU* is the combination of *-gAr* and *-U* (gerund).

There are also several suffixes that, unlike the ones above, do not contribute a constant meaning to the root. It seems that these suffixes only indicate that a denominal verb is being constructed, and the semantics is kept in the lexicon. These constitute one of the few classes of complex forms that resist compositional explanations for their semantics.

(45) Irregular Meaning

a. *-A*: *ata-* ‘to name’, *kürä-* ‘to disobey’, *bädzä-* ‘to paint’

b. *-I*: *biti-* ‘to write’, *tokı-* ‘to hit’, *yorı-* ‘to walk’

c. *-lA*: *başla-* ‘to start’, *illä-* ‘to establish a state’

d. *-U*: *yagut-* ‘to bring closer’

Erdal (2004) suggests that verbal derivation mostly occurs in the lexicon. We believe the reason for this claim is that the most productive suffixes of verb formation, given in the previous list, do not have a constant meaning; forcing the speakers to memorize the meaning of the resulting stems. On the other hand, there are plenty of examples where both the root and the affix have distinguishable meanings.

Finally, there are three more suffixes with few examples.

(46) Rare

a. *-(A)r*: *taŋlar-* > *tan ağar-*, *şafak sök-* (*taŋ* ‘tan’) ‘to dawn’

b. *-dI*: *udı-* > *uyu-* (*u* ‘uyku’) ‘to sleep’

c. *-(X)rkA*: *tokurka-* > *kendini tok say-* ‘to consider oneself satiated’

Erdal (2004) adds *-(A)r*, *-lAn*, *-kIr* and *-trI* to the list of suffixes forming denominal verbs.

Turning to MT, there are 12 derivational affixes that form denominal verbs: 6 denominal adverbs and 6 deadjectival adverbs.

Affixes forming denominal verbs

- VND_A
- VJD_AL
- VJD_AR
- VND_DA
- VND_ET
- VJD_IMSE
- VJD_LA
- VND_LA
- VND_LAN
- VJD_LAS
- VJD_SA
- VND_SA

VJD_SA and VND_SA seem to originate in the OT similarity suffix *-sIg*. Erdal (2004) presents many examples of *-sIg* taking phrasal scope, but we could find no such examples from MT.

VJD class affixes can be divided into two groups. Four affixes in the first group are used to indicate increases in the level of gradable adjectives. (Interestingly, none of them mark decreases.) The remaining two affixes indicate a change of the subject’s opinion of the object.

VND_LAN is composed of VND_LA with the reflexive voice marker.

(47) Interesting examples of denominal verbs

a. *koyu yeşilleşmek* ‘to become more dark green’

3.1.8 Deverbal Nominals

Erdal (2004) divides deverbal nominals into four groups. The first group denotes the subject when the base is intransitive and the object when the base is transitive or the action. The second group, *-çUk*, *-gUç* and *-gOk*, none of which are given in Tekin (2016), denotes instruments. The third group is composed of two suffixes *-(X)nçIg* (according to Erdal (2004) probably evolved from *-(X)nç-sIg*) and negative *-gUlXksXz* (probably also composite). The fourth group, the class of agentives, is described in more detail. Most of the suffixes listed under this group are also present in Tekin (2016); the ones missing are *-(X)nçU*, *-(U)t*, *-mA*, *-(X)z*.

Tekin (2016) lists 26 suffixes forming deverbal nominals. We classify these according to the characteristics of the resulting item. For instance, affixes that form the result of an act are given in a separate group.

(48) Closer to Noun

a. *-gUlUk*: *topulguluk* 'to pierce', *üzgülük* 'to break'

b. *-gUçI*: *ayguçı* 'spokesman', *itgüçi* 'builder'

c. *-mA*: *yälmä* 'discovery'

d. *-mAk*: *armakçı* 'fraud'

The first two affixes given in this list are clearly the combination of two simple affixes, that are mentioned among affixes making denominal nouns. Tekin (2016) claims *-gU* forms action nouns (infinitives) and *-çI* forms professional names. We believe *-lXk* has a semantics similar to *üçün* 'in order to', but it does not change the category of the stem. There have also been attempts at linking *-mA* and *-mAk* to each other through composition, but no consensus has been reached. Kuznetsov (1997) claims that the suffix *-mAk* is property of Oghuz dialects; it is either absent in other dialects or it is borrowed from the Oghuz.

(49) Closer to Adjective

a. *-(X)r* / *-(X)z*: *baz* 'subject'

b. *-sIk*: *açsık* 'becoming hungry', *tosık* 'becoming full'

c. *-DOk*: *bardok* 'where one arrives', *tägdök* 'where one attacks'

d. *-mIş*: *igidmiş* 'fed', *tägmiş* 'fought'

e. *-DAçI*: *kältäçi* 'will come'

f. *-gAn*: *korıgan* 'shelter'

g. *-(X)gmA*: *bitigmä*, sakınıgma

h. *-gA*: *bilgä* 'wise', *kısga* 'short', *tamga* 'seal'

i. *-(X)nçU*: *abınçu*, *inançu*

-(X)nçU may be the combination of *-(X)n* suffix denoting the result of an act, and *-çI* suffix forming professional names.

Tekin (2016) does not list *-r* as a derivational affix here, but he himself mentions *bar* as a lexicalized finite verb form constructed as *bar* from an ancient verb. Keeping in mind the /r/-/z/ alternation and this example, we believe *-r* must also be included in the list, along with its phonological sister *-z*.

(50) Closer to Adverb

a. *-A*: *ara*, *tapa*, *tägrä*, *yana*, *yämä*

b. *-p*: *kop*

There are quite a lot of ways to indicate the result of an act. Most of these suffixes are currently in use.

(51) Result of Act

a. *-(X)g*: *bilig* ‘knowledge’, *bitig* ‘writing’

b. *-gU*: *kürägü* ‘rebel’, *korıgu* ‘guard’

c. *-I*: *kalı* ‘deficit’, *takı* ‘as well’, *yazı* ‘lowland’

d. *-(U)k*: *bädük* ‘big’, *bulgak* ‘unclear’, *artuk* ‘surplus’

e. *-kun*: *buzkun*

f. *-(X)l*: *kısıl* ‘mountain pass’, *tükäl* ‘excellent’

g. *-(X)m*: *barım* ‘wealth’, *kädim* ‘dress’

h. *-mAn*: *tuman*

i. *-(X)n*: *bulun* ‘side’, *kälin* ‘bride’, *san* ‘number’

j. *-(X)nç*: *bulganç* ‘disorder’, *ärin* ‘undoubtedly’, *ötünç* ‘petition’

k. *-(X)ş*: *tägiş* ‘conflict’, *tokış* ‘to fight’

Turning to MT, the broad class of affixes forming deverbal nominals contains 49 affixes. It is represented in 3 subclasses in Bozşahin (2018): 27 deverbal nouns, 20 deverbal adjectives and 2 deverbal adverbs. The OT deverbal nominals we studied often serve multiple functions, such as forming a finite clause, forming a subordinate clause and deriving. Following the consensus among Western linguists presented in Kuznetsov (1997), we believe their main function is forming participles and gerunds.

Since the set of affixes falling into this group remained mostly the same, there is little reason to reanalyze it. Nevertheless, we include here the list of affixes forming deverbal nominals, for the sake of completeness.

Affixes forming deverbal nominals

- JVD_ACAK
- NVD_ACAK
- JVD_AGAN
- NVD_AK
- JVD_AK
- NVD_AMAK
- NVD_AN
- JVD_AN
- NVD_ANAK
- AVD_ARAK
- NVD_CA
- AVD_DIKCA
- NVD_GA
- NVD_GAC
- JVD_GAC
- JVD_GAN
- NVD_GAN
- NVD_GI
- NVD_GIC
- JVD_GIC
- JVD_GIN
- NVD_GIN
- JVD_I
- NVD_I
- JVD_ICI
- NVD_ICI
- JVD_IK
- NVD_IK
- JVD_ILI
- NVD_IM
- JVD_IN
- NVD_IN
- JVD_INC
- NVD_INC
- JVD_INTI
- NVD_INTI
- JVD_IR
- NVD_IT
- JVD_MA
- NVD_MA
- NVD_MACA
- JVD_MADIK
- NVD_MAK
- NVD_MAN
- JVD_MAZ
- NVD_MAZLIK
- JVD_MIS
- NVD_TI
- NVD_YIS

Notice how most affixes occur twice; once for the NVD category and once for the JVD category. Most affixes could be classified as deverbal substantives.

Grimshaw (1990) mentions the well-known dichotomy concerning result and process nominals, and puts forward another dichotomy between complex event nominals and others. According to her, the real difference between the two kinds of nominals is their having and lacking argument structures. In order to be able to construct meaning representations for such cases, we have to take into account the underlying argument structure, if one exists. We also take this position while devising the semantic representation of derivational processes.

All TAM markers and subordinate clause markers take phrasal scope. Because they head the phrase, they must wait until all arguments of the verb are fulfilled. This makes TAM markers agents of syntactic construction rather than derivation. Even highly lexicalized examples retain their argument structure, and suffixes continue to be productively applied on verbs almost without any semantic selection (s-selection). In other words, the existence of highly lexicalized words where the suffix seems to function as a derivational affix is not sufficient evidence to claim that it is indeed a derivational affix, especially when it still productively takes phrasal scope.

TAM markers include JVD_ACAK, NVD_ACAK, JVD_MAZ, JVD_MIS. These are capable of forming finite clauses, with the help of person markers, so they are obviously able to take phrasal scope. In this line of thinking, we must bear in mind Kuznetsov (1997) claim that Turkish predicates are invariably based on nominals. For now, we proceed with the conventional view on TAM markers.

Subordinate clause markers other than TAM markers include NVD_AN, JVD_AN, JVD_GAN, NVD-_GAN and JVD_MADIK. Like all syntactic constructions these are also capable of taking phrasal scope. JVD_MADIK is an interesting suffix; it is clearly composed of the negative marker and -DIK, but it still constructs relative clauses in the same format as in OT. It is not normally succeeded by a possessive / agreement marker.

Markers of action / manner nominals include JVD_MA, NVD_MA, NVD_MAK and NVD_YIS. These are all capable of applying on a verb whose arguments have been fulfilled, meaning they take phrasal scope.

JVD_AGAN, NVD_AN, JVD_AN, NVD_GA, JVD_GAN, NVD_GAN, JVD_GIN and NVD_GIN are likely phonological variations of a single *-gAn* suffix Nişanyan (2021). Also, NVD_GAC, JVD-_GAC, NVD_GIC, JVD_GIC, JVD_ICI and NVD_ICI may have originated from *-gUçI* (Nişanyan, 2021), which was in turn composed of the professional name forming *-çI* and deverbal nominal forming suffix *-gU*, still used in Modern Turkish in the form *-GI*. Perhaps they still carry out the same step-by-step operations during affixation.

(52) Interesting examples of deverbal nominals

a. *öngörü* ‘foresight’

b. *ilgi çekici* ‘interesting’

c. *uygulama geliştirici* ‘app developer’

d. *içe dönük* ‘introverted’

e. *eşgüdüm* ‘coordination’

f. *varsayım* ‘assumption’

g. *soğuk sıkım* ‘cold-pressed’

h. *ulusa sesleniş* ‘address to the nation’

i. *meydan okuma* ‘challenge’

j. *tepeden inme* ‘top down’

k. *kulaktan dolma* ‘hearsay’

l. *sonradan görme* ‘nouveau-riche’

m.*yekpare taştan yapılma* ‘built of solid rock’

n. *kuşkonmaz* ‘asparagus’

o. *kurşun geçirmez* ‘bullet-proof’

p. *rengi solmuş* ‘decolorized’

Most suffixes of the NVD class choose a thematic role of the verb and derive a noun for that role. Type of the resulting noun often depends on the arity of the verb. For instance, NVD_ACAK often forms nouns that denote patients, NVD_AK denotes locations, NVD_AMAK denotes patients, NVD_AN denotes agents, NVD_ANAK denotes agents for intransitives and theme for transitives, NVD_CA denotes names of acts for intransiitives and themes for transitives, NVD_GAC denotes agents, NVD_YIS denotes names of acts and a sense of manner. We further investigate this observation in Section 3.3.7. The remaining NVD suffixes form names of acts.

Deverbal adverbs may also be constructed from verbs whose arguments are fulfilled. Öner (2007) and Durmuş (2012) study the etymology of AVD_ARAK. Two important affixes forming deverbal adverbs are missing from the list, AVD_A and AVD_KEN, which are studied extensively in Yüceol Özezen (2008) and Özmen (2014). Finally, Yüceol Özezen (2018) makes an extensive review of gerunds from several Turkic languages.

The most important observation about deverbal nominals is that the argument structure of the verb is preserved during derivation. There is no way of removing the argument structure of the verb from the logical form (as in a CCG LF), one can at most fulfill the argument slots with skolem terms. Alternatively, if the affix has indeed taken phrasal scope, the arguments would have been fulfilled. If we have to distinguish between the different patterns of morphology, we would call the former pattern derivation, and the latter pattern syntactic construction. This is the basis on which we build a CCG grammar to represent Turkish morphology in Section 4.4.

3.1.9 Deverbal Verbs

Tekin (2016) lists 13 suffixes forming deverbal verbs. Most of these are voice markers, and more than half of voice markers are causatives.

(53) Voice markers

a. Causative: *-gUr / -Ur / -(X)t / -(X)z / -tUr / -tXz*

b. Passive: *-(X)l*

c. Reflexive-Passive: *-(X)n*

d. Reciprocal: *-(X)ş*

e. Middle: *-(X)r*

Except the middle voice, all voice types are present in Modern Turkish. Some causative markers dropped from use, but Göksel and Kerslake (2005) gives a long list of MT voice markers.

Three other suffixes conclude the presentation derivational suffixes in Orkhon Turkic.

(54) Others

a. Emphasis: *-d*: *ıd-*, *kod-* ‘to put’, *tod-* ‘to be satiated’

b. Frequency: *-lA*: *kunla-*

c. *-(X)k*: *kork-* ‘to fear’, *basık-* ‘to insert’

Erdal (2004) presents two more suffixes in this group, describing ‘types of inaction’, namely *-(X)gsA* and *-(X)msIn*.

In MT, there are 6 derivational affixes that form deverbal verbs.

Affixes forming deverbal verbs

- VVD_AKLA
- VVD_ALA
- VVD_DAR
- VVD_IKLA
- VVD_MAK
- VVD_USTUR

We believe classifying MAK as VVD is controversial. Its correct class should be NVD.

We could not find any cases where VVD affixes take phrasal scope. They leave the argument structure unchanged, but only modify the semantic content. Their application depends on the verbal class in Moens and Steedman (1988) given in Table 7.

Table 7: Event classification in Moens and Steedman (1988)

Class	Events		States
Consequence	Atomic	Extended	
+	Culmination: recognize, spot	Culm. Process: build a house	know, love
-	Point: hiccup, tap, wink	Process: run, swim, walk	

As the verb moves between classes, with the help of VVD affixes, it becomes qualified for markers and adverbs that are only compatible with the new class. Therefore, the purpose of the suffixes of this class might be to reshape the verb to fit certain TAM choices.

3.1.10 Syntactic Constructions

Syntactic constructions are typically the ones where the suffix takes phrasal scope over the base verb plus its arguments, and creates a nominal. Relative clauses are examples of this pattern. Syntactic constructions are generally considered to belong in the middle ground between inflection and derivation; because they change the syntactic category of the stem, but at the same time they are highly productive and their semantic contribution is highly regular.

Noun-Leaning OT Syntactic Constructions

- *-gUlUk*
- *-gUçI*
- *-mAkçI*

-gUlUk and *-gUçI* were also present among the affixes forming deverbal nominals in Section 3.1.8. *-mAkçI* was not in that list, but *-mAk* was; and it is highly plausible that *-mAkçI* is a complex affix composed of a deverbal nominal forming affix *-mAk*, and a denominal noun forming affix *-çI*. Although *-mAk* is one of the most frequently used affixes in its class in MT, there are very few examples of it in OT. This is probably because, as Kuznetsov (1997) suggests, *-mAk* was an affix of the Oghuz dialects, not in widespread use throughout the whole Turkic speaking population.

Adjective-Leaning OT Syntactic Constructions

- *-(X)r*
- *-mAz*
- *-sIk*
- *-DOk*
- *-mIş*
- *-DAçI*
- *-gAn*
- *-(X)glI*
- *-(X)gmA*

All affixes in the list above were included in deverbal nominals forming participles in Section 3.1.8, except *-(X)glI*. All affixes in Section 3.1.8 are also present in this one, except *-gA* and *-(X)nçU*. It is easy to imagine the two sets of suffixes are actually identical, but due to a lack of data demonstrating every affix in every function, the sets could not be completed.

There are quite a few affixes proposed by Tekin (2016) that form deverbal adverbs. If we focus on the simple affixes in the list, namely *-(y)X*, *-(X)p*, *-sAr*, *-kAn* and *-çA*, we observe that they are quite productive and still in use in MT.

Adverb-Leaning OT Syntactic Constructions

- *-A / -I / -U*
- *-yU*
- *-(X)p*
- *-sAr*
- *-kAn*
- *-çA*
- *-(X)pAn*
- *-(X)yIn*
- *-mAtI(n)*
- *-gAlI*
- *-gInçA*
- *-(X)glI*

According to Kuznetsov (1997), it is usually possible to find the first constituent of a complex verb appended with one of *-(y)a*, *-(y)e*, *-(y)ı*, *-(y)ıp*, with a few exceptions. These are often classified under TAM in grammars of MT (Göksel and Kerslake, 2005).

(55) Complex Verbs

a. *bekleyebilmek* 'to be able to wait'

b. *bakakalmak* 'to stand in wonder'

c. *gidivermek* 'to dash down'

d. *sorup durmak* 'to keep asking'

e. *takılıp kalmak* ‘to dwell on’

3.2 Issues

The previous section presented an overview of Turkish DM. This exercise clearly demonstrated how messy DM can be. On the other hand, there are many interesting cases that, with the help of a suitable method, just might hide previously unnoticed structure. This kind of hidden structure is what we are after.

The blurred boundaries of DM forces one to evaluate many related linguistic phenomena, before deciding to leave them outside the scope of a study. In this section, we examine the evidence regarding certain groups of affixes.

3.2.1 Person Markers

In Section 3.1.5, we presented the OT finite verb forms. This classification of finite verb forms demonstrates a hugely important fact: For the majority of finite verb forms, there is no person affix. They are instead accompanied by personal pronouns. There are only three exceptions: the voluntative-imperative, the perfective-past couple, and the future *-sIk*. Moreover, all nominal sentences are constructed with personal pronouns instead of person markers. Therefore, the proportion of sentences with personal pronouns is much higher than that of sentences with person markers. Following Şçerbak (1989), it is plausible to imagine that OT constitutes an evolutionary stage where personal pronouns are gradually replaced by person markers. Perhaps some time before the Orkhon period, older Turkic languages would have no person affixes.

In the grammars of MT, pronouns of the 1st and 2nd persons do have a person affix, while the 3rd persons’ affix counterparts are allegedly zero morphemes. OT lacks a personal pronoun for the 3rd person (Tekin, 2016); the demonstrative pronoun *ol* ‘this’ is used to fill this gap. This is a significant piece of evidence, that could help us answer why a 3rd person marker is missing from the person agreement paradigm in MT.

Many studies argue that Turkish person affixes indeed stem from personal pronouns and propose similar schemes to explain how it would be possible Buran (1996), Kuznetsov (1997), Demirci (2008), Yavuzarslan (2011), Başdaş (2014), Kürüm (2015), İlhan and Öz (2019), Ünal (2019), Alibekiroğlu (2019), Güven (2021). Kürüm (2015) also makes a survey of studies on the etymology of personal pronouns. Yavuzarslan (2011) provides tables showing the step by step evolution of Turkic person markers, including paradigms from the Orkhon, Uyghur, Karakhan, Khwarezm, Cuman, Seljuq and Ottoman periods.

Looking into phonology, we notice that the first phonemes of the 1st person pronouns are both /b/; similarly for 2nd person pronouns, both of which start with /s/. *bän* ‘I’ and *män* ‘I’ are used interchangeably; notice the similarity between /b/ and /m/. Perhaps the original 1st person pronoun was *män* which slowly turned into *bän* in some dialects. The same could be true for *biz* ‘we’. Also, both plural personal pronouns end in *-iz*. Nizam (2017) cites 16 studies that consider the possibility of a *-z* affix that contributes duality or plurality to the host.

We observe several intermediate stages from personal pronouns to person markers. Personal pronouns accompany not only finite verbs, but also nominals and subordinate clauses in OT. Demirci (2008) lists many examples as evidence that person markers evolved from personal pronouns.

(56) Personal Pronouns Turning into Person Markers

a. *bän türk bän* > *Ben Türküm* 'I am a Turk'

b. *edgü erür men* > *İyiyim* 'I am fine'

c. *edgü turur* > *İyidir* 'It is fine'

d. *özüm karı boltum* > *Ben yaşlandım* 'I grew old'

e. *biz eki bıŋ ärtimiz* > *Biz iki bin idik* 'We were two thousand'

We observe overt pronouns at the end of OT sentences (including the 3rd person pronoun *ol* in other examples). Most frequent TAM affixes *-dOk* and *-sIk* are likely to have fused with most frequent personal pronouns 1st and 2nd persons by the Orkhon period. Others continued to be applied individually. This observation brings about the possibility that the person marker does not just agree with the subject, but it is the proper subject of the sentence.

Öztürk (2001) makes the courageous claim that Turkish is not a pro-drop language. She looks into virtually all cases where an overt pronon may occur; such as matrix clauses, adjunct clauses, exceptionally case-marked constructions, genitive phrases and relative clauses. Öztürk (2001) agrees with Enç (1986) and Taylan (1986) that Turkish overt pronouns are not redundant, nor is their use optional. The central claim in Öztürk (2001) is that Turkish overt pronouns may be better analyzed as pragmatically conditioned topic pronouns, rather than subject pronouns. Recognizing that this view is in contradiction with the conventional verb-final analysis of Turkish, Öztürk (2001) also attempts to resolve the conflict using a grammar model where agreement morphology is cliticized to the inflected verb after vocabulary insertion at the Morphological Structure.

Good and Alan (2005) look into the morphosyntactic behavior of Turkish agreement markers, using diachronic facts. Out of the four suffixal paradigms of the subject pronominal IM, they consider only two: the k-paradigm which applies after *-(y)DI* or *-(y)sE* and the z-paradigm which applies on all other predicates except the optative and the imperative. Borrowing examples from Kaşgarlı (2005), Good and Alan (2005) argue that the past marker *-(y)DI* is due to the reanalysis of the nominalizer *-DIK* combined with possessive markers. We believe this may not be the case, since the *-DI* form can already be observed in the Orkhon inscriptions. Moreover, at that time, it could be followed by affixes from the z-paradigm. Pointing out that cliticized forms of pronominal subjects appear in the thirteenth century, they also seem to be unaware that OT made extensive use of these cliticized forms, which is called the z-paradigm.

Erdal (2000) argues for a clitic reading of the z-paradigm person markers. He presents criteria for identifying clitics and applies these criteria to several Turkish morphemes. In addition to person markers, he demonstrates that copular markers and several other morphemes are clitics. We believe that treating the auxiliary as an autonomous element results in a simpler grammar without losing any expressive power, but we formed this opinion only after carefully studying Erdal (2000) and Kornfilt (1996). A similar distinction is drawn by Miller (1992). Arguing that suspended affixation is a reliable test for identifying clitics, Miller (1992) demonstrates the difference between the two paradigms. Baker (2013) draws examples from Sakha and demonstrates that sometimes the apparent agreement may not be with the predicate itself, but with a tense item that cliticizes to the predicate. We believe this is also the case in Turkish, as nominal predicates require the usually hidden auxiliary verb.

Person markers are part of morphology, but their status is clearly controversial. The studies reviewed in this section establish that as a fact. On the one hand, we believe person markers are indeed vp-internal subjects (Öztürk, 2001) and they are clitics (Erdal, 2000); on the other hand, person markers definitely fall outside the scope of DM: Their semantics are perfectly regular and they are fully productive.

3.2.2 Possessive Markers

MT possessive markers are quite similar to person markers. The only exception is the 3sg/pl possessive marker, and it is semantically different also. As mentioned in Section 3.1.5, Tekin (2016) and Erdal (2004) describe paradigms for two primary finite verb forms, the perfect (*-DOk*) / the past tense (*-dX*) and the future (*-sIk*) as the combination of verbal nouns and possessive suffixes. Neither explains why the 3rd person possessive marker is overt, but the 3rd person marker on finite verb forms is not. On one hand, there is little functional similarity between person marking on finite verbs and possessive marking on noun phrases; these two paradigms must be considered separately. On the other hand, person markers and possessive markers are undeniably similar in form and semantics.

Şçerbak (1989) claims that ancient constructions of the form *men-kişi-men* 'I am a person' and *biz-kişi-biz* 'We are people' served to create connection between the constituents based on the repetition of personal pronouns. Over time, main phrases and possessive constructions evolved from this parent construction. During this separation, the genitive suffix emerged. Kürüm (2015) claims that the consensus on the etymology of possessive markers is their having evolved from personal pronouns. He cites several studies on the etymology of possessive markers and personal pronouns, reviews possessive markers in the disputed Altaic family and the mysterious pronominal N. The most controversial part in the debate seems to be the etymology of the 3rd person suffixes and the pronominal N. In order to explain this most problematic part of the paradigm, several studies refer to a proto-Altaic 3rd person pronoun, which was lost in Turkic and Mongolic languages but remained in Manchu. In Turkic languages, it is claimed, this pronoun could only leave its mark as the pronominal N. To aid in this investigation, Kürüm (2015) provides possessive paradigms from 27 Turkic languages.

Like person markers, possessive markers might have been evolved from personal pronouns, gradually fusing into the NP in the context of possessive constructions. Perhaps Proto-Turkic did not have any suffixes for person marking (It is quite probable it did not have any suffixes, at all) and speakers always expressed person using pronouns or names. Unlike person markers, which always indicate the subject of the sentence, possessive construction describes part-whole relation and otherwise affiliation, as well as possession. This might have been the reason why the two paradigms diverged. As an alternative hypothesis, we could imagine *-sI(n)* being a separate marker indicating relation between non-human entities (so, never 1st or 2nd person), which later merged with the main possessive paradigm.

Öztürk and Taylan (2016) make an extensive analysis of Turkish possessive constructions. They claim that there are three kinds of possessive constructions in Turkish and propose a syntactic architecture to explain the different structures. Based on the fact that Turkish verbs can take NP or PP arguments, while NPs normally cannot; they argue that it is POSS that makes it possible for NPs to take arguments. In this way, NPs can take other NPs or sentential complements as arguments.

Whether they mark valency or person or both, possessive markers' status as a distinct paradigm in MT is not controversial. Its being syntactic in nature is also evident due to its phrasal scope, full productiveness and semantic regularity. Only the discussion on the 3rd person possessive, regarding its etymology and range of functions is ongoing. Following Kürüm (2015) and Göksel and Kerslake (2005), we believe the 3rd person marker is etymologically and functionally different from the others. We consider it part of the compounding domain and exclude it in our study of DM.

3.2.3 TAM Markers

Korkmaz (1959) studies the morpho-etymology of the future tense marker, *-AcAk*. She reviews examples from many dialects and concludes that the suffix is composed of *-a* and *-çak*, semantics of both involving future. An alternative morpho-etymology could be *-tAçI +ok*, similar to the proposal of Nalbant (2002) for *-DIk*. *-tAçI* already covers the semantics of future Tekin (2016) and the addition of *ok* only serves to emphasize that the event will certainly take place (Nalbant, 2002). If a future event's taking place is uncertain, the present tense is used instead of the future tense. This might indicate that *-AcAK* has a specialized meaning involving certainty.

Kuznetsov (1997) believes that an agglutination-based account can be found for quite a few Turkish affixes, making proposals for 25 affixes. 9 of these proposals concern affixes reviewed in this subsection, namely voice markers and some TAM markers.To give one example, Kuznetsov (1997) claims that *-DIk* originated from the verb *tük-*, 'to finish', 'to end'. Analyzing phonological evidence from cognate languages, he also reaches the conclusion that *-DI* is the contracted form of *-DIk*. Therefore, the evolution of past tense would follow these steps: *kel tük män* > *keltük män* > *keltüküm* > *keltüm* > *geldim* 'I have come'. To reiterate, Kuznetsov (1997) explains that the accepted opinion among Western linguists is that Turkish uses participles and nominals instead of finite verbs in matrix clauses.

Tekin (1997) reviews several prior work on the possible connection between *-dOk* and the perfect marker. After examining evidence, he concludes that the most plausible scenario is *-dOk* being constructed out of the verbal nouns in **-(I)d* and the intensifying particle *ok/ök*. This combination is indeed frequent in OT texts. The particle can even apply twice on the same stem. We can add that the known / seen meaning involved in the past tense may be attributed to the intensifying marker.

Nalbant (2002) reviews the examples in Divanü Lugati't Türk with a focus on the uses of *-DUK*. He concludes that this suffix belonged to Oghuz dialects; it was used in all Oghuz dialects; it is composed of a past tense marker *-DI* and the intensification particle *ok*; and starting in the Khwarezm period, it replaced the 1st person plural marked form of *-DI*. Koç (2012) makes a similar analysis studying Old Anatolian Turkish.

There is indeed good evidence that *-DIK* and *-DI* in modern Turkish share the same roots and semantics. Their distributions are complementary: *-DIK* is exclusively used in gerunds and object RC; while *-DI* is used in direct complements and matrix clauses. Separating their distributions in this way might have served to disambiguate the kind of clause. Their difference in form and distribution is not sufficient evidence to deny them the same category and semantics.

Kornfilt (1996) analyzes the copular clitic forms in Turkish, and recognizes that the definite past and the conditional are the only "genuine verbal finite forms". Other tenses are merely combinations of participles and inflected copula sequences. Results from her synchronic analyses coincide with our diachronic investigations. Tekin (2016), who admits that the secondary finite verb forms are based on participles and gerunds, still classifies them as finite verbs. Erdal (2004) takes a similar position. We believe this position is untenable.

(57) Evolution of copular use

a. *ayıgması ben ärtim* 'I was his speaker.'

b. *Sözcüsü ben idim.* 'I was his speaker.'

c. *Sözcüsü bendim.* 'I was his speaker.'

d. *il tutsık yir ötükän yış ärmiş* 'The place to establish a state was evidently Ötüken.'

e. *İl tutacak yer Ötüken imiş.* 'The place to establish a state was evidently Ötüken.'

f. *Devlet kurulacak yer Ötükenmiş.* 'The place to establish a state was evidently Ötüken.'

g. *ilteriş kagan yok ärti ärsär türk bodun yok ärtäçi ärti* 'If there was no İlteriş Kagan, there would be no Turk people.'

h. *İlteriş kagan yok olsa idi Türk halkı yok olacak idi.* 'If there was no İlteriş Kagan, there would be no Turk people.'

i. *İlteriş kagan olmasaydı Türk halkı olmayacaktı.* 'If there was no İlteriş Kagan, there would be no Turk people.'

The auxiliary verb *är-* first contracts and becomes *i-*, and is later removed completely from the picture. Its presence can still be observed in the /y/ that goes between the final vowel of the stem and the initial consonant of the second TAM marker. This phoneme would otherwise be unnecessary.

In the light of arguments and examples from Kornfilt (1996) and Erdal (2004) among others, we describe the nature of the auxiliary in several claims.

(58) Understanding the auxiliary

a. The auxiliary is a defective verb. It may only occur in the present, perfective, evidential and conditional forms or as an adverbial with *-ken*. It is unmarked in the present tense. (Present tense in V-4 is probably the fossilized auxiliary *er-*.)

b. It may occur both in free form (*i-*) and bound form (*-y-*).

c. It always follows nominal predicates and participles (z-paradigm).

d. It follows TAM markers from the k-paradigm to form complex tense.

e. When the auxiliary is present, person markers from the z-paradigm are cliticized on it. The adverbial form does not accept person markers.

This description simplifies and resolves many conflicts found in the literature. Based on these generalizations, it is much easier to give a categorial account of verbal morphology. The auxiliary is a clitic, acting as the platform for additional tense information.

We believe that the TAM markers in MT subcategorize for V and produce a PredP. The first piece of evidence is that OT requires the auxiliary verb between two consecutive applications of TAM markers. MT still respects this rule, although the actual morphemes taking part in the operations may not be overt. For instance, MT allows constructing complex aspectual structures like *gelmiştiyse* 'if he had come' without overt auxiliary, but OT strictly follows the intended categorial application rules and use *är-* whenever a non-verbal category must be reconverted to a verbal category. Second, nominals cannot directly take TAM markers. Since TAM markers subcategorize for a V, an auxiliary verb is first required to construct a V out of the nominal. This rule is also still valid in MT. Yener (2018) provides plenty of examples validating these points. Third, we observe that for each TAM marker a derivational marker exists with the same form and semantics. These markers also subcategorize for V and produce predicates. We suggest that the two classes of markers have the same origin and the derivational markers were frozen in lexicalized constructions of TAM markers.

(59) Evolution of derivational suffixes out of TAM markers

a. *Bu bitkiye kuş konmaz.* 'Birds do not land on this plant.'

b. *Kuş konmaz bitki sağlığa faydalı.* 'The plant that birds do not land on is good for health.'

c. *Kuşkonmaz bitkisi sağlığa faydalı.* 'The kuşkonmaz plant is good for health.'

It seems the main, and probably the original, task assigned to these affixes is the formation of deverbal nominals. Their taking phrasal scope is critical here, as our categorial approach to finite verbs in this section demonstrate. If we summarize the idea here: Verbs bring argument structure and aspectual structure to the phrase. Deverbal affixes may be thought to subcategorize for verbs, but once they form nominals, they lose their argument structure and become categorially inert to argument fulfillment. To avoid this dead end, all deverbal affixes must subcategorize for verb phrases and apply only after all necessary arguments have been fulfilled.

Deverbal nominals formed in this way sometimes lexicalize, raising the status of their affixes from inflectional to derivational. Since the lexicalization process is continuous and arbitrarily long, it is not logical to try and separate inflectional TAM from derivational ones. Therefore, we include TAM markers in our analyses of DM.

All TAM markers in OT are candidates for constructing RC. Three additional markers, *-gAn*, *-(X)glI* and *-(I)gmA* can also take part in RC. It is possible that *-(I)gmA* and *-(X)glI* are complex forms, made out of *-(X)g*, *-mA* and *-lI*. *-(X)g* and *-mA* are already listed among verbal derivational affixes. *-lI* could be considered a contraction of the *-lIg* nominal derivational affix. Since *-(X)g* forms nominals and *-lIg* subcategorizes for nominals, the two affixes are compatible. However, since *-mA* subcategorizes for verbs, we cannot give a clean explanation on the composition of *-(I)gmA*. *-gAn* and the deverbal nominal forming suffix *-gA* probably had the same origin. Both denote the agent of an act. All these suffixes clearly take phrasal scope.

It is harder to define general rules for RC in MT. Only three affixes are usually recognized for forming RC, *-An*, *-DIK* and *-AcAK*. There are other TAM markers that are capable of clauses resembling subject relative clauses (SRC), but not object relative clauses (ORC).

We observe that the templates for RC in OT are generally quite similar to MT. One difference is due to heads of NOM and ORC in MT taking a suffix that resembles GEN. George and Kornfilt (1981) argues that this is a piece of agreement, and contributes finiteness to the clause. Kornfilt (1997) investigates why *-DIK* requires an agreement marker, while *-An* does not. She claims that it is the presence of the agreement marker that determines the choice of the particle.

Based on the results in Section 3.2.1 and in this section, we propose a simpler explanation. What we call agreement at the end of matrix clauses, as well as RC, is actually the subject. If this is the case, SRC cannot take a person marker (aka the subject), because the very purpose of the SRC is to modify the following NP, functioning as the subject of the clause. To avoid having two subjects, we cannot use person markers at the end of SRCs. This is clearly not the case for ORC and NOM.

There are plenty of controversies regarding subordinate clauses. We consider person markers at the end of RC and NOM fulfill the same role as person markers of the matrix clause.

3.2.4 Suspended Affixation

Suspended affixation (SA) offers strong evidence distinguishing inflectional processes from derivational ones, but it is often overlooked. True derived forms are not expected to share DM with coordinated items, as the DM is assumed to make a substantial and possibly irregular change in the host's semantics. Therefore, it should be possible to use SA as strong evidence for syntactic function of an affix, while the lack of it is only weak evidence for morphological function. We reviewed several studies to collect evidence for this claim.

Kornfilt (1996) argues that SA in Turkish verbal inflection is simply the coordination of participles. After the coordination is constructed, copula and the remaining IM cliticize to the rightmost item in the coordination. Basically, Kornfilt (1996) predicts that verbal predicates will not permit SA if the TAM slot is occupied by *-DI* or *-sA*, which are true tenses. Otherwise, the predicate base is a participle, followed by a cliticized copula, permitting SA at the clitic boundary. In our opinion, this is exactly the case.

Kabak (2007) borrows many ideas from Kornfilt (1996). He asserts that it is never possible for DM to be suspended. He emphasizes the notion of morphological word in order to explain why some affixes may be suspended, while others may not. In general terms, SA is permitted when the left-conjunct is a complete form, in other words a morphological word.

(60) SA with inflection on predicates

a. *Zengin ve ünlüydüm.* 'I was rich and famous.'

b. *Zengindim ve ünlüydüm.* 'I was rich and I was famous.'

c. *Gider, görür ve alırız.* ‘We will go, see and take.’

d. **Çalıştı ve başardık.* ‘We worked and succeeded.’

e. **Çalış ve başarırız.* ‘We (will) work and succeed.’

f. *Çalışır ve başarırız.* ‘We (will) work and succeed.’

g. *Çalışır ve başarırdık.* ‘We would work and succeed.’

Regarding DM, Kabak (2007) also observes that DM can never take part in SA. He demonstrates this on *-CI*, *-sAl* and *-lIK*.

(61) SA on derived forms

a. **fayans ve bacacı geldi* ‘tiler and chimney-man came.’

b. **ruh ve toplumsal açıdan* ‘spiritually and socially’

c. **güzel ve sadelik konusu* ‘regarding beauty and simplicity’

Akkuş (2016) specifically looks into SA possibilities for Turkish DM. He states that such occurrences are uncommon, but they are also too many to ignore. He remarks that many coordinated nouns that look like cases of SA, are actually co-compounds or cases of natural coordination. Since their constituents behave as a single unit, we shall not consider them as SA.

(62) Examples of co-compounds and natural coordination by Akkuş (2016)

a. *ana (ve) babalık* ‘mother (and) fatherhood’

b. *ay-yıldızlı bayrak* ‘crescent-star-bearing flag’

c. *sarı-kırmızılı takım* ‘yellow-red-wearing team’

Bozşahin (2007) gives the following example:

(63) SA example by Bozşahin (2007)

a. *tuz ve limonluk*

Under the non-SA reading, this construction can be translated as ‘salt and lemon squeezer’. With an SA reading, it would mean ‘salt shaker and lemon squeezer’. If the latter reading is valid, it would mean that a DM with two non-compositional or at least polysemous alternative interpretations is capable of distributing appropriate interpretations over a phrasal scope to multiple constituents. This is quite difficult to conceive. Kornfilt (2012) also discusses this example. She argues that changing the order of conjuncts eliminates the SA reading:

(64) SA example by Kornfilt (2012)

a. *limon ve tuzluk*

Based on a review of Japanese causative suffix, Akkuş (2016) joins Fukushima (2015) in questioning the adequacy of a lexical account for explaining DM. So far, the Japanese causative marker remains the sole well-researched example of SA in DM. Not having seen reliable examples for the existence of SA with MT DM, we conclude that it is not possible for a DM to be suspended, despite some DM sometimes taking phrasal scope.

3.2.5 Emphatic Reduplication

The kinds of duplication used in Turkish are emphatic left-reduplication (i.e. *sapsarı*), generic plural (m-reduplication) (i.e. *tabak mabak*) and adverbial doubling (i.e. *zaman zaman*) (Göksel and Kerslake, 2005). We leave the generic plural and the adverbial doubling aside, due to their little involvement with morphology. Perhaps Turkish emphatic reduplication (TER) deserves a special mention. TER always indicates a change in meaning or intensity of meaning, is never required by syntax and it cannot take phrasal scope. Thus, TER squarely falls into DM.

Wedel (1999) and Kılıç and Bozşahin (2013) present important findings on TER; the former looking into the phonological aspects of the phenomenon, while the latter prefers a data-driven approach. Wedel (1999) claims that reduplicative forms used for emphasizing adjectives are somewhat productive. Working within the framework of Optimality Theory, he shows that the schema for creating new forms is available to Turkish speakers.

Kılıç and Bozşahin (2013), on the other hand, show that in addition to the phonological constraints, speakers show two preferences in their linker type selection. First, they try to dissimilate the linker type from frequent consonant co-occurrences and word endings. Second, they prefer the linker type to be dissimilar from the existing root in their choice of linker type. Therefore, they claim, TER is a morpholexical process that depends on "a global lexical knowledge for selecting an appropriate linker". We exclude TER from further analyses, due to its narrow domain and morpholexical nature.

3.2.6 Position Classes

Position classes are slots around the root for bound morphemes to occupy. In terms of IM, each of these slots correspond to a grammatical feature. We briefly look into position classes in order to assess whether DM may (even if rarely) exhibit position classes.

Stump (1993) reviews four approaches to describe position classes. Below, we summarize the components of each approach:

(65) Four approaches to describe position classes

a. The subcategorization approach: Rewrite rules (organized in multiple layers) + Lexicon + Subcategorization restrictions

b. The pure PStr approach: A single rewrite rule (flat) + Lexicon

c. The linear ordering approach: Linear ordered rule blocks corresponding to each position class

d. The paradigm function approach: Morpholexical rules for each position class

Our position is closest to the first approach. Following categorial grammar, we assume layers of rewrite operations restricted by subcategorization rules.

Stump (1993) also presents the traditional assumptions about members of the same position class:

(66) Traditional assumptions about members of the same position class

a. They are in complementary distribution.

b. They share the same hierarchical relationship to other affixes.

c. They share the same position with respect to other affixes.

d. They offer alternative realizations of the same morphosyntactic feature.

There are certainly cases where derivational affixes are in a complementary distribution. Therefore, the first assumption is often valid. However, many DM can occur multiple times on the same base, within varying hierarchies and positions. This violates the second and third assumptions. DM cannot be considered realizations of the same morphsyntactic feature; in fact, DM contribute an often irregular and abstract meaning to the base. This violates the fourth assumption. All in all, position classes are not applicable for DM.

3.3 Observations

In order for an inventory of DM to be accurate or useful, several notions must be taken into consideration such as fusion, allomorphy, polysemy and synonymy. Perhaps the main difficulties in studying DM, the irregularities, the long list of categories and the complex semantics, in one way or another, stem from these notions permitting a flexible interaction between DM and the endless possibilities of linguistic expression. In this section, we explore these notions based on plenty of examples.

3.3.1 TrLex

We need an extensive dataset for populating the classes of DM with real-life examples. TrLex (Aslan et al., 2018) is a large morphological lexicon containing all entries from TDK Dictionary for Contemporary Turkish as well as some additions by the authors. It is a crucial resource for any computational study of Turkish morphology. We did not eliminate entries due to ambiguity in meaning or analysis, but relabeled suffixes based on the most prominent meaning and syntactic category. This way, we could build a dataset of 27954 base-lemma pairs out of 44549 derived forms in the original data. Out of 174 affix classes we identified, 90 affixes occur in 10 or more instances. 27 affixes occur in 100 or more instances. Only 10 affixes occur in more than 500 instances.

Since TrLex covers the whole dictionary, one could argue that it is not possible to construct a larger database of derivational relations. For derived forms, being included in the dictionary is the gold standard for lexicalization and widespread use. There might be reasonable cases of derived forms that have been for some reason left out of the dictionary, but their validity will always be open to debate. Even if we expanded the dataset with such cases, we would not be able to evaluate whether their usage is stable enough for their distributional representation (see Section 4.1) to be consistent and meaningful. For these reasons, it is a safe and satisfactory choice to base our dataset on TrLex.

Annotation of affixes in the TrLex data was based on their phonological form and similar forms were distinguished with upper/lower case. An example to this was the NND affix for similarity *sH* (*insansı* 'humanoid') and the 3rd person possessive marker *SH* (*başkası* 'other'). We relabeled the affixes according to an extended version of the coding scheme suggested in Bozşahin (2018). Affixes that were not present in the original paper were marked with a plus sign at the end. Including the categories of the base and lemma in the label made the category distinctions more explicit and prevented confusion due to the similarity between affix labels.

A second issue was that base words were not presented in TrLex in their original form. For instance, the base in *ayr-ım* 'distinction' was given as *ayr-* instead of *ayır-* 'distinguish'. For us to be able to find matching dictionary entries, we needed the original base form without any phonological modifications. We edited such cases manually.

Table 8: Syntactic categories of derived forms in TrLex

		Lemma Category				
		N	J	A	V	Total
Base Category	N	5506	4369	322	2135	12332
	J	2770	147	334	752	4003
	A	13	13	29	0	55
	V	10990	496	5	73	11564
	Total	19279	5025	690	2960	27954

Table 8 presents the statistics on derived forms found in TrLex. The largest number of derived forms fall into NVD, due to NVD_MA (7949) (*anlama* 'understanding', *küçülme* 'shrinking') and NVD_YIS (1149) (*dağılış* 'distribution', *dokunuş* 'touch'). Noun-based derived forms are common with all lemma categories and include many productive affixes. Adverb bases seem to resist derivation.

3.3.2 Interesting Cases

There are plenty of interesting cases that we came across during our reviews of Aslan et al. (2018), TDK (2019), Nişanyan (2021), Eyüboğlu (2017) and Ergin (2009). Some of these are non-standard uses of known affixes, while some are outright incorrect uses. But all are candidates for shedding a light on the underlying processes.

(67) Interesting uses of possessives

a. *açıkçası* 'to tell the truth'

b. *birisi* ‘someone’

c. *bencesi* ‘in my language’

At first sight, we analyzed açıkçası as *açık* (J) + *ça* (AJD_CA) + *sı* (NNI_POSS3s), and struggled with the categories. So did Nişanyan (2021). However, the correct analysis is probably *açık* (J) + *ça* (JJD_CA) + *Ø* (NJD_Ø)+ *sı* (NNI_POSS3s). *birisi* shows two consecutive uses of NNI_POSS1s. *Bencesi* is a recently popularized word that alters *bence* ‘in my opinion’ with the 3rd person possessive. In that use, *-CA* in *bence* gains a different interpretation as the affix forming language names, in analogy to the likes of *Türkçesi* ‘Turkish translation’ and *İngilizcesi* ‘English translation’.

(68) Interesting uses of XXD_LIK

a. *öncelik* ‘priority’, *ayrıcalık* ‘privilege’

b. *gündelikçilik* ‘being a wage earner’, *üstelik* ‘besides’, *yerindelik* ‘legitimacy’

c. *bizdenlik* ‘being one of us’, *kendiliğindenlik* ‘moving of one’s own accord’

d. *önemsemezlik* ‘inattentiveness’, *yılmazlık* ‘indomitableness’, *yürürlük* ‘enforcement’

e. *çekememezlik* ‘envy’, *görmemezlik* ‘connivance’, *duymamazlık* ‘pretending not to have heard’

f. *çaydanlık* ‘teapot’, *yağdanlık* ‘oilcup’, *iğnedenlik* ‘pincushion’

g. *güç beğenirlik* ‘finickiness’, *yüz görümlüğü* ‘price for seeing the bride’s face’

h. *benbencilik* ‘beadledom’, *beniçincilik* ‘egocentrism’, *vaybabamcılık* ‘clamorousness’, *vayvaycılık* ‘clamorousness’

i. *kaç yıllık* ‘how many years worth’

j. *bankayaonbinkoyupikiyılsonraellibinalangiller* ‘group of people who deposit ten thousand to the bank and collect fifty thousand two years later’

Önce ‘before’ and *ayrıca* ‘besides’ in the first group would normally be considered adverbs. This does not fit any accepted functions of *-lIK*, but the derived forms can be clearly understood. Whether they are analyzed by the hearer, or simply retrieved from the lexicon is a point of interest. The second group, *günde* ‘in a day’, *üste* ‘on’ and *yerinde* ‘in its place’ should also be unsuitable bases for *-lIK*. *Bizden* ‘one of us’ and *kendiliğinden* ‘automatically’ are again unexpected bases for *-lIK*, but both the base forms and the derived forms are quite common. Also, locative and ablative marked nominals clearly have a propensity to act as predicates.

Önemsemez ‘does not care’, *yılmaz* ‘does not yield’ and *yürür* ‘walks’ are conventionally considered finite verbs. But as bases for words in the fourth group, they take NJD_LIK. The same is true for the next group, but with a twist: *çekememezlik*, *görmemezlik* and *duymamazlık* involve not one, but two negative markers. This demonstrates that ungrammaticality does not always prevent lexicalization. The *-dan* affix apparent in the examples of *çaydanlık* and *yağdanlık* is actually from Persian, so it does not invalidate the ordinary use of NND_LIK. The way *güç beğenir* ‘hard to please’ and *yüz* *görüm* ‘seeing the face’ have been used as bases is interesting. These are clear examples of DM taking phrasal scope. The last two groups are also examples of phrasal scope.

(69) Interesting uses of XXD_CA

a. *olanca* ‘utmost’

b. *yeterince* ‘sufficiently’

c. *eğlence* ‘entertainment’, *dinlence* ‘leisure’, *düşünce* ‘thought’, *söylence* ‘myth’

Olanca is an interesting use of *-CA*, but is not impossible to analyze. *olan* is derived from *ol-* ‘to be’ by JVD_AN. The resulting adjective takes a JJD_CA that modifies its intensity. *Yeterince* can be analyzed as *yet-* ‘to suffice’ (V) + *er* (JVD_AR) + *Ø* (NJD_Ø) + *(s)in* (NNI_POSS3s) + *ce* (AND_CA).

(70) Interesting uses of TAM markers: *-DIK*

a. *tanıdık* ‘familiar’, *bildik* ‘familiar’, *görülmedik* ‘unusual’, *aramadık* ‘unsearched’

b. *gittikten sonra* ‘after one goes’, *geldiği sırada* ‘while one comes’, *gideceği için* ‘because one will go’, *istemediğine göre* ‘as one does not want’, *anlattığı gibi* ‘as one narrates’, *olduğu kadar* ‘as much as it goes’, *öldüğünden beri* ‘since one is dead’, *okuduğun sürece* ‘while you study’

c. *gittikçe* ‘increasingly’, *oldukça* ‘quite’

-DIK is generally considered a relative clause marker. Göksel and Kerslake (2005) gives both a relativizer account of this affix and lists it among the NVD. Ergin (2009) only considers its participle forming function. Examples in the first group and the relative ease with which one can derive new forms of this kind demonstrate that the affix is a productive one on its own. There is little reason to believe that this kind of *-DIK* is anything other than JVD_DIK. In all groups, templates include an NNI affix on the derived form. Such consistent application of NNI signals that the stem is a noun. Indeed, there is ample evidence (both snychronic and diachronic) for this kind of *-DIK* being of NVD_DIK class (or JVD_DIK + NJD_Ø).

(71) Lexicalized uses of TAM markers: Others

a. *alacak* ‘debt’, *verecek* ‘liability’, *gelecek* ‘future’, *giyecek* ‘clothes’, *silecek* ‘wiper’, *yiyecek* ‘food’

b. *bilmiş* ‘know-it-all’, *gelişmiş* ‘developed’, *okumuş* ‘educated’, *tanınmış* ‘well-known’

c. *dolmuş* ‘shared taxi’, *ermiş* ‘saint’, *geçmiş* ‘past’, *yemiş* ‘berry’

d. *alındı* ‘receipt’, *çıktı* ‘output’, *esinti* ‘breeze’, *girdi* ‘input’, *örüntü* ‘pattern’, *çalıntı* ‘stolen’

NVD_ACAK has an undisputed derivational affix label, due to examples in the first group. *-mIş* productively forms adjectives and rarely nouns in the next two groups. Even *-DI*, despite being almost exclusively associated with its role in finite verbal inflection, has its NVD_TI counterpart. NVD_INTI is considered as the combination of VVI_REFX with NVD_TI by Ergin (2009); we agree. The case of NVD_DIK is curious, because it leaves the noun forming function to NVD_TI and NVD_MIS, outside the lexicalized adverbial templates. In return, it is the only affix with past / perfective meaning usable in those templates. Therefore, they have complementary distribution.

The examples analyzed in this section demonstrate several important facts. First, ungrammatical constructions clearly may lexicalize. We have to find a way to accommodate that fact. Second, productive affixes of a certain base category may expand their domains to derive unexpected categories. We must be able to explain the fact that people can understand such forms and derive new forms by analogy. Third, some affixes may frequently be used together and lexicalize as a distinct morpheme (fusion). Fourth, a single morpheme may spawn several forms with slightly different form (allomorphy) or meaning (polysemy). We delve deeper into these facts in the next chapter.

(72) Linguistic facts regarding DM

a. Grammaticality is not required for lexicalization.

b. Categorial restrictions are not absolute.

c. Fusion is common.

d. Allomorphy and polysemy are common.

3.3.3 Fusion

Turkish is an agglutinating language, but groups of morphemes occasionally freeze together. In these cases, phonological changes and loss of productivity often conceals the complex nature of the morpheme. As a rule, we analyze a complex morpheme if and only if its contemporary function is recoverable from the contemporary function of its constituents. If constituents cannot be recognized, the underlying structure is not available to contemporary speakers and the analysis should be left to etymologists.

Analyzing composite affixes is no trivial task. First and foremost, these affixes are recognized in their complex form, because in MT, they consistently occur in a specific form with a specific function. Even if we can identify the components, we often cannot apply them to the base individually; either one of the components has lost its productiveness, or the meaning content of the complex affix is no longer compositional.

Unsurprisingly, the grammars we consult are not always in agreement. The decision to analyse or not to analyse a complex affix often comes to the definitions and assumptions adopted by the grammarian. Our aim in investigating fusion is to simplify the inventory of derivational affixes by eliminating complex affixes.

Besides composite affixes with phonologically realized components, perhaps an important question is whether we should attribute the variation in affix category to an underlying zero-derivation / conversion. Such variation frequently occurs between nouns and adjectives and between adjectives and adverbs. For instance, NVD_DIK could be recognized as JVD_DIK + NJD_Ø. In any case, these are not cases of fusion.

List of composite affixes involving -*CA*

- AAD_CACIK
- AJD_CASINA
- NVD_MAC+

It is not surprising that an extremely productive affix such as -*CA* has been involved in composite affixes. Both AAD_CACIK and AJD_CASINA are transparent in that respect. Ergin (2009) gives several examples on AAD_CACIK: *usulcacık* 'quietly', *yavaşçacık* 'slowly', *demincecik* 'just now', *ufacıcık* (< *ufacacık*) 'tiny'. We can add *hazırcacık* 'ready' and *hemencecik* 'straigtaway' to the list. As the combination of AAD_CA and AAD_CIK fails to account for AAD_CACIK, there is sufficient evidence for considering it as a distinct affix.

AJD_CASINA also assumes a function not entirely predictable from its parts. Göksel and Kerslake (2005) give three versions of this affix; one applied on adjectives with negative connotation (*salakçasına* 'foolishly'), one applied on JVD_IR adjectives (*nispet yaparcasına* 'just to spite') and one applied on JVD_MIS adjectives (*anlamazmışçasına* 'uncomprehendingly'). They add that it is sometimes possible to add a person marker to the second and third kinds (*konuşuyormuşumcasına* 'as if I was talking'), unlike any other subordinating suffix. With such specific uses, AJD_CASINA clearly deserves a separate place in the inventory.

NVD_MAC+ is a phonological variation over the combination of NVD_MA and NND_CA. We do not annotate -*mAcA* forms (*bilmece* 'riddle', *bulmaca* 'puzzle', *çekmece* 'drawer') as NVD_MACA, since the stems are still recognizable and the meaning change is attributable to NND_CA. However, the analysis of NVD_MAC+ is not so straightforward. This class of affixes also cover -*bAç* forms (*saklambaç* 'hide and seek', *dolambaç* 'meander') -*bAç* occurs when the base ends in /m/. So, the difference between -*mAç* and -*bAç* is phonological.

List of complex affixes involving voice

- VND_LAS
- VND_LAT
- VND_LAN
- NVD_INTI
- NVD_INC
- NVD_ANAK
- VND_SIN

Both Göksel and Kerslake (2005) and Ergin (2009) accept that the affixes VND_LAS, VND_LAT and VND_LAN are made of VND_LA with VVI_RECP, VVI_CAUST and VVI_REFX, respectively. However, there are cases where the stem derived by NVD_LA is unrecognized: *kirletmek* 'pollute', *kirlenmek* 'get dirty', but **kirlemek*. Also, there are non-compositional forms derived by these affixes, such as *dertleşmek* 'have a heart-to-heart talk'. With these pieces of evidence, we believe VND_LAS, VND_LAT and VND_LAN are lexicalized to some extent. Therefore, we include these affixes in the inventory.

The situaton is similar for the rest. Göksel and Kerslake (2005) do not recognize any of these as complex, but Ergin (2009) strongly claims that NVD_INC and NVD_INTI involve the VVI_REFX. He does not comment on NVD_ANAK and VND_SIN in a detailed fashion, but we suspect they involve VVI_REFX as well, in combination with NVD_AK and VND_SA (< VND_SI), respectively. For reasons similar to the ones given for VND_LA variants, they must be treated as valid affixes.

List of composite affixes involving -*CI*

- NVD_GAC
- NVD_GIC
- NVD_ICI
- JVD_ICI

According to Ergin (2009), NVD_GAC and NVD_GIC are allomorphs. Also, NVD_ICI and JVD_ICI are combinations of NVD_I and JVD_I with NND_CI.

The original form of NVD_GAC and NVD_GIC was -*gUçI*. NVD_ICI and JVD_ICI evolved partly from -*IgçI* forms and partly from -*gUçI* forms. All four affixes are complex, but synchronically unanalyzable.

List of composite affixes involving -*sI*

- JJD_MSAR
- JJD_MTRAK

VJD_IMSE, NJD_MSI, JND_SAL, VND_SA and many other affixes with the /s/ phoneme convey some sense of similarity. Şçerbak (1989) convincingly argues that such affixes have their origins in the verb *sımak* ‘to resemble’. According to Ergin (2009) the /m/ phoneme observed in some of these affixes is added later by analogy to bases ending in /m/.

We believe JJD_MSAR is the combination of VJD_IMSE and JVD_IR. However, the intermediate stems in the derived forms such as *iyimser* ‘optimist’, *kötümser* ‘pessimist’ and *karamsar* ‘pessimist’ are unrecognizable. JJD_MTRAK, on the other hand, perhaps formed with fusion of JJD_MSI and JJD_RAK+. Ergin (2009) points out that Anatolian Turkish included a -*mtI* variation of the former, adding credibility to this claim.

List of other composite affixes on verbal bases

- VVD_USTUR
- VVD_AKLA
- AVD_ARAK

Regarding VVD_USTUR, Ergin (2009) does not list the affix, but Göksel and Kerslake (2005) do. It is clearly the combination of VVI_RECP and VVI_CAUSD. However, again due to the intermediate stems sometimes being unrecognizable, we choose to include it in our inventory. Similarly, VVD_AKLA is made up of NVD_AK and VND_LA, but it deserves a place of its own.

Durmuş (2012) explains AVD_ARAK can be analyzed into AVD_A+ and a currently extinct AAD_RAK. AVD_A+ had the same meaning as AVD_ARAK in OT, while AAD_RAK only added emphasis. Gradually AVD_ARAK replaced AVD_A+. There are still a handful of examples for the latter: *nöbetleşe* ‘in turns’, *ortaklaşa* ‘jointly’, *bile bile* ‘knowingly’. In general, it would not be plausible to expect an average speaker to analyze this affix.

List of other composite affixes on nominal bases

- NND_GILLER+
- AND_LEYIN+

Göksel and Kerslake (2005) lists *-gil* among derivational affixes, but TDK (2019) does not have any entries derived with this affix. Instead, there are 182 items derived with NND_GILLER+ indicating animal and plant families. While *zürafagiller* 'giraffidae' is an acceptable word, **zürafagil* is not; so we cannot analyze the affix and we leave it in its complex form.

Ergin (2009) believes that AND_LEYIN+ is a combination of VND_LA, NVD_I and AND_IN+. The claim is reasonable, but there is no way an ordinary speaker could analyze this affix.

3.3.4 Allomorphy

Like Paster (2014), we use the term allomorphy to refer to both phonological allomorphy (also called morphophonology or non-suppletive allomorphy) and suppletive allomorphy. Phonological allomorphy occurs in a strictly rule-based manner. Phonological allomorphs are identical in their syntactic category and logical form. They also occupy same position class.

Vowel harmony and consonant rules make sure that Turkish morphology is full of phonological allomorphy. Backness harmony and frontness harmony together generate sets of /a/-/e/, /ı/-/i/-/u/-/ü/ pairs depending on the last vowel of the stem. Consonant assimilation generates pairs of /d/-/t/ and /g/-/k/. Generally, phonemes that are subject to variation in inflectional allomorphs are represented by capital letters for the first phonemes in each set. This means that, for vowels, back-unrounded alternatives are preferred, and for consonants voiced versions are preferred. Within Turkish IM, consonant assimilation seems to be restricted to plosives.

Past tense marker *-DI* constitutes a great example for demonstrating phonological allomorphy, exhibiting both harmony and assimilation with a full set of eight allomorphs: *-dı, -di, -du, -dü, -tı, -ti, -tu, -tü*. We have come across no controversy regarding the allomorphs of the past tense marker, or any other IM. Little suppletive allomorphy can found in Turkish IM. Perhaps one exception would be voice markers, if they are considered IM.

DM is much more complicated with respect to allomorphy relations. As examples in this section demonstrate, suppletive allomorphs can be found on many different grades of similarity. Some pairs of morphemes are really close to fitting the rules of phonological allomorphy. Variants due to vowel harmony, consonant assimilation and erosion are generally recognized as allomorphs. But in plenty of cases, slightly unconventional phonological variations (such as consonant deletion) suffice for an affix to be considered separate. Variations between open and close vowels (forming pairs of /a/-/ı/ and /e/-/i/) are almost never thought to be phonologically motivated; even when the resulting forms are identical in category and semantics. Although open-close variation is not generally considered as vowel harmony, perhaps such pairs of affixes should still be seen as allomorphs. Perhaps derivational affixes vary more freely in their phonological form.

Other allomorph candidates are very similar in form and meaning, as well as being in complementary distribution, but they can hardly be called phonological variations. A good compromise in the definition of suppletive allomorphs is put forward by Stockwell and Minkova (2001). They argue that a historically valid relationship between morphemes should be sufficient for them to be considered allomorphs. We believe this is an adequate restriction.

Two interesting cases need to be accounted for. First, even if the historical relationship and phonological similarity conditions are met, the two morphemes may have assumed different semantics. In that case, we must decide whether the alternative semantics can be considered polysemy. If semantics are substantially different, it is difficult to argue for allomorphy. Second, multiple allomorphs may derive different forms from the same stem. If such cases are systematic, we may be dealing with distinct affixes rather than allomorphs.

The NVD affix *-GAn* is a good example for suppletive allomorphy. In addition to the conventionally accepted allomorphs *-gan, -gen, -kan* and *-ken*, we have *-gın, -gin, -gun, -gün, -kın, -kin, -kun* and *-kün*, having similar functions and properties. There are also cases where the leading /g/ erodes and gives way to morphemes of the form *-An*, which should again be recognized as an allomorph of *-GAn*. Studying these affixes separately is not more reasonable than studying *-dı* and *-di* separately. For our purposes, they are the same affix.

NVD_AN, NVD_GAN and NVD_AGAN mostly form agent names, while NVD_GIN mostly forms patient names. Otherwise, the meaning of these four affixes seem to be quite similar.

Examples of agent forming NVD_AGXN affixes

- *alınan* 'who gets offended'
- *alıngan* 'easily offended'
- *yayvan* 'broad'
- *yaygın* 'widespread'
- *etken* 'factor'
- *etkin* 'effective'
- *kızan* 'who gets angry'
- *kızgın* 'angry'
- *gezegen* 'planet'
- *gezgin* 'traveler'

Ergin (2009) points out that NVD_GAN is exclusively applied to verbs with more than one syllable (*etken* is a counter-example), while NVD_GIN is mostly applied to verbs with a single syllable. If this is the case, there would be reason to believe that NVD_GAN and NVD_GIN are suppletive allomorphs, Ergin (2009) also suggests than NVD_AN and NVD_GAN are distinct forms, the former being a contraction of the OT NVD_GAN while the latter having evolved from NVD_KAN, the latter being a stronger version of the former. Ergin (2009) notes the similarity between NVD_GAN and NVD_AGAN.

Like Ergin (2009), Göksel and Kerslake (2005) presents all four affixes as distinct morphemes. Nişanyan (2021) distinguishes between NVD_GIN and NVD_GAN, but does not distinguish between NVD_AN and NVD_GAN or NVD_GAN and NVD_AGAN.

NVD_AK and JVD_AK form agent, patient and location names, while NVD_IK and JVD_IK+ form patient names.

Examples of location and patient forming NVD_XK affixes

- *yatak* 'bed'
- *yatık* 'horizontal'
- *konak* 'mansion'
- *konuk* 'guest'

- *batak* ‘swamp’
- *batık* ‘sunk’
- *kayak* ‘ski’
- *kayık* ‘boat’
- *kaydırak* ‘slide’
- *kaçak* ‘fugitive’
- *kaçık* ‘mental’
- *kıvrak* ‘agile’
- *kıvrık* ‘convoluted’

Ergin (2009) points out that many of the instances of NVD_AK have evolved from NVD_GAK forms. Ergin (2009), Göksel and Kerslake (2005) and Nişanyan (2021) list both affixes.

NVD_GA and NVD_GI are also similar phonologically and semantically.

Examples of NVD_GX affixes

- *yetke* ‘authority’
- *yetki* ‘authority’
- *bilge* ‘wise’
- *bilgi* ‘knowledge’
- *çizge* ‘diagram’
- *çizgi* ‘line’
- *dizge* ‘system’
- *dizgi* ‘typesetting’
- *örge* ‘motif’
- *örgü* ‘knitting’
- *tepke* ‘reflex’
- *tepki* ‘reaction’

Ergin (2009) posits that NVD_I and JVD_I cover three kinds of etymological roots. Some nominals with these affixes were originally formed with NVD_IG and lost their final consonant. Some of them were originally formed with NVD_GI and lost the initial /g/ of the affix. Yet, there are also the ones that formed after the NVD_I affix has lexicalized as a derivational affix, thus did not go through the loss of /g/.

Examples of NVD_I and JVD_I affixes

- *yazı* ‘writing’
- *ölü* ‘dead’
- *korku* ‘fear’
- *dolu* ‘full’

According to Ergin (2009) NVD_GAC and NVD_GIC are allomorphs. He mentions that the old forms of NVD_GIC only included round vowels. We could only find two stems where both NVD_GAC and NVD_GIC can be applied (alternative derived forms having the same meaning in both cases), otherwise they have complementary distributions. Göksel and Kerslake (2005) and Nişanyan (2021) list both affixes, but we agree with Ergin (2009) that they are allomorphs.

Yet another set of phonologically similar affixes are NVD_IM and NVD_AM+. Ergin (2009), Göksel and Kerslake (2005) and Nişanyan (2021) list both affixes. None of the sources comment on the possibility of a link between the two affixes.

Examples of NVD_XM affixes

- *tutam* ‘pinch’
- *tutum* ‘attitude’
- *kuram* ‘theory’
- *kurum* ‘institution’
- *biçem* ‘style’
- *biçim* ‘form’
- *dönem* ‘period’
- *dönüm* noktası ‘turning point’

According to Ergin (2009) the *-CAK* group, JJD_CAK+, AAD_CAK, AJD_CAK+, NND_CAK, and the *-CIK* group JJD_CIK, JND_CIK, NND_CIK along with NND_CAGZ all evolved from the *-CAK* group. However, in their current usage it is hard to establish a subgroup with sufficient similarity to be called allomorphs.

Allomorphy is extremely prevalent in Turkish DM. It is impossible to study DM without taking it into account. If we ignored allomorphy, some affixes would have so few derived forms that it would not be reasonable to represent them in a rule-based manner. When such affixes are considered with their allomorphs, we often obtain semi-productive DM affixes.

3.3.5 Polysemy

Another important notion regarding DM is polysemy. By polysemy, we refer to different uses of a morpheme, displaying non-identical but related semantics, often with different semantic selection criteria.

Rainer (2014) points out that the multiplicity of meaning issue has been present in morphology and there have been several approaches to its analysis. One of these is seeking a single abstract sense from which the specific senses are to be derived. Rainer (2014) believes that vagueness of the core sense, inexplicitness of the derivation processes and the over-generation problem are reasons to question the validity of this approach. Another point of view considered in Rainer (2014) is that specific senses may not be predicted by a single sense, but they may be motivated by one. In other words, a single sense may give rise to a constellation of other senses over time. Having compiled an extensive list of examples regarding polysemy, we realize that a single abstract sense to derive all the specific senses may not always be found. Therefore, we adopt the latter point of view.

Since inflectional morphemes realize morphosyntactic features, they are not free in assuming new semantics. This is not the case for DM. A DM may initially have served a single well-defined semantic function. With time, some derived forms may begin to assume related but different semantic properties. When a morpheme assumes multiple such meanings, we can speak of polysemy, and it is quite common.

This process is often triggered by the loss of productivity. When an affix loses widespread productivity, its semantics and form gain the freedom to vary. Usually the affix assumes more specific semantics and its scope narrows down to a smaller set of stems. We call this “referential narrowing” (see Section 4.3.1 for a distributional account of this process). During this process, the affix may branch into multiple niche uses, resulting in polysemy.

Niche uses do not only differ slightly in their semantics, they also differ in their semantic selection criteria. Certain semantic operations may only be compatible with bases having some specific feature.

The variation in semantics usually comes with a variation in selection properties. We believe that, among these two notions, polysemy is more basic and determines rule variation to a certain extent. This is why we will mainly use polysemy to describe this phenomenon.

The affix *-lIK* is a good example. All forms derived by *-lIK* describe some abstract semantic relation of "an object dedicated for a certain function". The two major categories NND_LIK and NJD_LIK constitute the vast majority of cases for XXD_LIK. JND_LIK is also a productive affix. Others can be considered non-standard uses.

(73) Examples and number of instances for the various categories of XXD_LIK

a. NJD_LIK: *acılık* 'bitterness', *bahtiyarlık* 'happiness', *çağdaşlık* 'modernness' (2761 across all polysemous uses)

b. NND_LIK: *abonelik* 'subscription', *arıcılık* 'beekeeping', *birlik* 'unity' (2637 across all polysemous uses)

c. NND_LIK: *gözlük* 'glasses', *boyunluk* 'collar' and *başlık* 'headdress'

d. NND_LIK: *odunluk* 'woodshed', *kitaplık* 'bookshelf' and *ayakkabılık* 'shoe rack'

e. NND_LIK: *analık* 'foster-mother', *evlatlık* 'adopted child'

f. JND_LIK: *adaklık* 'sacrificial', *yemeklik* 'reserved for food preparation', *kışlık* 'reserved for winter' (74)

g. NAD_LIK+: *biteviyelik* 'ceaselessness', *boşunalık* 'vainness', *kendiliğindenlik* 'spontaneity' (13)

The case with *-lIK* is not unique. As a result of our inventory work, we have found that 27 Turkish derivational affixes out of 62 (grouped by rules of allomorphy) have 3 or more polysemous uses (2.95 on average). In later sections, we argue that these different uses compete for salience.

In contrast, IM rarely displays polysemy. The Turkish perfective *-DI* is, like most inflectional affixes, wholly productive and acts as a member in an inflectional paradigm. These properties strictly tie it to its semantics and selection criteria. As a result, *-DI* has only one meaning within the paradigm of TAM. The second function of *-DI* as the past copular marker is easily distinguishable: The past copular always comes after a TAM marker; in other words, its position class is different than the perfective.

Despite its clearly inflectional nature, *-DI* appears also in lexicalized forms. *girdi* 'input', *çıktı* 'output' and *alındı* 'receipt' are such forms. In these cases, inflected forms containing *-DI* are lexicalized without assuming new semantics. They start to be used instead of the subject of the original sentence. This is not a productive use of *-DI* and the mechanism of lexicalization is quite different than DM in general. The likes of *kalıntı* 'ruins' and *sarsıntı* 'quake' should not be confused with these examples, as the latter pair contains the widely recognized derivational affix *-IntI*.

Many affixes have a single set of base-lemma categories, such as NND_CI, and JND_SI. These affixes are among the most frequent, but they are also very consistent in their choice of base category. Most low-frequency affixes also have a single category. TAM affixes such as VVI_TAORS, JVD_IR and

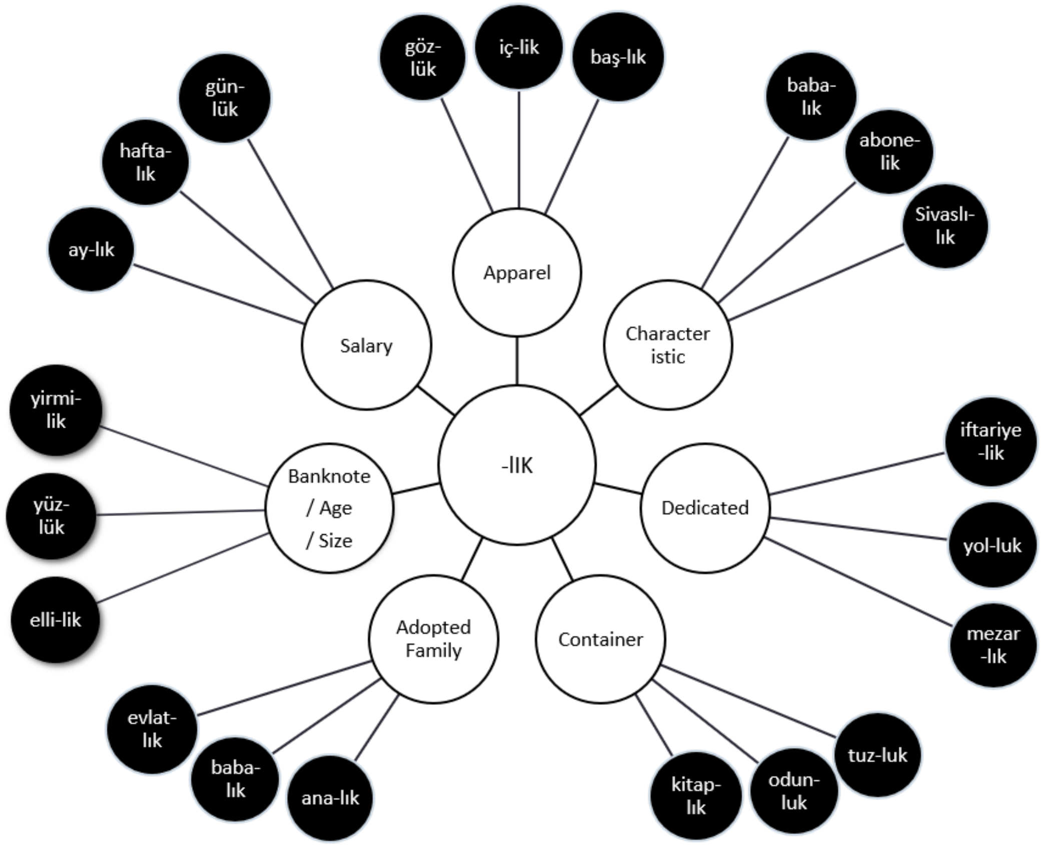


Figure 4: Uses of *-lIK*, a Turkish denominal nominal derivational affix

JVD_Z+ can be said to have only one category, if we accept that all predicates in Turkish are nominal, as per Kuznetsov (1997).

On the other hand, there are some affixes that apply on a variety of base categories, and produce a variety of lemma categories. Such affixes tend to have a large sample of affixes, too. A notable example is XXD_CA. The productive uses of *-CA* (AJD_CA, NND_CA producing language names, AND_CA, JJD_CA and AAD_CA) are semantically and etymologically related.

The *-CA* affix acted as the equative case marker in the OT (it would have been marked NNI_EQU), disappeared as a case marker, but spawned several productive derivational affixes. Semantically, all its variations clearly resemble the original. This does not mean that *-CA* has become an exceptional "category-independent affix", in the minds of the speakers; because different variations have different levels of productiveness and different semantics. Also, NVD_CA and JND_CA are stressed, while others are not.

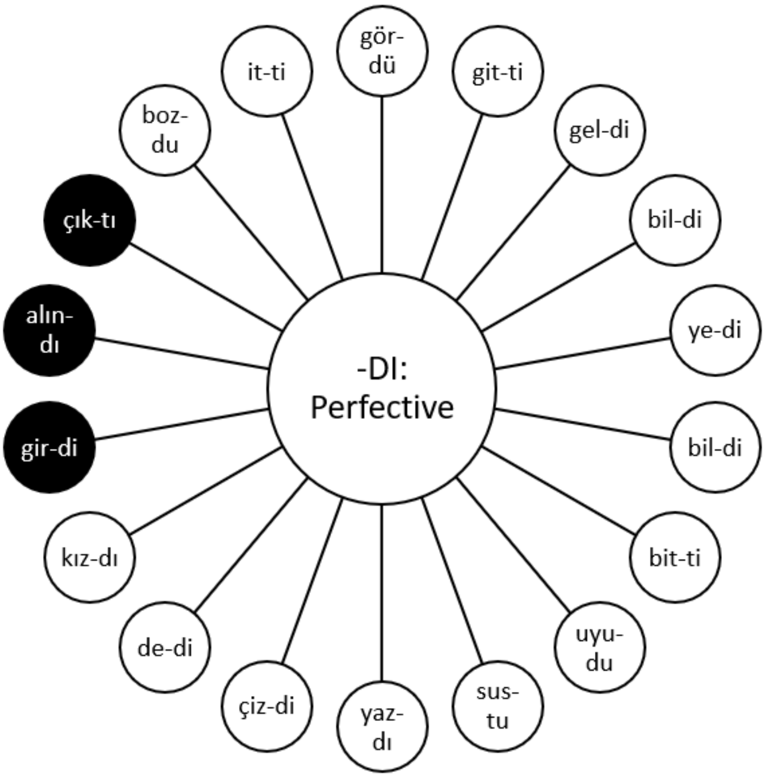


Figure 5: Uses of *-DI*, the Turkish perfective affix

(74) Examples and number of instances for the various uses of XXD_CA

a. AJD_CA: *bencilce* ‘egoistically’, *haksızca* ‘unfairly’, *sakince* ‘calmly’ (320)

b. NND_CA: *azerice* ‘azeri’, *cermence* ‘german’, *çekmece* ‘drawer’ (150) (several polysemous uses)

c. AND_CA: *askerce* ‘soldierly’, *boyca* ‘by length’, *insanca* ‘humanely’ (105) (several polysemous uses)

d. JJD_CA: *acıca* ‘somewhat bitter’, *irice* ‘largist’, *kızılca* ‘reddish’ (64)

e. NVD_CA+: *çekince* ‘reservation’, *dönence* ‘tropic’, *güvence* ‘guarantee’ (16) (productiveness questionable)

f. JND_CA+: *düzmece*, *kesmece*, *kurmaca* (13)

g. AAD_CA+: *beraberce*, *epeyce*, *öylece* (9)

Like allomorphy, polysemy is extremely prevalent in Turkish DM. This is probably the reason DM is generally considered as irregularly irregular. The regularity of DM is concealed by polysemy.

3.3.6 Synonymy

There are also cases where the same meaning is represented by a large variety of distinct affixes. Although such affixes often have their etymological origins in a single affix, they may not reasonably be considered allomorphs in their current form. An interesting example in this respect is the group of affixes indicating similarity. This group is quite large and covers many base-lemma categories.

There is a large number of Turkish affixes conveying a sense of similarity. Yıldırım (2011) reviews the denominal nominal forming affixes of this kind. Her review is not complete, but it is a nice starting point. It is noteworthy that most of these affixes involve the phoneme /s/. In Section 3.3.3 we review the composite affixes in this list and report the claim by Şçerbak (1989) that the /s/ is residual from the verb *sımak* 'to resemble'. We also cite Ergin (2009) for his suggestion that the /m/ morpheme appeared due to analogy.

(75) Examples and number of instances for the various affixes indicating similarity

a. JND_SAL: *açısal* 'angular', *anıtsal* 'monumental', *bitkisel* 'herbal' (192)

b. JND_SI: *ağaçsı* 'arboreous', *insansı* 'hominid', *süngersi* 'spongy' (135)

c. JND_IMSI: *demirimsi* 'ironish', *fiilimsi* 'verbal', *odunumsu* 'timbery' (76)

d. JND_IL+: *ardıl* 'successor', *birincil* 'primary', *ilkel* 'primitive' (47)

e. JJD_IMSI: *beyazımsı* 'whitish', *kekremsi* 'acrid', *tatlımsı* 'sweetish' (29)

f. VND_SA: *kapsa-* 'involve', *susa-* 'be thirsty', *umursa-* 'care' (19)

g. VJD_SA: *garipse-* 'find strange', *ıraksa-* 'diverge', *mühimse-* 'care about' (16)

h. JJD_MTRAK: *acımtrak* 'somewhat bitter', *kızılımtrak* 'reddish', *sarımtrak* 'yellowish' (15)

i. VJD_IMSE: *azımsa-* 'underestimate', *benimse-* 'embrace', *küçümse-* 'belittle' (13)

j. JJD_RAK+: *acırak* 'bitterish', *kısarak* 'fairly short', *ufarak* 'fairly small' (10)

k. JND_SIL+: *ağaçsıl* 'arborical', *otsul* 'herbaceous', *yoksul* 'poor' (9)

l. VND_SIN+: *gereksin-* 'require', *yüksün-* 'regard as burdonsome' (5)

m. JJD_SIN+: *akşın* 'albino', *sarışın* 'blonde' (4) (*-şIn* involves the /ş/ phoneme instead of /s/)

The semantic similarity among the elements of this group is hard to miss. One can also find clear etymological links between many of them. Therefore, we can confidently say that these affixes are tightly connected. While it would be too much to claim they are all synonyms, we can identify subgroups with very similar meanings. The group of similarity affixes are probably the largest constellation of affixes in Turkish.

3.3.7 An Independent Dimension: Thematic Relations

Following Bozşahin (2018), we have annotated derived forms in TrLex with XXD codes indicating the categories of the lemma and the base. This is a significant improvement over previous notations, as it allows us to track the categorial behavior of the affix and emphasize its syntactic properties.

When we look more closely at the behavior of some DM, even this analysis fails to adequately explain the semantic variety. For instance, *-(G)AÇ* forms a large set of dictionary entries. These entries consistently fall into several semantic subgroups. Unlike cases of simple polysemy, there is structure behind these subgroups. This structure follows thematic relations.

For instance, the base noun for a VND affix must have a thematic relation in the action denoted by the derived form. Conversely, an NVD affix derives a noun with a specific thematic relation with the base verb. VJD and JVD affixes have similar functions, but they deal with properties of nouns instead of nouns. VAD and AVD affixes have much narrower variety and scope, but some thematic relations can still be identified in these derivations.

Like all other linguistic categorizations involving prototypical notions, the set of thematic relations is the subject of an ongoing debate. We start with a short-list of relations acceptable to most researchers. Then, we build a larger, more complete set. We try to identify the underlying dimensions and proto-roles (Dowty, 1991) and fill the gaps in their combinations. Perhaps some of the relations we come up with can be eliminated by logical arguments, but we present the full list for the sake of completeness.

Table 9 presents the thematic relations that are widely circulated in the literature. Many other thematic relations have been identified, like constructum, destructum, topic, results, predicate, extent, product, material, asset, pivot and many others (Şahin (2018) describes the relations considered in Turkish PropBank.).

We define three basic properties for each relation. These properties are largely uncontroversial. First, volition is a property available to relations denoting initiators. It may take two values: deliberate and not deliberate. This property serves to distinguish true agency from other kinds of initiation. The second property, change of state, concerns the passive relations and may take two values: changed and unchanged. It is used to distinguish theme from patient, and more controversially, beneficiary from recipient. Finally, the third property serves to distinguish the relations of psychological verbs. It is only marked yes for stimulus and experiencer.

It is possible to identify some ill-justified decisions in every such list. For instance, including a directional or positional origin in the list, but excluding a time origin may be considered an inconsistency. Including stimulus and experiencer for psychological verbs, but not including constructum or destructum is similarly flawed. We believe that one has to justify these decisions by referring to a deeper structure within thematic relations. Dowty (1991) envisions proto-roles for this reason. As he elaborately explains, the value of such an effort is bound to be limited by the prototypical nature of the proto-roles themselves, but they still offer a way of furthering the analysis. Following Dowty (1991), we imagine three proto-roles and organize them in three dimensions. Our proto-roles are source, manner and goal. They may occur in the dimensions of causation, location and time. In principle, we would expect all combinations of these attributes to be represented on a list of thematic relations.

Table 9: Generally accepted thematic relations

Thematic Relation	Volition	Change of State	Psychological
agent	deliberate	-	-
force / nat. cause	not delib.	-	-
experiencer	not delib.	-	yes
stimulus	not delib.	-	yes
theme	-	unchanged	-
patient	-	changed	-
beneficiary	-	unchanged	-
recipient	-	changed	-
cause	-	-	-
source / origin	-	-	-
instrument	-	-	-
manner	-	-	-
location	-	-	-
time	-	-	-
purpose	-	-	-
direction / goal	-	-	-

In order to explain the sub-relations occurring in some combinations, we use four binary attributes: animacy, participant, psychological and change in state. Animacy is tied to volition, but it is more general. It applies on dimensions of causation and location, using a relaxed interpretation of location. Being a participant is only possible in the dimension of causation. We adopt the attributes psychological and change in state as they are described in the literature.

We annotated the properties of the widely accepted relations listed in Table 9. Then, in several iterations, we identified the unrealized combinations and revised the set of combinations where each property may apply. In this respect, dimension and proto role are the primary properties, while others are secondary, helping us generate subtypes for some prominent relations. Table 10 presents these thematic relations.

Based on the resulting list, it can be observed that some conventional thematic relations cover multiple possibilities. For instance, the initiator and receiver can be animate or inanimate; both are called stimulus. On the other hand, some combinations are not logically possible. For instance, we cannot expect a psychological reaction from an inanimate object, so an inanimate version of the experiencer does not exist.

Inanimate, non-participant relations of causation constitute a natural trio: cause, manner and purpose. We do not expect them to be concrete participants of an action, and indeed, they take the form of abstract modifiers for an action.

Animacy attribute is not meaningful in the time dimension, but as the recipient relation shows, it is relevant for the location dimension. We generate the full set of combinations for the location dimension, based on three proto roles and two values of animacy.

Table 10: Extended set of thematic relations

No	Thematic Relation	Dimension	Proto Role	Animacy	Participant	Psychological	Change in State	Desc
1	stimulus	causation	source	yes	yes	yes		1-stimulus (anim.)
2	agent	causation	source	yes	yes	no	-	2-agent
3	?	causation	source	yes	no	-	-	3-? (anim. cause)
4	stimulus	causation	source	no	yes	yes	-	4-stimulus (inanim.)
5	force / nat. cause	causation	source	no	yes	no	-	5-force / nat. cause
6	cause	causation	source	no	no	-	-	6-cause
7	instrument	causation	manner	yes	yes	-	-	7-instrument (anim.)
8	?	causation	manner	yes	no	-	-	8-? (non-part. İnstr.)
9	instrument	causation	manner	no	yes	-	-	9-instrument (inanim.)
10	manner	causation	manner	no	no	-	-	10-manner
11	experiencer	causation	goal	yes	yes	yes	-	11-experiencer
12	patient	causation	goal	yes	yes	no	yes	12-patient (anim.)
13	theme	causation	goal	yes	yes	no	no	13-theme (anim.)
*	*	causation	goal	no	yes	yes	-	*
14	patient	causation	goal	no	yes	no	yes	14-patient (inanim.)
15	theme	causation	goal	no	yes	no	no	15-theme (inanim.)
16	beneficiary	causation	goal	yes	no	-	-	16-beneficiary
17	purpose	causation	goal	no	no	-	-	17-purpose
18	source / origin	location	source	yes	-	-	-	18-source / origin (anim.)
19	source / origin	location	source	no	-	-	-	19-source / origin (inanim.)
20	?	location	manner	yes	-	-	-	20-? (anim. location)
21	location	location	manner	no	-	-	-	21-location
22	recipient	location	goal	yes	-	-	-	22-recipient
23	direction / goal	location	goal	no	-	-	-	23-direction / goal
24	?	time	source	-	-	-	-	24-? (source time)
25	time	time	manner	-	-	-	-	25-time
26	?	time	goal	-	-	-	-	26-? (goal time)

We combined the set of proto-roles with the time dimension and obtained two rarely recognized relations source-time and goal-time. We believe these are as valid as the location relations of source / origin and direction / goal.

In the next section, we present a structured system of thematic relations and motivate the use of thematic relations as a matrix for derivational processes to fill. We also discuss the consequences of adopting this point of view based on numerous examples.

3.4 Classification

Many sources on morphology simply present a list of affixes and provide short descriptions for each of them, along with a few examples. Usually a distinction is made between four major affix types: deverbal verbs, denominal verbs, deverbal nominals and denominal nominals. Any further classification is either completely missing or inconsistent. Homonymy and polysemy surely complicate such efforts.

At the top level, we divide derivational affixes into two: category-preserving and category-changing affixes. We only differentiate between nominals and verbs at this level. Category-preserving affixes fall into two groups: deverbal verbs (VVD) and denominal nominals (NND, JJD and AAD). Category-changing affixes divide into three groups: deverbal nominals (NVD, JVD and AVD), denominal verbs (VND, VJD and VAD) and denominal nominals (AND, AJD, NAD and JAD). thematic relations help a lot in organizing the deverbal nominals and denominal verbs. We encountered no cases of VAD.

In the following sections, properties of affix groups are summarized in tables. The set of properties are gathered from the points explored in Section 2.1.2. Their values are determined mainly based on data collected from TDK (2019).

First property, recursivity, indicates whether a morpheme can append multiple times on the same stem. This property is largely exclusive to DM, but it is certainly not required. Polysemy indicates the number of "sufficiently" productive uses of a morpheme. "Sufficiently" is an admittedly subjective metric; it indicates the existence multiple derived forms in the dictionary and the plausibility of new forms being derived with the same meaning.

Semantic selection describes the extent to which the morpheme selects for certain semantic features. IM usually do not have any semantic selection criteria, while DM often do. Number of dictionary entries are quantitative indicators of the productiveness of a morpheme, but inflected forms are rarely listed in the dictionary.

Changing the base argument structure and changing the base POS are two other properties of a morpheme, that can be objectively evaluated. Another one, phrasal scope, is mostly limited to IM and clitics.

Order of application and invariability of order of application are loosely related to the morpheme's position class. Since DM almost always apply before inflectional affixes and they may apply in various orders, they are all assigned the 0th position class. Others are assigned classes to indicate their relative position.

Finally, being a member of a paradigm and suspended affixation are properties exclusive to IM.

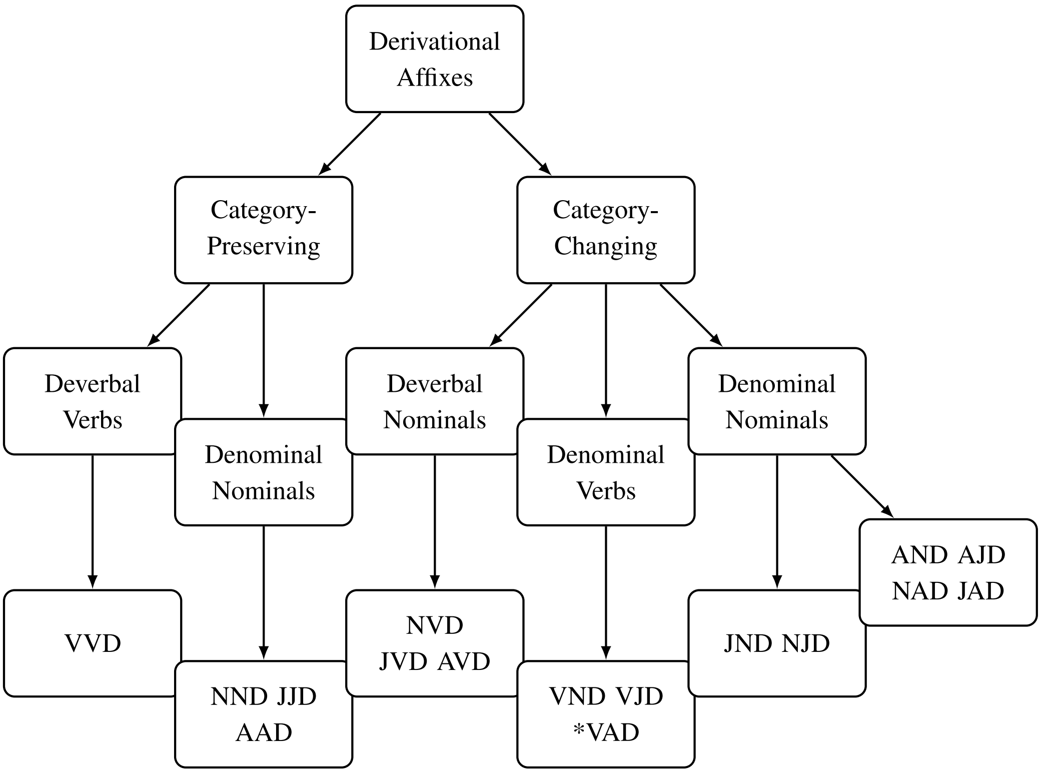


Figure 6: Classification of derivational affixes in terms of the changes in base POS and relevance of thematic relations

Based on these properties, each morpheme group is evaluated on scale from definitely IM, leans IM, leans DM to definitely DM. Clitics are kept outside this scale.

The following sections explore each group with respect to the properties of affixes, in order of deverbal verbs, deverbal nominals, denominal verbs and denominal nominals. These explorations are supported with a large number of examples from the dictionary, as well as matrices summarizing the properties of affixes. Tables containing the examples and property matrices are given in Appendix A to reduce the clutter. Phonological allomorphy is indicated with capital letters, while suppletive allomorphs are all listed as members of an affix group.

This section identifies the productive affix groups and their semantics. These constitute the basis of our analyses in the rest of this thesis.

3.4.1 Deverbal Verbs

Position classes given in Göksel and Kerslake (2005) makes the classification of deverbal verbs much easier. Table 38 in Appendix A summarizes the properties of these affixes.

Most deverbal verbs are without doubt part of DM, since they must apply directly on the base and they require extensive semantic selection. They all have derived forms represented in the dictionary. Verbal inflection is relatively easy to distinguish, due to its paradigmatic behavior.

We believe *-AklA*, *-IklA* and *-AlA* are allomorphs, exhibiting competition and rule variation. Similar items such as *durala-* 'to hesitate' and *durakla-* 'to pause', *itele-* 'to push' and *itekle-* 'to push' are in the dictionary, despite similarity in meaning and form. They are complex morphemes, but they have a specialized meaning that cannot be recovered from the sum of the meaning of their components. They are also quite productive in their complex form.

-A (*tıka-* 'to plug') , *-I* (*kazı-* 'to scratch'), *-p* (*kırp-* 'to clip', *serp-* 'to sprinkle') and *-ImsA* (*anımsa-* 'to recall', *gülümse-* 'to smile') basically have the same semantics, but only *-A* and *-I* can be considered derivation allomorphs. They diminish the intensity of the action denoted by the verb.

Voice markers are unique. They change the base argument structure without changing its POS. They do not take phrasal scope. They change the base category (in the sense of categorial grammar), unlike other VVI affixes. Moreover, they must occur after all VVD affixes, before any other VVI and NVD affixes (if the stem is converted to a nominal). These facts makes their status unusual, if not controversial.

Causative has recursivity, a very rare capability, thought to be reserved for DM. It also displays suppletive allomorphy. It applies very close to the root, but still comes after VVD affixes. Furthermore, derived forms involving causative are widely represented in the dictionary. We believe these properties are sufficient to call it a derivational affix.

One other interesting couple is *-(I)l* and *-(I)n*, together marking passive, reflexive, middle and anticausative voice (Gündoğdu, 2016). *-(I)l* and *-(I)n* are usually considered markers of passive and reflexive, respectively; but a quick run on the dictionary finds numerous cases where they are used interchangeably. Göksel and Kerslake (2005) list *-(I)l / -(I)n* as passive and *-(I)n* separately as reflexive, but does not make a detailed analysis of these voices. Gündoğdu (2016) gives a convincing account of how these two affixes signal the lack of an external argument. According to this account, *-(I)l / -(I)n* should be considered suppletive allomorphs.

Reciprocal voice is much less complicated. It lacks recursivity and suppletive allomorphy. It applies close to the root and changes its argument structure. It derives many forms represented in the dictionary. Like other voice markers, it bears many properties of DM, while mostly being considered an IM.

The negative marker is also an interesting affix. It realizes the grammatical feature of polarity, thus it has perfectly regular semantics. It is therefore generally accepted as an IM. However, it occurs before NVD affixes and it may occur either before or after the gerundium. It does not take phrasal scope, nor does it allow SA. These are not sufficient evidence to challenge the consensus that negative marker is IM.

The gerundium, specifically the one facilitating the Position 2 affixation in Göksel and Kerslake (2005), is rarely a subject of debate. It rarely appears in other functions. Its main job is to change base POS and license the application of a modal auxiliary. While the bare verb cannot be followed by clitics *dA* or *mI*, verbs that have received gerundium may be followed by those. It definitely achieves more than inflection and has a special status. Recognizing the gerundium helps us a lot with the categorial perspective adopted in Section 4.4.

All TAM markers should be considered IM, as they are required by syntax. They are members in a paradigm and invariably occur in the same position class. As many of our sources have suggested, members of the z-paradigm should definitely be considered participles, even when they are on matrix verbs. On the other hand, for members of the k-paradigm and imperative / voluntative paradigm, there is more evidence towards ordinary affixhood, since the inflected forms cannot act like adjectives on their own.

The question marker *mI* is a clitic Erdal (2000). It is listed so that the order of application can be complete.

We discuss the auxiliary in a more detailed fashion, in Section 3.2.3.

There is virtually no debate regarding the status of person markers, but there should be. We have already cited several sources in Section 3.2.1 claiming that person markers act as the real subject in a sentence (Öztürk, 2001) and that members of the z-paradigm are actually clitics (Erdal, 2000). We believe there is sufficient evidence for these claims. We have to admit that person markers of the k-paradigm are still affixes. Person markers are required by syntax and they are members of a paradigm, so we consider k-paradigm markers IM.

Epistemological copula is yet another interesting morpheme. It follows some selection criteria, and has a variable order of application. However, it applies very late and allows suspended affixation. We follow Erdal (2000) in calling it a clitic.

We do not try to analyze the function of VVD affixes in terms of thematic relations. The literature provides a reliable description of voice marking. We are satisfied with that explanation. Other affixes make no change on thematic relations, either. Selected derived / inflected forms of deverbal verbs are given in Table 39 in Appendix A.

3.4.2 Deverbal Nominals

Table 40 in Appendix A presents the properties of deverbal nominal affixes. Selected derived / inflected forms of deverbal nominals are given in Tables 41, 42 and 43 in the same Appendix.

Deverbal adverbs have almost no intersection with deverbal nouns and deverbal adjectives. (The only candidate to fall into this intersection is *-A*, but its status is speculative, as the number of derived forms is very low.) As the sets of NVD and JVD affixes largely coincide, we sometimes refer to them together as NVD.

The status of NVD markers is relatively uncontroversial. The affixes we have chosen to include in our list are productive, they all appear in multiple dictionary entries and most exhibit polysemy. They also change base POS and most exhibit suppletive allomorphy. With these properties, they are DM. On the other hand, these affixes occupy the same position as TAM markers. Like TAM, their position class is invariable. They also occur after all VVD operations and negation (which is an IM). Balancing the evidence, we label all deverbal nominals as DM despite some interesting properties.

Turning to deverbal adverbs, *-Ip*, *-(y)ArAK* and *-mAdAn* occupy the crowded V-4 position. They have no semantic selection criteria, no suppletive allomorphy and dictionary listings. They take phrasal scope and are required by syntax. Therefore, there is little reason to doubt their being IM.

-ken is not quite similar to others. It has a different order of application, it is not required by syntax and it is not a member in a paradigm. Some further investigation reveals additional clues. On verbs, *-ken* may only occur after TAM markers, it may not replace them like other deverbal adverbs. On nominals, it may occur either in bound form, sometimes preceded by a /y/, or in free form preceded by a /i/. These facts are sufficient to establish the presence of a previously overlooked auxiliary.

Unlike other occurrences of the auxiliary, person markers cannot follow *-ken*. This is because *-ken* forms an adverbial, not a participle like TAM markers. Participles subcategorize for NPs (pronouns or person markers are able to act in the same capacity), while adverbials subcategorize for verbs.

In the light of these observations, we can say that *-ken* is a unique affix that only applies on the auxiliary. Although it occupies the same position as copular markers, it is unique in its categorial behavior. As it does not exhibit semantic selection and it has uniform semantics, we cautiously label it as IM.

NVD affixes derive nouns that can fulfill a specific thematic relation of the base verb. Although the name of an action does not constitute a thematic relation by itself, we include it in the list of thematic relations for completeness, believing it is the only other possibility for a deverbal affix.

If an affix is able to derive multiple thematic relations, we interpret this as polysemy. NVD affixes show a great deal of polysemy, although some thematic relations do not exhibit as many examples as others.

We only consider the relatively productive NVD in our analysis, as the data on other affixes would be insufficient. The affixes we excluded are *-AmAK*, *-In*, *-It*, *-Av*, *-Al*, *-Ay*, *-(y)Ası* / *-(y)AsIcA*, *-tay*, *-AcAn*, *-sAl* and *-A*.

Among the productive affixes, patient forms of *-mAK* are quite few (*ekmek* 'bread', *çakmak* 'lighter', *yemek* 'food'), whereas it is a very productive former of action names. Similarly, there are few examples where *-(y)Iş* forms theme (*buluş* 'discovery', *gösteriş* 'show off') or location (*giriş* 'entry') names, but it is much more productive interacting with action and manner (*görüş* 'opinion', *bakış* 'point of view'). An affix may be productive with respect to one thematic relation, while it is not productive with respect to another.

The *-GA* group appear in examples of an unusually high number of thematic relations. While it does not seem to be highly productive with action, agent (*bilge* 'wise'), force (*etki* 'effect', *dalga* 'wave') and location (*yerleşke* 'campus'), it is quite productive in forming stimulus (*sezgi* 'intuition', *duygu* 'feeling', *coşku* 'enthusiasm', *sevgi* 'love'), instrument (*silgi* 'eraser', *sürgü* 'slide', *süpürge* 'broom'), theme (*vergi* 'tax', *içki* 'drink', *bilgi* 'knowledge') and patient (*dergi* 'magazine', *sömürge* 'colony', *çizge* 'diagram') names.

Similar observations can be made for other affixes. In general terms, it could be said that some affixes display an unusually wide variety of functions, while most restrict their interactions to just a few thematic relations.

If we focus on the other dimension, we see a different picture. Some thematic relations, such as stimulus (animate), cause and recipient are never realized by NVD affixes. This might be due to three reasons:

(76) Possible reasons why some thematic relations are never realized by NVD affixes

a. Incomplete data: It is virtually impossible to review and classify all derived forms.

b. Misclassifications: Some examples may have been misclassified.

c. Impossible thematic relations: Some thematic relations in the list may be logically unsound. We have eliminated at least one of these (the inanimate experiencer).

d. Convention: Speakers may not have felt it necessary to invent a new derivational relation for certain thematic relations. Thematic relations that occur only infrequently may not be deserving of a dedicated DM.

JVD affixes derive adjectives to describe an object able to fulfill a selected thematic relation.

Compared to 14 NVD affixes, we observe only 7 JVD affixes. The variety in thematic relations is also significantly smaller. For the 11 thematic relations realized by NVD affixes, there are only 6 thematic relations realized by JVD affixes. These are stimulus (possibilities for both animate and inanimate), agent, experiencer, theme and patient. The discrepancy is partly due to us considering most deverbal adjectives participles, thus outside of DM.

As most affixes interact with agent, experiencer, theme and patient thematic relations, it is hard to distinguish them based on their set of interactions. It also seems that, when realizing the same relation, different affixes make quite similar semantic contributions.

Perhaps a crucial difference lies in the hidden TAM sense exhibited by some affixes but not exhibited by others. For instance, *soğuk* ‘cold’, *bitkin* ‘exhausted’ and *yorgun* ‘tired’ describe an experiencer right after they went through the experience. On the other hand, *ürkek* ‘timid’ and *korkak* ‘coward’ describe the experiencer in general, or in present tense, so to speak.

We cannot associate these senses with particular affixes, because there are many counterexamples. When we look at theme properties, *büyük* ‘big’, *soluk* ‘pale’ and *kurak* ‘arid’ involve a sense of past tense (the event indicated by the verb has already happened), while *durağan* ‘static’ and *durgun* ‘calm’ do not. If we had to add one more dimension to the structure of JVD derivations, it would probably be based on TAM. Similar observations can be made for some NVD affixes.

There are rare cases where one derived form may be associated with multiple thematic relations. For instance, *iğrenç* ‘disgusting’ and *gülünç* ‘ridiculous’ may be interpreted to realize animate or inanimate stimuli, depending on context. If such cases systematically correspond to both thematic relations, we may have sufficient justification to merge the two thematic relations, as is the case in the literature.

AVD affixes produce adverbials that can fulfill time or manner related thematic relations. Compared to NVD and JVD, much fewer AVD affixes exist, realizing much fewer thematic relations. This is expected, as adverbial formation is mainly considered an inflectional process. The morphemes performing that task mostly have no semantic selection criteria. They are not recursive, they are required by syntax, and they may only occur in fixed positions on the stem. As far as we are concerned, these constitute sufficient evidence for labeling such affixes IM. Still we provide a small list of more prominent deverbal adverbs, for the sake of completeness. Thematic relations on the time dimension are only realized by adverbials.

Derived forms involving *-(y)ken*, *-(y)ArAK* and *-mAdAn* can be said to indicate the manner in which an action takes place. *-(y)ken* may also be associated with time. *-(y)Ip* is another unique morpheme that ties start time of an action to the end time of another. The ancient gerundium *-A* seems to live on between host verbs and modal auxiliaries, so we believe it deserves a place in this list.

3.4.3 Denominal Verbs

Table 44 in Appendix A presents the denominal verbal affixes. Selected derived / inflected forms of deverbal nominals are given in Tables 45 and 46 in the same Appendix. We did not come across any VAD affixes.

All denominal verbs exhibit extensive semantic selection, appear in the dictionary in reasonable numbers, change the base POS and may apply directly on the root. They are not required by syntax, nor are they members of a paradigm. Most exhibit polysemy. Based on these observations, we label all such affixes as DM with strong evidence.

VND affixes derive verbs from nouns that could fulfill a selected thematic relation of the derived verb. We only consider relatively productive affixes and exclude *-A*, *-At*, *-IK*, *-sIn* and the zero-morpheme. These affixes appear in fewer than 5 examples each.

There are two quite productive affixes whose behavior cannot be explained in terms of thematic relations: *-DA* (*çatırda-* ‘to crack’, *kıpırda-* ‘to move’) and *-KIr* (*püskür-* ‘to spew’, *hıçkır-* ‘to hiccup’). Both these affixes derive verbs from onomatopoeia.

-lA, *-lAn*, *-lAş* and *-lAt* are not suppletive allomorphs. *-lAn*, *-lAş* and *-lAt* are obviously complex morphemes, composed of *-lA* with voice markers. While most studies do not distinguish between compositional and non-compositional uses of these morphemes, we take care to analyze compositional cases and list the remaining ones under the complex form.

-lA realizes at least 9 thematic relations in addition to action and result. It even forms onomatopoeic verbs. Remarkably, it also interacts with time and goal time, which is quite rare. Furthermore, *-lA* seems to be productive in all these capacities.

-lAn, *-lAş* and *-lAt* each realize fewer thematic relations; 5, 3 and 1, respectively. *-ImsA* / *-sA* realize 4 thematic relations, in addition to forming result names. They are especially productive with theme names. *-Ar* and *-Al* are not really productive in the VND capacity.

Thanks to *-lA*, more than half of the thematic relations are represented in the NVD class. Stimulus (*kokla-* ‘to smell’), instrument (*taşla-* ‘to stone’, *zehirle-* ‘to poison’, *zincirle-* ‘to chain’), manner (*köpekle-* ‘to doggy paddle’), theme (*hedefle-* ‘to target’) and patient (*avla-* ‘to hunt’) examples are especially numerous, similar to other classes; while a large number of examples on location (*duraksa-* ‘to hesitate’, *aşağıla-* ‘to insult’) and time (*kışla-* ‘to winter’, *sabahla-* ‘to stay up all night’) thematic relations can also be found.

An interesting observation is that most forms derived by *-lA* have English counterparts derived by the zero-morpheme. It appears that *-lA* has flexible semantics, assuming new meanings in new contexts. One example is *şirinle-* ‘to smurf’ from the well-known comic franchise “The Smurfs”. Many different actions of the smurfs are described by the verb *to smurf*, and the Turkish translation is simply derived by *-lA* from the characters' Turkish name *şirin* .

VJD affixes derive verbs from adjectives that describe an object in a selected thematic relation.

Only 3 thematic relations are realized by JVD affixes: experiencer (*rahatsızlan-* 'to fall ill', *delir-* 'to go crazy'), theme (*hafifle-* 'to get lighter', *serinle-* 'to cool off', *şişmanla-* 'to get fat') and patient (*kurula-* 'to dry', *aydınlat-* 'to illuminate', *küçümse-* 'to belittle'). Only 2 affixes interact with experiencer, while 5 affixes interact with theme and patient. There are few examples on experiencer, but most affixes are productive with theme and patient.

To the best of our knowledge, there are no morphemes of type VAD.

3.4.4 Category-Preserving Denominal Nominals

There are many different denominal nominal DM. It is best to study these affixes in two groups, the ones changing the base POS and others. Since our analysis is based on the affix forms, each possibly covering multiple polysemous uses, a clean demarcation is not always possible. Still, we can underline the similarities within and differences across classes of NND affixes.

In Appendix A, Tables 47 and 48 present the DM-leaning and IM-leaning denominal nominal affixes. Tables 49 present the examples for category-preserving denominal nominals (NND, JJD and AAD).

First group consists of *-dAş* (*adaş* 'namesake', *arkadaş* 'friend', *vatandaş* 'citizen'), *-gil* / *-giller* (*Ahmetgil* 'Ahmet and his family', *amcasıgil* 'his uncle and his family', *baklagiller* 'leguminosae'), *-tI* (*takırtı* 'rattle', *gürültü* 'rumble') and *-gen* (*üçgen* 'triangle', *dörtgen* 'rectangle'). They do not display much polysemy or suppletive allomorphy, they do not even change base POS. but they exhibit extensive semantic selection and no indicators of IM. They are easily classified as DM. Perhaps the only irregularity in this group is *-gil* / *-giller* being able occur after the possessive marker in some cases such as *annemgil* 'my mother and her family'. This is a remarkable exception to the rule that DM applies before IM.

The next group of NND affixes is composed of *-lIK*, *-CI*, *-CA*, *-CAK*, *-ImsI* and *-Al*. All of them exhibit polysemy, extensive semantic selection and appear in many dictionary entries. Except *-lIK* and *-CI*, they all have several suppletive allomorphs. This group is undoubtedly part of DM.

Affixes in this group tend to exhibit one abstract core meaning, around which varying levels of polysemy can be observed. This is especially true for *-lIK* (*gözlük* 'glasses', *üyelik* 'membership', *mezarlık* 'cemetery') and *-CI*, as they are very productive derivational affixes with an extensive array of polysemous uses. *-lIK* is one of the most productive derivational affixes in Turkish with a total number of 5486 dictionary entries. As an NND affix, it has 7 at least polysemous uses, all of which describe some kind of purpose or dedication.

-CI (*aceleci* 'hasty', *yolcu* 'passenger', *Atatürkçü* 'kemalist') is similar to *-lIK* with its four polysemous functions, all expressing some kind of deliberate affinity towards an action or concept by an agent.

The *-CA* group (*-CA* / *-CAnA* / *-CAsI* / *-CAsInA*) (*almanca* 'german', *kovalamaca* 'game of tag', *acıca* 'somewhat bitter') covers quite different and very specific functions. The same affix derives language names, indicates repetition / playful insistence and reduces the intensity of an adjective. We cannot find the common semantics behind these uses, even if there is any.

Although the *-CAK* group (*-CAK* / *-CIK* / *-CAcIK* / *-CAğIz*) (*yavrucak* 'kiddie', *sıcacık* 'cozy') does not appear in many dictionary entries, they present a tightly connected small set of polysemous uses. They can apply on all nominal categories to form diminutives. Especially the JJD diminutives by *-CAK* and *-CIK*, and the AAD diminutives by *-CAK* and *-CAcIK* demonstrate a clear case of suppletive allomorphy. Interesting cases are *oyuncak* 'toy' and *salıncak* 'swing' make it harder to find a core semantics that is valid for all cases of *-CAK*. Perhaps *-CAK* in these cases originally denoted diminutives, but through semantic shift, started to denote names of tools.

The *-ImsI* group (*-ImsI* / *-ImtraK* / *-rAK* / *-sI*) (*sarımsı* 'yellowish', *tatlımsı* 'sweetish') has only one function, the JJD diminutive. All members have similar semantics. *-ImsI* and *-ImtraK* are the result of fusion, and *-ImtraK* is probably the fusion of three morphemes.

The *-Al* (*-Al* / *-Il* / *-sAl* / *-sIl*) (*ilkel* 'primitive', *birincil* 'primary') group derives very few forms in a category-preserving fashion and these are hard to classify. *yoksul* 'poor' and *varsıl* 'rich' are the only cases where it appears as a JJD affix and *belgesel* 'documentary' and *kumsal* 'beach' are the only cases where it appears as a NND affix. The latter could be said to broadly denote the name of a container. Regarding the shorter members of this group, perhaps it would be more appropriate to consider them as variants of *-sAl* and *-sIl*, where /s/ is allowed to erode. This is why this group requires further investigation.

Other category-preserving denominal nominals are not considered productive. For instance *-şIn* (*akşın* 'albino', *sarışın* 'blond'), falls into this group, but we believe it is not productive enough to make it on our list. Only four derived forms appear in the dictionary, *akşın*, *gökşin*, *karaşın* and *sarışın*.

3.4.5 Category-Changing Denominal Nominals

In Appendix A, Tables 50 and 51 present the examples for category-changing denominal nominals (first NJD and JND; then NAD, AND, JAD and AJD).

We identify 18 relatively productive category-changing NND affixes. We consider them in two groups: NJD-JND and NAD-AND-AJD-JAD.

Only one overt morpheme forms property names based on an adjective. *-lIK* (*acayiplik* 'strangeness', *iyilik* 'kindness', *çabukluk* 'quickness') is an almost fully productive NJD affix.

-CIl (*bütüncül* 'holistic', *barışçıl* 'peaceful', *kötücül* 'malicious', *balıkçıl* 'kingfisher') exhibits polysemy, extensive semantic selection and appears in many dictionary entries. *-lI* (*istanbullu* 'istanbulite', *abartılı* 'exaggerated', *beşiktaşlı* 'beşiktaş fan', *kısa saçlı* 'short-haired', *üniversiteli* 'university student', *köylü* 'villager') shares many properties of *-CIl*. *-sIz* (*ağrısız* 'painless', *eşsiz* 'unique', *sınırsız* 'endless') completely lacks polysemy and does not seem apply any semantic selection criteria.

-CI appears in only one case, *ablacı* 'lesbian', so its productivity as a category-changing NND marker cannot be established. It indicates an affinity towards the base noun, sharing this semantic relation with *-CIl*.

-IncI (*ikinci* 'second'), *-Iz* (*ikiz* 'twin') and *-şAr* (*ikişer* 'two each'), apply on number bases and form adjectives with completely regular semantics.

∅ is an interesting morpheme. Zero-derivation, or conversion, is widely accepted in the literature as part of DM. Researchers often recognize that Turkish nouns and adjectives are used interchangeably, but they avoid attributing this phenomenon to the existence of conversion. While it is outside the scope of this thesis, we include it in our list, for the sake of completeness. The zero-morpheme acts as a DM in NJD (*üniversiteli* 'university student') and AJD (*arabasız* 'without a car', parasız 'without any money') capacities, arguably subcategorizing for an adjective base. Generally, we can say that Turkish adjectives can be readily used as nouns, but conversion in the opposite direction is not straightforward. There is at least one case where an adjective seems to be converted into a verb, *kuru-∅-* 'to dry', but Nişanyan (2021) points out that *kuru-* was initially derived from the OT *kurug* 'dry' by an overt morpheme, which later eroded. There is no evidence to support a VJD_∅ morpheme.

The rest of category-changing NND affixes are in NAD, AND, AJD and JAD classes. Majority of the derivational relations are in the AND class, while NAD, AJD and JAD are only represented once, twice and once in the matrix, respectively.

-lIK can derive nouns from adverbs with plenty of interesting cases (*ayrıcalık* 'privilege', *farkındalık* 'awareness', *gündelik* 'daily pay', *öncelik* 'priority', *yerindelik* 'legitimacy'). It applies after inflectional affixes, but in most of these examples, the inflected form has been lexicalized with non-compositional meaning.

The *-CA* group forms adverbs from nouns with 4 productive meanings (*kafaca* 'mentally', *bence* 'in my opinion', *evce* 'with the whole family', *arkadaşça* 'friendly'). It is also possible to derive adverbs from adjectives, but we have only observed adjectives formed by *-CA* (*açıkça* 'openly', *bencilce* 'selfishly') and *-sIz* (*parasız* 'without money', *arabasız* 'without a car').

A large group of ancient and contemporary case markers continue to form denominal adverbs (*içeri* 'inside', *akşamleyin* 'in the evening', *yazın* 'in the summer', *ayakta* 'standing', *sıradan* 'ordinary'). It is possible to dismiss some of these as mere application of case markers. However, some items are listed in the lexicon for good reason. For instance, *sıradan* 'ordinary', *içten* 'sincere', *dünyada* 'on earth' and *güzellikle* 'gently' express non-compositional semantics that cannot be simply attributed to case marking. OT case markers *-Arı* / *-rA* and *-In* / *-lAyIn* only apply on a few stems and are not considered productive anymore.

Finally, *-ki* (*bugünkü* 'of today', *sonraki* 'next', *şimdiki* 'current', *deminki* 'of a moment ago) is the only morpheme forming adjectives from adverb bases. *-ki* has several interesting properties. It is a unique morpheme that takes phrasal scope, forms an adjectival phrase from a locative or temporal adverbial phrase, then is converted into a noun clause and starts to represent the head object. Even if the final form already contains a case marker, since it is a noun, it may take a second (or third...) one. We do not go into the many functions of *-ki*, but refer the reader to Göksel and Kerslake (2005) for further discussion.

The most interesting property of *-ki* is its resetting the application order to N-0. It acts like a cliticized adjective that is morphosyntactically independent, yet happens to have no free form. Essentially, its behavior is similar to the auxiliary. *-ki* does in nominal morphology, what the auxiliary does in verbal morphology.

Derivational operations forming adverbs may also be studied in terms of thematic relations, but so far we have observed very few examples. This is because the thematic relations most relevant to adverbs, thematic relations of the time dimension, are mostly realized by syntactic constructions. For instance, the two constructions *-DAn beri* (*sabahtan beri* 'since morning') and *-A kadar* (*akşama kadar* 'until evening') realize source time and goal time relations, respectively.

Unlike deverbal nominals and denominal verbs, where multiple morphemes derive forms with similar semantics, denominal nominals rarely overlap in meaning. Most semantic relations indicated by denominal nominals are realized by a single dedicated affix. This might be because deverbal nominals and denominal verbs operate within a smaller (perhaps also discrete) space structured by thematic relations, while denominal nominals operate in an unrestricted space.

3.5 Discussion

This chapter presents our review and reclassification of Turkish DM and how this effort lead us to CdS. Our contributions are fourfold.

Section 3.1 gives a conventional overview of Turkish grammar. This overview is complemented with insights from OT, so that the underlying form and meaning relations between morphemes can be discovered.

In Section 3.2, we collect evidence reviewing cases where morphemes take on interesting functions. In some of these cases, the line between IM and DM blurs. Reviewing these cases also helps us draw a boundary around the scope of this thesis. Interesting cases reviewed in this section demonstrate that contrary to popular belief, DM is mostly regular and compositional.

Section 3.3 discusses fusion, allomorphy, polysemy and synonymy to point out the form and meaning relations between different morphemes. For deverbal nominals and denominal verbs, thematic relations are considered. As a result of this investigation, we propose two dimensions for classifying Turkish DM.

Section 3.4 gives the actual classification of morphemes. This classification provides a cleaner basis for computational analysis.

The large number of morphemes per word results in many alternative ways for segmentation. The large number of homonymy and polysemy results in many alternative ways for lexical selection. These facts justify and require CdS for an adequate representation of Turkish DM.

CHAPTER 4

REPRESENTING MEANING

In this chapter, we look into two methods for meaning representation: distributional semantics and categorial grammar.

We study both a symbolic and a sub-symbolic architecture to see the extent to which DM can be represented on different settings. On one hand, the irregular nature of DM resists an entirely rule-based representation; on the other hand, we have to make use of rules to some extent, in order to make sense of the underlying structure of morphology.

4.1 Distributional Semantics

Distributional semantics (DS) has been a developing area of research for a long time. A wide variety of applications have been developed especially based on word embeddings. There has also been substantial work on generating vector representation on other levels, such as morphemes. In this chapter, we look for ways to apply the distributional approach to DM semantics.

Distributional semantics offers many benefits to a linguistic investigation. First, there is a considerable amount of data in the form of written corpora. Second, distributional semantics is not concerned with variations among individual speakers and over time; it simply aggregates a huge amount of data to construct a rough approximation of the semantics for nearly all lexical items. Third, it provides a numerical representation for semantic content which is convenient for computational studies. Finally, it relies on a single central hypothesis, which has been validated by the vast literature on word embeddings. The hypothesis was defined by Harris (1954):

> Distributional Hypothesis: Linguistic items with similar distributions have similar meanings.

The distributional hypothesis suggests that semantic similarity correlates with distributional similarity. It is a powerful claim, and has been shown to hold to a great extent. Based on this claim, word embeddings transform the sparse, discrete space of distributional semantics to a more compact, continuous one. They also offer a completely new perspective from which one can explore the nature of meaning. DS has been a revolutionary tool in many NLP applications.

Our aim is to produce vector representations for Turkish DM, assess their consistency and make use of this information in our study of the Conventionalized Structure (CdS). Turkish DM carry distinctive meaning content that blends with base semantics. It is possible to imagine Turkish affixes as little words with their own meaning and syntactic category, which happen to appear exclusively in bound form. (In fact, we have already observed in our study of Old Turkic that many affixes can be traced etymologically to clitics and separate word forms.) Turkish being an agglutinating language, it should be possible to find clear vector representations for individual affixes.

In the coming sections, we provide a short literature review on vector representation, our handling of the existing datasets and our assessment of the results. Crucially, we claim that it is possible to estimate affix embeddings by simple vector arithmetic. We argue the consistencty of our results by demonstrating that embeddings for the same affix, even if estimated from different base-lemma pairs, create compact clusters. This result also gives credit to our claim that affix semantics is generally regular.

4.1.1 Literature Review

After the seminal work of Mikolov et al. (2013), there have been numerous studies on the distributional semantics on the word-level. However, a much smaller number of studies investigate the distributional properties of items above or below the word-level following the same principles.

Attempts were made towards generating embeddings based on characters Bojanowski et al. (2017), morphemes Cotterell and Schütze (2019), and syllables Choi et al. (2017), Üstün et al. (2018), Şahin and Steedman (2018). While the results from these studies are competitive, character and morpheme-level embeddings present important drawbacks. Morpheme-based models require extensive pre-processing, manual annotation or morphological analysis (Choi et al., 2017). Moreover, character and syllable-based representation do not actually model distributional semantics, because these units do not really possess semantic content.

Some other studies analyze relations between word pairs using unsupervised learning. Zargayouna et al. (2017) examine analogy pairs to study semantic relations between word embeddings. They identify semantic relations (content-container and component-whole etc.) using similarity metrics. Gladkova et al. (2016) investigate morphological and lexical relations. They present a large test set for evaluating word embeddings. Musil et al. (2019) study Czech DM, obtaining word embeddings using skip-gram and neural machine translation models. With the help of clustering, they test the hypothesis that differences between word vectors reflect derivational relations. Rosa and Žabokrtský (2019) try to distinguish inflection from derivation based on character and word embedding similarity.

Cui et al. (2015) attempt to improve word embedding quality by introducing morphological and contextual info into the model. In a similar vein, Jurdzinski (2017) feeds both base forms and grammatical forms to the model and reports improvement on word analogy tests, compared to the benchmark by Pennington et al. (2014).

Musil et al. (2019) look closely into the Czech derivational morphology, using both skip-gram and neural machine translation models to obtain word embeddings. They experiment with the hypothesis that differences between word vectors showing the same derivational relations would cluster together. Botha and Blunsom (2014) also base their extensive computational analyses on the idea that addition could adequately model compositionality in morphological relations. They give the following examples to demonstrate:

$$\overrightarrow{imperfection} = \overrightarrow{im} + \overrightarrow{perfect} + \overrightarrow{ion}$$

$$\overrightarrow{perfectly} = \overrightarrow{perfect} + \overrightarrow{ly}$$

Lazaridou et al. (2013) review several composition methods used for constructing phrase representations from their parts and applies these to the problem of constructing representations for derived forms. The "fulladd" and "lexfunc" composition methods turn out to obtain the best results on the analogy tasks. For the "fulladd" method, component vectors are first multiplied with weight vectors and summed afterwards. During our exploration, we adopt the simpler method of Musil et al. (2019) and Botha and Blunsom (2014), and find it adequate.

4.1.2 Challenges

Using word embeddings for generating affix embeddings presents several challenges. The first and largest challenge is to construct a large enough dataset that contains base-lemma pairs for all affixes of interest. For instance, estimating the embedding for *-CI* requires pairs such as *kitap* 'book' - *kitapçı* 'bookseller'; *su* 'water' - *sucu* 'water seller'. We refer to TDK (2019) as the ultimate source for accepted words. Even with a very large dataset, there are several issues we have to contend with.

(77) Issues that prevent complete coverage and perfect clustering

a. Lexicalized derived forms that are not recognized by TDK (2019)

b. Derived forms recognized by TDK (2019) but missed by annotators

c. Misannotated derivational relations (Typos, incorrect categories etc.)

d. Homographs, alternative meanings

e. Some affixes simply having irregular / non-compositional meanings

f. Missing vector representations for some bases or lemmas

g. Vector embeddings for rare words being unreliable

h. Small sample size (especially with Gaussian Mixture Models)

The first two prevent us from achieving a complete coverage of Turkish derivational relations. The second and the third are issues are due to human error on our part. Examples from the fourth to the seventh are problems arising due to the nature of language and the method of word embeddings. The last one is related to the way some clustering algorithms work, explained in later sections.

Regarding the first issue, we believe TDK (2019) provides a reliable common ground. In order to minimize the effect of the second and the third issues, the initial annotated dataset was reviewed by a second annotator. As we moved to the clustering and visualization parts of our study, corrections were made whenever necessary.

The fourth and the fifth issues must be addressed effectively. Since word embeddings are created based on the orthographic form of a word, representations for derived words and some inflected forms can share the same word vector. Yüret and Türe (2006) give a good example to this. The word *masalı* can have three different morphological analyses: *masal-ı* 'tale-ACC', *masal-ı* 'his tale' and *masa-lı* 'possessing a table'.

This issue can be remedied by contextual embeddings. Contextual embeddings recognize and represent semantic differences of the same form in different contexts (Liu et al., 2020). Arora et al. (2020) demonstrate that contextual embeddings perform better than classical methods, and by a large margin in some tasks and datasets. Classical word embeddings are aggregations over all uses of a derived form. We leave that extension out of the scope of this thesis, because homographs of the kind described above are relatively rare, especially considering that the bases and lemmas we use are never inflected.

The fifth issue is not that problematic, in fact, it is actually useful in our theoretical analyses. We already expect some affixes to have irregular meaning. Derived forms may also assume additional meaning content in the lexicalization process. By estimating vector representations using a large number of base-lemma pairs, we are able to observe the variation and irregularity of affix semantics.

The sixth and the seventh issues both point to a possible deficiency of the data used for constructing word embeddings. For an agglutinating language such as Turkish the type-token ratio is much larger than in other languages. Therefore, we would not expect all possible forms to be present in the corpus. The large type-token ratio also reduces the number of instances for each individual affix, thus diminishing the reliability of the estimation. These issues are to a certain extent unavoidable. We explore two sources of pre-trained word embeddings and use the better one in further investigations.

Finally, the last issue describes the problem of working with affixes that are less frequent than others. Since we are essentially working on the distributional and statistical properties of affix vector representations, the data on affixes with small sample sizes is insufficient and unreliable. We avoid this problem by focusing on relatively more productive affixes.

4.1.3 Data Preparation

Before we could start constructing vector representations for affixes, we needed to prepare large datasets for both inflectional and derivational affixes. Our primary source for derivational relations is TrLex presented in Aslan et al. (2018) described in Section 3.3.1. It is based on an annotation effort over TDK (2019). We revised the annotated properties according to our own classification described in Chapter 3. This dataset supplies the base-lemma pairs and the corresponding affix. In total, we have 34588 base-lemma pairs. 6527 are the result of inflection, 28061 are the result of derivation. 174 unique derivational affixes are included in the list, while only 118 of them appear in 5 or more dictionary entries.

We consult TDK (2019) for the correct categories and meanings of words. We check the true stems with Nişanyan (2021) and Eyüboğlu (2017). For a few derived forms, Nişanyan (2021) or Eyüboğlu (2017) suggest bases not in contemporary use. It is impossible to find vector representations for such words, so these pairs are discarded.

While preparing the base-lemma dataset, we use the coding scheme suggested by Bozşahin (2018). This scheme was explained in Chapter 3.

We use two publicly available pre-trained word embedding datasets.The Polyglot project Al-Rfou et al. (2013) provide word embeddings for more than 100 languages based on Wikipedia. Grave et al. (2018) cover 157 languages including Turkish. They also use Wikipedia datasets and train on fastText with 300 dimensions.

For each base-lemma pair, we first obtain the embeddings for the two words. Then we subtract the base's vector from the lemma's. For instance, if we wanted to find the word vector of *-lIK*, we find it by vector subtraction of the embedding for *kuru* 'dry' from the embedding for *kuruluk* 'dryness'. We do this for a large number of base-lemma pairs for each affix. The collection of these samples constitutes the basis of our exploration.

$$\overrightarrow{lIK} \sim \overrightarrow{kuruluk} - \overrightarrow{kuru}$$

We expect that estimations from different base-lemma pairs for the same affix should be similar. This expectation relies on the well-accepted principle that Euclidean distance on DS space is a consistent reflection of semantic relationships.

$$\overrightarrow{woman} - \overrightarrow{man} \sim \overrightarrow{queen} - \overrightarrow{king}$$

$$\overrightarrow{woman} - \overrightarrow{queen} \sim \overrightarrow{man} - \overrightarrow{king}$$

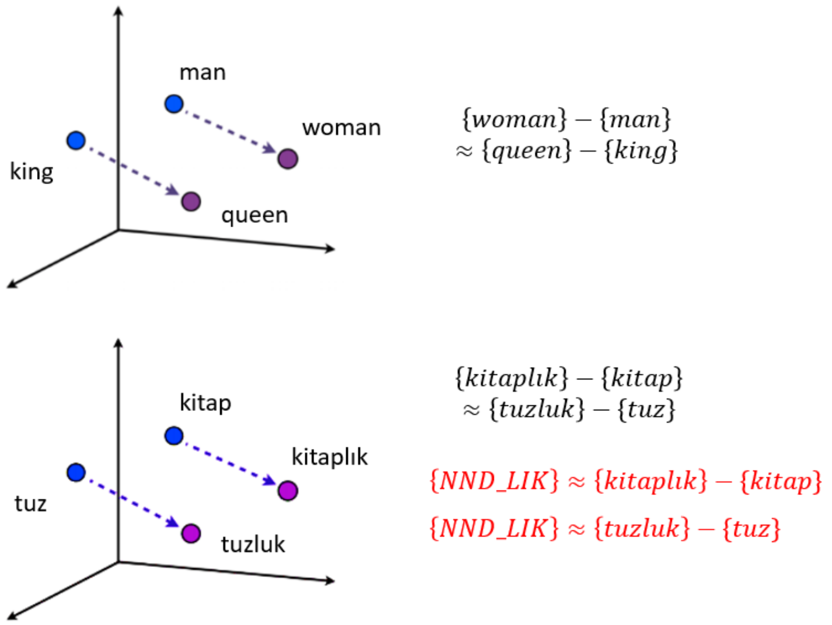
$$\overrightarrow{suluk} - \overrightarrow{su} \sim \overrightarrow{tuzluk} - \overrightarrow{tuz}$$

$$\overrightarrow{suluk} - \overrightarrow{su} \sim \overrightarrow{lIK} \sim \overrightarrow{tuzluk} - \overrightarrow{tuz}$$

As expected, this exercise does not obtain an estimation for some base-lemma pairs. We need to find both words in the embeddings datasets for the pair to be useful. Also, there are some affixes with fewer than 20 base-lemma pairs. Without a sufficient sample, the estimation cannot be considered accurate. We removed such cases.

For verbs, we use the infinitive form to avoid homography with 2nd person singular imperative. For instance, *ayır-* 'divide' was converted to ayırmak 'to divide'. Infinitive forms could usually be found in the pre-trained word embeddings and we expect them to be the most accurate representation of the underlying verb's distributional semantics.

For comparison, we manually prepared a dataset of base-lemma pairs with inflectional relations. This set is comprised of NNI and VVI class affixes (case, plural, possessive, voice, tense/aspect/modality and copular). It contains 3141 entries for 26 affixes, 24 of which have over 75 instances. Inflectional relations display more regular distributional properties and clearer clusters. Inflectional relations serve



$\{woman\} - \{man\}$
$\approx \{queen\} - \{king\}$

$\{kitaplık\} - \{kitap\}$
$\approx \{tuzluk\} - \{tuz\}$

$\{NND\_LIK\} \approx \{kitaplık\} - \{kitap\}$
$\{NND\_LIK\} \approx \{tuzluk\} - \{tuz\}$

Figure 7: Idea behind computing affix embeddings

to validate our methodology and the results from the "messy" derivational relations could be compared and contrasted with them.

4.1.4 Unsupervised Learning

In order to test Musil et al. (2019)'s hypothesis that similar morphological relations could be represented accurately by the difference between base and lemma word embeddings, we used three different clustering algorithms. Each algorithm has strengths and weaknesses depending on the distribution of items over the vector space. Hierarchical methods either start by assigning each item to its own cluster and joining two closest sets at each step (agglomerative clustering); or they start with all items in one cluster and divide it into two at each step (divisive clustering).

(78) Unsupervised learning algorithms to explore estimated affix embeddings

a. K-means algorithm (KM): We use cosine and Euclidean distance metrics. KM is successful in clustering non-intersecting convex groups that are similar in size.

b. Agglomerative Clustering (AGG): We use Euclidean distance and Ward linkage type. AGG is better with clearly separate groups of any shape, provided they show tight intra-cluster cohesion.

c. Gaussian Mixture Models (GMM): GMM estimates the distributional properties of each individual cluster, therefore handling possibly intersecting groups of varying sizes much better. The groups still need to be distributed according to the same family of distributions.

Since these methods take into account the distance between all pairs of items, they make good use of local information and work best with tightly connected clusters. The shapes of clusters do not matter much with hierarchical methods. Variation in the size of clusters may result in poor performance, depending on the preferred linkage type.

k-means and k-medoids algorithms choose k points (more specifically, k data points in the latter case) to act as spatial centers of clusters and assign each datapoint to the closest cluster. Clusters obtained this way are always convex (if unbounded edges are assumed not to violate convexity), so clusters of varying shapes may not be identified well. Using these algorithms, one also assumes cluster sizes to be similar, otherwise assigning items to the closest center would not be meaningful.

Agglomerative clustering is the more popular hierarchical clustering algorithm, since join operations are more transparent than division operations. They basically follow the local similarity chains.

Probabilistic clustering methods, as the name suggests, calculate the probability of assigning each datapoint to each cluster. In some methods, hard assignment is avoided if sufficient confidence cannot be achieved. Gaussian mixture models (GMM) in this group can be considered a generalization of the k-means approach.

Instead of simply assigning items to the closest center, GMM estimates the parameters for normally distributed subpopulations in the data. Means of these subpopulations act as cluster centers and their standard deviations determine the "range" of the cluster. After the estimation of parameters, GMM calculates the probability of each item falling in each cluster. Highest probability assignment is accepted.

GMM is robust against the variation of cluster size and it does not rely on within-cluster density. Moreover, there is a reasonable chance for correctly assigning datapoints in cluster intersections. It still assumes that subpopulations are normally distributed; consequently, clusters must be convex.

Since word embeddings are high-dimensional, there is no easy way of observing the patterns by which datapoints are distributed. t-SNE (t-Distributed Stochastic Neighbor Embedding) due to Van der Maaten and Hinton (2008) offers a way of dimension reduction and visualization. Unlike Principal Component Analysis, which transforms the whole vector space preserving as much information as possible, the objective in t-SNE is to place similar datapoints closer to each other and dissimilar ones distant from each other. It is a probabilistic algorithm. t-SNE has been employed in several studies on word embeddings with considerable success. We apply t-SNE mainly for visually evaluating the clustering results.

A number of studies visualize the results using t-SNE. Cotterell and Schütze (2019) point out that t-SNE shows different POS-tags in different clusters; plural forms and single forms are also separated. Hamilton et al. (2016) study the change in word meanings by a diachronic study of word embeddings and use t-SNE for visualizing the path of change for selected words. Peters et al. (2018) use t-SNE and heatmaps in order to visualize the syntactic category labels and pairwise similarities for their word embeddings. Bamler and Mandt (2017) work with time-stamped text data; they propose a dynamic word embeddings model and use dynamic t-SNE for dimension reduction. Gouws and Søgaard (2015) train a bilingual word embedding model for cross-language POS-tagging. They use t-SNE to demonstrate the clear clustering of POS classes. Liu et al. (2018) use several visualization techniques to make explicit the relations between analogy pairs.

For many of these methods there are efficient implementations on common platforms. scikit-learn made available by Pedregosa et al. (2011) is a Python library that offers implementations for many machine algorithms. The code is tested and verified by thousands of users around the world. We develop a pipeline that takes pre-trained word embeddings and base-lemma-suffix data and applies the selected clustering and/or visualization methods. In the process, data preprocessing is carried out and statistics are collected.

4.2 An Exploration on DS

The exploration is implemented in four stages. First, we start by importing base-lemma pairs (based on TrLex) and pretrained word embeddings (polyglot or fasttext). Second, we estimate affix embeddings. Third, we apply clustering algorithms (KM, AGG or GMM) to make explicit the similarity and dissimilarity relations between estimated datapoints. Finally, we obtain confusion matrices and apply visualization algorithms (t-SNE or heatmap) to assess the success of clustering, and indirectly, the success of estimated embeddings.

Since preliminary trials showed that fasttext embeddings performed significantly better than polyglot, we carry out all other trials using fasttext.

Each clustering algorithm requires several parameters to be set. Some of these parameters are common across many algorithms, while others only apply to a particular algorithm. To a great extent, sklearn uses the same set of parameters for many of its algorithms. Generally, we set the number of clusters equal to the number of affixes taking part in the trial. We allow large numbers of iteration and initialization for the algorithms to be able to escape local minima. We observe that convergence can be usually achieved.

We run GMM with four additional parameters: covariance type (full), initial parameters (kmeans), warm start (True) and metric (euclidean). For AGG, we prefer linkage type (ward). With t-SNE, we work with 2 dimensions and leave perplexity at its default value.

Confusion matrices show the quality of match between clustering results and true affix classes. Ideally, results of the clustering algorithm would produce clearly separated clusters, where each cluster only contains instances of a single affix. Due to the challenges explained in Section 4.1.2, this is not possible in practice. Still the homogeneity of each cluster is an indication of successful clustering, which in turn indicates successful estimation of affix embeddings. We track two metrics with respect to this: the entropy of clustering results and the ratio of clusters where a single affix achieves great majority.

Visualization helps us when qualitatively evaluating the success of clustering. Both heatmap and t-SNE are good ways to immediately assess the homogeneity of clusters with respect to true classes.

We conducted several trials with different algorithms and parameter settings. Table 11 presents the list of affixes and the number of base-lemma pairs for each affix in the baseline trial. We chose inflectional affixes with distinct semantic content in order to maximize the success of clustering.

For the baseline trial we run GMM with 6 clusters. The resulting confusion matrix is given in Table 12. Each row represents a cluster and each column a true affix class. Cells show the ratio of a specific affix to all instances in a specific cluster. For instance, 0.99 indicates that 99% of the items in the cluster

Table 11: Baseline trial for affix DS clustering

Affix	Number of Instances
NNI_GEN	100
NNI_PLU	100
VVI_ARAK	100
VVI_CAUS	88
VVI_NEG	87
VVI_PASS	79

Table 12: Confusion matrix for the baseline trial for affix DS clustering

NNI_GEN	NNI_PLU	VVI_ARAK	VVI_CAUS	VVI_NEG	VVI_PASS
1,00	0,00	0,00	0,00	0,00	0,00
0,00	1,00	0,00	0,00	0,00	0,00
0,00	0,00	1,00	0,00	0,00	0,00
0,00	0,01	0,00	0,95	0,02	0,04
0,00	0,00	0,00	0,01	0,99	0,00
0,00	0,00	0,00	0,09	0,00	0,91

belongs to the affix class of that column. Affixes in the baseline trial are divided into homogeneous clusters almost perfectly.

4.2.1 Most Frequent Affixes

The first trial is conducted on the 6 most frequent affixes: JND_LI, JND_SIZ, NJD_LIK, NND_CI, NND_LIK and NVD_MA. Each of these affixes possess a large sample of over 1000 instances for which word vector representations could be found in the pre-trained word embedding models. They are also arguably the affixes that first come to mind as examples of Turkish DM.

(79) a. Affixes with over 1000 instances

b. NVD_MA: *affetme* ‘forgiving’, *deneme* ‘trying’ (2534 instances)

c. JND_LI: *azimli* ‘resolute’, *başarılı* ‘successful’ (2009 instances)

d. NND_CI: *aristotelesçi* ‘follower of Aristo’, *kitapçı* ‘bookseller’ (1287 instances)

e. NJD_LIK: *anlaşmazlık* ‘conflict’, *barışseverlik* ‘pacifism’ (1046 instances)

f. NND_LIK: *arkadaşlık* ‘friendship’, *kitaplık* ‘bookshelf’ (1036 instances)

g. JND_SIZ: *bilinçsiz* ‘unconscious’, *çaresiz* ‘desperate’ (1003 instances)

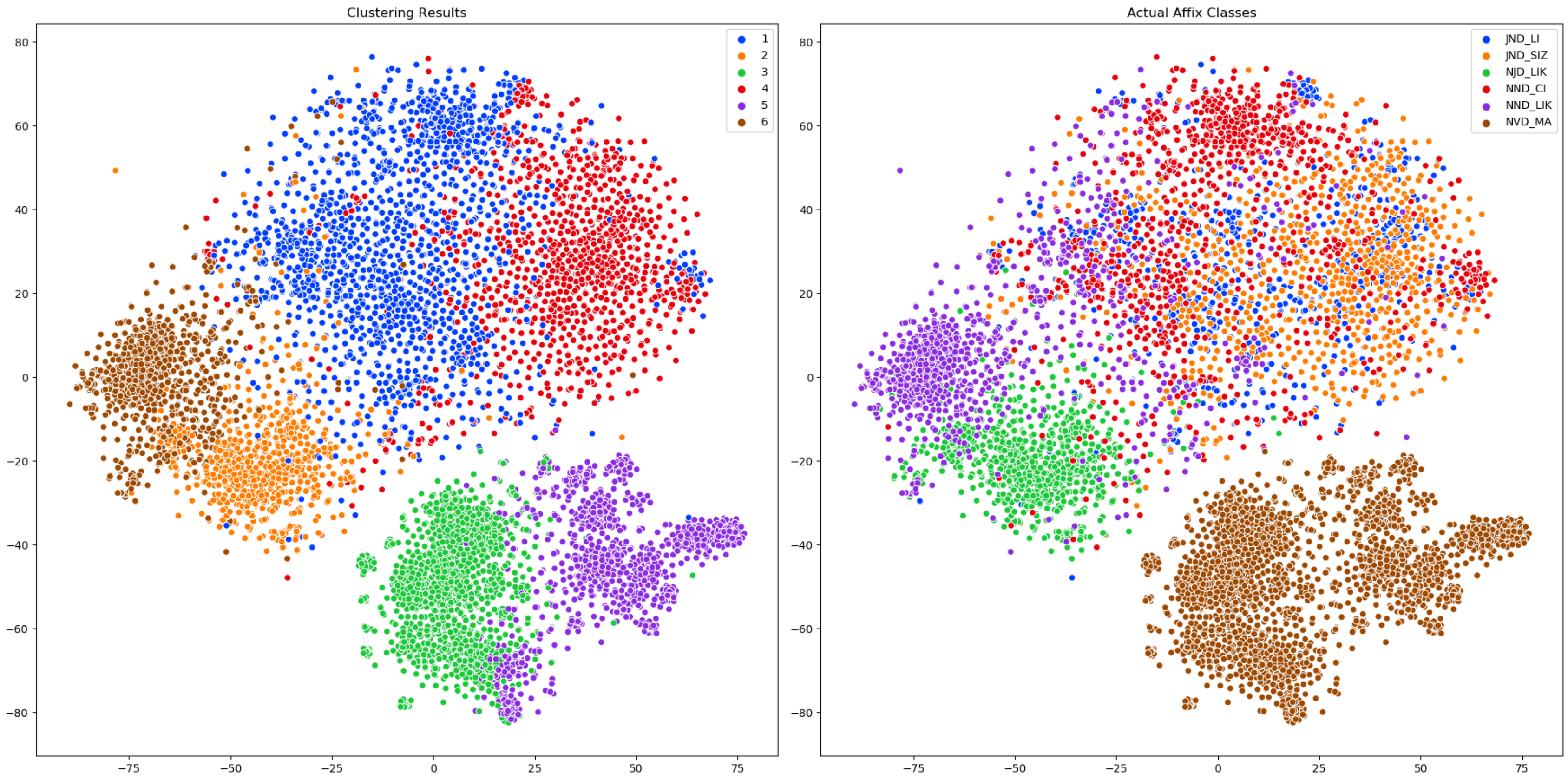


Figure 8: Distribution of affixes appearing in over 1000 instances

Table 13: Confusion matrix for affixes with over 1000 instances

	JND_LI	JND_SIZ	NJD_LIK	NND_CI	NND_LIK	NVD_MA
1	0,35	0,13	0,01	0,43	0,09	0
2	0,01	0	0,9	0,01	0,08	0
3	0	0	0	0	0	1
4	0,56	0,33	0	0,1	0,01	0
5	0	0	0	0	0	1
6	0,01	0,02	0,21	0,01	0,75	0

NVD_MA is the most frequent affix overall. It is productive over all Turkish verbs. Since it has such a large sample, it commands two clusters instead of one. It is perfectly separated from other affixes, as would be expected due to its NVD category. Another reason could be that it lacks a distinctive semantic content that is present in other affixes.

JND_LI is also a very productive affix, having over 2000 instances in our datasets. Along with JND_SIZ and NND_CI, it forms a large group, where the three affixes are not separated along clear lines. Nevertheless, comparing the two plots, we can observe that NND_CI could be mostly contained in one cluster, while another cluster is shared between JND_LI and JND_SIZ. This is a natural result, as JND_LI and JND_SIZ have the same base and lemma categories and they have very similar semantics. The only difference between the two is that they are on the opposite sides of polarity. Perhaps their similarity from the perspective of unsupervised learning algorithms could be evidence to the idea that a single dimension of a lexical item's semantic content (i.e. polarity) does not have a large effect on its representation distributional semantics. It is difficult, if not impossible, to find how exactly a single dimension is represented on the embedding vector, due to the distributed and superpositional representation of each dimension on word embedding vectors.

NND_LIK and NJD_LIK form another group, clearly separated from others and mostly separated from each other. Most studies in the literature do not distinguish between different categories of *-lIK*, as we did in the Section 3.3.5. Recognizing different classes of affixes in this way is syntactically motivated. The fact that the same morpheme consistently assumes different semantic content with different base and lemma categories lends more credibility to the view that affixes may occur in constellations of polysemous senses. For these two affixes, we can observe that their distributional semantics is distinct enough for them to largely occupy different clusters. In comparison to the pair JND_LI and JND_SIZ, which are different in form as well as in polarity, the clearer separation of NND_LIK and NJD_LIK reminds us that the form of a morpheme is not a factor in distributional semantics.

4.2.2 Inflection

Verbal inflectional affixes can be studied in several groups. We start with affixes that convey future or necessity modality.

We added VVI_XPAST and NND_LI+ to all remaining trials, so that the shapes and distributions on the plots can be compared against them. These three affixes display very distinguishable patterns, in reference to which the distribution of other clusters can be examined.

(80) a. Affixes of modality

b. VVI_XPAST: *anlamıştı* 'he had understood', *anlayacaktı* 'he would have understood' (278 instances)

c. NND_LI+: *karslı* 'from Kars', *hollandalı* 'Dutch' (178 instances)

d. NVD_ACAK: *giyecek* 'cloth', *içecek* 'drink' (21 instances)

e. VVI_TFUTR: *bilecek* 'he will know', *gidecek* 'he will go' (100 instances)

f. VVI_TNECE: *beklemeli* 'he should wait', *çalışmalı* 'he should work' (96 instances)

Table 14: Confusion matrix on affixes of modality

	NND_LI+	NVD_ACAK	VVI_TFUTR	VVI_TNECE	VVI_XPAST
1	0	0,17	0,83	0	0
2	0	0	0	0	1
3	1	0	0	0	0
4	0	0	0	0	1
5	0	0	0	1	0

VVI_XPAST and NND_LI+ form quite distinct clusters. Since the number of clusters is only five, three distinct groups of VVI_XPAST are consolidated into two clusters.

VVI_FUTR and NVD_ACAK largely overlap, which is not surprising considering that they have the same form and similar semantic content. Nevertheless, one could expect to see a clearer separation between the two, due to the difference in their base and lemma categories. The results may be pointing to a shortcoming of our methodology: Since the pretrained word embeddings only list the word forms and not their categories, it is not possible to distinguish between the finite verb *yiyecek* 'he will eat' and the noun *yiyecek* 'food'. Therefore, both VVI_FUTR and NVD_ACAK include *yiyecek* in their clusters, but the distribution of it is slightly different than VVI_TFUTR instances without a corresponding NVD_ACAK item.

VVI_TNECE is distinct but close to VVI_TFUTR, since they both convey modality. Still, VVI_TNECE has its own cluster.

A second trial on verbal inflection is concerned with the affixes indicating progressive aspect. Having observed from previous experiments that VVI_XADV_IP is relevant in this analysis, we include it as well.

(81) a. Affixes of progressive aspect

b. VVI_XPAST: *anlamıştı* 'he had understdood', *anlayacaktı* 'he would have understood' (278 instances)

c. NND_LI+: *karslı* 'from Kars', *hollandalı* 'Dutch' (178 instances)

d. VVI_XADV2: *anlayarak* 'understanding', *dokunarak* 'touching' (100 instances)

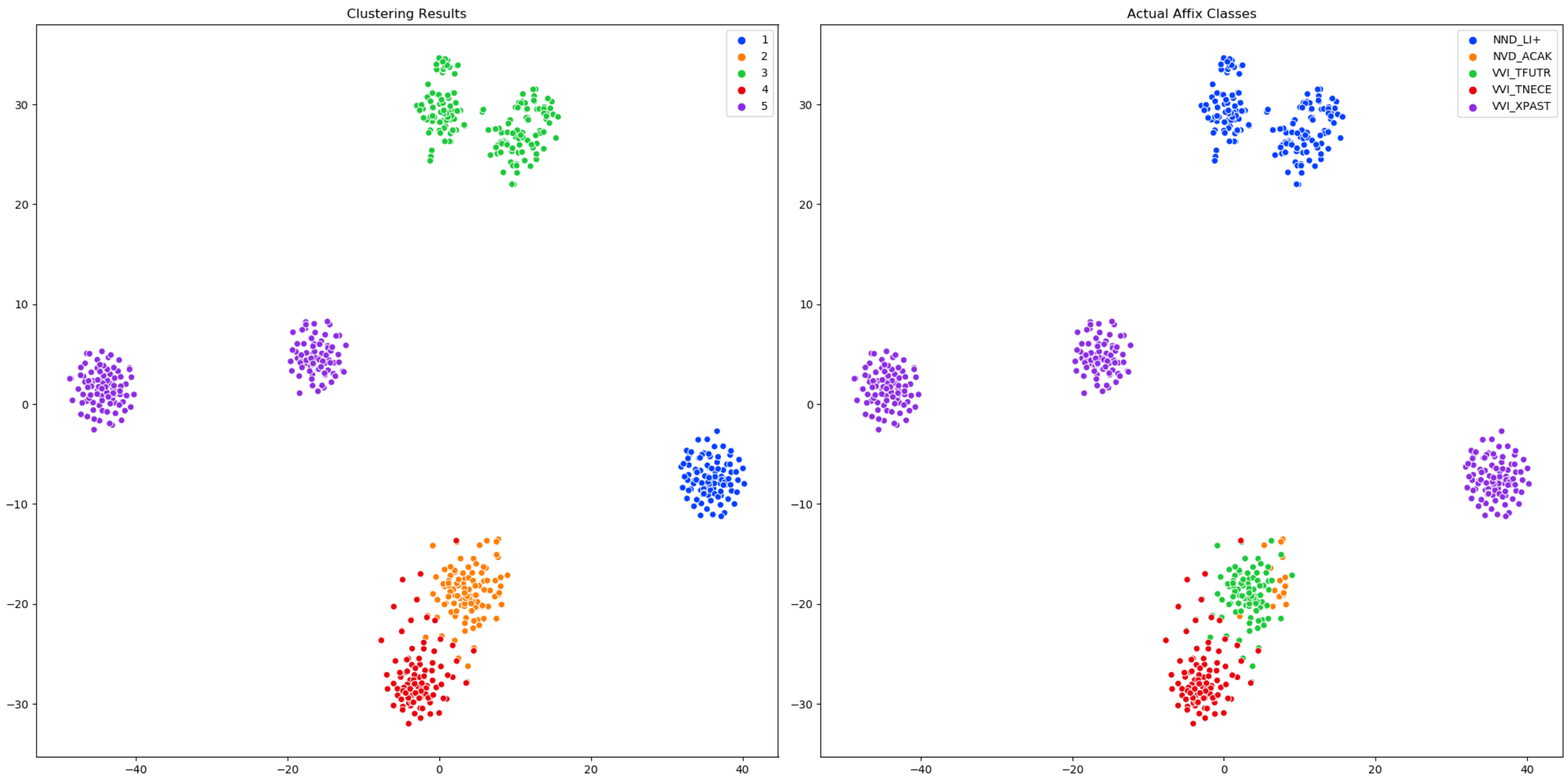


Figure 9: Distribution of affixes conveying future modality

e. VVI_XADV4_IP: *çalışıp* 'having worked', *deneyip* 'having tried' (100 instances)

f. VVI_TAORS: *cevaplar* 'he responds', *düzenler* 'he organizes' (100 instances)

g. VVI_TPROG: *akıyor* 'it flows', *giriyor* 'he enters' (100 instances)

Table 15: Confusion matrix for affixes of progressive aspect

	NND_LI+	VVI_TAORS	VVI_TPROG	VVI_XADV2	VVI_XADV4_IP	VVI_XPAST
1	0	0	0	0	1	0
2	0	0	0	0	0	1
3	1	0	0	0	0	0
4	0	0	0	0,99	0,01	0
5	0	0	0	0	0	1
6	0	0,45	0,55	0	0	0

Quite similar to the results of the previous experiment, we observe that VVI_XPAST and NNI_LI+ are clearly separated in their respective clusters. The aorist and the progressive fall into the same cluster. This is understandable considering the fact that these two aspects are often used interchangeably. Otherwise, all clusters are almost completely homogeneous.

There are several groups of affixes within nominal inflection. Case markers constitute one such group.

(82) a. Case markers

b. VVI_XPAST: *anlamıştı* 'he had understdood', *anlayacaktı* 'he would have understood' (278 instances)

c. NND_LI+: *karslı* 'from Kars', *hollandalı* 'Dutch' (178 instances)

d. NNI_OBJ: (100 instances)

e. NNI_GEN: (100 instances)

f. NNI_DAT: (100 instances)

g. NNI_LOC: (100 instances)

h. NNI_ABL: (100 instances)

i. NNI_INC: (98 instances)

NNI_LI+ and VVI_XPAST, as always, dominate their respective clusters. However, case markers do not seem to separate at all. Strangely, NNI_GEN and NNI_OBJ share the cluster number 4, while NNI_ABL, NNI_DAT, NNI_INC and NNI_LOC are distributed in mixed fashion among clusters 1, 6 and 7.

It is difficult to establish a definite reason for this clear difference between two groups of case markers. Although genitive case is often classified as an inherent case, it might have to do with the structural

Table 16: Confusion matrix for case markers

	NND_LI+	NNI_ABL	NNI_DAT	NNI_GEN	NNI_INC	NNI_LOC	NNI_OBJ	VVI_XPAST
1	0,02	0,19	0,24	0,09	0,19	0,15	0,12	0
2	0	0	0	0	0	0	0	1
3	1	0	0	0	0	0	0	0
4	0	0	0	0,53	0	0,04	0,44	0
5	0	0	0	0	0	0	0	1
6	0	0,16	0,09	0,08	0,27	0,34	0,06	0
7	0	0,3	0,31	0,01	0,19	0,13	0,06	0
8	0	0	0	0	0	0	0	1

vs. inherent case distinction. If structural cases are semantically empty and inherent cases are not, distributional semantics might have been able to capture the presence and lack of meaning in these case markers. However, in that case, we would still expect NNI_ABL, NNI_DAT, NNI_INC and NNI_LOC to be separated at least to some extent. The t-SNE plots show no significant concentration of instances for any of these affixes. Also, we observe that usually all case inflections of a word fall into one cluster, making one wonder whether case markers contribute just a small extension to the length of the base vector, in the same direction.

4.2.3 Deverbal Nominals

Another interesting group of affixes is the deverbal nominals. We present the analyses on two groups: affixes forming the agent of an action and affixes forming the name, patient or theme of an action.

Several affixes in the first group are similar etymologically, syntactically and semantically; their origins and similarities are discussed in Section 3.3.3 and Section 3.3.4.

(83) a. Affixes indicating the agent of an action

b. VVI_XPAST: *anlamıştı* ‘he had understdood’, *anlayacaktı* ‘he would have understood’ (278 instances)

c. NND_LI+: *karslı* ‘from Kars’, *hollandalı* ‘Dutch’ (178 instances)

d. JVD_AN: *bilinen* ‘known’, *geçen* ‘previous’ (13 instances)

e. NVD_AN: *açıklanan* ‘declared’ (31 instances)

f. NVD_GAC: *ayraç* ‘separator’, *bağlaç* ‘conjunction’ (40 instances)

g. JVD_GAN: *alışkan*, *bitişken*, *değişken* (58 instances)

h. NVD_GIC: *başlangıç* ‘beginning’, *dalgıç* ‘diver’ (5 instances)

i. JVD_GIN: *bitkin* ‘exhausted, *dalgın* ‘absent-minded’ (89 instances)

j. NVD_MAN: *araştırman* ‘researcher’, *yönetmen* ‘director’ (23 instances)

Table 17: Confusion matrix for affixes indicating the agent of an action

	JVD_AN	JVD_GIN	NND_LI+	NVD_AN	NVD_GAC	NVD_GAN	NVD_GIC	NVD_MAN	VVI_XPAST
1	0	0	0	0	0	0	0	0	1
2	0	0,53	0	0	0,03	0,45	0	0	0
3	0	0	1	0	0	0	0	0	0
4	0,27	0,02	0	0,62	0	0,08	0	0	0
5	0	0	0	0	0	0	0	0	1
6	0	0	1	0	0	0	0	0	0
7	0	0,51	0	0,01	0,01	0,21	0	0,26	0
8	0	0,1	0	0	0,74	0,02	0,1	0,04	0
9	0	0	0	0	0	0	0	0	1

NND_LI+ and VVI_XPAST form their own clusters, as always. JVD_GIN and JVD_GAN share two clusters, accompanied by NVD_MAN in one of those. Nearly all instances of JVD_AN and NVD_AN fall into one cluster, along with some JVD_GAN. Again nearly all instances of NVD_GAC and NVD_GIC fall into one cluster, along with some JVD_GIN and NVD_MAN. The difference between affixes' shares in a cluster is due to the difference in their numbers of available instances, so NVD_GAC and NVD_GIC could be said to have the same status in cluster 8. The same idea applies to clusters 2, 4 and perhaps 7. The consistency in these clusters cannot be attributed to coincidence. In Chapter 3, we devise groups of affixes based on their form, semantics and etymology. Clusters we observe in these trials largely coincide with our judgments.

The second group of NVD affixes we consider forms names, patients or themes of actions.

(84) a. Affixes indicating the name, patient or theme of an action

b. VVI_XPAST: *anlamıştı* 'he had understdood', *anlayacaktı* 'he would have understood' (278 instances)

c. NND_LI+: *karslı* 'from Kars', *hollandalı* 'Dutch' (178 instances)

d. NVD_GA: *bölge* 'area', *gösterge* 'indicator' (22 instances)

e. NVD_GI: *döngü* 'loop', *kurgu* 'fiction' (89 instances)

f. NVD_I: *anı* 'memory', *bildiri* 'notice' (110 instances)

g. NVD_IM: *akım* 'flow', *çağrışım* 'association' (265 instances)

h. NVD_MA: *hastalanma* 'getting sick', *ilaçlama* 'disinfestation' (2534 instances)

i. NVD_YIS: *atlayış* 'jump', *biniş* 'riding' (298 instances)

NVD_YIS indicates the manner in which an action is carried out, so it has a distinct character than both NVD_MA and NVD_IM. This is reflected on the clustering results.

It is difficult to analyze this group, because NVD_MA has a very large set of instances and it covers a wide area. Strangely, instances of NVD_MA seem to concentrate around several centers. The large area covered by NVD_YIS mostly coincides with NVD_MA, but it is distributed much more homogeneously. Patterns for other affixes are hard to identify, due to their smaller samples.

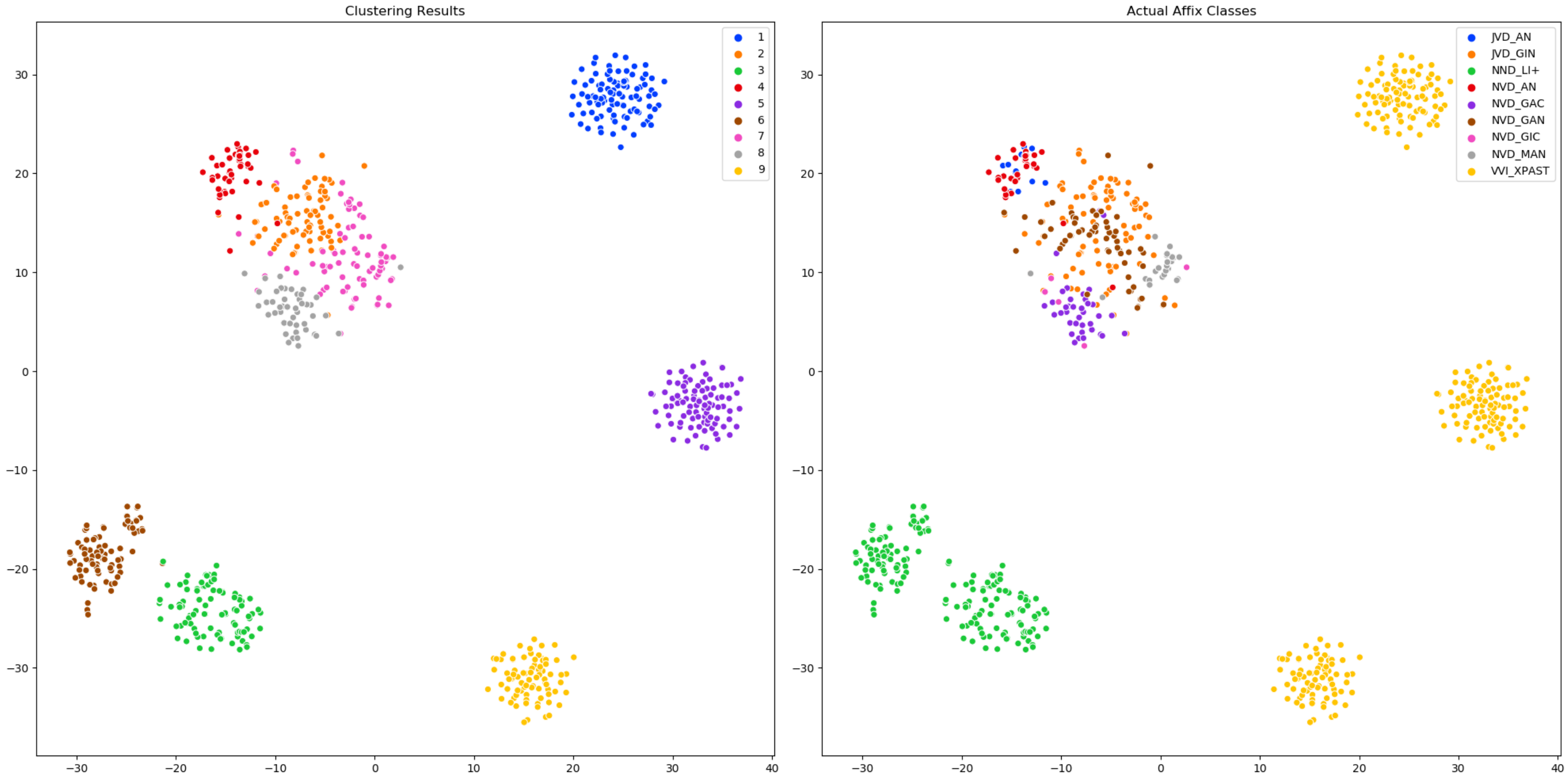


Figure 10: Distribution of affixes indicating the agent of an action

Table 18: Confusion matrix for affixes indicating the name, patient or theme of an action

	NND_LI+	NVD_GA	NVD_GI	NVD_I	NVD_IM	NVD_MA	NVD_YIS	VVI_XPAST
1	0	0,03	0,13	0,16	0,35	0,16	0,18	0
2	0	0	0	0	0,01	0,99	0	0
3	0	0	0	0	0	1	0	0
4	0	0	0	0	0	1	0	0
5	1	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	1
7	0	0	0	0	0	0	0	1
8	0	0,01	0,01	0,01	0,05	0,05	0,89	0

It is remarkable that NVD_MA and NVD_YIS can very clearly be separated from others; this is due to GMM's flexibility in identifying subpopulations. Smaller number of instances for the remaining affixes mean that they are made to share one cluster.

The extremely large number of NVD_MA make it harder to make sense of the t-SNE output.

4.2.4 Denominal Verbs

There are several affixes that form denominal verbs. In this subsection, we analyze the clustering results for these affixes.

(85) a. Affixes forming a denominal verb

b. VVI_XPAST: *anlamıştı* 'he had understdood', *anlayacaktı* 'he would have understood' (278 instances)

c. NND_LI+: *karslı* 'from Kars', *hollandalı* 'Dutch' (178 instances)

d. VJD_AL: *azalmak* 'to decrease', *kısalmak* 'to shorten' (21 instances)

e. VJD_LA+: *genişlemek* 'to widen', *serinlemek* 'to freshen up' (10 instances)

f. VJD_LAS: *acılaşmak* 'to go bitter', *çirkinleşmek* 'to get ugly' (53 instances)

g. VND_A: *boşamak* 'to divorce', *oynamak* 'to play' (9 instances)

h. VND_LA: *bıçaklamak* 'to stab', *ellemek* 'to touch' (430 instances)

i. VND_LAN: *ayaklanmak* 'to revolt', *çimlenmek* 'to germinate' (144 instances)

j. VND_LAS+: *antlaşmak* 'to conclude a treaty', *fakirleşmek* 'to get poor' (35 instances)

As always, NND_LI+ and VVI_XPAST demonstrate that a meaningful clustering took place. Besides those, there is only one clear line of separation: VJD vs. VND. VJD_AL, VJD_LA+ and VJD_LAS indicate a process of assuming the property denoted by the base adjective. Therefore, these three

Table 19: Confusion matrix for affixes forming denominal verbs

	NND_LI+	VJD_AL	VJD_LA+	VJD_LAS	VND_A	VND_LA	VND_LAN	VND_LAS+	VVI_XPAST
1	0,01	0,03	0	0,02	0,02	0,73	0,18	0,01	0
2	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0,01	0,83	0,14	0,02	0
4	1	0	0	0	0	0	0	0	0
5	0	0,11	0,07	0,33	0,01	0,3	0,13	0,05	0
6	0	0,01	0	0,02	0,02	0,39	0,51	0,05	0
7	0	0	0	0	0	0	0	0	1
8	0	0	0	0	0	0	0	0	1
9	0,01	0	0	0,01	0,01	0,75	0,11	0,12	0

affixes share a core meaning and the cluster 5. For the remaining four affixes, we cannot find any apparent subgroup or pattern. This might be due to the particularly irregular meaning contribution of VND_LA/VJD_LA+ and the constellation of affixes that can be linked to it (VJD_LAS, VND_LAN, VND_LAS+).

4.2.5 Denominal Nominals

There are four groups of NND affixes that are clearly linked. They are of the form *-CA*, *-CI*, *-lIK* or *-(m)sA*.

(86) a. Affixes of the form *-CA*

b. VVI_XPAST: *anlamıştı* 'he had understdood', *anlayacaktı* 'he would have understood' (278 instances)

c. NND_LI+: *karslı* 'from Kars', *hollandalı* 'Dutch' (178 instances)

d. AJD_CA: *bolca* 'amply', *cahilce* 'ignorantly' (203 instances)

e. AND_CA: *ahlakça* 'morally', *dostça* 'as a friend' (74 instances)

f. JJD_CA: *akça* 'whitish', *eskice* 'oldish' (42 instances)

g. JND_CA+: *binlerce* 'thousands of', *düzmece* 'false' (12 instances)

h. NND_CA: *almanca* 'german',*frenkçe* 'french' (138 instances)

i. AAD_CA+: *beraberce* 'together', *epeyce* 'quite' (8 instances)

j. NJD_CA: *kaplıca* 'hot spring', *kokarca* 'skunk' (2 instances)

Language names formed by NND_CA are assigned to a dedicated cluster. For two pairs, AJD_CA-AND_CA and AJD_CA-JJD_CA, we observe that there might be a common meaning that bring these affixes together in clusters 1 and 8. Otherwise, different categories of *-CA* are not distinguishable.

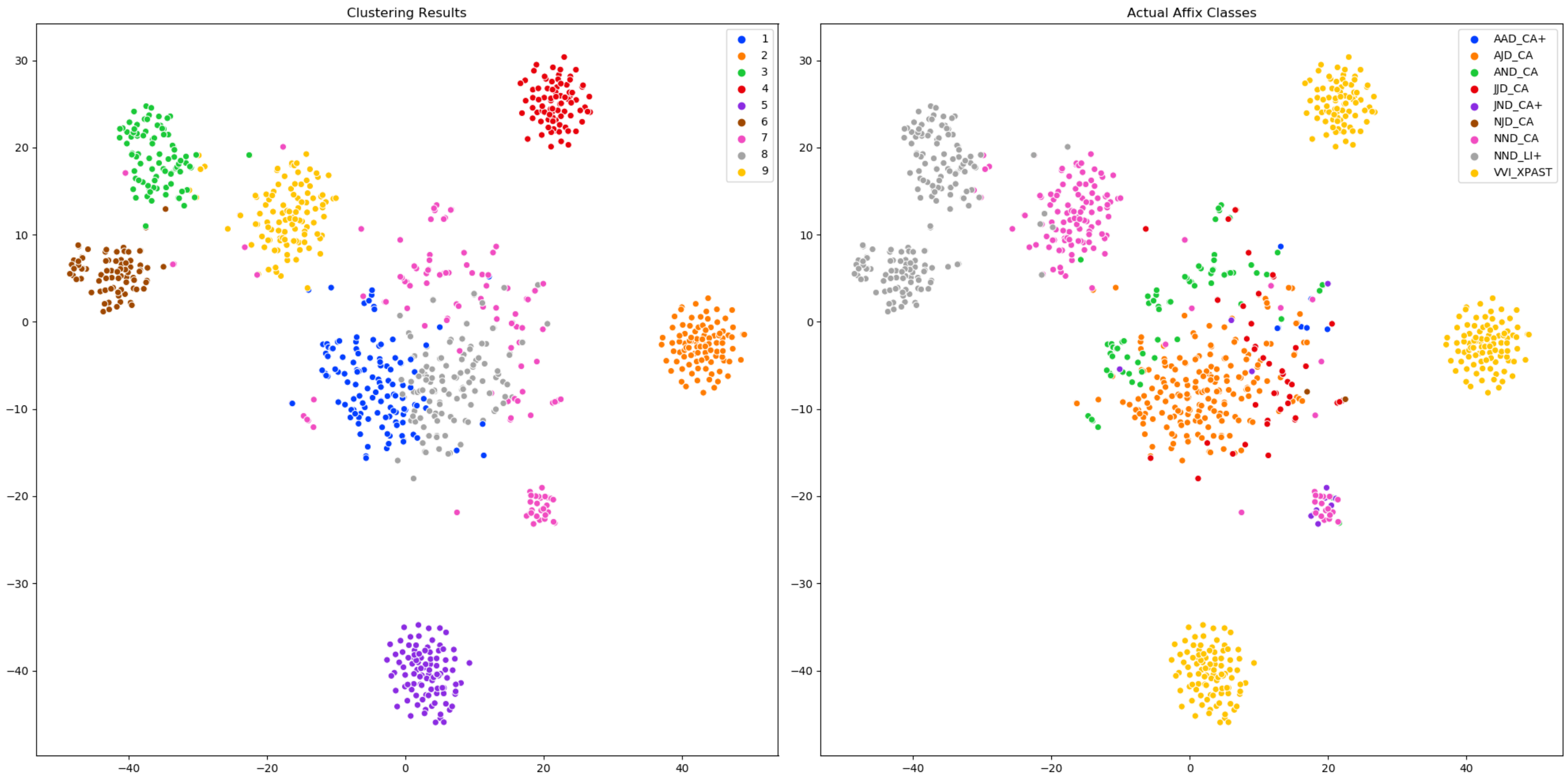


Figure 11: Distribution of affixes of the form *-CA*

Table 20: Confusion matrix for affixes of the form -*CA*

	AAD_CA+	AJD_CA	AND_CA	JJD_CA	JND_CA+	NJD_CA	NND_CA	NND_LI+	VVI_XPAST
1	0	0,71	0,24	0,03	0,01	0	0,01	0	0
2	0	0	0	0	0	0	0	0	1
3	0	0	0	0	0	0	0	1	0
4	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0	0	1
6	0	0	0	0	0	0	0	1	0
7	0,04	0,15	0,34	0,11	0,08	0,02	0,22	0,04	0
8	0,02	0,77	0,02	0,19	0,01	0	0	0	0
9	0	0	0,01	0	0	0	0,96	0,03	0

In Section 3.3.3 we looked into several complex affixes of involving -*CI*. Here we present their analysis along with NND_CI.

(87) a. Affixes related to -*CI*

b. VVI_XPAST: *anlamıştı* ‘he had understdood’, *anlayacaktı* ‘he would have understood’ (278 instances)

c. NND_LI+: *karslı* ‘from Kars’, *hollandalı* ‘Dutch’ (178 instances)

d. JVD_ICI: *akıcı* ‘fluid’, *çekici* ‘attractive’ (68 instances)

e. NND_CI: *antikacı* ‘antque dealer’, *bıçakçı* ‘knifer’ (1287 instances)

f. NVD_ICI: *bakıcı* ‘caretaker’, *delici* ‘piercer’ (251 instances)

g. NVD_GAC: *ayraç* ‘separator’, *bağlaç* ‘conjunction’ (40 instances)

h. NVD_GIC: *başlangıç* ‘beginning’, *dalgıç* ‘diver’ (5 instances)

Table 21: Confusion matrix for affixes related to -*CI*

	JVD_ICI	NND_CI	NND_LI+	NVD_GAC	NVD_GIC	NVD_ICI	VVI_XPAST
1	0	0,99	0	0	0	0	0
2	0	0	0	0	0	0	1
3	0,24	0	0	0	0	0,76	0
4	0	0,99	0	0	0	0	0
5	0	0	0	0	0	0	1
6	0	0,05	0,95	0	0	0	0
7	0,1	0,02	0	0,28	0,04	0,56	0

Although its meaning is linked to the remaining affixes (except NND_LI+ and VVI_XPAST, of course), NND_CI is firmly separated in its own dedicated clusters. Cluster 2 is populated only by JVD_ICI and NVD_ICI, but cluster 7 is a mix of JVD_ICI, NVD_GAC, NVD_GIC and NVD_ICI. These four affixes are composite; unlike NND_CI they involve a NVD component. Among them, their components have similar function. Clustering results support these observations.

There are four variations of *-lIK*, but two of them have much smaller number of instances compared to others.

(88) a. Affixes of the form *-lIK*

b. VVI_XPAST: *anlamıştı* ‘he had understdood’, *anlayacaktı* ‘he would have understood’ (278 instances)

c. NND_LI+: *karslı* ‘from Kars’, *hollandalı* ‘Dutch’ (178 instances)

d. JND_LIK: *anlık* ‘momentary’, *yemeklik* ‘cooking’ (46 instances)

e. NAD_LIK+: *boşunalık* ‘vanity’, *gündelik* ‘daily’ (10 instances)

f. NJD_LIK: *başarısızlık* ‘failure’, *ciddilik* ‘seriousness’ (1046 instances)

g. NND_LIK: *abonelik* ‘subscription’, *başlık* ‘headgear’ (1036 instances)

Table 22: Confusion matrix for affixes of the form *-lIK*

	JND_LIK	NAD_LIK+	NJD_LIK	NND_LI+	NND_LIK	VVI_XPAST
1	0	0,01	0,95	0	0,04	0
2	0	0	0	1	0	0
3	0	0	0	0	0	1
4	0,04	0	0,18	0	0,78	0
5	0	0	0	0	0	1
6	0	0	0	0,99	0,01	0

NJD_LIK and NND_LIK can be clearly separated. For others, sample sizes are small and the evidence is inconclusive.

Finally, we have 8 affixes that indicate similarity.

(89) a. Affixes indicating similarity

b. VVI_XPAST: *anlamıştı* ‘he had understdood’, *anlayacaktı* ‘he would have understood’ (278 instances)

c. NND_LI+: *karslı* ‘from Kars’, *hollandalı* ‘Dutch’ (178 instances)

d. JJD_IMSI: *acımsı* ‘bitterish’, *beyazımsı* ‘whitish’ (20 instances)

e. JND_IMSI: *abidemsi* ‘monumental’, *kekremsi* ‘acrid’ (24 instances)

f. JND_SAL: *bilişsel* ‘cognitive’, *dinsel* ‘religious’ (178 instances)

g. JJD_MTRAK: *mavimtrak* ‘bluish’, *sarımtrak* ‘yellowish’ (15 instances)

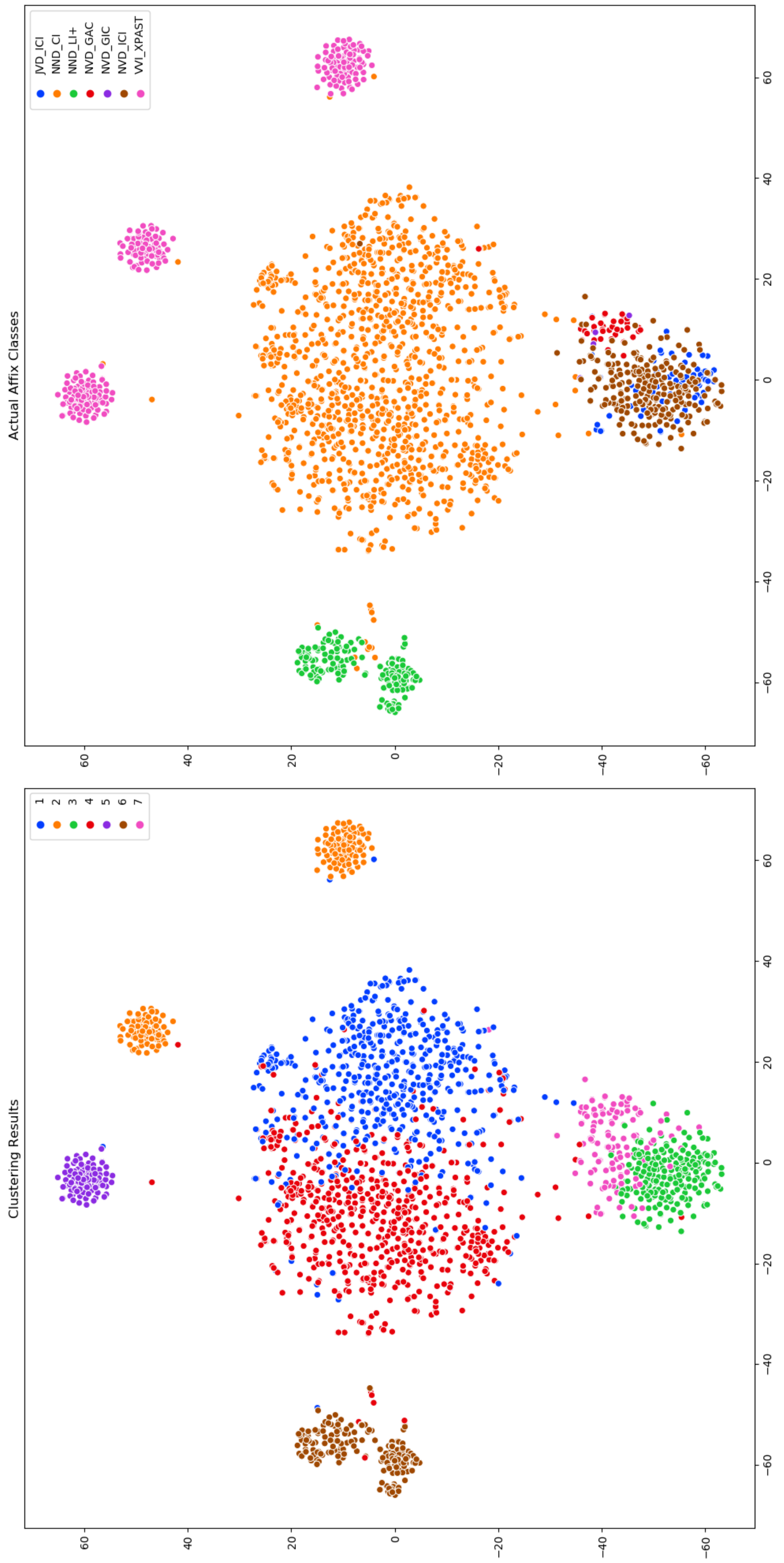


Figure 12: Distribution of affixes related to *-CI*

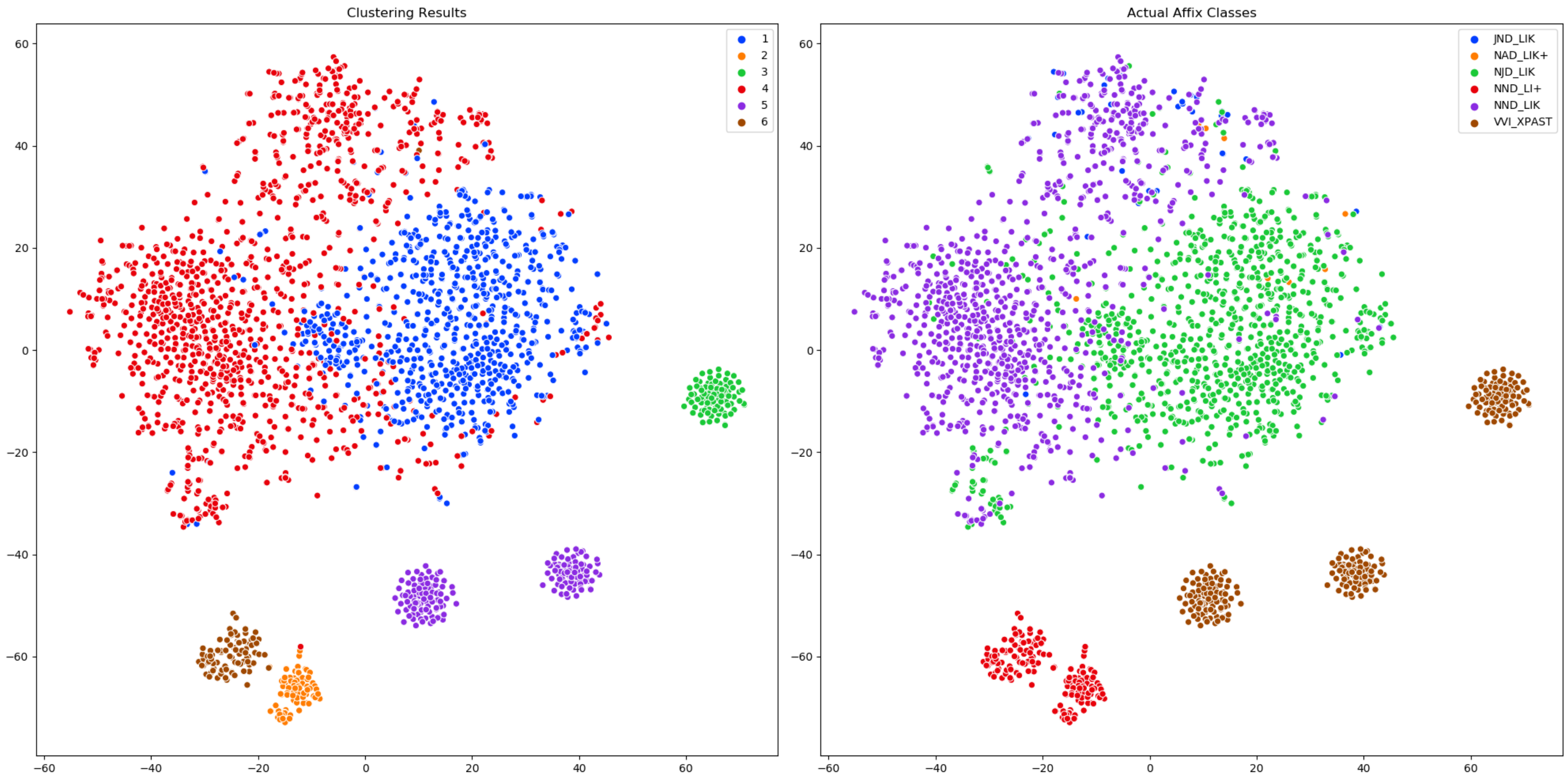


Figure 13: Distribution of affixes of the form *-lIK*

h. JND_SI: *çocuksu* ‘childlike’, *erkeksi* ‘manly’ (65 instances)

i. VJD_SA: *garipsemek* ‘to find strange’, *ıraksamak* ‘to diverge’ (2 instances)

j. VND_SA: *önemsemek* ‘to care’, *susamak* ‘to get thirsty’ (6 instances)

k. JND_IL+: *ardıl* ‘successor’, *birincil* ‘primary’ (44 instances)

l. VJD_IMSE: *ayrımsamak* ‘to notice’, *küçümsemek* ‘to belittle’ (4 instances)

Table 23: Confusion matrix for affixes indicating similarity

	JJD_IMSI	JJD_MTRAK	JND_IMSI	JND_SAL	JND_SI	NND_LI+	VJD_IMSE	VJD_SA	VND_SA	VVI_XPAST
1	0	0	0	0	0	0	0	0	0	1
2	0,26	0,02	0,23	0,06	0,43	0	0	0	0	0
3	0	0	0	0	0	1	0	0	0	0
4	0,02	0	0	0,95	0,02	0	0	0	0	0
5	0	0	0,09	0,76	0,13	0,02	0	0	0	0
6	0	0	0	0	0	0	0	0	0	1
7	0	0	0,03	0,82	0,14	0	0	0	0,02	0
8	0,03	0	0,03	0,49	0,27	0,04	0,05	0,03	0,07	0
9	0	0	0	0	0	1	0	0	0	0
10	0	0	0	0	0	0	0	0	0	1

Majority of instances for JJD_IMSI, JND_IMSI and JND_SI share cluster 2, demonstrating their similarity. JND_SI also has a large presence in that cluster. The same is true for VJD_IMSE, VJD_SA and VND_SA in cluster 8. JND_SAL and JND_SI cover a variety of meanings, occurring in all clusters. Still the majority of JND_SAL occurs in three clusters it dominates.

Overall, we observe three subgroups of affixes with related meaning: JXD (JJD_IMSI, JND_IMSI, JJD_MTRAK and maybe JND_SI), VXD (VJD_IMSE, VJD_SA and VND_SA) and JND_SAL. This grouping largely coincides with our classification in Chapter 3.

4.2.6 Distributionally Motivated Affix Classes

Having processed vector representations based on thousands of base-lemma pairs, we are aware that making point estimates for affix embeddings is not realistic. Estimates from different base-lemma pairs inevitably vary. We can use two simple metrics to describe the DS of an affix: mean and variance of vector lengths. Both depend on the assumption that token with similar semantics should be located close to each other on the DS space.

Mean vector length is expected to reflect the semantic content of the affix. In a perfectly compositional system of morphological relations, a more prominent semantic contribution would be represented by a longer vector. (Please note that vector lengths are affected by many other factors such as the number of tokens used to compute embeddings. We ignore such effects and focus on the difference with respect to semantic content.) A complete lack of semantic content would be represented by the zero vector.

Variance of vector length can be used to gauge the regularity of semantic contribution. Variance of vector length is expected to be lower for affixes with regular semantics and higher for affixes with irregular semantics. This is easily observed comparing IM estimations with DM estimations.

We calculate the “Mean Length” metric based on Euclidean distance. The calculation is carried out for each vector individually. Then mean lengths are averaged for each affix class.

$$\mu_{L2} = \frac{1}{m} \sum_{i=1}^{m} \sqrt{\sum_{j=1}^{n} \overrightarrow{v}_{ij}^{2}}$$

where $\overrightarrow{v}$ is the representative vector, n is the number of dimensions in the embedding model and m is the number of base-lemma pairs in the affix class.

An alternative “Mean Length” metric is based on the Manhattan distance. This time we sum the absolute values of vector components.

$$\mu_{L1} = \frac{1}{m} \sum_{i=1}^{m} \sum_{j=1}^{n} ||\overrightarrow{v}_{ij}||$$

where $\overrightarrow{v}$ is the representative vector, n is the number of dimensions in the embedding model and m is the number of items in the affix group.

For these two statistics that measure the mean lengths of vectors, we calculate the corresponding standard deviations: σ_{L2} and σ_{L1}.

For the analyses in this section, we avoid distorting the data with normalization or standardization. We aim to preserve vector lengths.

We plot these statistics to demonstrate correlation between metrics. Mean vs Std (standard deviation) plots indicate statistical similarity and dissimilarity between affix groups. Mean vs Mean and Std vs Std plots only tell us if the statistics based on these two distance significantly correlate (They are expected to correlate.). Statistics are more reliable for more frequent affixes, so Figure 14 demonstrates the results for the most frequent affixes (each over 100 instances).

L1 and L2 plots are quite similar, so we arbitrarily choose the L1 metric and prepare “Mean Length L1 vs Std Length L1” plots for three groups of affixes. Comparing these plots, we can observe the distributional properties of more and less frequent affixes.

“Mean Length L1 vs Std Length L1” and “Mean Length L2 vs Std Length L2” plots are almost identical. Unsurprisingly, affixes with longer vector representations tend to have a larger standard deviation on length. Correlation between means and correlation between standard deviations of the two distance metrics are close to perfect. This shows that the distance metric does not have a significant effect on our observations.

In Figure 15, we can focus on the Mean vs Std plot for the most frequent affixes. Several interesting clusters can be observed. The lower left corner is occupied almost exclusively by VVI affixes, with the exception of NVD_MA, which is arguably closer to being an inflection than most other derivational affixes. It always contributes the same meaning to the host and is applicable to the entire class of verbs. It should be noted that all voice markers and copular markers are firmly placed in this region, while TAM markers are firmly outside. Let us call this part of the plot the true VVI region.

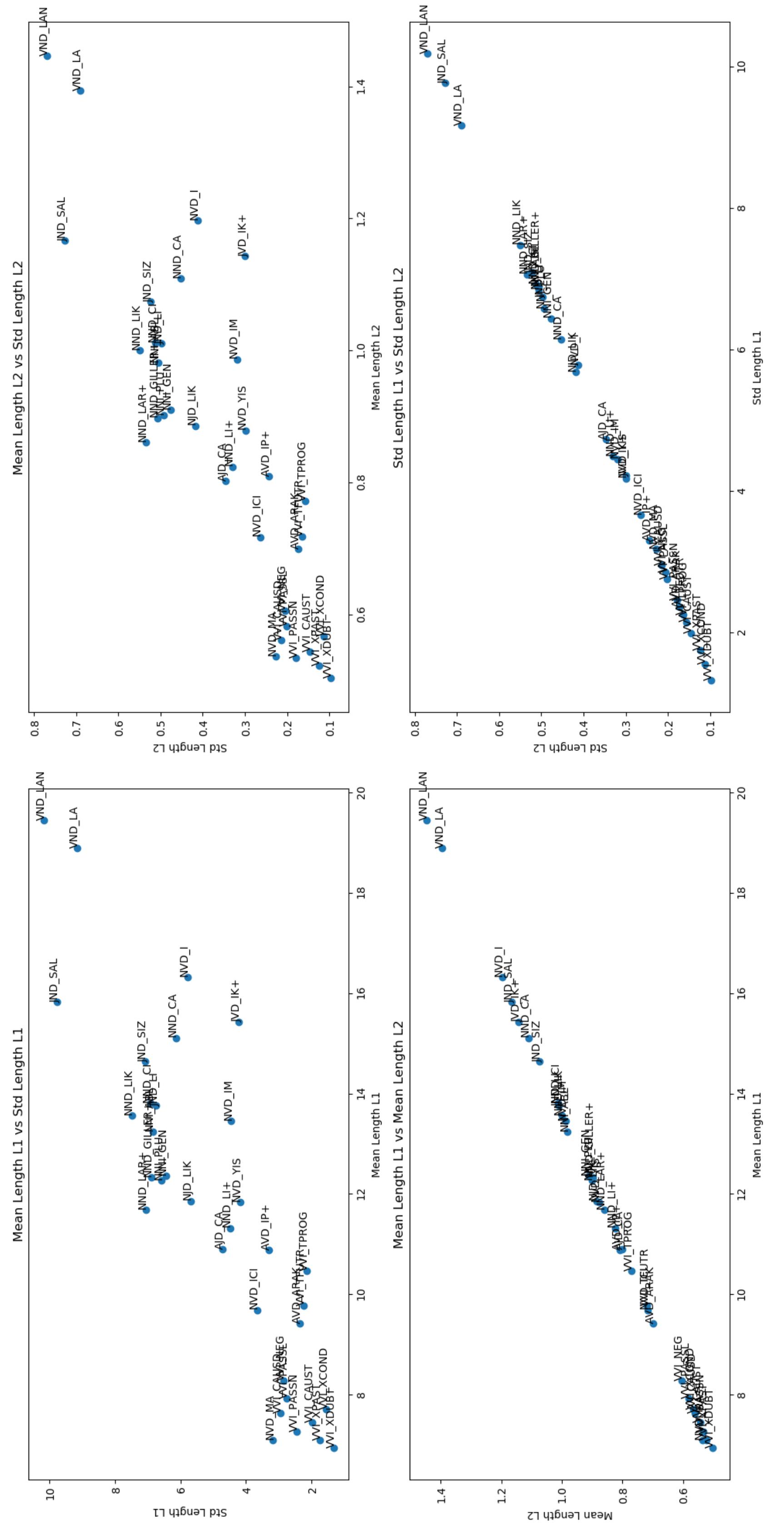


Figure 14: Statistics for affixes appearing in over 100 instances

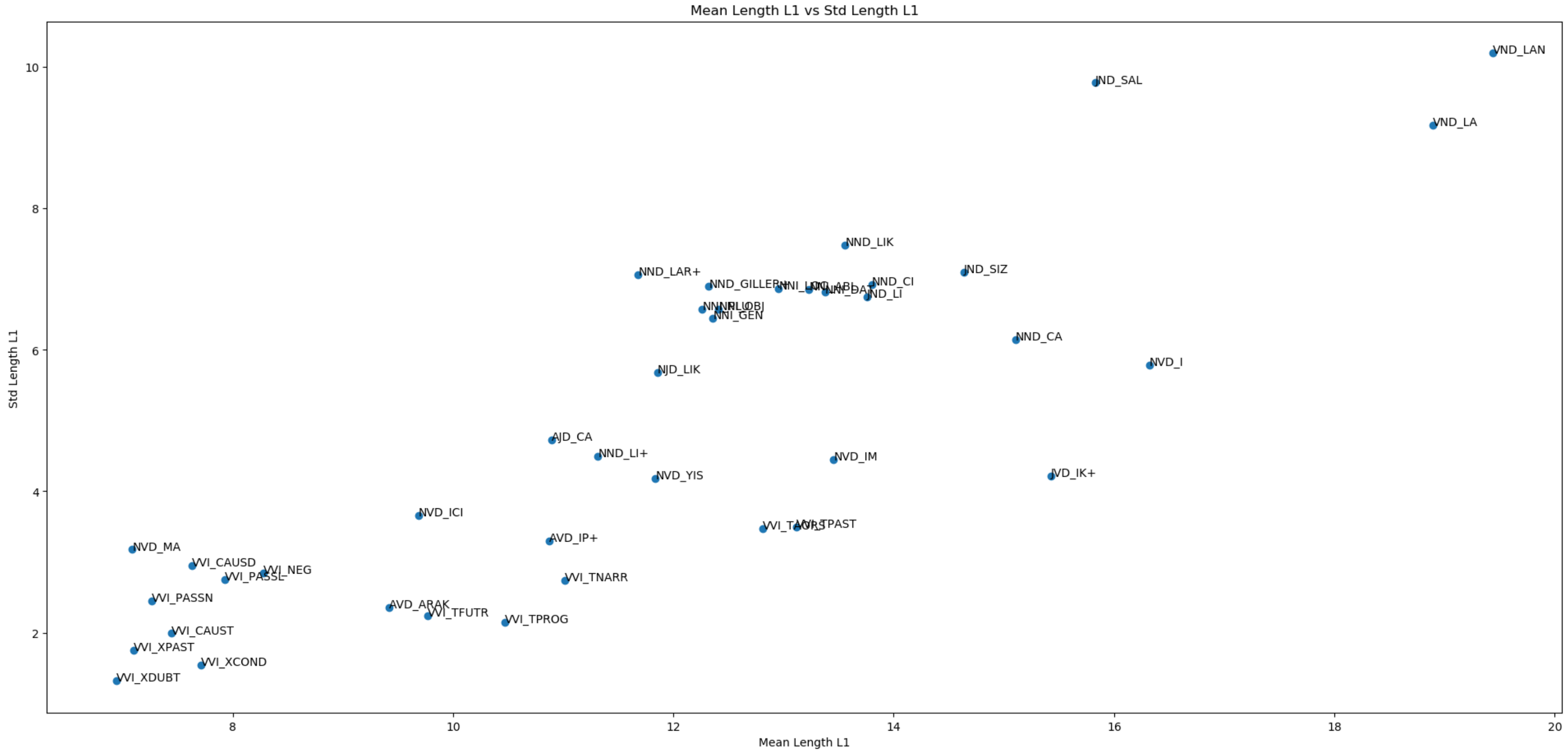


Figure 15: Statistics for affixes appearing in over 100 instances

TAM markers demonstrate significantly greater mean length but only slightly greater standard deviation. They are found with AVD_ARAK and AVD_IP+, which are again quite close to being inflectional affixes for the same reasons given for NVD_MA. We shall call this part the TAM region.

Around the center of the figure, we observe a number of NND and NNI affixes. NNI affixes are packed in a circle in the middle of the larger group. Four NVD affixes partially surround this larger group to the bottom and to the right.

On the upper right corner, we have two VND affixes, one possibly a variation (through fusion) of the other. Both these affixes are known for the irregularity of their semantics which is reflected in the large standard deviation of their length. This is the VND region.

The fact that categorically similar affixes easily separate into contiguous regions is very interesting.

The final plot is Figure 16 and it includes 32 additional affixes with 30-99 instances. Due to their smaller sample sizes, the statistical data for some of these affixes are less reliable than the ones considered previously.

Figure 16 is stretched due to affixes with high mean length and high standard deviation, such as NND_CIL, NND_DAS and JND_IL+. The true VVI, VND, NNI and NND regions are exactly the same. PRONOUN_3S+ and VND_LAS+ are placed between the NND and NVD regions.

JND_SAL, JND_SI, AND_CA and JJD_CA form a cluster of affixes that convey a meaning of similarity. We expected similar meaning content to be conveyed by similar vectors, but the extent to which this expectation holds is surprising.

NVD_AN, NVD_INTI, JVD_AK, NVD_IK, NVD_GAC are added to the NVD region. JVD_IR, NVD_TI, NVD_GAN and JVD_GIN are placed in the TAM region. With these final additions, we can now see a clearer picture suggesting the semantic similarity between NVD, JVD and TAM.

It is also surprising how VVI_XPAST, VVI_XDUBT and VVI_XCOND are so closely packed together but so far away from VVI_TAORS, VVI_TPAST and VVI_TCOND. Apparently, these two sets have quite dissimilar semantic content, despite their being of the same category. The latter set is semantically closer to AVD_ARAK, AVD_IP+, NVD_TI, NVD_GAN and NVD_AN, along with other TAM affixes VVI_TFUTR, VVI_TNECE, VVI_TPROG and VVI_TNARR.

The semantic similarity between TAM and XVD affixes and dissimilarity between TAM and copula lend credibility to the idea that Turkish predicates are always nominal. According to this idea, which is discussed in Chapter 3, the three sets of phonologically similar affixes marking finite verbs (*aldım* ‘I took’), relative clauses (*aldığım* ‘the one I took’) and derived forms (*alacak* ‘credit’) actually form deverbal nominals.

4.3 Discussion of DS

DM covers a wide variety of categories and semantics. The prevalence of suppletive allomorphy, polysemy and synonymy create an extremely complex network of affixes. This complexity is often interpreted as irregularity.

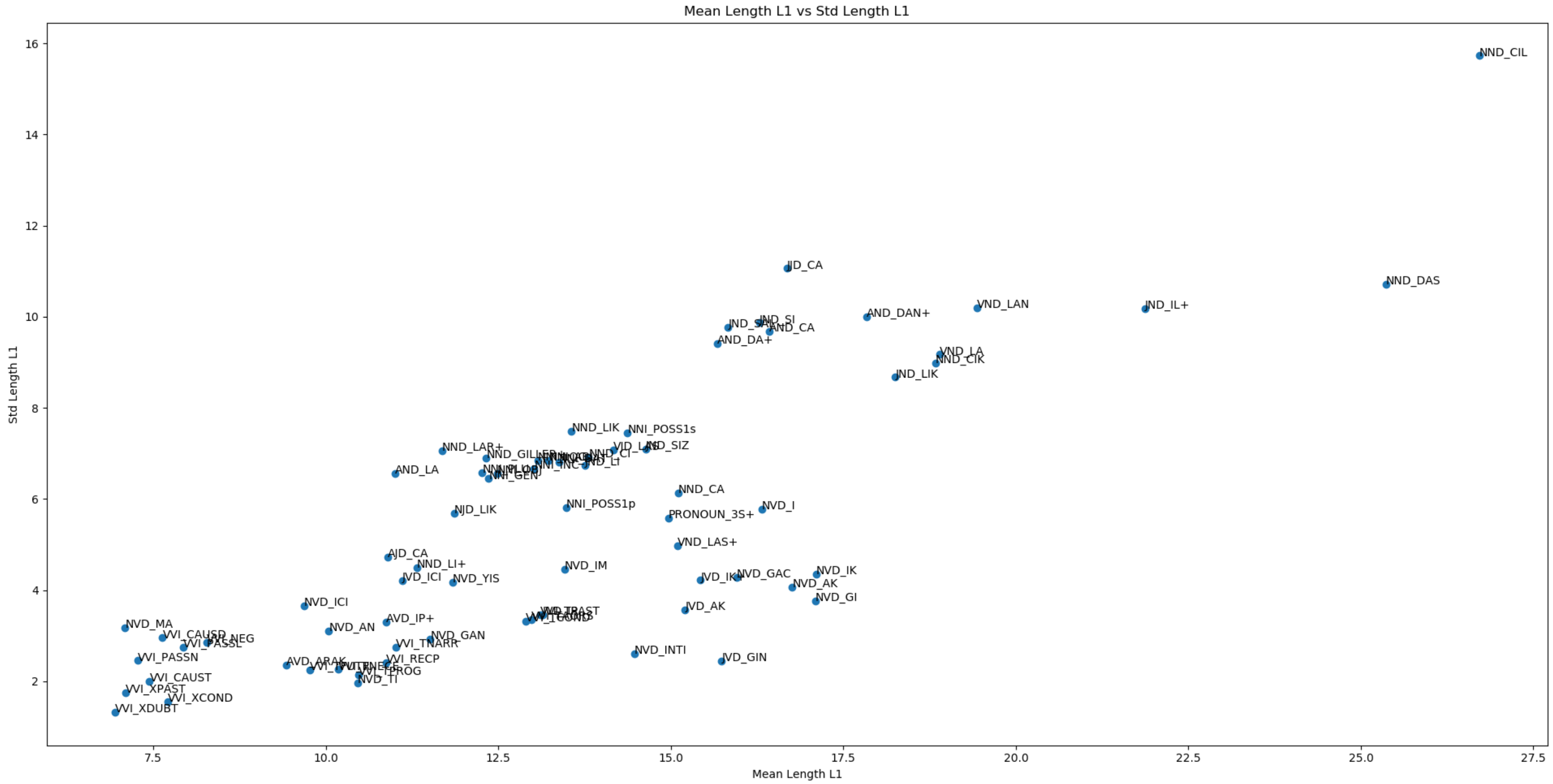


Figure 16: Absolute values of means vs standard deviations for lengths of affixes appearing in over 30 instances

However, there is structure behind this complexity. In Chapter 2, we reviewed the evidence of speakers' awareness of DM. This was the first piece of evidence that DM has a structure that is discoverable. In Chapter 3, we created a new organization for Turkish DM. Relying on a large number of examples, we showed that most of Turkish DM is regular and can be represented by a set of rules.

Evidence from psycholinguistics may be discounted by suggestions that awareness of DM is restricted to a very small number of processes. Our efforts towards organizing DM may not be considered proof that most DM is regular. We needed an independent piece of evidence to demonstrate this regularity.

The exploration given in this section shows beyond reasonable doubt that our premise is correct. Estimations of affix embeddings create clear clusters based on corresponding affixes. Synonymous affixes share clusters, while polysemous affixes are represented in multiple clusters. Based on a completely unsupervised method, we are able to successfully map the semantic space covered by Turkish DM. These observations are true for not only the most productive affixes, but also fairly unproductive affixes.

Results of this exploration show us that even simple vector arithmetic gives consistent estimations for the DS of DM. We do not look into estimations using other functions, in order not to lose focus, but future work could compare the the overall homogeneity of clusters constructed with different functions.

4.3.1 Contextual Embeddings and Referential Narrowing

Contextualized embeddings, while outside the scope of this thesis, are even more promising in generating accurate affix embeddings. Having observed that stems and derived forms often exhibit homonymy and polysemy, our pre-trained embeddings constituted merely an aggregation of the underlying distinct semantics. Arguably, contextual embeddings would obtain more accurate results by distinguishing between polysemous uses.

Classical word embeddings are aggregations over all uses of derived form. No matter how many different polysemous uses occur, the word embedding is a single vector. Where there is a single vector, one must imagine a collection of vectors within its neighborhood, each arising from a specific context. Contextual embeddings aggregate over subsets of these uses, possibly distinguishing between different semantics. The result is several vectors, indicating the presence of several neighborhoods. This time residents of a neighborhood arise from the same context, but different subcontexts.

This coincides with what we discovered in this chapter about affix embeddings. Affix embeddings vary to some extent, according to the base-lemma pairs used to approximate them. For each affix, we calculated many approximations across many base-lemma pairs and found that embeddings often resist being approximated by a single vector. Otherwise, it is easy to miss different clusters of embeddings within a single affix. VVI_XPAST, which clearly divides into three clusters, is the best example of this.

The referential narrowing concept in Section 2 finds a simple, visual representation with a DS perspective. To repeat the idea, Swadesh (1938) argues that derived forms assume more special semantics than what would result from the combination of their constituents. The Figure 17 shows the imaginary distributional semantics for two derived forms (*kitaplık* 'bookshelf', *gözlük* 'glasses') and their

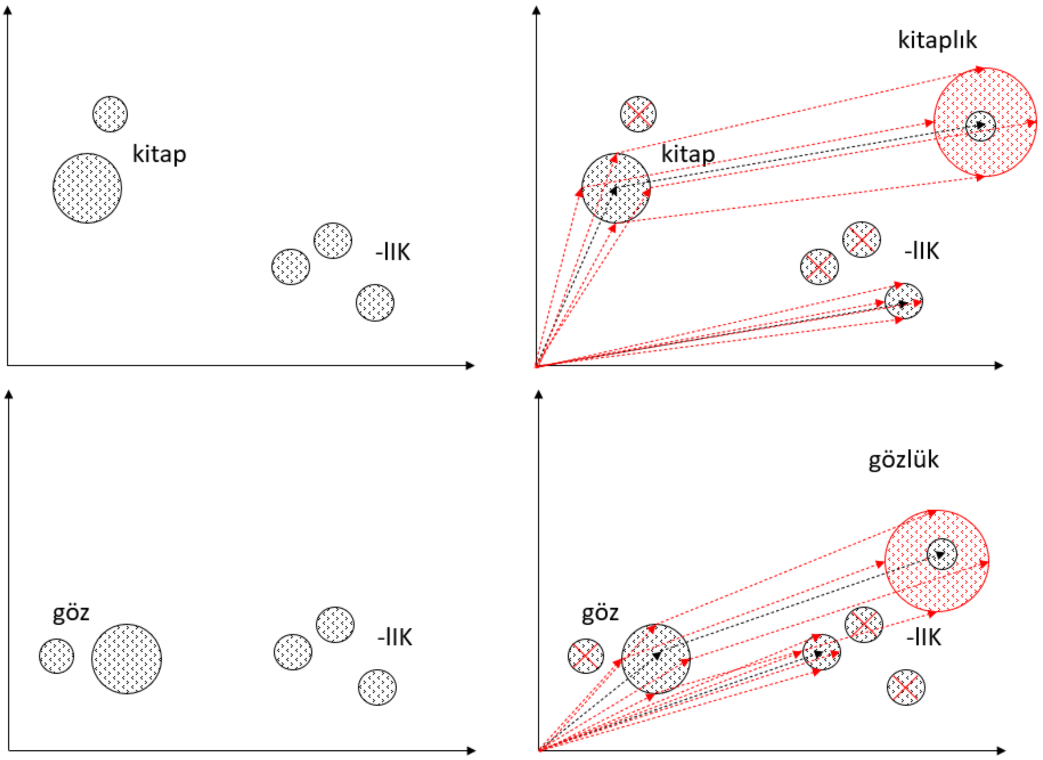


Figure 17: Illustration of distributional semantics of referential narrowing

constituents (*kitap* ‘book’, *göz* ‘eye’, *-lIK*). Embeddings for polysemous uses are given in different neighborhoods. In this case, *kitap* and *göz* have two each, while *-lIK* has three.

Within DS, referential narrowing simply means that the embeddings cluster of a derived form is expected to show less variation than the combination of its constituents’ clusters. The figure shows the derivation of *kitaplık*. Even though a single meaning of *kitap* and a single meaning of *-lIK* are used, the combination of their embeddings gives rise to a large cluster of semantic possibilities (represented by the red circle). The claim is that kitaplık as a whole, does not exhibit this level of semantic variation. Rather, it assumes a narrower meaning. A similar case can be shown for *gözlük*, using a different cluster from *-lIK* embeddings.

We do not pursue this idea further within the scope of this thesis. Perhaps a later study could numerically demonstrate referential narrowing, based on this line of thinking.

4.4 Categorial Grammar

In order to represent the semantics of morphemes and model the derivation process, we need a tool with adequate expressive power. The tool must be flexible enough to allow a comprehensive analysis of a wide range of phenomena, but restrictive enough to limit our hypothesis space meaningfully.

Categorial Grammar (CG) has over decades accumulated an impressive track record proving its capacity for such an analysis. An important feature of CG is its handling linguistic processes with an equal regard for syntax and semantics. As derivational processes are heavily involved with semantics, this proves to be a major advantage of CG over alternative frameworks.

Chomskian theories of language have been dominant in the literature for decades. Constraint-based grammars, which follow a completely different strategy than the Chomskian generative grammars, have been developing rapidly. Among these, CG has received considerable attention, due to its adequate expressive power. Proponents of CG prefer envisioning syntax and semantics simultaneously being "projected from the lexicon". By representing syntax and semantics in lockstep, CG provides a much more suitable platform to study DM, which is much more involved with semantics than inflectional morphology and syntax.

The kind of CG we take inspiration from is Combinatory Categorial Grammar (CCG) (Steedman, 1996), (Steedman, 2000), (Steedman and Baldridge, 2011). Its emphasis is on constituency rather than dependency. Three elements of this representation are surface form, category and logical form. CCG also introduces combinator mechanics into CG derivation processes and prefer lambda calculus for meaning representation. As a result, CCG is a powerful tool that has been developed and used to study many languages. It allows for complicated constructions and is able to model many different linguistic phenomena.

In Chapter 5 we use this framework in two ways. First, we use CG to store and use semantic knowledge along with surface forms and categories. Second, we assume that context is often strong enough for the hearer to infer the semantics and category of a novel observation. In such cases, we represent that context in the same form as a traditional CG lexical item. The assumption that the context can be strong enough to enabled such an approach is related to the Semantic Bootstrapping Hypothesis (Abend et al., 2017).

Each lexical item contains the surface form, the category and the logical form of the expression. The surface form is the written form of a lemma. The category is defined according to the ordinary CG formalism. The logical form (LF) is written in lambda-calculus. This provides a suitable structure for consistently representing and deriving the semantic structures of derived forms, as well as stems and affixes. A short list of examples are given below:

(90) a. *göz* ⊢ N: λx1.be eye x1

b. *gör* ⊢ V\N: λx1λx2.see x1 x2

c. *gözlük* ⊢ N: λx1λx2.and (be eye x2) (wear (on x2) x1 anon)

d. *gözlükçü* ⊢ N: λx1λx2λx3.and (and (be eye x3) (wear (on x3) x2 anon)) (sell x2 x1)

e. *-lük* ⊢ N\N: λx1λx2λx3.and (x1 x3) (wear (on x3) x2 anon)

f. *-çü* ⊢ N\N: λx1λx2λx3λx4.and (x1 x3 x4) (sell x3 x2)

Combinatory Categorial Grammar (CCG) employs several combinatory rules to cover a wide range of linguistic phenomena. Three combinators combine adjacent constituents to derive larger expressions: application, composition and substitution (Steedman and Baldridge, 2011), (Bozşahin, 2012). Kruijff and Baldridge (2004) define modalities to distinguish between the different interactions between constituents. For the purposes of this study, we only implement application (both forward and backward), as the adjacency principle can be assumed to always hold for Turkish derivational morphemes. As a result, the representation scheme we are using could be simply considered CG and not CCG. There is also no example of non-concatenative DM in Turkish. Therefore, the ordinary CG mechanics based on segments should be appropriate.

We take inspiration from sources on CCG for insights on formulation, too. Steedman and Baldridge (2011) is a complete introductory level discussion of CCG's capabilities and basic grammar-building strategies. It is a primary resource in the matters related to CCG. Bozşahin (2002) argues that setting aside inflectional morphology as a word-internal process and "designating words as minimal units of the lexicon" is too restrictive. Instead, in this study, inflectional morphology and syntax are integrated within the framework of CCG and a morphemic lexicon. We also follow this perspective when trying to extend syntax into the sub-word level. Bozşahin (2017) is the manual for CCGlab, the computational tool that makes rapid prototyping possible for complex CG rule sets. It implements the CKY parser and reduces the time required for testing by several orders of magnitude. We use this engine to test our grammars and validate our custom-built CKY parser.

We build a baseline grammar to observe the interactions between lexical items of different categories. This grammar contains a representative sample from the affixes reviewed in Chapter 3 and constitutes the data basis for our exploration in Chapter 5. In this section, we present the principles we follow building the baseline grammar and explain the reasons behind our decisions.

4.4.1 Representing Free Forms

Achieving an adequate representation for all syntactic categories is crucial, because the success of morphological operations depends on the stems' compatibility with a wide range affixes in terms of syntactic category and logical form.

We need to account for four major grammatical categories: nouns, verbs, adjectives and adverbs. We leave prepositions aside for the moment. While noun phrases seem to be a simple, even trivial class in terms of their logical form representation, Grimshaw (1990) presents quite detailed arguments against this default position. Verbs are more complicated, assuming different categories for different stages of inflection, as well as hosting the argument and event structures. For adjectival semantics we rely on the ideas of Paoli (1999), Paradis (2001), Kennedy and McNally (2005) and Kennedy (2007). For adverbs, the literature is not as generous, but we modify and reuse the principles from adjectival semantics.

Syntactic category for a noun is N. In their simplest form, nouns can be represented without any bound variable. Most studies in the literature take this path. On the syntactic level, words are only arguments for the sentence structure. As long as all arguments fall into correct places with correct interactions, the sentence structure can be correctly interpreted. On the other hand, when morphology is under focus, we must keep track of word-internal operations by always clearly identifying the variable that denotes the object represented by the lemma.

Let us examine the difference on examples (91) and (92). When we use the conventional free-variable-representation for a noun, the former derivation is used. *gözlük* 'glasses' can be easily derived from *göz* 'eye'. One complication is that, while converting the stem into a completely different object, the affix has to introduce new structure to the semantic representation. Since the affix must be able to apply on other stems, the object denoted by the new representation must be generic. Therefore, the affix introduces a free variable along with the structure explaining its relation to the original object. Therefore, the final object is represented by a bound variable, while the original one is represented by a free variable. Further derivations must also introduce new bound variables.

(91) Conventional formulation of nouns

a. *göz* $\vdash$ N: eye

b. *gözlük* $\vdash$ N: λx1.wear (on eye) x1 anon

Dowty (1981) employs Skolem functions as "existentials taking narrow scope with respect to the operator". Also, the set of variables taken by the Skolem function may be empty, in which case the Skolem term becomes a Skolem constant, "behaving like a wide-scope existential". We use Skolem constants like *anon* to fulfill some thematic relations. This method allows us to represent word-internal structure without modifying the existing framework. For instance, *gözlük* 'glasses' is derived by a DM indicating that the new object functions as the patient in some action. The agent or any other thematic relation is irrelevant. Therefore, we fulfill the irrelevant roles with Skolem constants.

We prefer to treat nouns as properties. For common nouns, this is easily justifiable. Common nouns are used for specific objects based on their belonging to a group of objects with a common property. In this way, even an object denoted by a simple noun is represented by a bound variable. While carrying out derivations, we always introduce the new bound variable to the outermost position; therefore, the lexical-conceptual structure is not just started by an affix, it is present all the way including the root. As an added benefit, the object denoted by the full lemma is always easily identified.

(92) Property formulation of nouns

a. *göz* $\vdash$ N: λx1.be eye x1

b. *gözlük* $\vdash$ N: λx1λx2.and (be eye x2) (wear (on x2) x1 anon)

c. Lemma: *gözlükçü* $\vdash$ N: λx1λx2λx3.and (and (be eye x3) (wear (on x3) x2 anon)) (sell x2 x1)

In their bare form, nouns are not taken to be cased and must get a case to assume their role within a sentence, as prescribed by the case filter Cowper (1992). Named entities are no exception; they may take any case marker, and they must take case. In order to reduce clutter, we assume all nouns to have nominal case. We never mark case on categories or features, nor do we use any other case than nominative, in examples.

All in all, the template we use for building lexical items of category N is as follows:

(93) Template for nouns

a. *Surface Form* $\vdash$ N: λx1.be *Denoted Property* x1

Verbs have the additional task of organizing sentence structure. Turkish verbs are inflected on multiple positions for voice, polarity, tense-aspect-modality (TAM), copula and person. As discussed in Chapter 3, there is ongoing debate on finiteness, the auxiliary and the person marker. Here, we present a simple formulation, which still does justice to the complexities discussed earlier.

Given the category for the whole sentence is S, the finite verb must be able to form S by taking at least one N as argument. When its valency is 1, the single argument corresponds to the subject. When it is more than 1, the additional arguments are objects.

There are two additional facts. First, the overt subject is optional (Öztürk, 2001) and serves as the topic. Therefore, the person marker fulfills the role of subject, when there is no topic shift. Second, the bare verb cannot take part in any syntactic operation before it takes TAM and becomes finite. Bare verbs can take part in morphological operations, not only with IM such as voice, polarity and TAM, but also with DM.

We label the bare verb, the finite verb and person marked verb with different categories. Verbs may be intransitive, transitive or ditransitive. A verb's arity is reflected in both its syntactic category and its logical form, with an appropriate number of bound variables. Some verbs may need to be represented by multiple lexical items, if they assume different valency in different contexts.

Representing TAM and copula is only possible with the help of an adequate model. Moens and Steedman (1988) provide an outstanding analysis of universal temporal categories and temporal relations between events. The claims made in the paper are not based on purely temporal primitives, but on causation and consequence. Following the classifications and strategies in Moens and Steedman (1988), quite complex event structures may be analyzed in a robust and elegant way. Also, thanks to the framework given in the paper, the vast variety of possibilities from Turkish TAM markers can be easily represented in logical forms. We find it sufficient to demonstrate the principles on the time dimension, and leave out the world dimension. In order to be able to experiment with tense markers, we sometimes include the event time indicator. Speech time is indicated by t_0 (or $tref$ more generally) and the event time by a bound variable.

The event time is the innermost one in all verb LF. The subject (the external argument) is represented by the second innermost bound variable. Object slots (internal arguments) are fulfilled first.

(94) Formulation of verbs

a. *gel* ⊢ V: λx1λx2.come x1 x2

b. *geldi* ⊢ S/N: λx1λx2.and (x2 < tref) (come x1 x2)

c. *geldim* ⊢ S: λx1λx2.and (be speaker x1) (and (x2 < tref) (come x1 x2))

d. *gör* ⊢ V\N: λx1λx2λx3.see x1 x2 x3

e. *seni gör* ⊢ V: λx1λx2λx3.and (be hearer x1) (see x1 x2 x3)

f. *seni gördü* ⊢ S/N: λx1λx2λx3.and (x3 < tref) (and (be hearer x1) (see x1 x2 x3))

g. *seni gördüm* ⊢ S: λx1λx2λx3.and (be speaker x2) (and (x3 < tref) (and (be hearer x1) (see x1 x2 x3)))

Syntactic processes fulfill argument slots by introducing new lambda terms. More specifically, applications on the sentence level operate on the θ-structure, while IM fulfills grammatical features and DM changes the semantics altogether. All these groups can be observed on verbs. For instance, the lambda term (come x1 x2 x3) for *gördü* 'saw', represents both the θ-structure, the argument structure and the event structure. (In this case, the θ-structure and the argument structure coincide.) The lambda terms (be speaker x2) and (be hearer x1) fulfill the θ-structure, while (x3 < tref) fulfill the tense feature.

All in all, the templates we use for building intransitive and transitive verbs are as follows:

(95) Template for verbs

a. *Surface Form* ⊢ V: λx1λx2.*Denoted Action* x1 x2

b. *Surface Form (TAM Marked)* ⊢ S/N: λx1λx2.and (x2 *Relation between Speech Time and Event Time* tref) (*Denoted Action* x1 x2)

c. *Surface Form (person marked)* ⊢ S: λx1λx2.and (be *Denoted Subject* x1) (and (x2 *Relation between Speech Time and Event Time* tref) (*Denoted Action* x1 x2))

d. *Surface Form* ⊢ V\N: λx1λx2λx3.*Denoted Action* x1 x2 x3

e. *Surface Form (Argument Fulfilled)* ⊢ V: λ x1λ x2λ x3.and (be hearer x1) (*Denoted Action* x1 x2 x3)

f. *Surface Form (TAM Marked)* ⊢ S/N: λx1λx2λx3.and (be hearer x1) (and (x3 *Relation between Speech Time and Event Time* tref) (*Denoted Action* x1 x2 x3))

g. *Surface Form (Person Marked)* ⊢ S: λx1λx2λx3.and (be *Denoted Subject* x2) (and (x3 *Relation between Speech Time and Event Time* tref) (and (be *Denoted Object* x1) (*Denoted Action* x1 x2 x3)))

Adjectives take nouns as arguments and act like predicates. Their LF contain bound variables to be fulfilled by the noun. If we did this in the simplest way, using a single lambda term, receiving the noun would complete the term. As a result, we would not have any bound variables left to link the entity denoted by the noun to the remainder of the sentence. This is the same problem that led us to adopt a property representation for nouns. Again, we prefer separate lambda terms for each property; one for the noun and one for the predicate.

Adjectives come in two groups. Many adjectives are gradable. For a noun to be modified by some gradable adjective, it must have the property at some level. This level might be modified by later derivations. For example, for a car to be fast, it must be fast at some level of fastness. This level is stored in each speaker's context and somehow implicitly conveyed to the hearer. There are also adjectives which are not gradable, or which occur on a binary scale. Paoli (1999) calls the former kind relative adjectives, and the latter absolute adjectives. Formulations for gradable adjectives require additional lambda terms and bound variables.

Derived adjectives must reflect the word-internal structure in their LF. This is accomplished in a similar way to derived nouns. The semantic contribution of the affix is represented by an additional lambda term.

(96) Formulation of adjectives

a. *araba* ⊢ N: λx1.be car x1

b. *kırmızı* ⊢ N/N: λx1λx2.and (x1 x2) (be red x2)

c. *kırmızı araba* ⊢ N: λx1.and (be car x1) (be red x1)

d. *hak* ⊢ N: λx1.be right x1

e. *anayasa* ⊢ N: λx1.be constitution x1

f. *anayasal* ⊢ N/N: λx1λx2λx3.and (x1 x2) (and (be constitution x3) (relate (to x3) x2))

g. *anayasal hak* ⊢ N: λx1λx2.and (be right x1) (and (be constitution x2) (relate (to x2) x1))

The templates we use for building lexical items for adjectives with two different levels of arity are as follows:

(97) Templates for adjectives

a. *Surface Form* ⊢ N/N: λx1λx2.and (x1 x2) (be *Denoted Property* x2)

b. *Surface Form* ⊢ N/N: λx1λx2λx3.and (x1 x2) (and (be *Root Concept* x3) (*Relation to Root Concept* x3 x2))

Adverbs specify when, how or why an action takes place. Time adverbs require us to represent the time relation explicitly in the logical form. A specification of manner can be conveyed in the way adjectives modify nouns. We experiment with only a few adverbs, as the number of deadverbial DM is quite low. Again, arity of the adverb depends on the verb. We present examples with the lowest number of arguments.

(98) a. *iç* ⊢ N: λx1.be inside x1

b. *içeri* ⊢ V/V: λx1λx2λx3.and (x1 (towards x3) x2) (be inside x3)

c. *demin* ⊢ V/V: λx1λx2λx3.and (x1 x2 x3) (x3 = tref minus eps)

d. *demincek* ⊢ V/V: λx1λx2λx3.and (and (x1 x2 x3) (x3 = tref minus eps)) (be small eps)

As well as modifying verbs, adverbs may modify adjectives and other adverbs, adding another level of complexity to their representation. Degree adverbials should perhaps be considered a wholly different class, while time adverbials can be assumed to only modify verbs etc. Such constraints seem to be grounded in the way lexical items' argument structures fit with each other. For instance, adverbs of time, expect a grammatical category that carries a time variable, and that is only possible with verbs. Verbs happen to be the only grammatical category that fulfills the criteria for being modified by an adverb. For the purposes of this thesis, we do not delve deeper into the LF formulation of all kinds of adverbs.

The templates we use for building lexical items for adverbs are as follows:

(99) Templates for adverbs

a. *Surface Form* ⊢ V/V: λx1λx2λx3.and (x1 (*Relation to the Denoted Place* x3) x2) (be *Denoted Place* x3)

b. *Surface Form* ⊢ V/V: λx1λx2λx3.and (x1 x2 x3) (x3 *Relation to the Denoted Time Denoted Time*)

4.4.2 Representing IM

IM consists of morphological operations following syntactic rules (the C region in the partition by Dowty (1979)). Fully productive, IM rarely involves semantic selection criteria. On the other hand, category selection and affix positioning is strict. Chapter 3 discusses these issues in detail.

Constructing lexical items for IM, we follow the same general rules as Section 4.4.1. First, each morpheme introduces a new lambda term to the LF. Second, the innermost variable represents the object / action / property denoted by the root, and the outermost variable represents the object / action / property denoted by the lemma.

Turkish IM divides into two: nominal inflection and verbal inflection. Nominal inflection is fairly simple. There are only three main classes of markers: plural, possessive, case and the relative marker. We ignore the relative marker, due to its high complexity and little overlap with thesis claims.

For surface forms of bound morphemes, we use capital letters to denote meta-phonemes: A {a,e}, C {c,ç}, D {d,t}, G {g,k}, I {ı,i,u,ü}, K {ğ,k}.

(100) Nominal inflection

a. *-lAr* ⊢ N\N: λx1λx2.and (x1 x2) (be plural x2)

b. *-I* ⊢ nACC\N: λx1λx2.x1 x2

c. *-A* ⊢ nDAT\N: λx1λx2.x1 x2

d. *-X* ⊢ nNOM\N: λx1λx2.x1 x2

e. *-Im* ⊢ N\N: λx1λx2.and (own x2 speaker) (x1 x2)

f. *-In* ⊢ N\N: λx1λx2.and (own x2 hearer) (x1 x2)

g. *-I* ⊢ N\N: λx1λx2.and (own x2 3rdperson) (x1 x2)

The plural marker introduces a lambda term indicating plurality of the denoted object. It acts like a predicate in bound form. Possessives are represented in a similar way. Possessive markers indicate the ownership / possession of the denoted object by the corresponding person.

If we accept there is a zero marker indicating nominative case, case marking is obligatory. Case marking ensures that each noun phrase assumes its intended place within the thematic structure.

The fact that case markers have no semantic content presents some complications. Most importantly, it becomes possible to falsely identify case markers inside other morphemes. For instance, the accusative marker *-I* does not exist inside the 1st person possessive *-Im*, but a segmental discovery mechanism would find it. Perhaps this is indeed the reason why the acquisition of case is so late.

The templates we use for building nominal IM are as follows:

(101) Templates for nominal IM

a. Plural: *-lAr* $\vdash$ N\N: λx1λx2.and (x1 x2) (be *Plural* x2)

b. Case: *Surface Form* $\vdash$ n*Case*\N: λx1λx2.x1 x2

c. Possessive: *Surface Form* $\vdash$ N\N: λx1λx2.and (own x2 *Possessor*) (x1 x2)

Verbal inflection is a little bit more involved. As discussed in Chapter 3, verbs can be inflected on quite a few positions. These positions largely correspond to grammatical features one-to-one; therefore, it remains possible to devise a well-organized set of of lexical items. Within the scope of this thesis, we only work with a subset of Turkish finite verb inflection, but the interested reader is referred to Kunter and Bozşahin (2018) for a more complete analysis.

Turkish verbal inflection divides into four groups: voice, TAM, copula and person. While TAM, copula and person markers fulfill grammatical features, voice markers change the argument structure of the verb. Although it is hard to formulate the change in the argument structure without an explosion in the number of lambda terms, we carry out some trials involving voice markers.

TAM markers convert a bare verb to a finite verb. Each marker must have variations with different levels of arity, so that they may fit verbs with different valency. All three copular markers share their form with TAM markers, but their semantics are different. Using the framework laid out by Moens and Steedman (1988), we represent reference time with tref, and speech time with t0. When there is only reference time, it is equal to the speech time by default. This approach is necessary, because the copular marker may or may not apply; semantics of the finite verb must be complete either way.

At least one TAM marker is obligatory in Turkish finite verbs. Based on this fact, we use different syntactic categories for the bare verb and the TAM-marked verb. The bare verb is a function onto V. Transitive verbs are V\N. The finite verb is a function onto S/N, able to receive a subject to form a complete sentence.

If we assume the existence of a third-person singular marker, person markers are also obligatory. Since the overt pronoun is optional in Turkish, person marker is capable of fulfilling the role of subject.

Since we assume the hearer freely examines segmentation alternatives of a word, it is possible to pick up multiple morphemes in one segment. This results in many alternative analyses for the word. As shown in Chapter 5, the hearer is expected to prefer the least flexible alternative; which tends to be full decomposition for IM.

(102) TAM markers

a. *-DI* ⊢ (S/N)\V: λx1λx2λx3.and (x3 < tref) (x1 x2 x3)

b. *-DI* ⊢ (S/N)\V: λx1λx2λx3λx4.and (x4 < tref) (x1 x2 x3 x4)

c. *-m* ⊢ S\(S/N): λx1λx2λx3.and (be speaker x2) (x1 x2 x3)

d. *-m* ⊢ S\(S/N): λx1λx2λx3λx4.and (be speaker x3) (x1 x2 x3 x4)

e. *-DIm* ⊢ S\V: λx1λx2λx3.and (be speaker x2) (and (x3 < tref) (x1 x2 x3))

f. *-DIm* ⊢ S\V: λx1λx2λx3λx4.and (be speaker x3) (and (x4 < tref) (x1 x2 x3 x4))

g. *-yDI* ⊢ (S/N)\(S/N): λx1λx2λx3.and (tref < t0) (x1 x2 x3)

h. *-yDI* ⊢ (S/N)\(S/N): λx1λx2λx3λx4.and (tref < t0) (x1 x2 x3 x4)

i. *-DIydI* ⊢ (S/N)\V: λx1λx2λx3.and (tref < t0) (and (x3 < tref) (x1 x2 x3))

j. *-DIydI* ⊢ (S/N)\V: λx1λx2λx3λx4.and (tref < t0) (and (x4 < tref) (x1 x2 x3 x4))

k. *-DIydIm* ⊢ S\V: λx1λx2λx3.and (be speaker x2) (and (tref < t0) (and (x3 < tref) (x1 x2 x3)))

l. *-DIydIm* ⊢ S\V: λx1λx2λx3λx4.and (be speaker x3) (and (tref < t0) (and (x4 < tref) (x1 x2 x3 x4)))

The templates we use for building verbal IM are as follows:

(103) Templates for verbal IM

a. TAM: *Surface Form* ⊢ (S/N)\V: λx1λx2λx3.and (x3 *Temporal Relation of Event Time to Reference Time* tref) (x1 x2 x3)

b. Copula: *Surface Form* ⊢ (S/N)\(S/N): λx1λx2λx3.and (tref *Temporal Relation of Reference Time to Speech Time* t0) (x1 x2 x3)

c. Person: *Surface Form* ⊢ S\(S/N): λx1λx2λx3.and (be *Person* x2) (x1 x2 x3)

4.4.3 Representing DM

We only study DM that can be considered productive to some extent. Speakers of a language cannot be expected to discover obscure affixes that apply on one or two distinct stems. Therefore, we expect each lexical item to represent the rule for a derivation process. Using Aslan et al. (2018) as a baseline in Chapter 3, we took inventory of derived forms in Turkish. In this section, we present a sample lexicon of Turkish DM. We aim for this lexicon to be representative of Turkish DM, spanning across different syntactic categories and lexical relations.

Most IM only change the LF. Some, like voice and TAM, also change the argument structure, but keep to the same Part-of-Speech (POS). DM goes beyond the limits of IM and is capable of making much more substantial changes on the stem, both in terms of category and semantics. For consistency, we have to represent these processes using the same rules described earlier.

DM can be considered as a shortcut for syntactic alternatives. In fact, many affixes originate from free forms occurring in a syntactic relation to stems. This is evidenced by etymological studies of Turkish affixes such as Alibekiroğlu (2019) and İlhan and Öz (2019). After they become bound forms, derivational morphemes may assume new, often related, semantics. Morphemes that remain productive to some extent continue to reflect the syntactic structure of the underlying derivation. The logical form of a morpheme is identical to the one that arises from the equivalent syntactic construction.

Ideally, the complexity of a constituent would have no effect in its interaction with other constituents. Including a simple, separate lambda term in the LF for each new semantic contribution makes this possible.

NVD and JVD affixes can be represented easily, thanks to the property formulation of nouns. Most suffixes of the NVD class choose a thematic role of the verb and derive the name of an entity that could assume that role. Others derive the name of an act.

(104) Patient NVD

a. *alacak*: The object to be taken (NVD_ACAK)

b. *verecek*: The object to be given (NVD_ACAK)

c. *basamak*: The object that is stepped on (NVD_AMAK)

d. *tutamak*: The object that is grabbed (NVD_AMAK)

(105) Location NVD

a. *durak*: The place where one stops (NVD_AK)

b. *yatak*: The place where one lies (NVD_AK)

(106) Agent NVD

a. *bakan*: The one who looks (NVD_AN)

b. *kapan*: The object that catches (NVD_AN)

c. *süzgeç*: The object that filters (NVD_GAC)

d. *büyüteç*: The object that magnifies (NVD_GAC)

(107) Agent for intransitives / Theme for transitives

a. *gelenek*: The habits that have come (NVD_ANAK)

b. *yetenek*: The skills that suffice (NVD_ANAK)

c. *tutanak*: The object that is held (NVD_ANAK)

d. *ödenek*: The amount that is paid (NVD_ANAK)

(108) Name of the act for intransitives / Theme for transitives

a. *eğlence*: The act of having fun (NVD_CA)

b. *dinlence*: The act of resting (NVD_CA)

c. *sakınca*: What is to be avoided (NVD_CA)

d. *düşünce*: What is thought (NVD_CA)

(109) Name of the act + A sense of manner

a. *yürüyüş*: The act of walking (NVD_YIS)

b. *uçuş*: The act of flying (NVD_YIS)

These examples demonstrate how NVD processes select the agent, patient, location, theme or the action itself as the content of the new noun. This selection is often made according to the arity of the host verb. The fact that the choice of thematic relation is quite consistent is an important clue on how derivation interacts with the argument structure.

An important observation here is how the argument structure of a verb must be preserved even after a noun is derived from it. Once the argument structure of a verb is represented in a logical form, it is there to stay. We can fulfill individual arguments, for instance, we may reduce the number of bound variables by replacing them with Skolem terms. However, the argument structure is kept till the end, no matter what operations take place during the derivation process.

Grimshaw (1990) mentions the generally accepted dichotomy between result and process nominals, but suggests another dichotomy between complex event nominals and others. She claims that the real distinction between the two kinds of nominals comes from their having or lacking argument structures. She points out that if a nominal lacks aspectual analysis, it will also lack an argument structure; otherwise, it will have an argument structure.

This dichotomy has a clear parallel in our representation of deverbal nominals. When we derive a complex event nominal from a verb, it seems we keep the argument structure of the host verb. On the other hand, when the result is not a complex event nominal, we are not allowed to keep the argument structure. Perhaps this is an indication that complex event nominals are actually derived every time and never lexicalized, but other nominals are lexicalized and the argument structures from host verbs are bypassed. As an added benefit, nominals that are locally ambiguous in terms of denoting complex events are automatically disambiguated during parsing, because the existence of an argument forces the reading with the argument structure, and vice versa.

In the following examples, we present both our formulation and an alternative Neo-Davidsonian formulation for comparison. In our formulation the outermost variable is the default indicator of the derived form's reference. The Neo-Davidsonian formulation lacks a direct mechanism for establishing reference.

(110) NVD and JVD Affixes

a. Stem: *bul* ⊢ V\N: λx1λx2λx3.find x1 x2 x3

b. Lemma: *bulgu* ⊢ N: λx1λx2λx3.and (find x1 x2 x3) (be anon x2) (be unspecific x3)

c. Affix: *-GI* ⊢ N\(V\N): λx1λx2λx3λx4.and (x1 x2 x3 x4) (be anon x3) (be unspecific x4)

d. Stem: *yığ* ⊢ V\N: λx1λx2λx3.pile x1 x2 x3

e. Lemma: *yığıntı* ⊢ N: λx1λx2λx3.and (pile x1 x2 x3) (be anon x2) (x3 < tref)

f. Affix: *-IntI* ⊢ N\(V\N): λx1λx2λx3λx4.and (x1 x2 x3 x4) (be anon x3) (x4 < tref)

g. Stem: *sına* ⊢ V\N: λx1λx2λx3λx4.and (action test) (agent x1) (patient x2) (instrument x4) (time x3)

h. Lemma: *sınav* ⊢ N: λx1λx2λx3λx4.and (and (unspecified x1) (unspecified x2) (unspecified x3)) (and (action test) (agent x1) (patient x2) (instrument x4) (time x3))

i. Affix: *-v* ⊢ N\(V\N): λx1λx2λx3λx4λx5.and (and (unspecified x2) (unspecified x3) (unspecified x4)) (x1 x2 x3 x4 x5)

j. Stem: *yaz* ⊢ V: λx1λx2.write x1 x2

k. Lemma: *yazman* ⊢ N: λx1λx2.and (write x1 x2) (and (be professional x1) (unspecified x2))

l. Affix: *-mAn* ⊢ N\V: λx1λx2λx3.and (x1 x2 x3) (and (professional x2) (unspecified x3))

m. Stem: *yaz* ⊢ V: λx1λx2.and (action write) (agent x1) (time x2)

n. Lemma: *yazman* ⊢ N: λx1λx2.and (and (action write) (agent x1) (time x2)) (and (professional x1) (unspecified x2))

o. Affix: *-mAn* ⊢ N\V: λx1λx2λx3.and (x1 x2 x3) (and (professional x2) (unspecified x3))

The templates we use for building NVD and JVD affixes are as follows:

(111) Templates for NVD and JVD affixes

a. Patient NVD (Present) ⊢ N\(V\N): λx1λx2λx3λx4.and (x1 x2 x3 x4) (be anon x3) (unspecified x4)

b. Patient NVD (Past) ⊢ N\(V\N): λx1λx2λx3λx4.and (x1 x2 x3 x4) (be anon x3) (x4 < tref)

c. Instrument NVD (Neo-Davidsonian) ⊢ N\(V\N): λx1λx2λx3λx4λx5.and (and (unspecified x2) (unspecified x3) (unspecified x4)) (x1 x2 x3 x4 x5)

d. Agent NVD (Neo-Davidsonian) ⊢ N\V: λx1λx2λx3.and (x1 x2 x3) (be professional x2) (unspecified x3)

VND and VJD affixes are also relatively straightforward in such an analysis. Derivation introduces the argument structure of the verb. By far the most productive VND / VJD affix is *-lA*, followed by *-lAn* and *-lAş*. *-lA* contributes a generic meaning by indicating a generic action takes place that is somehow related to the object denoted by the stem. *-lA* is so productive that its combination with voice markers form *-lAn*, *-lAş* and *-lAt*, which in turn assumed slightly different meanings.

While there are many other affixes, most of them derive fewer than 10 distinct lemmas. Other than VND_DA and VND_KIR forming verbs from onomatopoeia, only VND_AL, VND_AR, VND_IMSA and VND_SA can be considered productive.

VJD class affixes can be divided into two groups. The first group (four affixes) is used to indicate increases in the level of gradable adjectives. The second group (two affixes) indicate a change in the subject's view of the object. We present only one example, *hazırla-* 'prepare' from this second group.

(112) VND and VJD Affixes

a. Stem: *el* ⊢ N: λx1.be hand x1

b. Lemma: *elle* ⊢ V\N: λx1λx2λx3.and (be hand x3) (do (with x3) sth x1 x2)

c. Affix: *-lA* ⊢ (V\N)\N: λx1λx2λx3λx4.and (x1 x4) (do (with x4) sth x2 x3)

d. Stem: *miyav* ⊢ N: λx1.be meow_sound x1

e. Lemma: *miyavla* ⊢ V: λx1λx2λx3.and (be meow_sound x3) (sound (like x3) x1 x2)

f. Affix: *-lA* ⊢ V\N: λx1λx2λx3λx4.and (x1 x4) (sound (like x4) x2 x3)

g. Stem: *hazır* ⊢ N/N: λx1λx2.and (x1 x2) (be ready x2)

h. Lemma: *hazırla* ⊢ V\N: λx1λx2λx3λx4.and (and (x3 x4) (be ready x4)) (make (like x3) x1 x2)

i. Affix: *-lA* ⊢ (V\N)\(N/N): λx1λx2λx3λx4λx5λx6.and (x1 x5 x6) (make (like x5) x2 x3 x4)

j. Stem: *cesaret* ⊢ N: λx1.be courage x1

k. Lemma: *cesaretlen* ⊢ V: λx1λx2λx3.and (be courage x3) (obtain x3 x1 x2)

l. Affix: *-lAn* ⊢ V\N: λx1λx2λx3λx4.and (x1 x4) (obtain x4 x2 x3)

m. Stem: *su* ⊢ N: λx1.be water x1

n. Lemma: *susa* ⊢ V: λx1λx2.and (be water x3) (need x3 x1 x2)

o. Affix: *-sA* ⊢ V\N: λx1λx2λx3λx4.and (x1 x4) (need x4 x2 x3)

The generic nature of *-lA* and its variations make it especially hard to formulate a precise representation of their semantics. The thematic relations dimension discussed in Chapter 3 come to the rescue. Instead of the above formulation, we can adopt a Neo-Davidsonian (Parsons, 1995) formulation for generic actions.

(113) Neo-Davidsonian formulation of generic actions

a. Lemma: *elle* ⊢ V: λ x1λ x2λ x3λ x4.and (be hand x4) (and (agent x1) (patient x2) (instrument x4) (time x3))

b. Affix: *-lA* ⊢ V\N: λ x1λ x2λ x3λ x4λ x5.and (x1 x5) (and (agent x2) (patient x3) (instrument x5) (time x4))

Although it sacrifices the clarity of reference for the derived form, this formulation is a more precise alternative. It also represents oblique objects and adverbs in a much more straightforward manner. On the other hand, it increases the number of lambda terms by a large amount. Templates and derivations become a lot more complicated when multiple levels of derivation are present. In order to keep our focus on the thesis claims, we avoid oblique objects and generally use the simpler formulation as a shortcut for the Neo-Davidsonian formulation.

The templates we use for building VND and VJD affixes are as follows:

(114) Templates for VND and VJD affixes

a. Generic VND ⊢ V\N: λx1λx2λx3λx4.and (x1 x4) (do (with x4) sth x2 x3)

b. Onomatopoeia VND ⊢ V\N: λx1λx2λx3λx4.and (x1 x4) (sound (like x4) x2 x3)

c. Generic VJD ⊢ V\(N/N): λx1λx2λx3λx4λx5λ x6.and (x1 x5 x6) (make (like x5) x2 x3 x4)

d. Specific VND ⊢ V\N: λx1λx2λx3λx4.and (x1 x4) (*Action* x4 x2 x3)

e. Davidsonian VND ⊢ V\N: λx1λx2λx3λx4λx5.and (x1 x5) (and (agent x2) (patient x3) (instrument x5) (time x4))

VVD affixes are few and do not display a wide variety. The productive affixes we identified in Chapter 3 were repetition and diminutive. The Davidsonian formulation is slightly more preferable due to the specification of manner.

VVD suffixes leave the argument structure unchanged, but only modify the semantic content. Their application depends on the verbal class in Moens and Steedman (1988) given in Table 7. Most of these suffixes convert an atomic event to a process, or at least contribute a sense of repetition and inconclusiveness of action.

(115) Productive VVD Affixes

a. *durakla* (VVD_AKLA)

b. *itekle* (VVD_AKLA)

c. *eşele* (VVD_ALA)

d. *şaşala* (VVD_ALA)

e. *itiştir* (VVD_USTUR)

f. *veriştir* (VVD_USTUR)

We do not go into much detail on the temporal ontology of these affixes. We present only one example on VVD affixes:

(116) VVD Affixes

a. Stem: *kır* ⊢ V: λx1λx2λx3.and (action break) (agent x1) (patient x2) (time x3)

b. Lemma: *kırp* ⊢ V: λx1λx2λx3.and (manner diminutive) (and (action break) (agent x1) (patient x2) (time x3))

c. Affix: *-p* ⊢ V\V: λx1λx2λx3λx4.and (manner diminutive) (x1 x2 x3 x4)

The template we use for building VVD affixes is as follows:

(117) Templates for NVD and JVD affixes

a. Manner VVD ⊢ V\V: λx1λx2λx3λx4.and (manner *Manner*) (x1 x2 x3 x4)

Denominal nominal derivational morphemes display a wide variety of functions. Polysemy is also more prevalent with denominal nominal them. Each instance of polysemy must be represented by a different lexical item, containing the appropriate LF. We cover a fraction of those, but care to obtain a representative sample. When introducing the event structure for the DM semantics, we do not represent the time by a dedicated bound variable, because the time is always unspecified.

(118) Polysemous uses of *CI*

a. CI_1: Names of agents with an affinity towards a specific action (selecting names of actions or nouns expressing action due to semantic shift)

b. CI_2: Names of agents engaged in a specific activity (selecting names of actions or nouns expressing action due to semantic shift)

c. CI_3: Names of professionals (selecting names of instruments and locations associated with a profession)

d. CI_4: Names of proponents of an ideological position (selecting names of people to whom the ideology is attributed)

(119) Polysemous uses of *-lIK*

a. *-lIK*$_1$: Names of apparel (selecting names of body parts)

b. *-lIK*$_2$: Names of professions (selecting names for professionals)

c. *-lIK*$_3$: Names for objects or areas dedicated for some use (selecting names for relevant use)

d. *-lIK*$_4$: Names for adopted family members (selecting names for close relatives)

e. *-lIK*$_5$: Names for banknotes (selecting round numbers)

f. *-lIK*$_6$: Names for types of periodic income (selecting time intervals)

g. *-lIK*$_7$: Names of containers (selecting durable items)

The case of *-lIK* is an interesting one, as it clearly shows the extent of polysemy. Since each of these uses are productive in their own right, with relatively clear semantic selection criteria, it is possible to write rough categorial rules to represent them. Still, they cannot be expected to be fully productive and be able to apply on every base that fits the selection criteria. Therefore, categorial rules for derivational morphemes are inevitably prone to deriving nonsensical forms. On the other hand, natural use routinely generates new forms that cannot be derived by the "ordinary" rules. In these two ways, there will always be a mismatch between the forms derivable by our rules and the forms observed in the data. Still, this is not a problem. Even if there are few lexicalized forms derived by an affix, we often observe that new forms can easily be invented and understood, provided that the general categorial rules are observed.

In our inventory, all JJD affixes except one modify the intensity of the adjective. Thus, they apply on gradable adjectives. The exception is JJD_MSAR that makes the derivations *iyimser* 'optimist', *karamsar* 'pessimist'. To reflect the change in the grade of the adjective, we introduce the threshold. For instance, when JJD_MSI is applied, the grade of the adjective attained by the noun is thought to be lower than the threshold.

(120) Formulations for denominal nominal DM

a. Stem: *kitap* ⊢ N: λx1.be book x1

b. Lemma: *kitapçı* ⊢ N: λx1λx2.and (be book x2) (sell x2 x1)

c. Affix: *-CI* ⊢ N\N: λx1λx2λx3.and (x1 x3) (sell x3 x2)

d. Stem: *servis* ⊢ N: λx1.be shuttle x1

e. Lemma: *servisçi* ⊢ N: λx1λx2.and (be shuttle x2) (drive x2 x1)

f. Affix: *-CI* ⊢ N\N: λx1λx2λx3.and (x1 x3) (drive x3 x2)

g. Stem: *Atatürk* ⊢ N: λx1.be Atatürk x1

h. Lemma: *Atatürkçü* ⊢ N:λx1λx2.and (be Atatürk x2) (believe (in x2) x1)

i. Affix: *-CI* ⊢ N\N: λx1λx2λx3.and (x1 x3) (believe (in x3) x2)

j. Stem: *Ankara* ⊢ N: λx1.be Ankara x1

k. Lemma: *Ankaralı* ⊢ N: λx1λx2.and (be Ankara x2) (be (from x2) x1)

l. Affix: *-lI* ⊢ N\N: λx1λx2λx3.and (x1 x3) (be (from x3) x2)

m. Stem: *doktor* ⊢ N: λx1.be doctor x1

n. Lemma: *doktorculuk* ⊢ N: λx1λx2.and (be doctor x2) (be (to_impersonate x2) x1)

o. Affix: *-CIlIK* ⊢ N\N: λx1λx2λx3.and (x1 x3) (be (to_impersonate x3) x2)

p. Stem: *tat* ⊢ N: λx1.be sweetness x1

q. Lemma: *tatlı* ⊢ N/N: λx1λx2λx3λx4.and (x1 x2) (be sweetness x4) (have (x3 x4) x2)

r. Affix: *-lI* ⊢ (N/N)\N: λx1λx2λx3λx4λx5.and (x2 x3) (x1 x5) (have (x4 x5) x3)

s. Lemma: *tatsız* (N/N): λx1λx2λx3λx4.and (x1 x2) (be sweetness x4) (lack (x3 x4) x2)

t. Affix: *-sIz* ⊢ (N/N)\N: λx1λx2λx3λx4λx5.and (x2 x3) (x1 x5) (lack (x4 x5) x3)

u. Stem: *bu* ⊢ N/N: λx1λx2.and (x1 x2) (be close x2 speaker)

v. Lemma: *bura* ⊢ N: λx1λx2.and (x1 x2) (be close x2 speaker) (be place x1)

w. Affix: *-rA* ⊢ N\(N/N): λx1λx2λx3.and (x1 x2 x3) (be place x2)

x. Stem: *mavi* ⊢ N/N: λx1λx2λx3.and (x1 x2) (be (x3 blue) x2)

y. Lemma: *mavimsi* ⊢ N/N: λx1λx2λx3.and (x3 < threshold) (and (x1 x2) (be (x3 blue) x2))

z. Affix: *-msI* ⊢ (N/N)\(N/N) : λx1λx2λx3λx4.and (x4 < threshold) (x1 x2 x3 x4)

The templates we use for building denominal nominal DM are as follows:

(121) Templates for denominal nominal affixes

a. NND ⊢ N\N: λx1λx2λx3.and (x1 x3) *Lambda Term to Indicate DM Semantics*

b. NJD ⊢ N\(N/N): λx1λx2λx3.and (x1 x2 x3) *Lambda Term to Indicate DM Semantics*

c. JND ⊢ (N/N)\N: λx1λx2λx3λx4λx5.and (x2 x3) (x1 x5) *Lambda Term to Indicate DM Semantics*

d. JJD ⊢ (N/N)\(N/N): λx1λx2λx3λx4.and (x1 x2 x3 x4) *Lambda Term to Indicate DM Semantics*

Arity of DM must always be accurately determined. DM follows different templates for creating the argument structures of intransitive and transitive verbs. This is also true for derivation of morphemes with transitive meanings.

(122) Arity of DM

a. Stem: *kitap* ⊢ N: λx1.be book x1

b. Lemma: *kitapçı* ⊢ N: λx1λx2.and (be book x2) (sell x2 x1)

c. Affix: -*CI* ⊢ N\N: λx1λx2λx3.and (x1 x3) (sell x3 x2)

d. Stem: *gözlük* ⊢ N: λx1λx2.and (be eye x2) (wear (on x2) x1 anon)

e. Lemma: *gözlükçü* ⊢ N: λx1λx2λx3.and (and (be eye x3) (wear (on x3) x2 anon)) (sell x2 x1)

f. Affix: -*CI* ⊢ N\N: λx1λx2λx3λx4.and (x1 x3 x4) (sell x3 x2)

Since we base our computations on written-form, we cannot present a straightforward representation of zero-derivation. Perhaps the inclusion of clues based on intonation and category could be sufficient for the hearer to deduce that zero-derivation has taken place. For instance, the use of *kırmızıya* 'to the red one' instead of *kırmızı arabaya* 'to the red car' implies zero-derivation by case marking an adjective. This is a non-segmental, categorial clue; which is outside the scope of our approach.

Deadverbials and adverb deriving affixes are rarer. Only 3 AJD affixes exist. There is only 1 affix that applies on adverbs, and that affix is not a productive one. Possibly, deriving from Turkish adverbs is too complex or counter-intuitive for some reason.

(123) Formulations for AND and AAD affixes

a. *çocuk* ⊢ N: λx1.be child x1

b. *çocukça* ⊢ V/V: λx1λx2λx3.and (x1 (like x3) x2) (be child x3)

c. -*CA* ⊢ (V/V)\N: λx1λx2λx3λx4.and (x2 (like x4) x3) (x1 x4)

d. *iç* ⊢ N: λx1.be inside x1

e. *içeri* ⊢ V/V: λx1λx2λx3.and (x1 (towards x3) x2) (be inside x3)

f. -*ArI* ⊢ (V/V)\N: λx1λx2λx3λx4.and (x2 (towards x4) x3) (x1 x4)

g. *demin* ⊢ V/V: λx1λx2λx3.and (x1 x2 x3) (x3 = t0 minus eps)

h. *demincek* ⊢ V/V: λx1λx2λx3.and (and (x1 x2 x3) (x3 = t0 minus eps)) (be small eps)

i. -*CAK* ⊢ (V/V)\(V/V): λx1λx2λx3λx4.and (x1 x2 x3 x4) (be small eps)

The templates we use for building denominal nominal DM are as follows:

Table 24: Category statistics of the baseline grammar

Category	Count
N	120
N/N	14
N (Case-Marked)	18
S	117
S/N	12
S\N	3
V	22
V/V	8
V\N	14

(124) Templates for denominal nominal affixes

a. AND ⊢ (V/V)\N: λx1λx2λx3λx4.and (x2 (like x4) x3) (x1 x4)

b. AAD ⊢ (V/V)\(V/V): λx1λx2λx3λx4.and (x1 x2 x3 x4) *Lambda Term to Indicate DM Semantics*

4.4.4 Baseline Grammar

We build a baseline grammar to support our efforts. Part of this grammar is given in the examples of the previous section. The trials in Chapter 5 are based on the baseline grammar.

Statistics regarding the lexical items included in the lexicon is given in Table 24. We pay close attention to represent a wide variety of syntactic categories in the lexicon.

The baseline grammar makes it possible to discover several dozen new morphemes. Statistics regarding these morphemes are given in Table 25.

4.5 Syntax vs. Morphology

In Section 2.1.1, we cite Sezer (1991) for his insights on the asymmetry between syntactic and morphological processes. He demonstrates this asymmetry primarily on how the two processes interact with the thematic structure and the argument structure respectively. In this section, we review Sezer (1991)'s assumptions and claims. We justify our decisions regarding CG representation based on his premises and possible counterexamples. We also present an asymmetry, but rely on different assumptions and claims.

Sezer (1991) mainly relies on the definitions of Grimshaw (1990) when setting up the structure of his theory of DM. Grimshaw (1990) explains that sentences have both a thematic structure and an

Table 25: Category statistics of the morphemes discovered by processing the baseline grammar

Category	Count
N/N\(N/N)	1
N/N\N	2
N/N\V	1
N\(N/N)	1
N\(V\N)	3
N\N	13
N\V	2
S/N\(S/N)	2
S/N\V	4
S\(S/N)	3
S\(S\N)	12
S\(S\N)\N	3
S\N	27
S\S	1
S\V	3
S\V\(S/N\V)	1
V/V\(V/V)	1
V/V\N	2
V\N	3
V\N\(N/N)	1
V\N\N	2
V\V	1
N (Case Marked)\N	6
N (Case Marked)\N (Case Marked)	2
N\N (Case Marked)	2
S\N (Case Marked)	5

argument structure. These structures often coincide, but they do not have to. Therefore, there is benefit in considering these two structures separately.

(125) Definitions of Grimshaw (1990)

a. The thematic structure establishes the conceptual relations between the participants of an event.

b. The argument structure reflects the hierarchy between a head's arguments.

c. The hierarchy of arguments is determined by the hierarchy of thematic relations, with Agent as the most prominent and the Theme as the least prominent.

The logical forms we constructed to represent sentence-level structures follow these definitions. Verbal LF invariably have a thematic structure. In that structure, the agent is represented by the outermost bound variable. We do not go into oblique objects to avoid clutter, but direct objects are more embedded than the Agent.

One complication arises due to the explicit representation of tense. The additional variable that represents time must be inserted somewhere in the hierarchy. This could be avoided with a slightly different representation scheme (Davidsonian, or partially Davidsonian?), but we keep the current scheme as a shorthand.

Sezer (1991) makes several claims on top of these definitions.

(126) Claims of Sezer (1991)

a. Every lexical item has a thematic structure associated with it.

b. DM is completely confined to the lexicon. This is true for both productive and unproductive derivational processes.

c. Agent is always the external argument. DM operates on the argument structure, and not on the thematic structure.

The first two claims can be considered as premises, while the last one is more like a conclusion. The validity of the first two claims are argued at length by Sezer (1991), based on a number of examples. The examples are drawn from act and fact nominals and voice marking.

(127) Examples of nominalization from Sezer (1991)

a. *Düşmanın şehri istilası* 'enemy's invasion of the city'

b. *Düşmanın şehri istila etmesi* 'enemy's invading the city'

c. *Doktorun hastayı muayenesi* 'doctor's examination of the patient'

d. *Doktorun hastayı muayene etmesi* 'doctor's examining the patient'

e. *Sekreterin yazıyı daktilo etmesi* 'secretary's typing the script'

f. **Sekreterin yazıyı daktilosu* '*secretary's type of the script'

As can be seen from these examples, some words allow constructions implying an event structure, even if they are not verbs or nominalizations. This is evidence towards the first assumption. However, it is equally interesting that constructions of this sort are not possible for all words of comparable status. *daktilo* 'typewriter' does not imply an event structure in the same way.

This irregularity is one reason why at least some properties of derivational operations must be kept in the lexicon. However, it is not sufficient to entirely restrict DM to the lexicon.

(128) Examples of JVD from Sezer (1991)

a. *Çalışkan* 'hard-working' (External Argument)

b. *Solgun* 'faded' (Internal Argument)

In both examples, the derivation process operates on the only argument of the stem. When the verb licenses several thematic relations, it is indeed hard to match affixes with dedicated functions.

The third claim aims to explain the semantic asymmetry between examples such as *çalışkan* 'hard-working' and *solgun* 'faded'. From the first two assumptions, we know that all lexical items have a thematic structure, but DM does not give us regular rules to explain derived forms with respect to individual thematic relations. However, there is still an apparent asymmetry between agent / theme derivations; which can be attributed to their positions in the argument structure. The third assumption follows: DM does not refer to the thematic structure, but it refers to the argument structure.

We acknowledge that particularly NVD and VND affixes express a variety of functions with respect to the thematic structure. The many-to-many relationship between affixes and thematic relations makes it difficult to devise clear-cut rules. For instance, *bil-* 'to know' pairs with many different affixes to derive a variety of forms.

(129) Forms derived from *bil-*

a. *bilge* 'wise' (External Argument)

b. *bilgin* 'scholar' (External Argument)

c. *bilgiç* 'know-it-all' (External Argument)

d. *bilgi* 'knowledge' (Internal Argument)

e. *bilim* 'science' (Internal Argument)

f. *bilinç* 'consciousness' (Internal Argument)

g. *bilmece* 'riddle' (Internal Argument)

In Section 3.3.7, by classifying NVD and VND operations with respect to thematic relations, we try to discover a structure among all this irregularity. We argue that the variety of functions for each affix can be explained by polysemy. Moreover, we observe that some thematic relations concentrate on specific affixes. For instance, *-GI* derives many stimulus NVD.

(130) Stimulus NVD

a. *sezgi* 'intuition'

b. *duygu* 'emotion'

c. *coşku* 'enthusiasm'

d. *sevgi* 'love'

e. *kaygı* 'concern'

f. *saygı* 'respect'

g. *tutku* 'passion'

h. *övgü* 'praise'

i. *algı* 'perception'

j. *yergi* 'ridicule'

From our review in Chapter 2, we know that speakers have an awareness of DM and they use this knowledge during comprehension. Examples of derivation taking phrasal scope (see Section 3.3.2) also suggest that at least some DM can be represented with syntax-like rules. Therefore, we argue that DM should not be confined to the lexicon, at least not totally. This is the crucial justification behind our investigation of DM.

Regarding the asymmetry between morphology and syntax, we recognize that they differ in the way they operate on thematic / argument structures. In this chapter, we presented in detail the self-imposed rules for a consistent CG formulation. One consequence of these rules is a slightly different kind of asymmetry.

(131) Self-imposed rules for a consistent CG formulation

a. An object denoted by a noun is represented by the outermost bound variable in the noun's logical form.

b. No object can be denoted by a free variable, not even named entities. LF for all lexical items represent both an argument structure and a thematic structure.

c. If a bound morpheme causes the lemma to indicate a different object than the stem; it introduces a new bound variable. This variable becomes the new outermost variable.

The first rule ensures that the next operation will always find the relevant object in the same position. The second rule ensures that a lexical item always has a thematic structure associated with it. In the case of nouns and simple adjectives, the thematic structure is trivial; because there is only one object and a property. However, for verbs and adverbs, a true thematic structure must be provided. For verbs, this is required for the sentence to be constructed. For adverbs, this is required for the verb's structure to be accommodated. The second premise is not simply an assumption in our case, because we need it for consistency across multiple levels of derivation. These rules are parallel to the premises by Sezer (1991).

The third rule is required for consistency. Using the CG framework, we need to work with bound variables that establish reference to an object. On the sentence level, constituents come together to fill the thematic structure licensed by the verb. Each constituent refers to a certain object (no matter its specificity or definiteness) and is marked by a case that indicates its thematic relation. Adjective clauses and noun clauses may occur, but they only establish a relation between their constituents; the constituents do not cease to exist.

On the other hand, morphological processes do not just combine constituents in this way. Constituents of a DM operation cease to exist and end up creating a reference to an entirely new object. This is most obvious in NND operations. *kitap* 'book' and *kitaplık* 'bookshelf' refer to completely different objects. We keep a reference to *kitap* within the LF of *kitaplık* to reflect its semantics, but there is no actual reference to a book inside *kitaplık*. A similar explanation can be made for other derivation classes. For instance, JND operations create reference to a property by abolishing reference to an object; while VJD operations create reference to an event by abolishing reference to a property.

As a result, morphological operations create a structure to represent the reference to a particular object. When a new affix brings new structure, a new outermost variable takes over the reference. After all morphological operations are completed, the outermost variable represents the object / property / event denoted by the lemma on the sentence level. Constituents of the sentence level participate as equal members in the matrix thematic structure. (We pass over the phrase level for simplicity. In essence, the phrase level also operates on principles of syntax.)

In order to apply these principles consistently, we have to introduce a slight but important asymmetry in the templates we use for constructing LF for new morphemes. Bound morphemes process previous references following two principles. First, they process the variables of the stem in the same order as the stem. Second, the innermost variable of the stem remains as the innermost variable of the lemma. In this fashion, the lemma takes over the internal structure of the stem without making any changes.

On the other hand, free morphemes process their arguments differently. They operate on the sentence level, where only the outermost variable from each constituent is able to participate on the matrix thematic structure. (Again we pass over the phrase level. Phrase level applications make it possible for a free morpheme to process multiple bound variables from its argument. These variables must still be the outermost ones.)

Morphological operations process all the variables of the stem, with the innermost one remaining the innermost variable of the lemma. Syntactic operations process the variables of the argument starting from the outermost one.

We develop our arguments and baseline grammar based on examples from Turkish. Therefore, when discussing rules, we follow the argument structure in Turkish. The innermost / outermost distinction should be considered with this limitation in mind.

We talk about free morphemes and bound morphemes in order to avoid confusion. What we mean by morphology and syntax corresponds to morphological and syntactic operations in Dowty (1979)'s typology. (The difference between morphological and syntactic rules in the same typology correspond to lexical and grammatical rules, respectively.)

As Sezer (1991) suggests, syntactic processes operate on the thematic structure. Syntactic processes' interaction with the argument structure is indirect, via the thematic structure. Regarding morphological processes, we argue that they also convey a thematic structure. Some morphological processes preserve the existing thematic structure; simply fulfilling the arguments according to the pre-existing hierarchy. In other words, these processes only interact with the argument structure. However, DM can also introduce entirely new thematic structure. NVD and VND affixes offer plenty of evidence for this. Therefore, morphological processes are able to interact with both structures.

In Section 5.2.3, we explain how we implement these principles in an algorithmic setting.

4.6 Discussion of CG

In this section, we demonstrated that the CG framework can be effectively used to represent a wide variety of lexical items, including derivational morphemes. We discussed the difficulties regarding the representation of word-internal structure and proposed rules to ensure consistency across morphological and syntactic operations. We justified these rules based on a large number of examples.

Later, we discovered an asymmetry between the CG representation of syntax and morphology. We compared this asymmetry with the findings of Sezer (1991), and contrasted our assumptions regarding the structural properties of syntax and morphology. We found that while Sezer (1991)'s claims are valid in specific cases, they need to be modified in order to be valid in the general case.

The semantic representation framework and the baseline grammar constructed in this section constitutes the basis for the trials carried out in Chapter 5.

CHAPTER 5

BAYESIAN LEARNING OF DERIVATIONAL MORPHOLOGY

In this chapter, we define our computational model, explain its interaction with meaning representation schemes, analytical tools and simplifying assumptions. Our primary aim is to demonstrate how the Conventionalized Structure (CdS) devised in Section 2.3.4 may be represented on a Bayesian Belief Network (BBN) and to flesh out a computational model that reflects the dynamics between linguistic exposure and the lexicon.

5.1 A Bayesian Model of CdS

We first discuss the BBN framework, its advantages and its operating principles. Later, we discuss CdS's representation on a BBN.

5.1.1 Bayesian Belief Networks

BBN are graphical probabilistic models (Pearl, 1988; Koller and Friedman, 2009). The structure of a BBN is a directed acyclic graph. Each node in the graph represents a variable which may take multiple values, and each edge represents dependencies between their end nodes. The graphical structure of a BBN encodes conditional independence assumptions between variables. Every variable in a BBN is independent of its non-descendants given their parents. This enables BBNs to model the joint probability distribution of a set of variables in a compact and factorized manner based on the local probability distributions of directly connected nodes.

For a start, a simple BBN is given in Figure 18 (adapted from a well-known example):

In this example, there are three probabilistic variables (or nodes), conditionally dependent on each other: rain, sprinkler and grass (being wet):

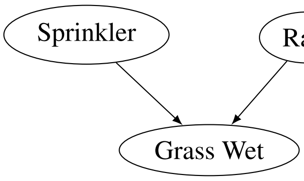
(132) a. Rain: T (raining) / F (not raining)

b. Sprinkler: T (working) / F (not working)

c. Grass Wet: T (wet) / F (dry)

Sprinkler	
T	F
0.2	0.8

Rain	
T	F
0.3	0.7



		Grass Wet	
Sprinkler	Rain	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Figure 18: A simple BBN representing the relationships statistically determining the probability of grass being wet

The grass being wet depends on both Rain and Sprinkler variables. When one event is known, we can update probabilistic information about the others. The calculations are based on the joint probability function and the chain rule of probability.

$$P(GW,S,R) = P(GW|S,R)P(S)P(R) \tag{1}$$

The same example could be represented by a joint probability table. However, such a table would not exploit the conditional independence relations (in many cases, causal relations) between different elements. It would also be much larger to accommodate all the possible configurations. Figure 26 presents the joint probability representation of the sprinkler example.

As can be observed from this example, a BBN representation is advantageous in several ways. First, it allows us to reflect the causal structure between different variables. The causal interactions are not kept implicit in the joint probabilities of certain configurations; they are explicitly represented on the BBN by conditional dependence and independence relations.

Second, BBN allows for a compact representation of probabilistic interactions. Joint probability tables grow exponentially with the number of states for each variable. As a result, they may grow so large that inference becomes intractable. In contrast, by exploiting independence relations between pairs of variables, BBN stores information in much smaller chunks.

Third, there are efficient algorithms for computing Bayesian inference on a BBN. When a value is set for a variable, Bayesian inference ensures that probabilities of all dependent variables are updated

Table 26: Joint probability table for the sprinkler example

Raining	Sprinkler Working	Grass Being Wet	Probability
T	T	T	0.0594
T	T	F	0.0006
T	F	T	0.192
T	F	F	0.048
F	T	T	0.126
F	T	F	0.014
F	F	T	0
F	F	F	0.56

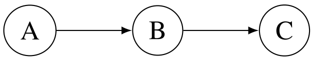


Figure 19: Junction pattern: chain

accordingly. This update mechanism becomes much more computationally demanding as the size of the network grows; therefore, the existence of algorithms tailored for BBN is a crucial advantage for computational studies. The simple structure of BBN is also a factor in this. BBN does not require parameters to be set for fine-tuning; the user just has to define the network structure and probability distributions.

These advantages of BBN are demonstrated on examples in later sections. For now, we continue discussing the structural properties of BBN. BBN being directed acyclic graphs, there are three possible junction patterns between nodes: chain, fork and collider. These junction patterns share the same skeleton; in other words, their graphs have the same nodes and links. However, the directions of arrows between nodes are different. This constitutes a crucial difference in terms of conditional dependence and inference.

When the chain junction pattern is present, as in Figure 19, we can speak of conditional dependence between A and B, and between B and C. Neither of these relations are affected by the state of the third node. On the other hand, the nature of the relation between A and C depends on B. When B is given, A and C are conditionally independent.

When the fork junction pattern is present, as in Figure 20, we can again speak of conditional dependence between A and B, and between B and C. Also, when B is given, A and C are conditionally independent. Due to conditional dependence relations being identical, the chain pattern and the fork pattern are indistinguishable.

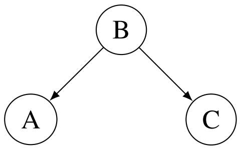


Figure 20: Junction pattern: fork

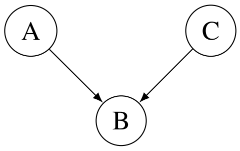


Figure 21: Junction pattern: collider

The collider pattern in Figure 21 is a substantially different configuration. This time A and C are independent given B.

Defining the network is a problem in itself. In large applications, the conditional dependence and independence relations may be too complex for a human to define. Even with complex applications, the BBN literature offers algorithms that learn the network structure, in addition to the probability distributions. As we demonstrate in Section 5.1.3, defining the network is not the issue in our case. CdS constitutes the theoretically motivated structure for our BBN implementation. We do not use BBN for discovering structure; we use it to model the structure we have in mind for morphology processing. Section 5.1.3 also points out the junction patterns we devise for representing CdS on a BBN.

Before moving on to the representation of CdS, we review the relevant literature.

5.1.2 Relevant Literature on Bayesian Learning

There has been major advances in the algorithms to efficiently learn and compute BBNs since 1980's, enabling tractable computation of large and complex BBN models (see Heckerman (2008) for a concise tutorial. As a consequence, the last few decades have seen application of BBN on a number of domains. Notable examples range from medical diagnostics to environmental modelling, system diagnostics and linguistics.

Gopnik et al. (2004) and Gopnik and Schulz (2004) explore causal learning and theory formation by children in a visual setting. Their experimental results demonstrate that causal maps are suitable representations of children's learning. These maps can be accurately modeled by BBN.

Lake et al. (2015) models Bayesian learning of visual concepts. According to their results, Bayesian Program Learning models are shown to be much more successful than deep learning algorithms. While there are considerable differences between the two fields, their promising results lend further credibility to this approach.

There are also studies that adopt a neuroscience perspective. Takahashi and Ichisugi (2017) and Ichisugi and Takahashi (2018) model cortical areas with the help of BBN. Ichisugi and Takahashi (2018) reproduce disorders specific to syntax and semantics with the help of a Bayesian network model.

Other studies such as Moghimifar et al. (2020) extract causal relationships from textual data and form a Causal Bayesian Network. They analyze the data in three stages: They extract linguistic data; they identify causal relations and they calculate conditional probabilities. They demonstrate that the Bayesian model significantly outperforms comparable methods.

These studies, while they cannot be readily adapted to the principles and purposes in this thesis, demonstrate a general tendency towards modeling a variety of mental processes on Bayesian learning algorithms. A particularly relevant line of research focuses on language / concept processing. Tenenbaum (1999) sets the stage with models for concept learning. Xu and Tenenbaum (2007) model word learning as well as generalization of word meaning. They conduct experiments with both adults and children. Their results are consistent with model predictions.

Frank et al. (2009) model learning lexical semantics by PCFG. Unlike most conventional applications of PCFG, they take into account the context and the speaker's referential intention as separate layers. However, their best lexicon includes many incorrect matches between words and objects, such as bottle-bear, hiphop-mirror and laugh-cow. No experimental results are provided to justify these errors.

Tenenbaum et al. (2011) propose using Hierarchical Bayesian models to explore learning of abstract concepts. Both visual and verbal concepts are handled in the article. Their approach resembles Xu and Tenenbaum (2007) in trying to present an effective model of generalizing concepts. Perfors et al. (2011) argue that the hierarchical phrase structure rules can be learned based on statistical clues. They consider this evidence against the argument from poverty of stimulus.

While Xu and Tenenbaum (2007), Frank et al. (2009), Tenenbaum et al. (2011) and Perfors et al. (2011) use the Bayesian learning framework, they do not reflect incremental learning nor do they allow alternative interpretations of the same surface form. In such studies, the main idea is often to create a virtual set of possible lexicons, and choose the one with the highest probability. In essence, this method is also used in most applications of PCFG, reviewed in Section 2.4.3.

Piantadosi et al. (2008) and Piantadosi (2011) are also applications of Bayesian learning in a linguistic setting. Both employ CCG for meaning representation. Piantadosi et al. (2008) model the acquisition of compositional semantics, while Piantadosi (2011) models the acquisition of number words and quantifiers. They report successful results and believe that their model can be scaled up to explain sentence production, complex syntax and semantics. For modeling the inference of Language of Thought, Piantadosi (2011) uses BBN.

Statistical approaches such as these have been implemented frequently on language learning. Most literature on psycholinguistics, as well as connectionist approaches are essentially based on statistical learning. Effect of statistical phenomena on acquisition is very well-documented. Furthermore, Bayesian learning provides an effective method for implementing the learning process. However, studies in this field often make conflicting assumptions.

The application of unsupervised learning constitutes one such issue. Frank et al. (2009), Piantadosi et al. (2008) and Piantadosi (2011) carry out statistical learning by matching a list of surface forms with a list of logical forms, both of which are provided in advance. Essentially, their algorithms are tasked with learning to match the surface forms with the correct logical forms. It is hard to understand labeling these methods as "unsupervised learning" approaches, while most linguistic information is provided to the algorithm in advance or during observations. If surface forms, logical forms and categories of observations are available, why not use an explicitly supervised setting?

Some studies such as Piantadosi et al. (2008) focus on picking the most probable logical form for a surface form. Candidate logical forms are generated by enumeration from atomic blocks (connectives, quantifiers, variables etc.), without any regard to context-based semantic clues. On the other hand, observations provide the contextual clues necessary for learning. One must ask: If context is available for entire sentences; why is it missing for individual words?

There are many other examples of the unsupervised learning approach, as reviewed in Section 2.4.3. Crucially, these studies ignore the fact that most linguistic observations, especially the ones encountered by young children, are accompanied by a rich set of contextual clues. Often, these clues may be sufficient for inferring the meaning of an expression. The main obstacle to this is the Gavagai problem Quine (1960). Perhaps it would be more appropriate to provide several alternative interpretations (meanings displaying metonymy and meronymy relations) for each form, rather than the 1-to-1 form-meaning pairs used by most "unsupervised" learning studies.

Most of these studies also do not take into account that language learning seems to be incremental. One-shot learning is obviously not possible every time, but statistical learning from processing an entire corpus cannot be deemed naturalistic, either.

Another issue is that the focus of these studies are on the statistical learning capabilities of BBN. They do not make use of much linguistic theory, despite modeling language learning. For instance, the basic algorithm in Frank et al. (2009) is almost completely domain-general. The only exception is that context-intention and intention-expression relations are modeled in addition to expression-meaning relations. The extra steps from context to expression feel artificial and not well-justified. Piantadosi et al. (2008) also does not really look into the interaction between linguistic objects, but focus on solving the statistical problem of matching a set of objects with another set of objects. We believe that a psychologically plausible model of language processing must exploit linguistic structure.

There are other common issues in the Bayesian learning literature. Studies often rely on the principle of Minimum Description Length (MDL) and its variants to justify their results. Considering the vastness of human mental capacity, it is hard to imagine why acquisition and processing would be restricted by the description length of linguistic knowledge. MDL can be used as a method of hypothesis selection, but it cannot be used as justification for a specific hypothesis. The justification may only come from an adequate linguistic structure, based on which the hypothesis selection method operates.

Another common issue is the use of free parameters. Many studies require free parameters to be set, but have no way of reliably determining a value for those parameters. For instance, Piantadosi et al. (2008) have the free parameter K, which determines the proportion of true atomic propositions in a context. The value of this parameter cannot be justified externally; K=0.8 is just shown to work well for a specific example. Piantadosi (2011) uses 3 parameters α, β and γ to control the inference algorithm. While we also make use of learning parameters, as explained in 5.2.5, it may be possible to estimate these parameters in a future experimental setting. Our parameters can be more concretely tied to whether a speaker can or cannot recognize a specific derivation process.

In the next section, we propose a model that avoids the shortcomings discussed above.

5.1.3 Representing CdS

Whether we take the interactive activation approach or the connectionist approach, meaning is built over many interacting nodes. The strengths of interactions between these nodes, symbolic or sub-symbolic, has to be accounted for on a network. CdS is one way of organizing such a network.

CdS first identifies and builds a suitable BBN for the observation. We call this inward direction, because we imagine the meaning node to be in the center of the network and the last node to be generated. The network is built starting from the segmentation and lexical nodes, proceeding towards the center, passing through derivation nodes and finally the meaning node.

At this first stage, the inputs are the lexicon and the observation's surface form. The lexicon supplies the lexical items and previously identified segmentation alternatives (unlike conventional approaches). The proposed algorithm builds all the alternative interpretations of the observation and organizes them on a BBN. The outputs are the updated segmentation alternatives list and the BBN representation of the observation.

At the second stage, the correct logical form is instantiated on the BBN, and the BBN is propagated to compute the posterior probabilities of its nodes, given the correct logical form. These probabilities are added to the previous observation counts of segmentation and lexical alternatives. The outputs of this stage are the updated segmentation alternatives list and the updated lexicon. Inputs to this stage are the lexicon, the segmentation alternatives list, the BBN and the observation's logical form.

These two stages follow one another for each observation. Typically, an observation is received, the BBN is built (inward direction) and Bayesian inference is carried out (outward direction). The updated segmentation alternatives list and the updated lexicon are used in the next cycle with a new observation. Now we shall look into these processes in more detail and define the graphical network. The inward stage starts with the generation of alternative segmentations (i.e. segmentation). The segmentation process takes into account categorial compatibility of adjacent segments; therefore, all alternatives are expected to be grammatical in the derivation stage (but they do not have to be felicitous). Details of the process are given in Section 5.2.2. Segmentation alternatives constitute the states of Segmentation Node (SN).

Second, lists of matching lexical items are generated for each unique segment, creating a set of Lexical Nodes (LN). Each LN holds the lexical items whose surface forms match one segment identified during segmentation.

Third, interpretations licensed by each segmentation alternative are derived and stored in Derivation Nodes (DN). It is possible to derive multiple interpretations from the same sequence of segments. Each interpretation is kept as a state in the relevant DN. There is a causal relation between LNs and DNs, as the selection of a lexical item may be required for a specific interpretation. DNs act as logical gates between lexical alternatives (LNs) and the set of all interpretations (MN).

Finally, all possible interpretations are gathered in a Meaning Node (MN). There is a causal relation between interpretations in MN and the segmentation node. Some interpretations are only available if a specific segmentation alternative is chosen. MN acts as a logical gate between segmentation alternatives (SN) and derived interpretations (DNs).

The links between nodes represent causal relations. SN determines the possible segmentations of the expression. It is directly linked to MN. MN operates based on input from SN and DNs.

LNs connect to DNs and provide the building blocks for derivation. When there are alternative lexical items for a segment, DNs are expected to store multiple derivations. Therefore, in the inward direction, the link is from LNs to DNs.

MN is where interpretations are checked against contextual evidence. Therefore, derivations from all DNs are collected and stored in MN. The links are from DNs to MN. Once all licensed interpretations are generated in MN, their validity can be checked against contextual information.

In the inward direction, the aim of the network is to generate all interpretations licensed by the current lexicon. Therefore, for the basic network, the inputs are the expression and the lexicon. If we did not track probability information, CdS would only build the lexicon, but would not develop any preferences between segmentation alternatives or lexical items.

Evidence reviewed in Section 2.2.1 demonstrates that neither segmentation alternatives (in the simplest case, retrieval and decomposition), nor lexical alternatives are treated equally. Speakers and hearers develop preferences towards and against some alternatives, presumably based on previous exposure. We model this effect on probabilities.

For this, we consider each node a discrete random variable. Total probability of states in each node is 1.00. States of each node compete for salience (higher probability); success of a state is evaluated based on its taking part in context-appropriate interpretations. In order to keep track of past performance, we assign each lexical item and each segmentation alternative a counter. Each time an item contributes to a correct interpretation, its counter is increased by the posterior probability calculated by Bayesian inference. Before processing the next observation, prior probabilities of each item are recalculated according to these counters. In order to keep this section manageable, we focus on the structure and leave implementation details to Section 5.2.4.

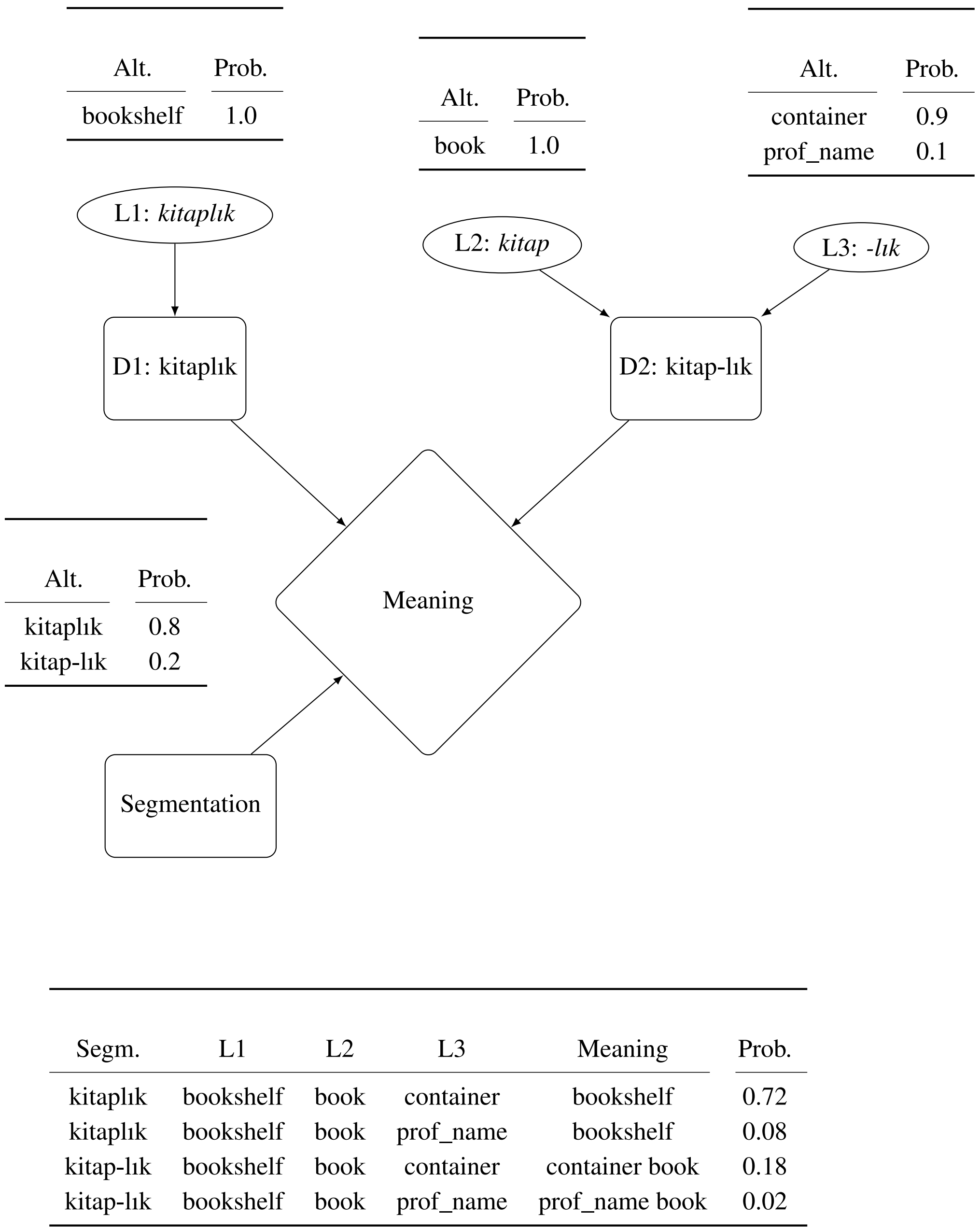
In the outward direction, information flows from MN towards SN, and LNs. After evaluating the correctness of alternative interpretations, MN rewards some segmentation alternatives and some derivations for their participation in correct interpretations. In turn, DNs propagate rewards to some lexical alternatives in a similar fashion. The reward is simply the increase on the corresponding states' counters, ensuring higher prior probability for these states in future observations. Again, we leave the details on hypothesis selection to Section 5.1.6 and focus on the structure.

BBN provides a suitable framework for representing CdS. SN and its states represent the Segmentation Layer discussed in Section 2.3.1, LNs and their states represent the Lexical Selection Layer in Section 2.3.2, and DNs and their states represent the Derivation Selection Layer in Section 2.3.3. MN collects all interpretations to carry out evaluation and triggers the feedback process. Probabilistic nature of these three nodes correspond to three levels of ambiguity that must be resolved by the hearer. Salience of each alternative is tracked by its probability.

To illustrate the structure, we present two example networks. Figure 22 demonstrates the BBN for *kitaplık* 'bookshelf':

Two junction patterns are exemplified in Figure 22. When there is only one segment in a segmentation alternative, we observe the chain junction pattern: L1 $\rightarrow$ D1 $\rightarrow$ Meaning. When there are multiple segments, we observe the collider junction pattern: L2 $\rightarrow$ D2 $\leftarrow$ L3. We always observe the collider pattern around the Meaning Node: Segmentation Node $\rightarrow$ Meaning Node $\leftarrow$ Derivation Nodes.

Because a segment might match multiple lexical items, a segmentation alternative may have multiple interpretations. These interpretations are held in MN and supplied by DNs. This means that given the observation's meaning, SN and DNs are conditionally independent. This is a strong restriction, because



Alt.	Prob.
bookshelf	1.0

Alt.	Prob.
book	1.0

Alt.	Prob.
container	0.9
prof_name	0.1

Alt.	Prob.
kitaplık	0.8
kitap-lık	0.2

Segm.	L1	L2	L3	Meaning	Prob.
kitaplık	bookshelf	book	container	bookshelf	0.72
kitaplık	bookshelf	book	prof_name	bookshelf	0.08
kitap-lık	bookshelf	book	container	container book	0.18
kitap-lık	bookshelf	book	prof_name	prof_name book	0.02

Figure 22: A toy BBN to represent the conventionalized structure of *kitaplık*

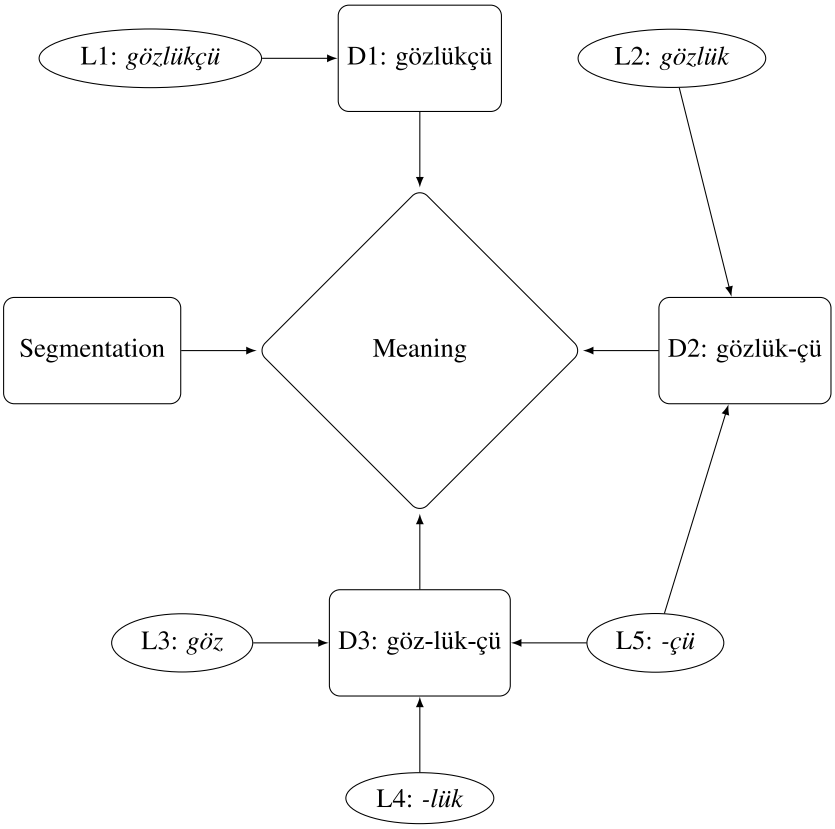


Figure 23: A BBN to represent the derivational structure of *gözlükçü*

it decouples the segmentation module from the lexical selection module; significantly reducing the problem complexity. We return to this point in Section 5.1.4.

A similar case is true between lexical nodes. Once we select the correct interpretation, this information is propagated to DNs. Next, Bayesian inference assigns posterior-probabilities to parents nodes, which are LNs. The collider pattern between LNs and DN means that once we observe the correct interpretation, lexical selection on different segments are independent from each other. This is another important restriction on the problem space, and massively reduces complexity. We also return to this point in Section 5.1.4.

With every observation, a BBN is created from scratch based on the latest available lexicon. Therefore, its contents depend on what has been learned up to the observation at hand. Figure 22 is assumed to occur after the hearer has learned one lexical item for *kitaplık* and *kitap*, each; and two lexical items for *-lık*. Consequently, the segmentation alternative kitap-lık is available for processing. If *kitaplık* were encountered earlier, there could be just one segmentation alternative and one lexical alternative. If it were encountered later, there could be multiple lexical alternatives for each segment.

Another example is *gözlükçü* ‘optician’. It offers more segmentation alternatives and unique segments. Figure 23 demonstrates its BBN:

As can be seen in L5: *-çü*, a lexical node can supply multiple derivation nodes. This is an example of the fork junction pattern: D3 ← L5 → D2.

In the inward direction, the lexicon provides prior probabilities for each segmentation and lexical alternative. These prior probabilities are used for the calculation of prior probabilities in MN. In the outward direction, Bayesian inference distributes the probability of occurrence to each segmentation and lexical alternative. Depending on their being context-appropriate, each alternative is awarded a posterior-probability, which is tallied by the lexicon. During the processing of the next observation, the lexicon recalculates all prior probabilities and the BBN is built from scratch.

Since probability values are determined by past observations, different linguistic exposure results in different lexical preferences. Long-term experience in comparable linguistic environments must still produce similar steady-state results, as can be observed in adult level performance. This behaviour is naturally produced by the proposed structure and model.

Unlike most Bayesian learning studies of language learning, we are not looking for a model that chooses the correct meaning of a form from a set of alternatives. The context decides the appropriate meaning, and the context is ever-changing. As a result, form-meaning pairs are dynamic and appear in many-to-many relations. From the individual's perspective, we cannot really speak of correct or incorrect interpretations; we can only speak of possible and impossible ones. Even if a candidate interpretation is obviously wrong, it continues to be processed in CdS. Eventually, it is expected to be crowded out by more appropriate alternatives. Once an alternative's probability drops below some level, it stops playing a significant role anyway.

Following this approach, we process all alternative readings of an observation simultaneously, in parallel.

In essence, CdS reflects the structure of morphology and BBN is a model of the lexicon. It is dynamically built and maintained on the basis of the subject's linguistic encounters. Details of this process is given in Section 5.2.

The structure (CdS), the framework (BBN) and the model are independent of language-specific features of Turkish; except that linguistic observations are assumed to be segmental. The model itself should be applicable on other agglutinating languages (more generally, languages whose morphology can be studied in terms of Item-and-Arrangement) with minimal to no adjustment.

At the implementation level, we make some simplifying assumptions relying on our understanding of Turkish DM. These assumptions are merely algorithmic choices and do not disrupt the generality of the structure.

5.1.4 Assumptions

The adult lexicon presumably contains information on a huge number of items, as well as information on their likeliness in certain contexts. It is obvious that humans process linguistic information in an efficient way. This is not to say that the principle of parsimony should be our primary guide; that would be too simplistic. After all, the brain also possesses a huge storage capacity, which is not yet demonstrated to impose any limit on language learning. Our preference towards efficiency is only meaningful if we remain within psychologically plausible boundaries.

Conventional applications of BBN aim to improve efficiency without imposing theoretical limits on the solution space. Most importantly, the network structure of BBN eliminates statistical redundancies by explicitly modeling conditional independence relations. However, this does not mean that a BBN would be computationally tractable regardless of its configuration. The computational complexity of a BBN is primarily dependent on the node with the largest number of parent nodes.

In our case, the segmentation node and lexical nodes on the CdS do not have any parent nodes. Derivation nodes have lexical nodes as parents; therefore, the number of parents for a derivation node is the number of segments in the corresponding segmentation. The meaning node has the segmentation node and derivation nodes as parents; therefore, the number of parents for the meaning node is the number of segmentation alternatives plus one. The maximum number of parent nodes in a BBN grows linearly with the number of morphemes in an observation.

Trials in Section 5.3 demonstrate that with wordforms and clauses made of four or five segments, Bayesian inference can be carried out using reasonable time and computational resources.

The number of morphemes we work with is sufficient for exploring DM, but if we were to focus on more complex clausal structures, we would either need substantially larger computational power or tighter selection criteria to eliminate parts of the segmentation tree. We make some simplifying assumptions in order to reduce the complexity of the algorithm, without sacrificing much generality. On the BBN framework, these simplifying assumptions translate into conditional independence relations.

The first simplifying assumption is that segmentation takes place before lexical selection. This should not be controversial, as we have discussed in Section 2.3.1, because lexical selection is only possible after individual segments are identified. SN represents the segmentation layer.

Availability of the decomposition path, by itself, proves that processing must incorporate a segmentation stage. The only question is whether it is automatically triggered at the beginning of every processing effort, or only when necessary. If the retrieval path were always prioritized, segmentation would occupy the second stage. However, the evidence is against such a universal preference. To the contrary, more recent studies favor a parallel processing view, where retrieval and decomposition paths are traversed concurrently. For this to be possible, segmentation must be the first stage of processing.

Putting segmentation to the first stage breaks it off from any influence of lexical selection. This stage takes the latest lexicon and the observation's surface form as input, but otherwise it is completely autonomous. For simplicity, we assume that segmentation is completed before any lexical selection is carried out.

Alternatively, we could start lexical selection immediately after identifying a segmentation alternative. We could also follow a more incremental process, starting from the left of an observation and immediately carry out lexical selection on segments. These changes to the process would not obtain a different end result, because we already process all possible segmentation alternatives.

The second simplifying assumption is that the probability distribution among states of an LN is independent from other nodes. In other words, probabilities of states of an LN do not change by which states are selected in other nodes. Practically, selection of a lexical item does not depend on the surface forms or meanings of neighboring morphemes.

Similarly, the probability distribution among states of the SN is independent from other nodes. Segmentation is directly determined by the contents of the lexicon. A particular segmentation alternative does not become more or less probable due to some qualities of its constituent morphemes.

Since morphemes can be interpreted in completely new contexts, we know that there cannot be strong dependence between LNs and the SN. Whether there is still significant dependence is open to debate. Proponents of interactive activation models would believe the interaction to be significant. Proponents of connectionist models would believe this is the wrong question to ask. We believe that until evidence can be found to the contrary, we must focus on the simplest possible model. It would be a bigger assumption to establish dependence between LNs and the SN without adequate experimental data to determine the strengths of these relations. Existing experimental data is tainted by too many confounding factors to strongly suggest significant dependence between nodes. If, in the future, the need arises for additional dependence relations to be represented, BBN allows this via extra links between nodes.

This assumption results in a simpler and more transparent web of connections than the ones presented in interactive-activation approaches and connectionist approaches. Representation of information on CdS is not distributional or superpositional. Information on each individual linguistic item is stored in a dedicated node. Parts of the network can be altered without direct effect on other parts. Therefore, CdS offers a more transparent, symbolic representation.

These two assumptions form the backbone of CdS. Once the segmentation and lexical selection layers are taken care of, the derivation layer simply calculates derivations for segmentation alternatives. Representing morphology in this way reduces complexity considerably.

The third simplifying assumption is that CdS only includes nodes that are used during processing. Given a lexicon, a full BBN representation of the lexicon would include LNs for all known segments and DNs for all possible combinations of segments. Every possible observation would be carried out on that same representation. As the lexicon expands, new LNs, new DNs and new states would be added wherever necessary.

Such a BBN would not only be impractical, it would be unnecessary. Our focus is on finding a suitable representation for morphological processing, not on finding an efficient way to carry out a very complex computation. The enormous size of the full network would also bring about a prohibitive level of computational complexity.

The fourth simplifying assumption is that the BBN is constructed from scratch for each new observation. Since we do not work with the full BBN discussed above, this is more of a requirement than an assumption. Each observation licenses different segmentation alternatives, requiring different DNs and LNs. Connections between nodes are also different. Keeping the BBN built for each observation, and modifying it in later observations is simply impractical for our current purposes, but future implementations could find it desirable to conserve computational power by a caching and adapting mechanism for previously built BBNs.

This second pair of assumptions ensure that the lexicon and the BBN are decoupled. In other words, throughout all observations, the lexicon stores all the values necessary for processing and BBNs are disposable. It could be said that we simulate the evolution of the lexicon; BBN is just a model for temporary representation of the current state of the lexicon.

Finally, the fifth assumption is that the lexicon does not leak. Once they are learned, segmentation and lexical alternatives are never forgotten. The Bayesian inference algorithm ensures that more specific interpretations increase in probability, while others are statistically crowded out by more successful alternatives. Even large sequences such as multi-word entities, phrases and sentences can be stored in the lexicon.

The fact that phrases and sentences can be lexicalized to form idioms is enough evidence for this. Nevertheless, using brute force for retrieving entire sentences from the lexicon would not be plausible. Bayesian Occam's Razor (see Section 5.1.6 makes sure that segmentation alternatives with shorter segments gain preference in the long-term.

5.1.5 Compact Representation

Using the assumptions in Section 5.1.4, we greatly reduce the complexity of the problem. As a result, items and their interactions can be represented in a more compact way. In this section, we compare alternative methods for representing knowledge of morphology and illustrate their performance on minimal examples.

We take *kitaplık* 'bookshelf' as an example. The number of ways this form can be interpreted is determined by the contents of the lexicon. Let the lexicon contain the following items:

(133) a. *kitaplık* 'bookshelf'

b. *kitap* 'book'

c. *kitap* 'holy book'

d. *-lIK* 'container'

e. *-lIK* 'apparel'

f. *-lIK* 'profession'

There are two segmentation alternatives: kitaplık and kitap-lık. The first alternative licenses a single interpretation. The second alternative contains two segments with 2 and 3 lexical matches, respectively. For the overall meaning to be constructed, we first carry out lexical selection for all segments and derive the interpretations for each combination of lexical alternatives. An unstructured representation of the resulting interpretations is a simple list, as demonstrated in Table 27.

Since we adopt a probabilistic representation of salience, we assign each row in Table 27 a probability value. These probabilities constitute a probability distribution governing the likeliness of specific interpretations. For each distinct surface form, we have a separate probability distribution serving this purpose.

In order to have a mental model of the variety of interpretations for an observation, the hearer has to keep track of these probabilities. Sum of the interpretation probabilities for a surface form must always be 1. Therefore, the number of independent parameters is equal to the number of interpretations minus 1; in this case, 6.

Table 27: Unstructured representation for *kitaplık*

Observation	Segmentation	Lexicon		Probability
		Segment1	Segment2	$\sum_i p_i = 1$
kitaplık	kitaplık	bookshelf		p_1
kitaplık	kitap-lIK	book	container	p_2
kitaplık	kitap-lIK	book	apparel	p_3
kitaplık	kitap-lIK	book	profession	p_4
kitaplık	kitap-lIK	holy book	container	p_5
kitaplık	kitap-lIK	holy book	apparel	p_6
kitaplık	kitap-lIK	holy book	profession	p_7

Table 28: Factorized representation for segmentation alternatives

Observation	Segmentation	Probability
kitaplık	kitaplık	p_{11}
kitaplık	kitap-lIK	p_{12}

As pointed out in Section 5.1.4, speakers' ability to analyze novel derived forms constitutes evidence that understanding a morpheme does not strongly depend on adjacent segments. There might be a weak and indirect statistical relationship between frequently co-occurring segments, but a this kind of relationship does not require an explicit and direct representation. The BBN-based model already has the statistical infrastructure to handle such relationships.

If we consider such effects as a consequence of the model rather than part of it, it becomes possible to represent morphological structure in a much more compact manner. In that case, once we choose a segmentation alternative, probability distributions governing lexical selection are independent from each other. The hearer does not have to keep track of the huge probability distribution over the combinations of lexical alternatives of individual segments; probability distributions for segmentation alternatives and individual lexical items are sufficient to run the model.

In other words, we decouple morphological information from contextual information, thus avoid replicating morphological information for different contexts. We sacrifice the shortcuts embedded in the context-dependent unstructured representation, but achieve a more compact representation. The frequency effects so often cited in the psycholinguistics literature are relegated to being statistical consequences of the model.

Tables 28 and 29 demonstrate how the same example featuring *kitaplık* can be represented in a factorized way. This time the segmentation layer and the lexical layer are represented separately. We have one probability distribution for the segmentation alternatives of *kitaplık* and one probability distribution for the lexical alternatives of each segment. Probability distributions are all independent from each other. The number of independent parameters is equal to the number of alternatives (both segmentation and lexical) minus the number of probability distributions (subtracting 1 redundant parameter for each distribution). As a result, factorized representation of *kitaplık* requires just 4 independent parameters.

Table 29: Factorized representation for lexical alternatives

Observation	Meaning	Probability
kitaplık	bookshelf	p_{21}
kitap	book	p_{31}
kitap	holy book	p_{32}
-lIK	container	p_{41}
-lIK	apparel	p_{42}
-lIK	profession	p_{43}

Table 30: Unstructured representation for *kitaplık*

Observation	Segmentation	Lexicon		Probability
		Segment1	Segment2	$\sum_i p_i = 1$
kitaplık	kitaplık	bookshelf		p_1
kitaplık	kitaplık	library		p_2
kitaplık	kitap-lIK	book	container	p_3
kitaplık	kitap-lIK	book	apparel	p_4
kitaplık	kitap-lIK	book	profession	p_5
kitaplık	kitap-lIK	book	dedicated_to	p_6
kitaplık	kitap-lIK	holy book	container	p_7
kitaplık	kitap-lIK	holy book	apparel	p_8
kitaplık	kitap-lIK	holy book	profession	p_9
kitaplık	kitap-lIK	holy book	dedicated_to	p_{10}

It is easy to see that the size of the unstructured representation grows much faster than the size of the factorized representation. The number of interpretations grows in proportion to the number of combinations across segmentation and lexical layers. In contrast, the number of lexical alternatives grows with homonymy and polysemy. Factorized representation is much more scalable and compact.

The adequacy of BBN in modeling CdS is also apparent in this respect. Complexity of a BBN model is determined by the node with the largest joint probability table. The largest joint probability table is often constructed for the node with the maximum number of parent nodes. By separating segmentation and lexical layers, CdS eliminates the statistical dependence between many node pairs. As a result, the number of parent nodes is expected to be small. SN and LNs do not have any parent nodes. The number of parent nodes for each DN is equal to the number of segments in the corresponding segmentation alternative. The number of parent nodes for MN is equal to the number of segmentation alternatives (the number of DNs) plus 1 (SN). As long as the segmentation tree can be pruned effectively, these numbers are not expected to climb very high.

To illustrate the scalability of factorized representation, we add two more items to the lexicon. A coding related meaning for *kitaplık* is 'library'. *-lIK* also has another meaning 'dedicated to'. When we add these items to the lexicon, unstructured representation of *kitaplık* grows more than 2. The difference is due to the multiplying effect of lexical selection of the other segment *kitap*. The number of independent parameters is now 9. Elements of the unstructured representation are given in Table 30.

Table 31: Factorized representation for segmentation alternatives

Observation	Segmentation	Probability
kitaplık	kitaplık	p_{11}
kitaplık	kitap-lIK	p_{12}

Table 32: Factorized representation for lexical alternatives

Observation	Meaning	Probability
kitaplık	bookshelf	p_{21}
kitaplık	library	p_{22}
kitap	book	p_{31}
kitap	holy book	p_{32}
-lIK	container	p_{41}
-lIK	apparel	p_{42}
-lIK	profession	p_{43}
-lIK	dedicated_to	p_{44}

Updated factorized representation of *kitaplık* is given in Tables 31 and 32:

The new items do not make a new segmentation alternative available; therefore, the only change is the list of lexical alternatives. The multiplication effect is not present in the factorized representation, so the number of independent parameters grows only by 2. The small difference between the two methods in this example translates to very large differences for observation lists of realistic size.

In order to measure and compare the compactness of the two methods of representation, we calculate the representation sizes for the entire lexicon (as opposed to a specific lexical item). Since a larger lexicon is bound to require a larger representation, we standardize measurements by calculating the average number of independent parameters required for a lexical item.

Let N_U be the number of independent parameters tracked for an unstructured representation. Let N_I be the number of alternative interpretation of a form. Let N_P be the number of alternatives in the lexical selection layer, given a particular segmentation alternative. Let N_L be the number of lexical items that match the surface form of a segment. Let $\mathscr{F}$ be the set of distinct free forms in the lexicon. Using these terms, we calculate the number of paths a hearer may take through the segmentation and lexical selection layers. Since we are after the number of independent parameters, we subtract 1 for each distinct free form ($|\mathscr{F}|$ in total). We do not take into account alternative interpretations of bound forms, because they are not expected to be encountered in isolation; only free forms are subject to segmentation.

$$
\begin{aligned}
N_U &= \sum_{i \in \mathscr{F}} N_{I_i} - |\mathscr{F}| \\
&= (N_{I_{kitaplık}} + N_{I_{kitap}}) - |\mathscr{F}| \\
N_{I_{kitaplık}} &= N_{P_{kitaplık}} + N_{P_{kitap-lIK}} \\
&= N_{L_{kitaplık}} + N_{L_{kitap}} * N_{L_{-lIK}} = 2 + 2 * 4 = 10 \\
N_{I_{kitap}} &= N_{P_{kitap}} = N_{L_{kitap}} = 2 \\
\mathscr{F} &= \{kitaplık, kitap\} \\
N_U &= (10 + 2) - 2 = 10
\end{aligned}
\tag{2}
$$

Let N_F be the number of independent parameters tracked for a factorized representation. Let N_S be the number of segmentation alternatives for an observation. Let $\mathscr{L}$ be the set of lexical items. Let $\mathscr{S}$ be the set of distinct surface forms in the lexicon. This time, we separately consider segmentation and lexical selection layers. The number of independent parameters for the segmentation layer is the total number of segmentation alternatives minus $\mathscr{F}$. The number of independent parameters for the lexical selection layer is simply the number of homonymy and synonymy relations.

$$
\begin{aligned}
N_F &= \sum_{i \in \mathscr{F}} N_{S_i} - |\mathscr{F}| + |\mathscr{L}| - |\mathscr{S}| \\
&= (N_{S_{kitaplık}} + N_{S_{kitap}}) - |\mathscr{F}| + |\mathscr{L}| - |\mathscr{S}| \\
N_{S_{kitaplık}} &= 2 \\
N_{S_{kitap}} &= 1 \\
\mathscr{F} &= \{\text{kitaplık}, kitap\} \\
|\mathscr{L}| &= 8 \\
\mathscr{S} &= \{\text{kitaplık}, kitap, \text{-}lIK\} \\
N_F &= (2 + 1) - 2 + 8 - 3 = 6
\end{aligned}
\tag{3}
$$

We keep independent parameters only for SN and LNs, because CdS only requires probability distributions for segmentation and lexical selection layers. DNs and MN are deterministic nodes acting as gates between other nodes. The joint probability distributions embedded in DNs and MN are directly determined by the probability distributions of SN and LNs.

Addition of *kitaplık* 'library' and *-lIK* 'dedicated to' to the lexicon affect the size of the two representations differently. The number of independent parameters in the unstructured representation grows from 7 to 10. On the other hand, the number of independent parameters tracked by the factorized representation grows from 4 to 6.

We define C_U and C_F as the average number of independent parameters tracked for an item. Lower values indicate a more compact representation. As expected, C_U is significantly larger than C_F.

$$
\begin{aligned}
C_U &= 10/3 = 3.33 \\
C_F &= 6/3 = 2.00
\end{aligned}
\tag{4}
$$

If the BN is an I-MAP of the data generating distribution, that is the conditional independencies assertions in the BN are equivalent to or a subset of the true conditional independencies in the data generating mechanism, this can be considered a lossless compression of an unstructured joint probability distribution.

The most important difference between the conventional unstructured representation and CdS is that the former allows the speaker to encode contextual information. It tracks independent parameters for each combination of segmentation alternative and lexical selection. This way, the conventional representation binds together the segmentation and lexical layers. Certain lexical alternatives could be more salient with some segmentation alternatives, while less salient in others. However, this information would not be generalizable. In order to interpret even small sequences of morphemes, the speaker would need to learn the preferred interpretations in the context of countless segmentation alternatives. This contradicts with the most basic principle of generative grammar. The speaker must be able to learn a lexical item and put this knowledge into use in previously unseen contexts.

CdS only tracks the competition between alternatives within the same layer. The 'book' interpretation of *kitap* may be dominant overall, but the 'holy book' interpretation may have to be used with a specific derivation (for instance with *kitapsız* 'blasphemous'). In such cases, the decomposition path in CdS can still produce the correct interpretation, but with a lower confidence. The retrieval path would produce the correct interpretation with much more confidence. This is exactly what is observed in human data. If human speakers can interpret such a word with some difficulty, it can be attributed to the decomposition path being more demanding. If human speakers can interpret such a word quickly and with confidence, it can be attributed to the retrieval path.

This is not to say that interaction between lexical items is unimportant. It is clear that, to some extent, lexical selection depends on the context. Previously encountered morphemes are part of that context. We only argue that contextual information is of secondary importance; it cannot be required for interpretation.

Now, let us work with a more complicated example: *gözlükçülük* 'profession of an optician'. This word is quite rare, but can be interpreted by Turkish speakers. The lexicon contains the following items:

(134) a. *gözlükçülük* 'profession of an optician'

b. *gözlükçü* 'optician'

c. *gözlük* 'glasses'

d. *göz* 'eye'

e. *göz* 'drawer'

f. *-lIK* 'container'

g. *-lIK* 'apparel'

h. *-lIK* 'profession'

i. *-lIK* 'adopted family'

j. -*CI* ‘seller of an item’

k. -*CI* ‘follower of a person’

l. -*CI* ‘having an affinity towards’

m.-*CIlIK* ‘role-play’

$$\begin{aligned}
N_U &= \sum_{i \in \mathscr{F}} N_{I_i} - |\mathscr{F}| \\
&= (N_{I_{gözlükçülük}} + N_{I_{gözlükçü}} + N_{I_{gözlük}} + N_{I_{göz}}) - |\mathscr{F}| \\
N_{I_{gözlükçülük}} &= N_{P_{gözlükçülük}} + N_{P_{gözlükçü-lIK}} + N_{P_{gözlük-CI-lIK}} + N_{P_{göz-lIK-CI-lIK}} + N_{P_{gözlük-CIlIK}} + N_{P_{göz-lIK-CIlIK}} \\
&= N_{L_{gözlükçülük}} + N_{L_{gözlükçü}} * N_{L_{-lIK}} + N_{L_{gözlük}} * N_{L_{-CI}} * N_{L_{-lIK}} + N_{L_{göz}} * N_{L_{-lIK}} * N_{L_{-CI}} * N_{L_{-lIK}} \\
&+ N_{L_{gözlük}} * N_{L_{-CIlIK}} + N_{L_{göz}} * N_{L_{-lIK}} * N_{L_{-CIlIK}} \\
&= 1 + 1 * 4 + 1 * 3 * 4 + 2 * 4 * 3 * 4 + 1 * 1 + 2 * 4 * 1 = 122 \\
N_{I_{gözlükçü}} &= N_{P_{gözlükçü}} + N_{P_{gözlük-CI}} + N_{P_{göz-lIK-CI}} \\
&= N_{L_{gözlükçü}} + N_{L_{gözlük}} * N_{L_{-CI}} + N_{L_{göz}} * N_{L_{-lIK}} * N_{L_{-CI}} \\
&= 1 + 1 * 3 + 2 * 4 * 3 = 28 \\
N_{I_{gözlük}} &= N_{P_{gözlük}} + N_{P_{göz-lIK}} \\
&= N_{L_{gözlük}} + N_{L_{göz}} * N_{L_{-lIK}} \\
&= 1 + 2 * 4 = 9 \\
N_{I_{göz}} &= N_{P_{göz}} = N_{L_{göz}} = 2 \\
\mathscr{F} &= \{gözlükçülük, gözlükçü, gözlük, göz\} \\
N_U &= (122 + 28 + 9 + 2) - 4 = 157
\end{aligned} \tag{5}$$

$$\begin{aligned}
N_F &= \sum_{i \in \mathscr{F}} N_{S_i} - |\mathscr{F}| + |\mathscr{L}| - |\mathscr{S}| \\
&= (N_{S_{gözlükçülük}} + N_{S_{gözlükçü}} + N_{S_{gözlük}} + N_{S_{göz}}) - |\mathscr{F}| + |\mathscr{L}| - |\mathscr{S}| \\
N_{S_{gözlükçülük}} &= 6 \\
N_{S_{gözlükçü}} &= 3 \\
N_{S_{gözlük}} &= 2 \\
N_{S_{göz}} &= 1 \\
\mathscr{F} &= \{gözlükçülük, gözlükçü, gözlük, göz\} \\
|\mathscr{L}| &= 13 \\
\mathscr{S} &= \{gözlükçülük, gözlükçü, gözlük, göz, \text{-}CI, \text{-}lIK, \text{-}CIlIK\} \\
N_F &= (6 + 3 + 2 + 1) - 4 + 13 - 7 = 14
\end{aligned} \tag{6}$$

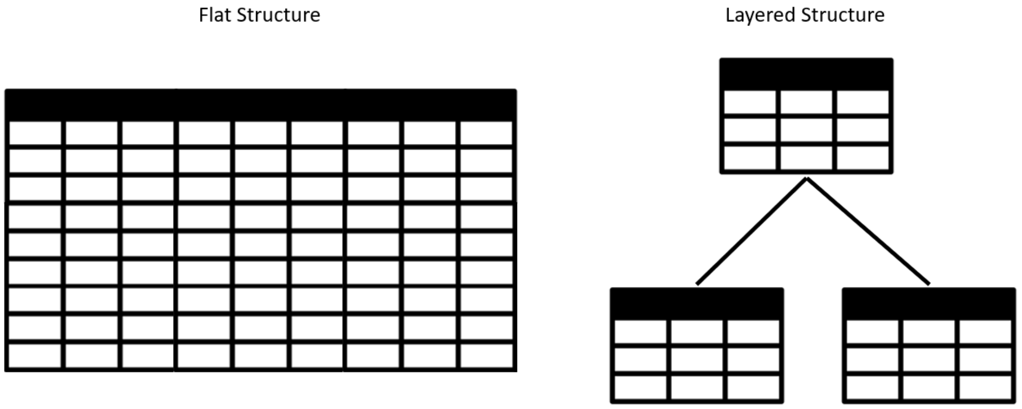


Figure 24: How information is compressed with a factorized representation

$$\begin{aligned} C_U &= 157/13 = 12.08 \\ C_F &= 14/13 = 1.08 \end{aligned} \tag{7}$$

With unstructured representation, we need 157 independent parameters to represent this lexicon, while CdS only requires 14. C_U and C_F metrics indicate a huge difference in terms of compactness. Admittedly, this is an extreme case where polysemy and homonymy occupy a large portion of the lexicon. Even if it is not so dense with homonymy and polysemy, the enormous adult lexicon is bound to license a large number of interpretations for words with multiple morphemes. In such cases, the benefit of a compact representation is evident. Perhaps an unstructured representation would be adequate for a language that mostly lacks polysemy or homonymy, but it is certainly not practical for Turkish. During our classification work in Chapter 3, we observed that the average number of polysemy for Turkish DM is 2.95.

So far, we only presented examples on DM. Turkish words regularly contain 3-4 IM on top of the DM. These IM also contribute a degree of ambiguity to the comprehension problem. The same line of thinking can easily be extended to the word-external realm. If sentence comprehension also starts with segmentation, the number of alternative interpretations may be very high.

Considering the immense size of an adult lexicon, and the word-internal complexity in agglutinating languages, the necessity for a compact representation is clear. By following a hierarchical architecture, CdS ensures that information is represented efficiently. Figure 24 illustrates our strategy.

5.1.6 Hypothesis Selection

Most studies in the computational linguistics literature cite principles such as Minimum Description Length (MDL) and parsimony, in order to justify their efforts towards a lexicon compressed as much as possible. Even some psycholinguistics studies fall into this category. To the best of our knowledge, there is no experimental evidence to suggest that human processing simply pursues a compact lexicon.

A compact lexicon could be desirable for computational reasons, but we have seen no proof that human mental capacity is a limiting factor during the acquisition process.

The child does not have access to the entire corpus; therefore, his goal cannot be to find the most compact lexicon to describe it. Besides, linguistic exposure continues indefinitely. Whatever the intermediate lexicon is at a certain time, it is bound to change with new observations. Since the most compact lexicon at time t and the most compact lexicon at time t+1 may be different, some lexical items may have to be discarded. We have not come across evidence to suggest that there's a mental mechanism for discarding lexical items when they violate parsimony.

MDL and equivalent methods can only be considered tools for hypothesis selection. Contrary to most studies in the literature, we believe that a principle for hypothesis selection cannot be the source of any insight by itself. Instead, one must first examine the structure underlying the linguistic process. Hypothesis selection is only meaningful after an adequate structure is laid out to represent the process. Only hypotheses generated by an adequate structure can be candidates for meaningful hypothesis selection. Without an adequate structure, employing MDL constitutes a stretch of the capabilities of this principle.

There are lines of research that make use of MDL in conjunction with an underlying structure. PCCG is one such area. Generally, the linguistic structure implicit in these studies generate alternative hypotheses with respect to lexical (and sometimes derivational) ambiguity. Therefore, their trade-offs and hypothesis selection mechanisms take place on lexical (and derivational) layers. So far in this section, we presented a new, more comprehensive structure for morphological processing. We have incorporated segmentation ambiguity into our structure; therefore, hypothesis selection must take place on the interaction between the three layers of processing. In the rest of this section, we discuss the operating principles of this hypothesis selection mechanism.

It is an observable fact that all valid interpretations do not have the same salience. Some lexical items are more probable to be retrieved in certain contexts. Computational studies often attribute this to MDL. Since we argue against discarding lexical items, we propose another mechanism to produce this difference. This mechanism is the Bayesian Occam's Razor (BOR) (Blanchard et al., 2018), a statistically motivated method for hypothesis selection. Our approach is closer to the general consensus in psycholinguistics, namely statistical effects and priming effects.

Occam's Razor is the idea that among hypotheses of equal explanatory power, one should prefer the simpler one. BOR is both an extension of Occam's Razor and evidence for it. As an extension of Occam's Razor, BOR states that less flexible hypotheses should be preferred, because less flexible hypotheses tend to have fewer degrees of freedom, thus are simpler. As evidence for Occam's Razor, BOR states that more complex hypotheses tend to be more flexible and accommodate a wider range of data (Blanchard et al., 2018); reducing their usefulness in explaining the intended range of data.

The hypotheses we deal with in this thesis are the hearer's interpretations of an observation. Typically, there are multiple interpretations and possibly several of these interpretations match the observation's true logical form. Therefore, the main issue is deciding how to choose the best hypothesis among alternatives with the same explanatory power. This is exactly the kind of problem Occam's Razor and BOR may help us navigate.

In its essence, BOR is a by-product of Bayesian inference. Bayesian inference predicts the probabilities for alternative explanations of a given result. When these alternatives are hierarchically organized, as in CdS, asymmetries occur in these predictions. If a segmentation licenses a single interpretation, it is said to be less flexible than a segmentation that licenses ten interpretations. Bayesian inference automatically develops a preference towards less flexible explanations.

Polysemy and homonymy amplify this effect. When there are multiple lexical items with the same form but different meanings, a segmentation may license both correct and incorrect interpretations. In that case, the more flexible segmentation is penalized both for being more flexible and for sometimes being incorrect.

Less flexible segmentation alternatives and lexical alternatives are selected in this way. However, BOR does not work in these layers separately. It operates on the whole network, namely on the interaction between segmentation and lexical selection. BOR prefers a simpler solution space, not just a lower number of parameters.

As a result, BOR serves as a force to avoid ambiguity. More ambiguity means more alternatives, and a more flexible solution space. To reduce lexical ambiguity, BOR prefers forms with fewer alternative meanings. Since whole forms are often less ambiguous than the sum of their parts, BOR often prefers retrieval. This is especially true in cases where the number of morphemes and homonyms are large. Therefore, BOR tries to expand the lexicon, by preferring memorization of whole forms.

This force is countered by constituent recognition. This is the idea underlying the entire literature on generative grammar. If large linguistic observations can be broken up into smaller parts and put back together based on certain rules, a tiny lexicon and a tiny grammar are capable of producing an infinite number of distinct expressions. This is obviously true and evidenced by the very fact that language acquisition is possible.

The ability to break up novel observations into smaller parts is what we call constituent recognition. By recognizing constituents, the hearer is free to work with a smaller lexicon, rather than memorizing whole expressions. Therefore, constituent recognition tries to compress the lexicon by making it possible to generate whole forms from their parts.

Both BOR (expanding the lexicon) and constituent recognition (compressing the lexicon) are easily justifiable, mechanistic processes. When ambiguity is higher, BOR dominates and rewards whole forms. When ambiguity is lower, constituent recognition generates better alternatives and whole forms are penalized by BOR. Our claim is that the lexicon evolves based on the interaction of these two forces. In the next section, we describe a CdS-based model that is able to represent both forces concurrently.

5.2 Model Architecture

The previous section laid out the theoretical structure for our investigation, as well as the principles and assumptions behind our choices. We have discussed the adequacy of BBN as a modeling framework and the mechanism of hypothesis selection. In this section, we present how this structure is represented on a model. We explain the data structures and the algorithms used in this effort.

In essence, we process the set of observations sequentially. Each observation triggers a set of operations to generate segmentation patterns and construct a BBN based on the latest lexicon. If the observed expression cannot be interpreted with the current lexicon, it is said to be learned by explicit instruction. If it can be interpreted, prior probabilities of segmentation and lexical alternatives are updated by Bayesian inference to introduce a bias towards alternatives that contributed to the correct interpretation.

A higher level view of the algorithm for processing an observation is given in Figure 25.

5.2.1 Method of Supervision

The model we propose is, at its core, a supervised learning model. Unlike the computational models extracting forms for candidate morphemes from a large corpus, we work with individual observations that provide information on both form and meaning.

Observations consist of three parts: form, meaning and category. In our case, observations take place in written form, for convenience. It is much easier for a computational algorithm to operate on the written form, but ultimately, there is no reason why the sound form could not be used for the same purpose. All the principles we have discussed so far would be applicable to a sound-based observation.

Supervision is carried out by us providing the algorithm with the logical form that represents the observation's meaning. This is analogous to the context providing some clues for the child to deduce the observation's meaning. In other words, we assume the role of context, by providing the semantic component of the observation. It can be said that our position is close to the semantic bootstrapping hypothesis of Abend et al. (2017).

There are a few important caveats regarding this approach. We should not simply provide the algorithm with the dictionary definitions of observations; this would not be psychologically plausible. If the child already knows the meaning, the observation would be redundant. If the child does not already know the meaning, it would be too much to expect the context to always be strong enough to make it possible for the child to make a perfect deduction.

On the other hand, if the context does not provide information to some extent, children would not be able to acquire language. Even if it is deficient or incomplete, contextual information must be enough for the child to deduce the meaning of at least some observations. The exact meaning of a lexical item may take years to determine, or the exact semantic selection criteria of an affix may never be completely learned; but the essential meaning of some items must be deducible from the context. Surely, there are contexts where no useful information can be deduced for the observation. Such observations are simply discarded. All observations we use in trials contain some contextual information.

We model this process by providing incomplete and alternative logical forms for the same form. We assume that different contexts may allow for different interpretations for the same form; even if the intended meaning is the same. In Section 5.3.1, we discuss how alternative meanings are organized by Bayesian Occam's Razor. Ultimately, this process gradually favors more specific meanings.

The third component of an observation is its category. An observation may be the name of an object (a noun phrase) or a full sentence (S). We assume that these are the only two possible categories for direct observations. Other categories often invoked in the categorial grammar literature (adjectives, adverbs, verbs, prepositions etc.) can only be learned as a result of their function within a phrase. For instance,

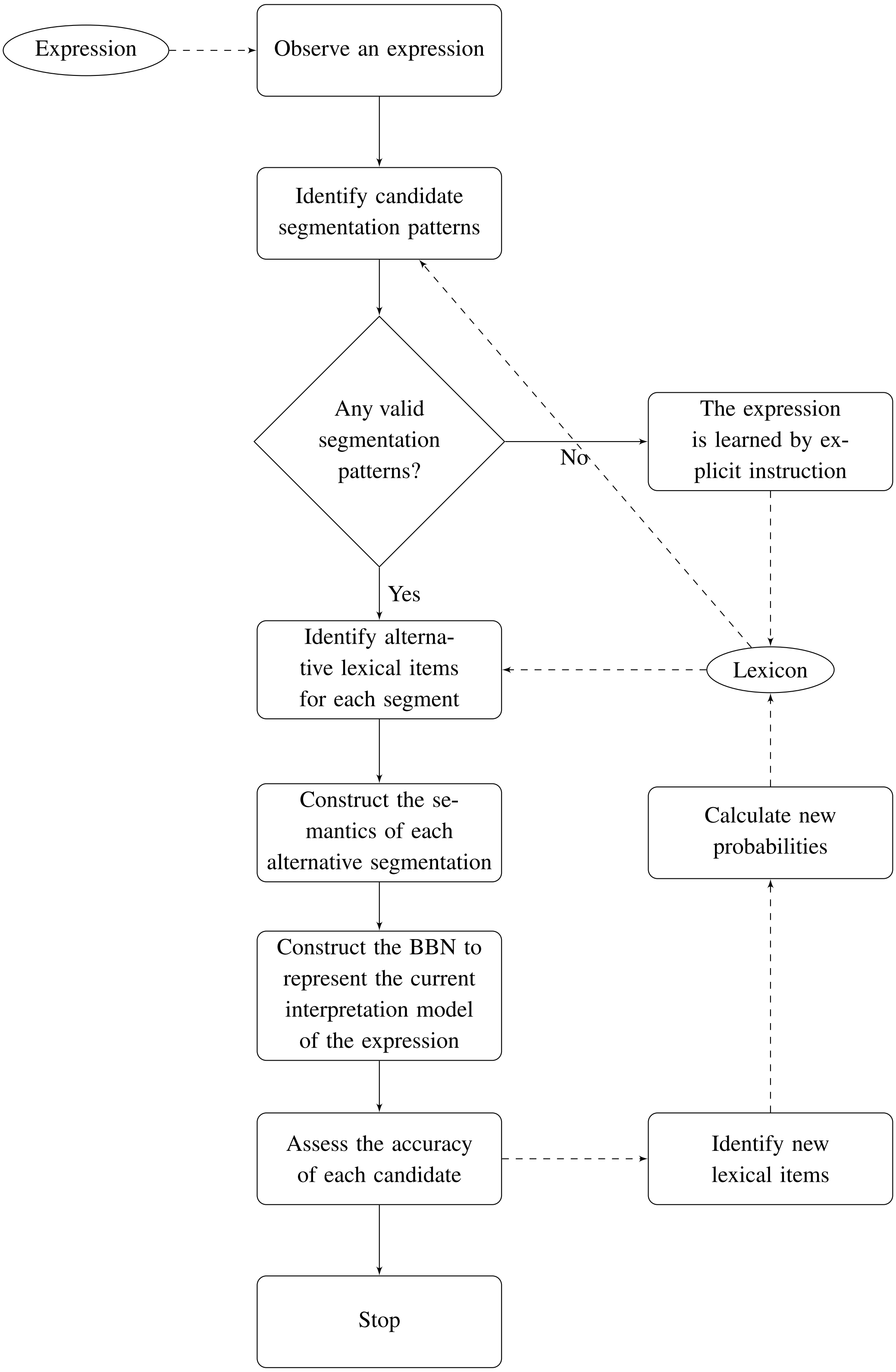

Expression
Observe an expression
Identify candidate segmentation patterns
Any valid segmentation patterns?
No
The expression is learned by explicit instruction
Yes
Identify alternative lexical items for each segment
Lexicon
Construct the semantics of each alternative segmentation
Calculate new probabilities
Construct the BBN to represent the current interpretation model of the expression
Assess the accuracy of each candidate
Identify new lexical items
Stop

Figure 25: Steps of processing an observation

kırmızı ‘red’ can be learned as the name of a color, but the adjective *kırmızı* ‘red’ cannot be learned in isolation. It must be interpreted within an adjectival phrase (or a sentence) such as *kırmızı elma* ‘red apple’. If *elma* ‘apple’ is previously known, the category of *kırmızı* will be deduced as NP
NP.

We need categories, because categories help us significantly reduce the huge number of possible segmentations. As discussed in Section 2.3.1, alternatives in the segmentation layer may grow exponentially with the number of morphemes in an observation. Checking the categorial compatibility of adjacent segments helps eliminate the majority of these alternatives. In our approach, categories are only useful because they keep the segmentation layer tractable.

Using categories in this setting does not hurt psychological plausibility. In Section 2.2.1, we reviewed evidence that categories are available to children. In fact, evidence shows that children are aware of the difference between nouns and verbs. They demonstrate this awareness by consistently applying affixes with correct subcategorization. We use this knowledge in a restricted fashion, by externally providing only the N-S difference. The rest of the categories, such as NP\NP and S\NP are discovered in relation to these two basic categories.

Crucially, we assume that the child is capable of working with categories, but this assumption can also be justified with the same evidence towards awareness of categories. Children not only avoid making mistakes applying the first affix, they avoid making mistakes applying the second one, too. Thus, not only they must be aware of the categories of the root and the first affix, they must also be aware of the category resulting from the application of the affix.

With these three components, we supervise the model and incrementally build a lexicon. Figure 1 demonstrates the pseudo-algorithm.

```
Data: Initial Lexicon
Data: Observation List
Result: Final Lexicon
Initialize Lexical Prior Probabilities
Initialize Segmentation Prior Probabilities
for Observation in Observation List do
    Identify Segmentation Alternatives Construct BBN
    if Correct Interpretation in Derivation List then
        Update Lexical Prior Probabilities
        Update Segmentation Prior Probabilities
        Add Missing Segmentation Alternatives to the Lexicon
    else
        Add New Lexical Item to the Lexicon
        Add New Segmentation Alternative to the Lexicon
    end
    Attempt to Recognize Affixes on the Observation
end
```

Algorithm 1: Pseudo-algorithm for processing observations

Observation lists are organized essentially as sequences of variation sets, in the sense used by Küntay and Slobin (2014). The observation list is a sequence of utterances, where an item (a phrase, a word or a morpheme) is repeated in combination with different constituents. The learning process focuses on the constituent that remains constant, while the variety of contexts makes it possible to match the recurring meaning with the recurring form.

According to Küntay and Slobin (2014), variation sets are characterized by:

(135) a. Lexical substitution and rephrasing

b. Addition and deletion of specific reference

c. Reordering

For instance, *kitap* 'book' can first be used on its own, with a valid reference. When this observation is followed by *kitaplık* 'bookshelf' and *kitapçı* 'bookseller', the new observations create an opportunity for recognizing the common base. On the other hand, the sequence *kitaplık* 'bookshelf' and *odunluk* 'woodshed' repeat the same affix and create an opportunity for recognizing it.

(136) A toy observation list

a. *kitap* 'book'

b. *odun* 'wood'

c. *kitaplık* 'bookshelf'

d. *odunluk* 'woodshed' (The affix *-lIK* is recognized.)

Since Turkish DM almost exclusively relies on suffixation, we only simulate segmentation from the right; emphasizing the learning of affixes, not bases. We do not think this is a deficiency, because we assume affixes cannot be learned in isolation. Inference of base semantics from known affixes should be pretty rare for children, if at all possible.

(137) A toy observation list cont'd

a. *kitapçı* 'book seller'

b. *oduncu* 'wood seller' (The affix *-CI* is recognized.)

In order to somewhat alleviate the lack of segmentation from the left, we use reordering. Reordering full sentences makes it possible for the algorithm to learn phrases and words within the sentence.

(138) a. *Ahmet* 'Ahmet'

b. *Ayşe* 'Ayşe'

c. *Ahmet geldi* 'came Ahmet'

d. *Ayşe geldi* 'came Ayşe' (The verb *geldi* is recognized.)

(139) a. *geldi* ‘came’

b. *gitti* ‘went’

c. *geldi Ahmet* ‘Ahmet came’

d. *gitti Ahmet* ‘Ahmet went’(The noun *Ahmet* is recognized.)

5.2.2 Segmentation

Listing alternative segmentations of an expression is no trivial task. There are several ways to approach the problem. Some unsupervised methods reviewed in Section 2.4.3 use complete enumeration of a string’s possible segments. During this process, most studies do not use a prior lexicon to judge whether individual segmentation alternatives are valid or invalid.

Other studies use a limited lexicon or a set of morphosyntactic rules to filter out some alternatives. Mostly, these studies aim to infer valid segments directly from the form of the expression. As discussed earlier, we pursue a theoretically motivated approach in this thesis. Therefore, we find it more preferable to apply a method that integrates form and meaning.

At the other end of the spectrum, it is possible to base the segmentation process on a prior lexicon. Segmentation can be carried out recursively on smaller and smaller segments identified in the original expression. During this process, validity of the segment may be determined by the latest lexicon containing a lexical item with the same form. Going a step further, it could be enforced that adjacent segments have compatible syntactic categories.

This is the approach we adopt. In this section, we explain the steps of the segmentation algorithm, its complexity and the effect of allomorphy.

Our strategy is to construct a large tree of alternative partitions. At every step, only one segment border is inserted. At every leaf node of the tree, two rules are strictly followed:

(140) a. Each segment must have at least one match in the lexicon.

b. Adjacent segments must be categorially compatible for at least one lexical match.

Categorial compatibility is ensured only between adjacent segments, because we do not implement the several combinators of CCG. As long as adjacent segments are compatible, the whole segmentation alternative is guaranteed to be categorially valid. (This of course does not guarantee that the meaning interpretation would be contextually appropriate.)

Categorial selection is directly implemented in this way. Feature implementation is taken into account during this check. If categories in the lexicon are very specific, such as N(plural), or N/N(uncountable), segmentation algorithm both respects category requirements and applies feature unification when necessary.

If none of the segmentation alternatives are found to be valid according to these two rules, segmentation algorithm stops and prompts direct supervision.

Algorithm 2 summarizes the segmentation algorithm.

```
Data: Lexicon
Data: Observation
Result: List of Segmentation Alternatives for Observation
for Character c in Observation do
    Set Segment1 as the part of observation from the beginning to the character c, including c
    Set Segment2 as the part of observation from character c, not including c
    if Segment1 is NULL then
        RETURN
    else
        if Segment2 is NULL then
            if Segment1 Exists in the Lexicon then
                Add Segment1 to the List of Segmentation Alternatives
            end
        else
            Generate the List of Segmentation Alternatives for Segment2
            for Segmentation Alternative SA for Segment2 do
                if SA is Categorially Compatible with Segment1 then
                    Add Segment1-SA to the List of Segmentation Alternatives
                end
            end
        end
    end
end
```

Algorithm 2: Pseudo-algorithm for segmentation

In Section 2.3.1, we presented a simple example (*kitaplık*) for the segmentation process. Here, we discuss a more complex example (*gözlükçü*) to illustrate the details of the algorithm.

The prior lexicon is given as the following:

(141) a. *gözlükçü* ⊢ N: λx1λx2λx3.and (and (be eye x3) (wear (on x3) x2 anon)) (sell x2 x1) (1)

b. *gözlük* ⊢ N(object): λx1λx2.and (be eye x2) (wear (on x2) x1 anon) (2)

c. *göz* ⊢ N: λx1.be eye x1 (3)

d. *-lük* ⊢ N(object)\N: λx1λx2λx3.and (x1 x3) (wear (on x3) x2 anon) (4)

e. *-lük* ⊢ N\N: λx1λx2λx3.and (x1 x3) (contain x3 x2) (5)

f. *-çü* ⊢ N\N: λx1λx2λx3.and (x1 x3) (sell x3 x2) (6)

g. *-çü* ⊢ N\N: λx1λx2λx3.and (x1 x3) (drive x3 x2) (7)

h. *-çü* ⊢ N\N(person): λx1λx2λx3.and (x1 x3) (believe (in x3) x2) (8)

The segmentation algorithm follows these steps:

(142) a. gözlükçü: Attested
gözlükçü exists in the lexicon. (1) N - Valid

b. gözlükç-ü: Unattested
ü does not exist in the lexicon.

c. gözlük-çü: Attested
Both gözlük and çü exist in the lexicon.
(2) N(object) + (6) N\N - Valid
(2) N(object) + (7) N\N - Valid
(2) N(object) + (8) N\N(person) - Incompatible categories

d. gözlük-ç-ü: Unattested
ü does not exist in the lexicon.

e. gözlü-kçü: Unattested
kçü does not exist in the lexicon.

f. gözlü-kç-ü: Unattested
ü does not exist in the lexicon.

g. gözlü-k-çü: Unattested
k does not exist in the lexicon.

h. gözlü-k-ç-ü: Unattested
ü does not exist in the lexicon.

i. gözl-ükçü: Unattested
ükçü does not exist in the lexicon.

j. ...

k. göz-lükçü: Unattested
lükçü does not exist in the lexicon.

l. göz-lükç-ü: Unattested
ü does not exist in the lexicon.

m. göz-lük-çü: Attested
göz, lük and çü exist in the lexicon.
(3) N + (4) N(object)\N + (6) N\N - Valid
(3) N + (4) N(object)\N + (7) N\N - Valid
(3) N + (4) N(object)\N + (8) N\N(person) - Incompatible categories
(3) N + (5) N\N + (6) N\N - Valid
(3) N + (5) N\N + (7) N\N - Valid
(3) N + (5) N\N + (8) N\N(person) - Valid

n. göz-lük-ç-ü: Unattested
ü does not exist in the lexicon.

o. ...

The number of segmentation alternatives grows exponentially with the number of morphemes. If the number of morphemes in an expression is m, the number of segmentation boundaries is $m-1$ and the number of subsets of these boundaries is 2^{m-1}.

This way of segmentation results in an interesting consequence: Sometimes simple forms can be segmented based on pseudo-morphemes. For instance, *sirküler* ‘communique’ is a foreign form that is originally complex but hardly analyzable by Turkish speakers. It is also singular in the sense used in Turkey. However, if a speaker somehow believes that *sirküler* is a plural form, it becomes possible to find the *-lAr* affix inflecting **sirkü*. For the sirkü-lAr segmentation alternative to be possible, both errors (believing that *sirküler* is plural and that **sirkü* form exists) must be committed. Over time, some speakers may commit both errors and share this false knowledge with others. In reality, *sirkü* has become a well-established form in spoken Turkish, although dictionaries refuse to include it.

Allomorphy often prevents common segments from being recognized, as the surface forms are phonologically different. In order to observe its effect, we included a small method that represents allomorphy. Only regular allomorphy is modeled in the method and suppletive allomorphy is ignored, as the latter is often irregular. We also ignore consonant deletion and epenthesis.

We start by identifying the phonemes that allow allomorphy. Below, we list phonemes in lower case and meta-phonemes in upper case. Not all phonemes create allomorphy anywhere inside the morpheme; some create allomorphy only if they are at the beginning of the morpheme, some only at the end. This is also taken into account. (k creates allomorphy both at the beginning and at the end, but behaves differently according to its position.)

(143) Allophony

a. A: a,e (anywhere in the morpheme)

b. C: c,ç (only at the beginning and at the end of the morpheme)

c. D: d,t (only at the beginning and at the end of the morpheme)

d. G: g,k (only at the beginning of the morpheme)

e. I: ı,i,u,ü (anywhere in the morpheme)

f. K: ğ,k (only at the end of the morpheme)

In this manner, *-lük* and *-lığ* are converted to *-lIK*; *-çu* and *-ci* are converted to *-CI* etc. This ensures that the morpheme is represented by a single lexical item. The same item takes part in all probability inferences.

When we generate allomorphs for a morpheme, we respect vowel harmony (both backness and flatness harmony). For instance, decomposing *doktorculuk* ‘role play as a doctor’, a valid segmentation alternative is doktor-culuk. If allomorphy is assumed, this alternative is represented as doktor-CIlIK. We generate *-cılık*, *-cilik*, *-culuk* and *-cülük* as allomorphs of *-CIlIK*, but we do not generate *-cıluk* or *-cilık*.

Allomorphy only affects bound forms. If a common segment includes both a bound form and free forms, only the bound part is represented by its meta-phonemes. For instance, in *göresi gelmek* 'to miss', a valid segmentation alternative is gör-esi gelmek. If allomorphy is assumed, this alternative is represented as gör-AsI gelmek.

We modeled allomorphy in a binary fashion, assuming or not assuming its existence. However, this is not the only way. Allomorphy could itself be recognized based on experience. This is presumably what children must be doing. In order to do that, allomorphs could be learned individually and then bundled together due to their common logical forms. Having carried out this operation on many affixes, children may be recognizing a general pattern for allomorphy, leading them to assume the existence of allomorphy in future affix recognition.

We have no evidence for or against this mental mechanism. Therefore, we do not make any claims on whether allomorphy assumption takes hold early or late, if it does at all. In any case, our simple model serves to highlight the effect of allomorphy on learning and on the lexicon. It can be interpreted as a late stage module of language processing, or it can be interpreted as an algorithmic shorthand to compress information.

5.2.3 Learning Morphosemantics by Latent Syntax

Surface forms and categories of lexical items direct the search for segmentation. In order to construct the BBN, we must also derive the overall meaning of the expression based on attested segments. Section 4.4 demonstrated the principles we follow creating lexical items. With a few self-imposed rules, it is possible to obtain a lexicon very consistent across different categories. In this section, we go over these rules and present the templates we used for constructing new affixes based on an observed stem-lemma pair.

(144) a. An object denoted by a noun is represented by the outermost bound variable in the noun's logical form (LF). No object can be denoted by a free variable, not even named entities.

b. LF for adjectives, verbs and adverbs reflect the argument structure and the thematic structure.

c. Bound morphemes express new content by introducing a lambda term.

d. If a bound morpheme only modifies a property of the stem, it does not introduce a new bound variable.

e. If a bound morpheme causes the lemma to indicate a different object than the stem; it introduces a new bound variable. This variable becomes the new outermost variable.

f. Morphological operations process all the variables of the stem, with the innermost one remaining the innermost variable of the lemma.

g. Syntactic operations process the variables of the argument starting from the outermost one.

h. Different logical forms are constructed for contexts imposing different arity.

The first principle is uncommon but crucial. It is quite possible to represent syntactic operations on CG without dedicated bound variables for every object. In fact, additional bound variables do nothing but create clutter inside LF. Because, at the level of syntax, each word / construction denotes an argument at the same level as others. In other words, constituents of the sentence are all atomic. As long as categorial compatibility is ensured, it does not matter whether an object is denoted by a bound variable or a free one.

On the other hand, morphological operations introduce a new kind of complexity to the picture. If the root of a word is denoted by a free variable, bound morphemes that contribute new meaning to the root often have no way of manipulating or replacing this variable. This is especially true for DM. It is often the case that DM modifies the stem in such a way that the resulting lemma denotes a completely different object. In such cases, DM may introduce new lambda terms to the LF, but it cannot replace or duplicate existing ones. Representing all objects with free variables provides crucial flexibility for representing DM. This method slightly benefits representation of IM, too.

(145) An object denoted by a noun is represented by the outermost bound variable in the noun's logical form.

a. *doktor* ⊢ N: λx1.be doctor x1

b. *Ankara* ⊢ N: λx1.be Ankara x1

LF for other syntactic categories, namely adjectives, verbs and adverbs (among others) are modified to accommodate the extra bound variables in noun LF. This is accomplished by adding a lambda term to the predicate LF, which is only composed of bound variables. This lambda term serves to receive the noun and process its bound variables in the correct order.

(146) LF for adjectives, verbs and adverbs reflect the argument structure and the thematic structure.

a. *hazır* ⊢ N\N: λx1λx2.and (x1 x2) (be ready x2)

b. *gel* ⊢ V: λx1λx2.come x1 x2

c. *geldi* ⊢ S/N: λx1λx2.and (x2 < t0) (fall x1 x2)

d. *geldin* ⊢ S: λx1λx2.and (be hearer x1) (and (x2 < t0) (come x1 x2))

e. *içeri* ⊢ V/V: λx1λx2λx3.and (x1 (towards x3) x2) (be inside x3)

Some bound morphemes introduce new content, some do not. For instance, case markers do not require us to modify the logical form of the stem. If the bound morpheme introduces new content, this is accomplished by adding a lambda term to the LF.

(147) If a bound morpheme causes the lemma to indicate a completely new object; it introduces a new bound variable.

a. Stem: *bul* ⊢ V: λx1λx2.find x2 x1

b. Lemma: *bulgu* ⊢ N: λx1λx2.and (find x2 x1) (be anon x1)

c. Affix: *-GI* ⊢ N\V: λx1λx2λx3.and (x1 x2 x3) (be anon x2)

On the other hand, if the bound morpheme just modifies a property of the stem, it does not introduce a new bound variable. It contributes to the semantics by introducing a lambda term. IM generally follows this pattern.

(148) If a bound morpheme modifies a property of the stem, it does not introduce a new bound variable.

a. Stem: *kitap* ⊢ N: λx1.be book x1

b. Lemma: *kitaplar* ⊢ N: λx1.and (be book x1) (be plural x1)

c. Affix: *-lAr* ⊢ N\N: λx1λx2.and (x1 x2) (be plural x2)

Some bound morphemes completely change the object denoted by the stem. This is accomplished by an additional bound variable. For instance, if a morpheme represents a product-seller relations, it introduces a new bound variable to denote the seller. This new variable becomes the outermost variable in the derived form. DM generally follows this pattern.

(149) Bound morphemes express new content by introducing a lambda term.

a. Stem: *kitap* ⊢ N: λx1.be book x1

b. Lemma: *kitapçı* ⊢ N: λx1λx2.and (be book x2) (sell x2 x1)

c. Affix: *-CI* ⊢ N\N: λx1λx2λx3.and (x1 x3) (sell x3 x2)

We can construct logical forms for all syntactic categories and carry out derivations successfully. On the surface, syntactic operations seem to apply without any changes all the way to the morphological level. However, structure of bound morphemes are different than free morphemes in one crucial way. Bound morphemes are concerned with the word-internal structure; therefore, they process the bound variables that denote objects used and modified in the word-internal structure. The lambda term to receive the stem LF and handle its bound variables must be able to preserve the status of the outermost variable.

(150) Morphological operations process all the variables of the stem, with the innermost one remaining the innermost variable of the lemma.

a. Stem: *iç* ⊢ N: λx1.be inside x1

b. Lemma: *içeri* ⊢ V/V: λx1λx2λx3.and (x1 (towards x3) x2) (be inside x3)

c. Affix: *-ArI* ⊢ V/V\N: λx1λx2λx3λx4.and (x2 (towards x4) x3) (x1 x4)

Syntactic operations process bound variables in the opposite way. At the level of syntax, all bound variables are expected to be arguments to the verb (for now, disregarding nominal constructions). The lambda term that handles incoming LF is more concerned with placing constituents in the correct positions in the argument structure. Due to the adjacency assumption, we expect each derivation step to fulfill one slot in the argument structure.

(151) Syntactic operations process the variables of the argument starting from the outermost one.

a. Part: *geldi* ⊢ S\N: λx1.came x1

b. Construction: *geldi kitapçı* ⊢ S: λx1λx2.and (and (be book x2) (sell x2 x1)) (came x1)

c. Affix: *kitapçı* ⊢ S\(S\N): λx1λx2λx3.and (and (be book x3) (sell x3 x2)) (x1 x2);

In a morphology context, the last lambda term for *kitapçı* would be (x1 x3) instead of (x1 x2). Syntactic operations are concerned with the outermost bound variables. At the sentence level, constituents are only the objects denoted by the constituents of the sentence. Word-internal operations only serve to establish the meaning of the lemma; only the object denoted by the lemma participates in the sentence-level argument structure.

(152) Different logical forms are constructed for contexts imposing different arity.

a. Stem: *gözlük* ⊢ N: λx1λx2.and (be eye x2) (wear (on x2) x1 anon)

b. Lemma: *gözlükçü* ⊢ N: λx1λx2λx3.and (and (be eye x3) (wear (on x3) x2 anon)) (sell x2 x1)

c. Affix: *-CI* ⊢ N\N: λx1λx2λx3λx4.and (x1 x3 x4) (sell x3 x2)

We construct LF for new lexical items, based on existing stem and lemma LF. Algorithm 3 demonstrates the steps of constructing a LF for a new affix. The same process applies for syntactic structures, too; we must only slightly modify the mechanism for constructing the LF. This difference is discussed later in this section.

We must emphasize an important property common in all the example sets presented in this section. Candidate morphemes are not simply generated as the string difference between some stem-lemma pairs; they have semantic content. Once we identify common segments in several lemmas, we use templates to construct tentative LFs (and syntactic categories) for the candidate morpheme. Candidate morphemes are accepted into the lexicon only if their LFs (and categories) are able to derive the stem to obtain the original lemma.

We set such a high bar for morpheme candidates, because reducing morphemes to their forms inevitably result in the morphological processing problem being intractable. Morphemes must be learned and represented with their form, category and LF, altogether. Otherwise, a form-based algorithm could pick up pseudo-morphemes such as *-ba* from collections of semantically unrelated lemmas such as *baba*, *araba*, *akraba*, *torba*, *soba* and *lamba*.

Despite the large variety of possibilities regarding LF structure, our two templates are able to consistently generate LFs for candidate morphemes. These templates are only skeletons for LF, they do not add or subtract semantic substance. The candidate LF must convey the semantics present in the lemma, but not in the stem. It must also be able to derive the lemma LF, when applied on the stem LF.

The p_similarity metric we use is inspired by but not the same as Gaussier (1999). We make comparisons between an observed form and the entries in the lexicon. Only the lexical items that match the first p of the observed form are qualified for further processing. p must be greater than 0.

```
Data: Lexicon
Data: Embedding Matrix
Data: Observation
Result: New Lexical Item
Initialize Empty List of Stem Candidates for Lexical Item in Lexicon do
    if p_similarity between Lexical Item and Lemma $> 0$ then
        Add Lexical Item to Stem Candidates
    end
end
for Stem Candidate in Stem Candidates do
    Define Affix Candidate such that Observed Form $\equiv$ Stem Candidate $\frown$ Affix Candidate
    for Lexical Item in Lexicon do
        Construct LF for Affix Candidate Based on the Observed Form and the Stem
          Candidate if Observed LF $\equiv$ Affix Candidate LF (Stem Candidate LF) then
            if Euclidean Distance between Stem and Lemma Embeddings $< 1.5$ then
                Add New Item to the Lexicon
            end
        end
    end
end
```

Algorithm 3: Pseudo-algorithm for constructing LF for a newly recognized affix

Üstün (2017) also constructs LF for new lexical items based on existing ones, in a similar, supervised setting. He calls this process "learning morphosemantics by latent syntax" due to the interaction of lambda terms with categorial information. We adopt this term.

We derive wholes from constituents using the CKY-algorithm. We developed a custom-built Python library for this purpose. Forward and backward application are implemented, as well as feature unification. Our focus being on morphological operations; CCG combinators are not implemented, because the adjacency assumption is expected to hold inside word-internal structure. When a new LF is constructed for a candidate lexical item, we test its correctness by deriving the stem. If the observed lemma's LF can be obtained, the new LF is validated.

We use flexible templates to construct candidate LF for new affixes. An algorithm searches points in both lemma and stem LF, where common lambda terms start. The new affix is constructed around that point. A new lambda term is introduced to receive the stem LF. The new lambda term is connected to the rest of the LF by an AND operator.

We do not use static templates such as the ones in Zettlemoyer and Collins (2007) and Zettlemoyer and Collins (2012). If we used static templates, morphological processes would require a very a large number of arity-category combinations. Instead, we specify two scenarios, each of which are handled with slightly different procedures. Of course, careful construction of LF helps a lot; but it is not the critical factor determining the success of these templates.

Since they do not presume any arity, any category can be handled by these templates. Syntactic category of the new item is directly determined based on the stem and lemma categories. For the templates to work, the number of bound variables in the lemma must be greater than or equal to the number of bound variables in the stem.

(153) a. Morphology
Initialize the new LF as equal to the lemma LF
Introduce $\lambda x0$ at the beginning
Find the common lambda terms between lemma and stem LF
Let n be the number of bound variables in stem LF
Let sx be the set of the last n bound variables from stem LF
Replace the common segment in the new LF with a lambda term of the form $(x_0\ sx)$

b. Syntax: Initialize the new LF as equal to the lemma LF
Introduce $\lambda x0$ at the beginning
Find the common lambda terms between lemma and stem LF
Let rx be the reverse of bound variables in stem LF
Let n be the number of bound variables in stem LF
Let sx be the set of the last n bound variables from rx
Replace the common segment in the new LF with a lambda term of the form $(x_0\ sx)$

The fact that we use different templates for morphology and syntax is not an algorithmic choice. When it comes to manipulating LF, there is a crucial asymmetry between morphological and syntactic operations. Recognizing this asymmetry is crucial for consistently representing both word-internal and word-external syntax, as well as constructing appropriate LF for new lexical items.

5.2.4 Implementation of BBN

As in Section 5.1.3, implementation of BBN should be discussed in two operational stages: the inward direction (construction) and the outward direction (inference).

In the inward direction, we build the graph for the BBN, based on a specific linguistic observation. Even if the same observation has been encountered before, the model is constructed from scratch; because the lexicon could have evolved in the meantime.

The Figure 4 demonstrates the pseudo-algorithm for constructing the BBN.

In order to process new observations, the algorithm needs to keep track of two tables: the lexicon Λ and the segmentation alternatives list Γ. As both sources evolve after every observation, it is most precise to keep track of them with a reference to the observation index, such as Λ_t and Γ_t, meaning the lexicon after observation t. To reduce the clutter, we present the following definitions for a static lexicon.

We define the lexicon Λ, including n items, as follows:

$$\Lambda = \{\lambda_1, \lambda_2, ...\} \tag{8}$$

```
Data: Lexicon
Data: Observation
Result: BBN
Create Segmentation Tree for Observation;
Generate Segmentation Node (SN)
for Valid Segmentation in Segmentation Tree do
    Add Valid Segmentation as a State of SN
end
for Segment in List of Unique Segments do
    Generate Lexical Node (LN#)
    for Lexical Item that Matches Segment do
        Add Lexical Item as a State of LN#
    end
end
for Segmentation in Segmentation Tree do
    Generate Derivation Node (DN#)
    for Node in List of Relevant Lexical Nodes do
        Create an Edge from Node to DN#
    end
    for Derivation in List of Unique Derivations do
        Add Derivation as a State of DN#
    end
end
Generate Meaning Node (MN)
Create an Edge from SN to MN
for Node in List of Derivation Nodes do
    Add an Edge from Node to MN
end
for Derivation in List of Unique Derivations do
    Add Derivation as a State of MN
end
```

Algorithm 4: Pseudo-algorithm for constructing the BBN for processing an observation

$$\lambda_i = \{\phi_i, \kappa_i, \mu_i, \omega_i, \pi_i\} \quad \forall i \in \{1..|\Lambda|\} \tag{9}$$

where ϕ stands for form, κ for category, μ for logical form, ω for observation count and π for prior probability. No two lexical items may contain the same values for the first three elements.

Prior probabilities π in Λ are calculated based on observation counts ω of competing lexical items. At moment of calculation, they which is after an observation, they are technically posterior probabilities. At the start of the next observation, they act as prior probabilities.

$$\pi_i = \frac{\omega_i}{\sum_{k:\phi_k=\phi_i} \omega_k} \quad \forall i \in \{1..|\Lambda|\} \tag{10}$$

As a result, the following rule holds for lexical prior probabilities:

$$\sum_{k:\phi_k=\phi_i} \pi_k = 1 \quad \forall i \in \{1..|\Lambda|\} \tag{11}$$

We define the list of segmentation alternatives Γ as follows:

$$\Gamma = \{\gamma_1, \gamma_2, ...\} \tag{12}$$

$$\gamma_j = \{\psi_j, \sigma_j, \mathrm{o}_j, \rho_j\} \quad \forall j \in \{1..|\Gamma|\} \tag{13}$$

where ψ stands for form, σ for segmentation pattern, o for observation count and ρ for prior probability.

Again keeping prior probabilities ρ in Γ is redundant. We calculate them after processing each observation, according to updated observation counts o:

$$\rho_j = \frac{\mathrm{o}_j}{\sum_{k:\psi_k=\psi_j} \mathrm{o}_k} \quad \forall j \in \{1..|\Gamma|\} \tag{14}$$

As a result, the following rule holds for segmentation prior probabilities:

$$\sum_{k:\psi_k=\psi_j} \rho_k = 1 \quad \forall j \in \{1..|\Gamma|\} \tag{15}$$

At the beginning, both Λ and Γ are initialized as empty tables Λ_1 and Γ_0. If it is preferable to initialize them as nonempty tables, presumably for simulating learning from a specific configuration, initial tables Λ_1 and Γ_0 can be supplied to the algorithm. If these tables violate Equations 11 or 15, initial prior probabilities should be recalculated before the first processing stage. Λ is initialized as Λ_1, while Γ is initialized as Γ_0; because the lexicon is directly input to the processing stage, while new segmentation alternatives can be discovered immediately before processing, during the segmentation stage.

Following this notation, we can define the nodes in an alternative way. This time, we reincorporate the observation index t for clarity. The observation list Θ is composed of individual observations θ, which are in turn made of 3 components:

$$\Theta = \{\theta_1, \theta_2, ...\} \tag{16}$$

$$\theta_t = \{\varphi_t, \chi_t, \nu_t\} \quad \forall t \in \{1..|\Theta|\} \tag{17}$$

where φ denotes the surface form, χ the category and ν the logical form of the observation. First, segmentation is carried out based on the available lexicon Λ_t. Missing entries are added to the Γ_{t-1}, resulting in Γ_t. Each γ_t constructed for this observation hold the same surface form, but different segmentation patterns. From now on, all equations should be assumed to hold for all values of t, from 1 to $|\Theta|$. SN has the following states:

$$SN_t = \{\gamma_{j,t} | \psi_{j,t} = \varphi_t\} \tag{18}$$

SN only needs the segmentation patterns σ_t and prior probabilities ρ_t. A segmentation pattern is composed of individual segments ς:

$$\sigma_{j,t} = \{\varsigma_{j,t,1}, \varsigma_{j,t,2}, ..\} \tag{19}$$

We create a lexical node (LN) for each unique segment that takes part in the segmentation alternatives relevant for the current observation. Ξ denotes the set of unique segments for the current observation:

$$\Xi_t = \bigcup_{j:\psi_{j,t}=\varphi_t} \sigma_{j,t} \tag{20}$$

Each LN is defined in relation to a particular segment:

$$LN_{l,t} = \{\lambda_{i,t} | \phi_{i,t} = \Xi_{l,t}\} \quad \forall l \in \{1..|\Xi_t|\} \tag{21}$$

A DN is created for each segmentation alternative. Derivations are carried out according to the rules described in Section 4.4. We take derivation Δ as a function of the segments taking place in a segmentation alternative. Naturally, the sequence of segments is important. LEX collects the lexical selection results for a particular segmentation alternative. DN computes the derivation on every combination in LEX.

$$LEX_{s,t} = \bigotimes_{l:\lambda_{i,t} \in LN_{l,t} \bigwedge \phi_{i,t} \in \sigma_{j,t} \bigwedge \sigma_{j,t} \in SN_{s,t}} LN_{l,t} \quad \forall s \in \{1..|SN_t|\} \tag{22}$$

$$DN_{s,t} = \{(LEX_{k,s,t}, \Delta(LEX_{k,s,t})) \quad \forall k \in \{1..|LEX_{s,t}|\} \quad \forall s \in \{1..|SN_t|\} \tag{23}$$

If a segmentation alternative or a lexical item has been observed / recognized for the first time, it must be absorbed into the lexicon. Observation counts in this case are incremented by 0.01, in order to prevent the new alternative from dominating others. Right after a new item is added to either Λ or Γ, prior probabilities are recalculated.

While BBN nodes are being created, in parallel, conditional probabilities are calculated. Probabilities in SN and LNs are directly drawn from Γ and Λ. Conditional probabilities in DNs and MN are trivial. DNs are used as gates between derivations and lexical alternatives; while MN is used as a gate between segmentation alternatives and derivations. The conditional probability of a DN state is 1, if and only if the corresponding derivation is linked to its constituent LN states, and 0 otherwise. Similarly, the conditional probability of an MN state is 1, if and only if the corresponding segmentation alternative is linked to a matching derivation, and 0 otherwise.

This concludes the inward direction (construction) part of the algorithm. The outward direction (inference) part serves to update the probabilities of alternatives. This process takes place in two parts.

In the first part, the Bayesian inference algorithm runs. There are three possibilities at this stage: a BBN could not be constructed, at all; the observed meaning ν_t cannot be found among the states of MN; ν_t can be found among the states of MN.

If a BBN could not be constructed, at all, no valid segmentation alternatives were found. This means that the observation is completely novel. In some cases, the hearer can deduce the meaning of the observation from contextual clues; therefore, a new item should be added to the lexicon. In other cases, the hearer cannot deduce the meaning of the observation; therefore, the observation must be discarded. The cases in-between, such as partially understanding, are further explored in later sections.

If the observed meaning cannot be found among the states of MN, the hearer is able to interpret the observation, but cannot reach the correct interpretation. Similar to the first case, either the lexicon should be expanded, or the observation must be discarded.

If the observed meaning can be found among the states of MN, Bayesian inference algorithm (Lauritzen and Spiegelhalter (1988)) runs to calculate the posterior probabilities of each state at each node, given the observed meaning ν_t. Essentially, the algorithm determines the probability that a state has contributed to the correct interpretation. The exact algorithm is based on the Bayes' Theorem:

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)} \tag{24}$$

In the second part, we take posterior probabilities as individual contributions of corresponding items. We use these values to increment observation counts of each segmentation and lexical alternative.

$$\omega_{i,t+1} = \omega_{i,t} + \omega'_{i,t} \quad \forall i \in \{1..|\Lambda_t|\} \tag{25}$$

$$o_{j,t} = o_{j,t-1} + o'_{j,t-1} \quad \forall j \in \{1..|\Gamma_{t-1}|\} \tag{26}$$

Finally, we recalculate the prior probabilities of segmentation and lexical alternatives. This calculation is linear on observation counts.

$$\pi_{i,t+1} = \frac{\omega_{i,t+1}}{\sum_{k:\phi_{k,t+1}=\phi_{i,t+1}} \omega_{k,t+1}} \quad \forall i \in \{1..|\Lambda_{t+1}|\} \tag{27}$$

$$\rho_{j,t} = \frac{o_{j,t}}{\sum_{k:\psi_{k,t}=\psi_{j,t}} o_{k,t}} \quad \forall j \in \{1..|\Gamma_t|\} \tag{28}$$

The algorithm moves on to the next observation, with Λ_{t+1} and Γ_t at hand. We call the base structure Conventionalized Structure (CdS), due to the ever-changing nature of the lexicon and the reconstruction of BBN for every observation. This is not to be confused with Dynamic BBN. CdS does not require a Dynamic BBN implementation.

5.2.5 Learning Parameters

Three learning parameters are used in the algorithm: the learning threshold (LT), the initial observation count (IOC) and the maximum embedding dissimilarity (MED). These are essentially free parameters for which it is impossible to set a universally appropriate value. We choose and work with a value for each of these parameters, and leave their experimental justification to later research.

LT determines the minimum number of distinct lemmas an affix must derive for the hearer to recognize it. Its value can be any whole number. There are two extreme conditions for LT: When LT is 1, an affix is recognized even if it is encountered only on one stem. When LT is ∞, no affix can be recognized; because it is impossible to encounter an infinite number of distinct stems derived by the affix.

Both cases are implausible. It is impossible for a hearer to recognize an affix that derives only one lemma. This is the reason many derived forms come to be considered as simple forms. There are even cases where an affix is not recognized, despite multiple derived forms: bura ‘here’, şura ‘there’, ora ‘there’, nere ‘where’. LT must also be lower than ∞, because we know that people are able to recognize affixes. We can predict that LT is above 2 and it is low enough to allow fairly less productive affixes to be recognized.

Nevertheless, we cannot claim the existence of a universal learning threshold for affix recognition. It may be different for different languages, for different individuals, as well as for different kinds of lexical items. Quite possibly, there are multiple interacting factors deciding whether an affix can be recognized or not. Perhaps a minimum value of 2 is accompanied by the requirement for a strong similarity between stem and lemma embeddings. Perhaps the minimum value changes based on stem and lemma categories. We have no evidence in those regards.

It is still possible to make an informed guess for this threshold. There are many affixes that are frequently used, but unrecognized, *-ArI* and *-rA* are among these. These are valid affixes with consistent use, but they are not productive. Nişanyan (2021) lists 8 words as derived with *-ArI*, but only two of their stems are still recognizable to the speakers of Modern Turkish. Similarly, Nişanyan (2021) lists 9 words formed with *-rA*. In that list, 5 stems are still recognizable, but only 3 of them are of the same category. There are several other suffixes such as -Am (*biçem* ‘form’, *dönem* ‘period’) *-A* (*dize* ‘verse’, *doğa* ‘nature’ with fewer than 7 derived forms in the latest dictionary. These affixes can be verified etymologically, but they are no longer productive enough to be listed in grammar books. On the other hand, affixes with more than 7 derived form *-DIK* (*alışıldık* ‘familiar’, *tanıdık* ‘acquaintance’), *-lAT* (*aydınlat-* ‘illuminate’, *kirlet-* ‘pollute’) *-tI* (*bulantı* ‘nausea’, *çığırtı* ‘yell’) are usually considered proper affixes. A plausible value for LT may be set around 7. In order to keep observation lists short, we take LT as 3 in our proof-of-concept trials in Section 5.3.

IOC is a parameter required by the algorithm. When we encounter a previously unseen segmentation alternative or a lexical item, we need to create a new entry in the corresponding matrix. Both matrices include the observation count, which keeps track of previous encounters of the item. The item’s prior probability in the next observation depends on this observation count. We have to initialize this value for the new entry.

Practically, there are two possibilities for a new entry: Either it has a unique form (therefore, it does not share prior probability with any other forms), or it does not. If the former is true, any nonzero IOC gives the new item full prior probability, while a zero observation count creates a contradiction. If the latter is true, a zero observation count ensures that the new item can never participate in the BBN. Logically, a non-zero IOC must be assigned to new entries.

If too large an IOC is assigned, the new entry may immediately dominate its alternatives. Since observation counts of alternatives can be high or low, this is an unnecessary risk to take. On the other hand, setting a small IOC does not change the pre-existing probability distribution too much. The new entry may still gradually gain probability over alternatives. In essence, IOC must be non-zero, and its value determines the speed at which new entries can gain salience.

Both for segmentation and lexical alternatives, we set IOC to 0.2. We have no evidence to suggest whether new items should more rapidly or more slowly gain salience. This value serves well in keeping the probability distribution more stable and easy to interpret.

MED is concerned with the largest allowed difference between the distributional semantics of the stem and lemma. In other words, if the word embedding for the lemma is very dissimilar to the word embedding for the stem, we expect the affix to not be recognized inside the lemma.

Several dissimilarity metrics can be considered for this purpose. For instance, Üstün and Can (2016) use cosine distance and report that the most feasible cosine distance threshold for Turkish word segmentation task is 0.25. If we follow Kunter et al. (2020), Euclidean distance should also be appropriate for this task, because vector magnitudes for affix embeddings must also be taken into account.

Word embeddings are shown by Kunter et al. (2020) to demonstrate consistent similarity/dissimilarity with respect to derivational processes. Lemmas derived with the help of the same affix appear in a similar distance and direction to their stem embeddings. In other words Euclidean distance between stem and lemma embeddings can be used as estimations for affix embeddings.

Figures 26 and 27 demonstrate the distribution of cosine and euclidean distances between stems and lemmas for each affix. Embeddings used in these calculations are taken from the pre-trained datasets of Grave et al. (2018).

We prefer a restriction on Euclidean distance, in order not to ignore the effect of vector magnitudes. We set the MED to 1.5, in the trials in Section 5.3.

These figures also show an asymmetry between IM and DM. On average, semantic contribution of an inflectional morpheme is significantly less pronounced than a derivational morpheme. It also has much less variation than that DM. We have seen evidence for that in Section 4.1. Figure 27 presents another take on the same kind of evidence.

NNI (nominal inflection) and VVI (verbal inflection) affixes show a much more regular semantic contribution. For VVI especially, and for NNI to some extent, the estimated embeddings are shorter. This means that, as expected, stems and inflected forms are semantically closer. If semantic closeness has an effect on the recognition of an affix, then IM must be easier to recognize than DM. In fact, this is documented by psycholinguistic experiments. While there are other reasons for this asymmetry, such as children's encountering more types with the same IM due to IM's productiveness, effect of semantic closeness might also be playing a role.

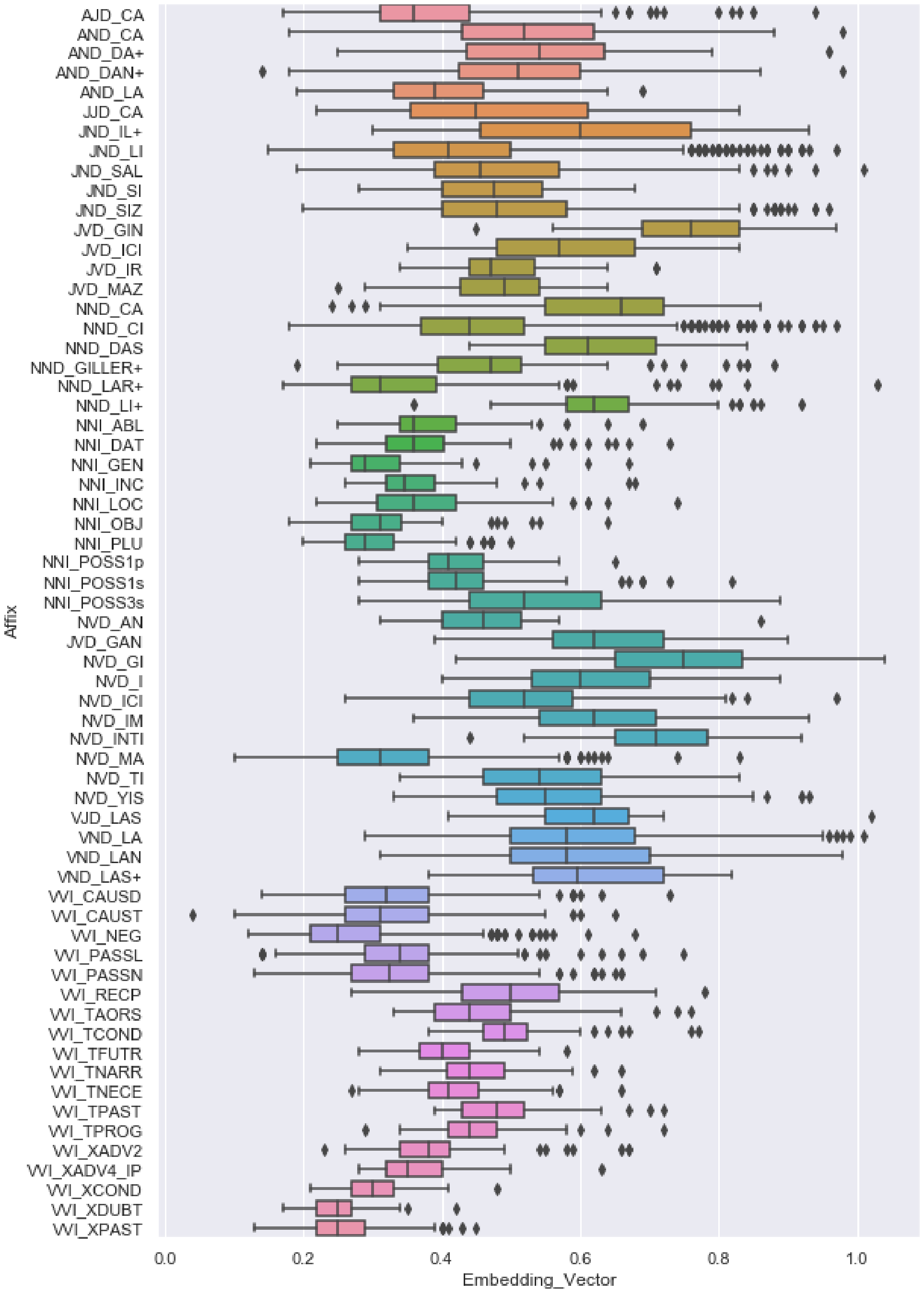


Figure 26: Boxplot for the cosine distance between stem and lemma embeddings for affixes with at least 30 lemmas

5.2.6 Baseline Trial

The toy example involves 3 simple and 3 complex words: *diz* ‘knee’, *dirsek* ‘elbow’, *bilek* ‘wrist’, *dizlik* ‘kneepad’, *dirseklik* ‘armrest’, *bileklik* ‘wrist strap’, respectively. *-lik*, which derives names for

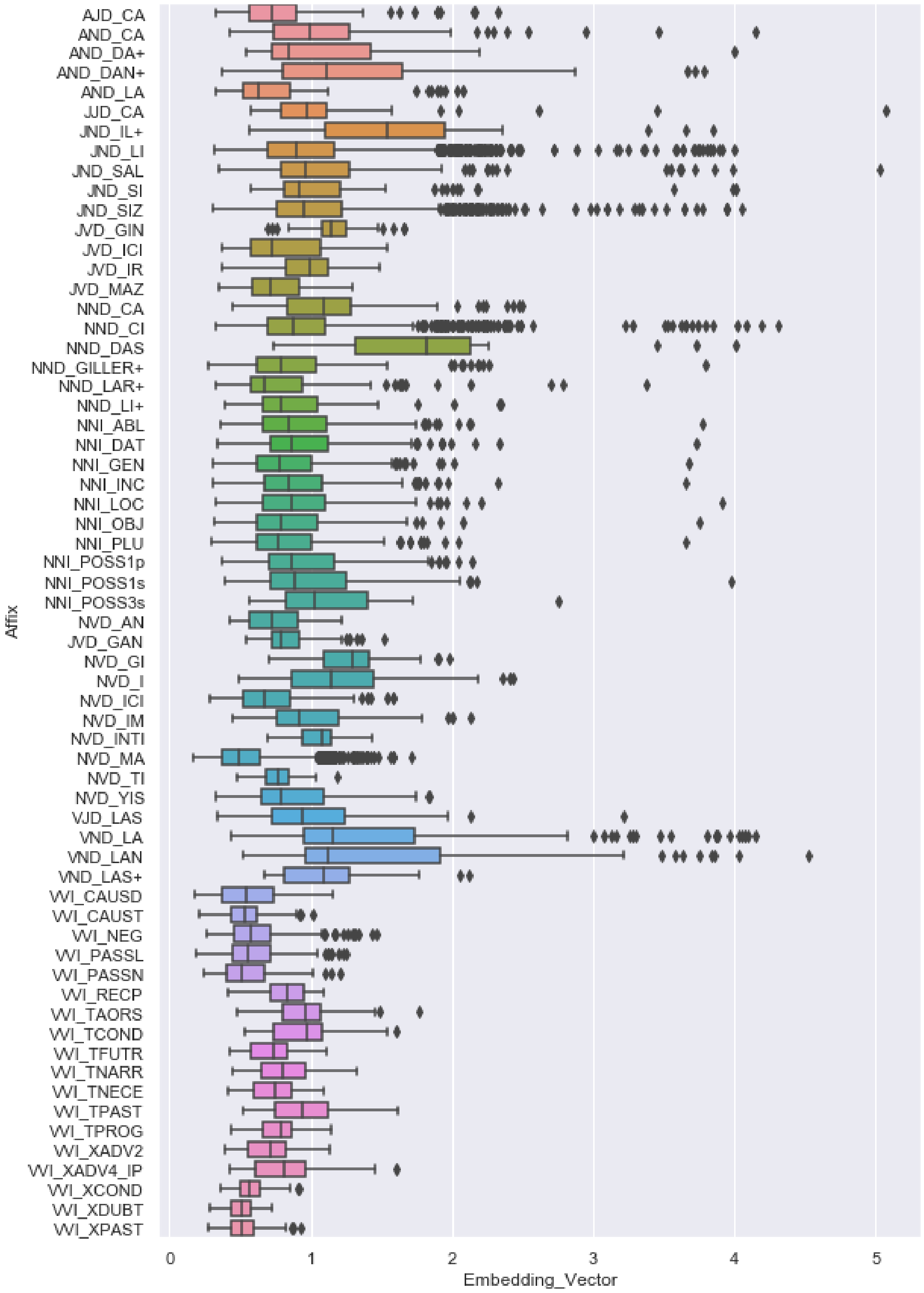


Figure 27: Boxplot for the distribution of Euclidean distance between stem and lemma embeddings for affixes with at least 30 lemmas

apparel worn on a specific body part, is the common suffix in the last three observations. The learning threshold is set as 2.

(154) Observation list for the baseline trial

a. *diz* ⊢ N: λa.(be knee a)

b. *dirsek* ⊢ N: λa.(be elbow a)

c. *bilek* ⊢ N: λa.(be wrist a)

d. *dizlik* ⊢ N: λaλb.and (be knee b) (wear (on b) a anon)

e. *dirseklik* ⊢ N: λaλb.and (be elbow b) (wear (on b) a anon)

f. *bileklik* ⊢ N: λaλb.and (be wrist b) (wear (on b) a anon)

g. *dizlik* ⊢ N: λaλb.and (be knee b) (wear (on b) a anon) (4 times)

h. *bileklik* ⊢ N: λaλb.and (be wrist b) (wear (on b) a anon) (4 times)

Going through the list, the algorithm first learns the 3 simple words. As they are simple words, and the lexicon is initially empty, there are no valid segmentations. The fourth word is a complex one, but there is no valid segmentation such as diz-lik, since *-lik* is not in the lexicon yet. *dirseklik* is also learned without segmentation, because the affix has not been recognized yet.

If there is no choice, there is no preference; if there is no preference, Bayesian inference is unnecessary. This is the case when SN and LN do not contain multiple states. With the first few observations, the algorithm cannot even interpret the expression. If the simple observations were encountered again, there would not be multiple alternatives for segmentation and lexical selection. In those cases, BBN is trivial.

Right after the fifth observation, the encounter threshold for *-lik* is passed. As a result, it is recognized as a distinct lexical item, according to the rules explained in Section 5.2.3. When *bileklik* is observed, it is finally possible for the segmentation stage to produce the segmentation alternative bilek-lik.

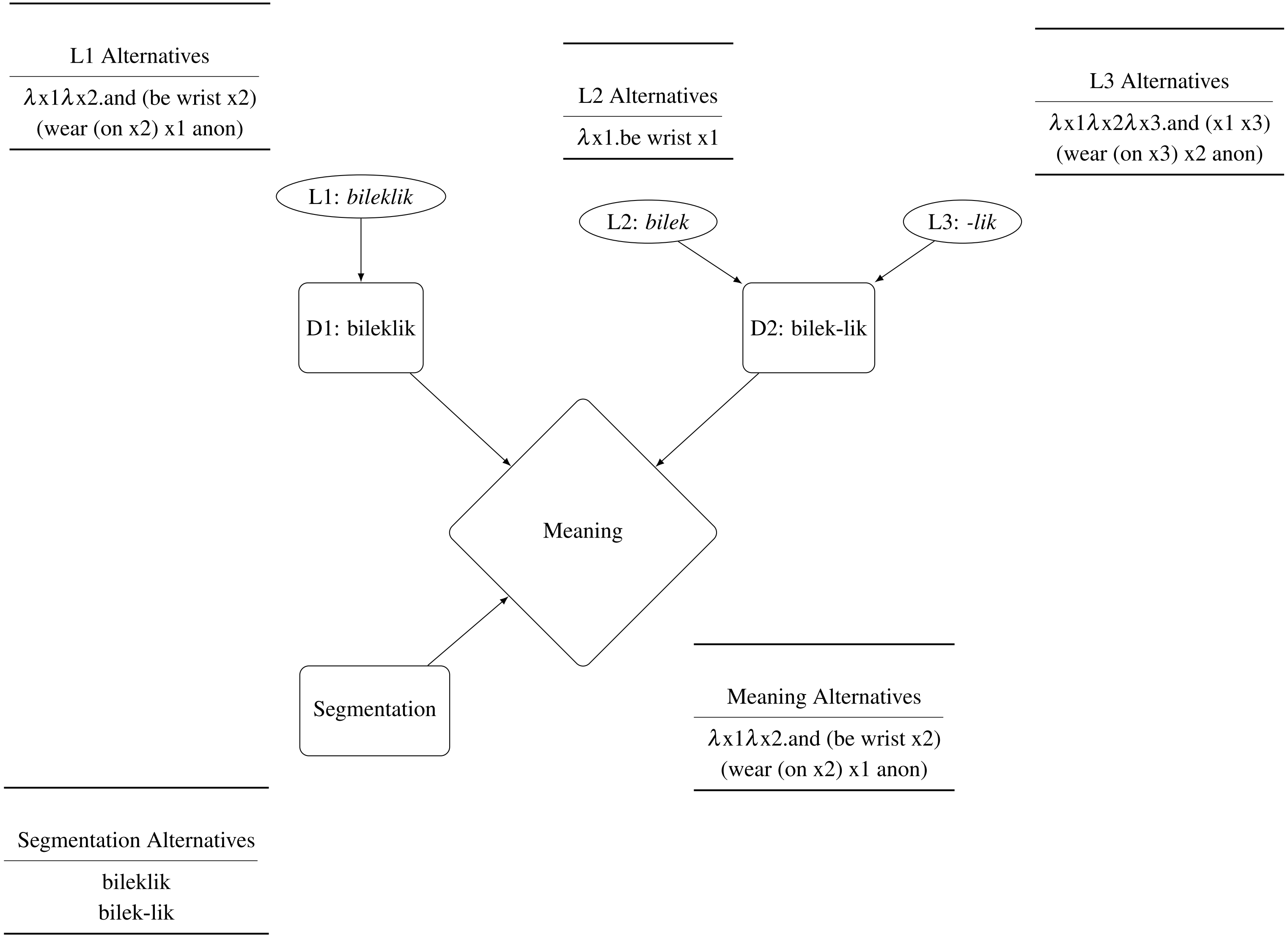
To illustrate the details, we present and comment on the nodes and states of the BBN built for the observation *bileklik*. We do not use set notation in order to improve readability. Figure 28 demonstrates the BBN for *bileklik* ‘wristband’.

(155) Observation 6

a. *bileklik* ⊢ N: λx1λx2.and (be wrist x2) (wear (on x2) x1 anon)

Since none of the segments match multiple lexical items, the only competition in this BBN is between segmentation alternatives. Figures 29 and 30 illustrate the evolution of probabilities of segmentation alternatives for *dizlik* and *bileklik*, respectively.

Since *dizlik* was first encountered while decomposition was not possible, the retrieval path was the only available option, hence probability 1. This was the case until it is encountered again, after which the decomposition path became more prominent (0.69 against 0.31). IOC determines the initial bias due to the order of exposure; a larger probability is assigned for the alternative that is observed first.



L1 Alternatives
λx1λx2.and (be wrist x2) (wear (on x2) x1 anon)

L2 Alternatives
λx1.be wrist x1

L3 Alternatives
λx1λx2λx3.and (x1 x3) (wear (on x3) x2 anon)

Meaning Alternatives
λx1λx2.and (be wrist x2) (wear (on x2) x1 anon)

Segmentation Alternatives
bileklik
bilek-lik

Figure 28: A toy BBN to represent the derivational structure of *bileklik*

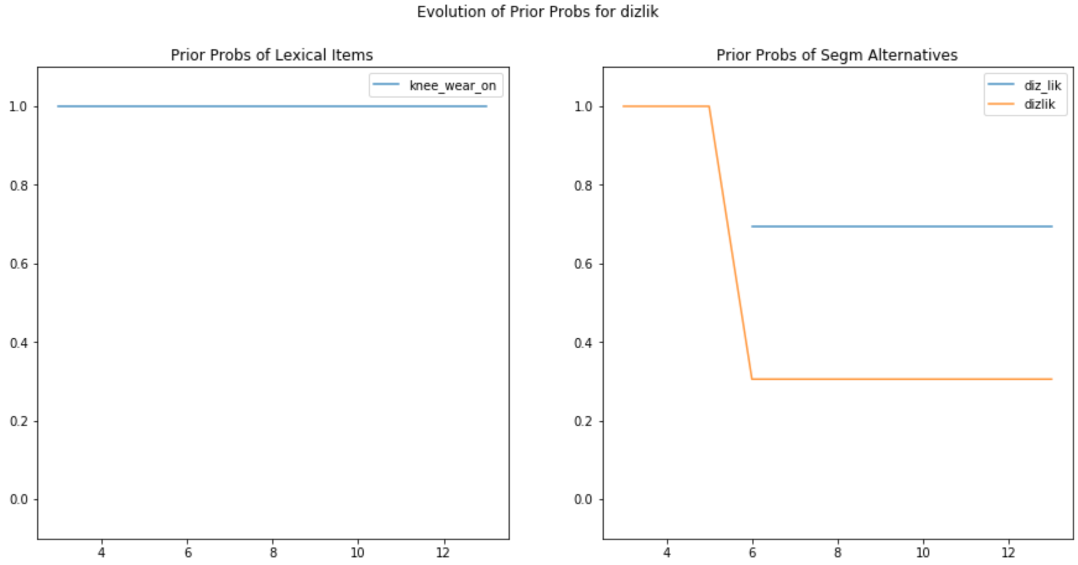


Figure 29: Lexical and segmentation probabilities for *dizlik*

Repeated encounters with the same observation does not change the probabilities in this case, because in the absence of further asymmetries (such as polysemy), there is no statistical effect to tip the balance. Segmentation alternatives quickly reach a steady-state.

The case for *bileklik* is similar. The only difference is that when *bileklik* is first encountered, the decomposition path is already available. Therefore, IOC does not have a significant effect; BBN distributes the observation counts to the alternative segmentations.

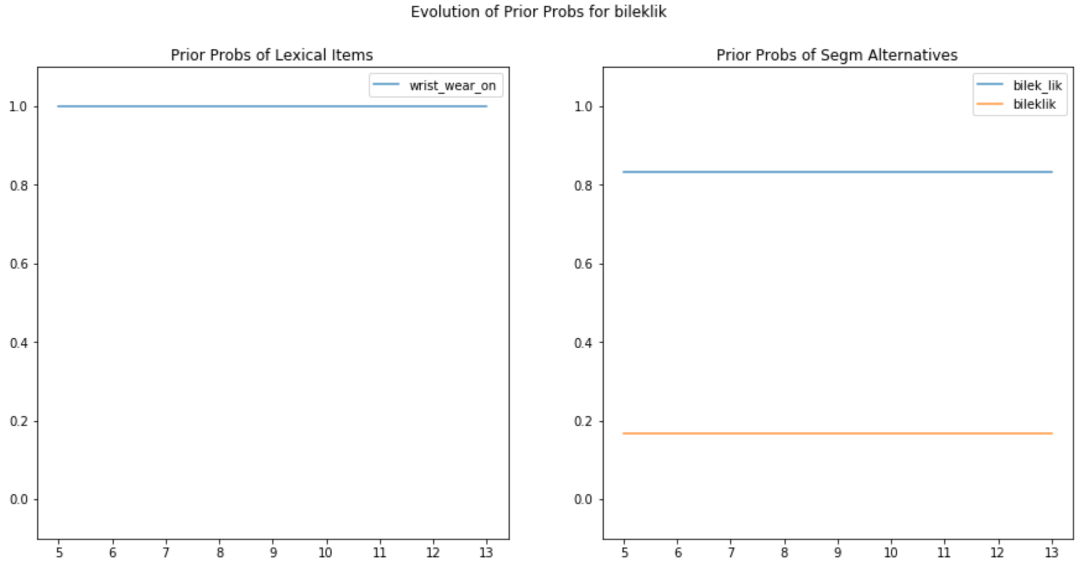


Figure 30: Lexical and segmentation probabilities for *bileklik*

5.3 Results

We conducted two kinds of trials. In the first group (core trials) we focus on the core functions of the algorithm in order to demonstrate the operating principles in an applied setting. We use minimal observation lists, so that subtle differences due to small changes in the properties of observations and lexical items can be clearly observed. These trials serve as a proof-of-concept for the possible applications of the model and the algorithm.

For the second group (pilot trials), we use larger datasets. In these datasets, our focus is more on pointing out larger patterns emerging from the structural properties of CdS and BBN. These trials aim to demonstrate the model's degrees of freedom and the structural differences between linguistic modules such as IM, DM and syntax.

In this section, we discuss the results from these trials. In order to keep the discussion simple, we always start with empty grammars; the algorithm has nothing to work with, except the observation list.

Our setup follows the steps laid out in Figure 1. An observation list and an empty lexicon are provided containing surface forms, categories and logical forms of several expressions. These expressions range from simple words to phrases and full sentences. The algorithm is expected to populate the lexicon with learned items. These items may be phrases, sentences, as well as words or bound forms, depending on the observations.

Trials were carried out on a Windows 10 computer with 16 GB RAM, using Python 3.7. BBN is implemented using the pomegranate library (Schreiber, 2018). Minimal custom libraries were built to implement segmentation, CCG parsing, derivation, allomorphy, affix recognition and visualization operations.

For meaning representation, we use the grammar built in Section 4.4. We focus on DM that are especially frequent in child acquisition data. We scanned the transcripts used in Avcu (2014) and Slobin (1982) to select suitable affixes for experiments:

(156) a. *-CI*: *kitapçı* 'bookseller', *kamyoncu* 'truck driver', *Atatürkçü* 'follower of Atatürk'... Denominal nominal derivational affix that indicates a product-seller, driver-vehicle or leader-follower relationship between the root word and the lemma.

b. *-lIK*: *kitaplık* 'bookshelf', *gözlük* 'glasses', *insanlık* 'humanity'... Denominal nominal derivational affix that indicates an object-container, body part-apparel or property relationship between the word and the lemma.

c. *-lI*: *tuzlu* 'salty', *Ankaralı* 'from Ankara'... Denominal nominal derivational affix that indicates ownership or quality.

d. *-lA*: *bağla* 'tie', *mayınla* 'mine'... Denominal verbal derivational affix that indicates an action that somehow relates to the stem noun or adjective.

e. *-sIz*: *evsiz* 'homeless', *parasız* 'money-less'... Denominal nominal derivational affix that indicates the lack of an object. Roughly the opposite of the primary meaning of *-lI*.

Table 33: Summary of core trials

Trial	$\lvert\Theta\rvert$	$\lvert\theta\rvert$	$\lvert\Lambda_{t+1}\rvert$	C_U	C_F	Theme
CT0	14	6	7	0.29	0.29	Baseline
CT1	14	6	7	0.29	0.29	Order of Exposure
CT2A	25	9	12	0.67	0.67	Multiple Affixes/Segmentation Ambiguity
CT3A/B/C	28	20	23/21/24	0.43	0.43	Allomorphy (Without/With/Without then With)
CT4	111	21	24	1.21	0.46	Polysemy/Lexical Ambiguity

f. *-ArI*: *içeri* 'inwards', *dışarı* 'outwards'... Denominal nominal derivational affix that indicates the direction towards an object. Although it is frequent and used with multiple stems, it is not generally recognized as an affix.

g. *-rA*: *bura* 'here', *şura* 'there', *ora* 'there', *nere* 'where', *sonra* 'afterwards', *ücra* 'remote'. Denominal nominal derivational affix that indicates the location of an object. Although it is frequent and used with multiple stems, it is not generally recognized as an affix.

Combination of these affixes may produce multiply-derived forms such as göz-lIK-CI-lIK (*gözlükçülük* 'the profession of an optician') and göz-lIK-lI-lIK (*gözlüklülük* 'the state of wearing glasses'). We do not expect the pre-trained word embedding datasets to provide reliable embeddings for such rare forms, but we want the algorithm to operate on some highly complex forms. Therefore, we ignore the distributional semantics restriction during trials.

5.3.1 Core Trials

An overview of the core trials, including the baseline trial discussed in Section 5.2.6 is given in Table 33.

The baseline trial (CT0) shows that the order of exposure has an effect on the prior probabilities of certain segmentation alternatives; but this is a difference in degree, not in kind. Moreover, the gap between the treatment of earlier and later learned expressions close rapidly.

Still CT0 only looks at the case where complex expressions are encountered after their stems. This is not a requirement, but a preference for simplicity. CT1 probes the other extreme, by putting all stems after the corresponding derived forms. The observation list for CT1 is given in Appendix B.1.3.

Having encountered derived forms before stems, the algorithm does not recognize *-lik* in the first six observations of CT1. Affix recognition is only triggered when there are lexical items in the lexicon that match a segment of the derived form. The lexicon has sufficiently expanded after the sixth observation, so that *dizlik* and *bileklik* trigger the recognition of *-lik*. This produces a small and temporary difference in the probability distribution among segmentation alternatives, compared to the results of CT0.

Again, the difference is only in degree; full BBNs can be constructed for these items in any possible order of exposure. The only requirement is for the algorithm to encounter a derived form after the lexicon is large enough. Even if just once, such an encounter triggers affix recognition.

Real-life analogies for this process is easy to find, albeit virtually impossible to experimentally verify. If CDS only includes single-morpheme words, it is obviously not possible for the child to learn affixes. If it only includes derived forms (with different roots), it is again not possible for affixes to be learned. Unlike free forms that can be learned individually, bound forms require the hearer to establish a connection between two free forms. Without knowing these two forms (stem and lemma) separately, the hearer cannot learn the affix.

Of course, this is not the only requirement for affix learning. The hearer must be able to construct an appropriate logical form for the candidate affix. This is only possible when a valid logical relationship exists between the stem and the lemma. The logical relationship is tested by the application of LF templates described in Section 5.2.3.

After the affix is recognized, the strength of preference between retrieval and decomposition depends on previous exposure. If the hearer encounters a derived form many times before decomposition becomes available, retrieval may be preferred. If the derived form is encountered only a few times before decomposition becomes available, decomposition may gain more preference. This is reflected in CT0 and CT1. (If we increase the number of encounters for *dizlik* to 3 from 2, before *-lik* is available, final prior probabilities of retrieval and decomposition change from 0.79 to 0.90 and from 0.21 to 0.10, respectively.)

In both CT0 and CT1, four encounters are provided for *dizlik* and *bileklik* to observe the evolution of probabilities, but without any lexical ambiguity, the trade-off between segmentation alternatives immediately reaches steady-state.

The minimal trials of CT0 and CT1 demonstrate the working principles of the algorithm, and results can be logically justified. However, a child encounters thousands of forms per day, in various sequences and combinations; therefore, it is virtually impossible to extract such principles from experimental data. The richness of CDS is so vast, that even if such principles are actually in effect during acquisition, they would be quickly covered up by confounding factors and processes. In the long term. steady-state lexicons of children are expected to converge to similar lexicons, forever erasing the effect of these simple trade-offs.

CT2 looks into the case where multiple affixes are applied on a stem. The slightly expanded observation list used for CT2 is given in Appendix B.1.4. In this trial, we demonstrate two extensions to the baseline: affix arity and consecutive affixes being recognized in one piece. In addition to the observations in CT0, *dizlikçi* ‘kneepad seller’, *dirseklikçi* ‘armrest seller’, *bileklikçi* ‘wrist strap seller’ are added to the observation list.

The number of segmentation alternatives grows exponentially with the number of morphemes, as discussed in Section 5.2.2. Therefore, with the full lexicon, there are $2^{3-1} = 4$ segmentation alternatives for the three-morpheme derived forms in CT2. These are quickly discovered after affix recognition is complete. Similar to CT0 and CT1, the absence of lexical ambiguity leads the probability distribution to quickly stabilize.

The first highlight of this trial is the possibility of learning complex morphemes. The algorithm treats all segments in the same way in terms of affix recognition. A segment found to be common across multiple lexical items triggers affix recognition. This is true if the segment contains a single morpheme, or if it contains multiple morphemes. The same mechanism applies. There is no rule to encourage or prevent the recognition of larger or smaller segments. Once a complex segment fits the rules of recognition, it is recognized all the same.

This could be the very process that enables fusion, which is discussed in Section 3.3.3. For the speaker, there is no other way than statistical inference for developing a preference between competing segmentation alternatives.

(157) a. *-likçi* ⊢ N\N : λx1λx2λx3λx4.and (and (x1 x4) (wear (on x4) x3 anon)) (sell x3 x2)

Normally, we expect individual morphemes to be more frequently used than combinations of morphemes. When this is the case, individual morphemes gradually gain prominence at the expense of complex forms (provided that complex forms do not provide a less flexible meaning). In the extreme case where two morphemes only occur together, they can only be recognized together. In the cases in-between, the proportion of times two morphemes occur together versus they occur separately, determine the salience of the complex segment. Fusion may be the consequence of this statistical asymmetry.

If we follow this idea to its limit, fusion does not have to be the consequence of phonological changes binding two morphemes together. In fact, phonological changes cannot occur before the complex morpheme is recognized, because the phonological change would make the expression non-interpretable. In our view, fusion emerges due to statistical effects and phonological variations follow, further concealing the individual morphemes.

The second highlight is the necessity to keep track of the arity of an affix. As discussed in Section 4.4, affixes have to accommodate the number of variables in their stem. This is a requirement of categorial grammar. For instance, an affix may derive from both a simple noun containing a single variable, or a derived noun containing multiple variables. It is impossible to derive both nouns using the same LF. Therefore, multiple lexical items must be constructed for the same affix, but with different number of bound variables. We call this affix arity.

A simple example can be found comparing CT2A and CT2B. *-çi* learned in CT2A is an affix that derives already complex forms such as *dizlik*. It contains four variables: three variables denoting the input variables and one variable denoting the seller. In contrast, *-çi* learned in CT2B is an affix that derives simple forms such as *et* ‘meat’. It contains three variables: two variables denoting the input variables and one variable denoting the seller. In both cases, logical forms of the derived forms contain one variable for each object and derivation operation.

(158) a. *-çi* (CT2A) ⊢ N\N : λx1λx2λx3λx4.and (x1 x3 x4) (sell x3 x2)

b. *-çi* (CT2B) ⊢ N\N : λx1λx2λx3.and (x1 x3) (sell x3 x2)

Learning multiple lexical items due to affix arity is easily justified. Just like how multiple lexical items must be learned for a verb’s uses with different arity, affixes must also be able to handle different numbers of arguments in different contexts. A single affix producing multiple lexical items is simply a consequence of using categorial grammar as a meaning representation framework.

In trials CT3A-B-C, we study the effect of allomorphy. The algorithm does not require the existence or non-existence of allomorphy. Allomorphy can be easily represented as the interchangeability between characters. We ignore suppletive allomorphy and only represent regular allomorphy for simplicity Details of this method are presented in Section 5.2.1.

When allomorphy is not assumed (CT3A), the algorithm operates in the same way as explained in previous trials. Given 10 place names (*Amerika* 'the USA', *Ankara* 'Ankara', *İstanbul* 'İstanbul' etc.) and 10 adjectives indicating one's hometown (*Amerikalı* 'from the USA', *Ankaralı* 'from Ankara' etc.), the algorithm learns three different affixes with the same function. If we provided additional observations such as *Üsküp* 'Skopje' and *Üsküplü* 'from Skopje', all allomorphs of the affix would be learned, but this is not necessary.

(159) Allomorphs of *-lI*

a. *-lı* ⊢ N\N : λx1λx2λx3.and (x1 x3) (be (from x3) x2)

b. *-li* ⊢ N\N : λx1λx2λx3.and (x1 x3) (be (from x3) x2)

c. *-lu* ⊢ N\N : λx1λx2λx3.and (x1 x3) (be (from x3) x2)

When allomorphy is assumed (CT3B), only the meta-morpheme *-lI* is learned.

(160) a. *-lI* ⊢ N\N : λx1λx2λx3.and (x1 x3) (be (from x3) x2)

Meta-morphemes can be used in the same way as allomorphs; they take part in derivations and they are represented in segmentation alternatives. If allomorphy is not assumed, the segmentation alternatives for *Ankaralı* are Ankaralı and Ankara-lı. If allomorphy is assumed, the segmentation alternatives are Ankaralı and Ankara-lI.

The effect of assuming allomorphy is two-fold. First, by binding together the allomorphs of an affix, the hearer is able to recognize common segments in a larger pool of observations. For instance, without allomorphy, *Ankaralı* and *İstanbullu* have no common segment; but with allomorphy, these words are derived by the same affix. Due to the learning threshold, assuming allomorphy facilitates learning.

Second, allomorphy acts as a way of information compression. An affix can be represented by a single lexical item, instead of several allomorphs. When regular allomorphy and suppletive allomorphy are combined, this kind of compression may have a considerable effect on the size of the lexicon.

It may also be the case that allomorphs are first represented individually, but are gradually replaced by the meta-morpheme. This is exactly what happens in CT3C, where we carry out CT3A and CT3B back to back. We add *Muğlalı* observation 40 times at the end, in order to observe how segmentation alternatives converge to the steady-state.

Figure 31 shows this process. The trial starts without allomorphy. The decomposition path for *Muğlalı* is more prominent at the beginning, as the affix *-lı* has been previously recognized. At t=28, allomorphy is introduced. Subsequently, *-lI* is also recognized and generates a new segmentation alternative Muğla-lI. As expected, this alternative gradually gains prominence at the expense of Muğla-lı. Retrieval path gains even more prominence. The change is not due to Bayesian inference; it is simply because the *-lı* is ignored during segmentation in favor of *-lI*.

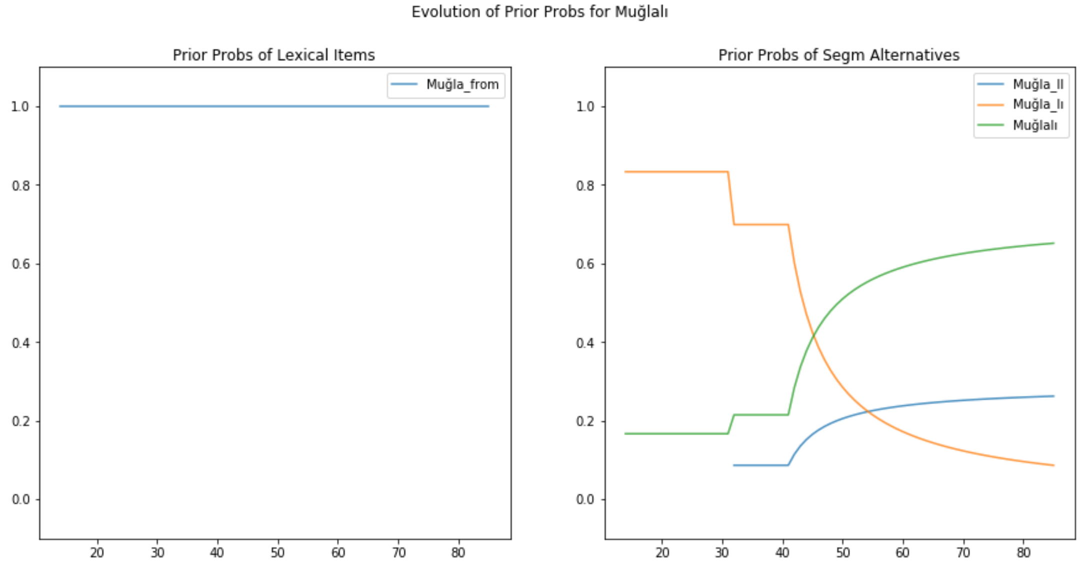


Figure 31: CT3C lexical and segmentation probabilities for *Muğlalı*

Of course the full force of Bayesian inference emerges when there is ambiguity in all layers of processing. So far, we have mostly looked into segmentation ambiguity. CT4 includes several examples of homonymy and polysemy. Observation list of CT4 is given in Appendix B.1.6.

Items with homonyms are *servis* ‘shuttle’ / ‘cutlery’, *servisçi* ‘shuttle driver’ / ‘cutlery seller’ and *kamyoncu* ‘truck seller’ / ‘truck driver’. 3 variations of *-CI* is expected to be learned, denoting object-seller, vehicle-driver and person-follower relations. Having the same form and related meanings, *-CI* is said to have 3 polysemous uses.

Due to homonymy and polysemy, prior probability graphs for lexical items are no longer trivial. Also, for the same reason, prior probabilities of segmentation alternatives do not immediately converge to a steady-state. Graphs for *kitapçı*, *gözlükçü*, *Atatürkçü*, *Aristocu*, *kamyoncu*, *taksici*, *servisçi* and *-CI* are presented in Appendix B.1.6. Here, we only go over a few of the graphs to illustrate the trade-offs emerging from this observation list.

The most eye-catching difference of this trial compared to the previous ones is the retrieval path’s tendency to gain prominence over time. In previous trials, the opposite case was true. The preference towards retrieval can be observed for all derived forms without exception. The effect is more significant after all meanings of *-CI* are recognized.

This is what we anticipate due to Bayesian Occam’s Razor (BOR) discussed in Section 5.1.6. When -*CI* has only one meaning, it does not produce flexible hypotheses (interpretations). The decomposition path produces only a single interpretation. The retrieval path does the same. Therefore, segmentation alternatives are equally flexible; no preference develops towards either alternative. Once *-CI* acquires 3 meanings, it starts to license multiple interpretations. This means the decomposition path becomes more flexible. On the other hand, the retrieval path still produces a single correct interpretation. Due to BOR, Bayesian inference develops a preference towards less flexible alternatives. In this case, the less flexible alternative is retrieval.

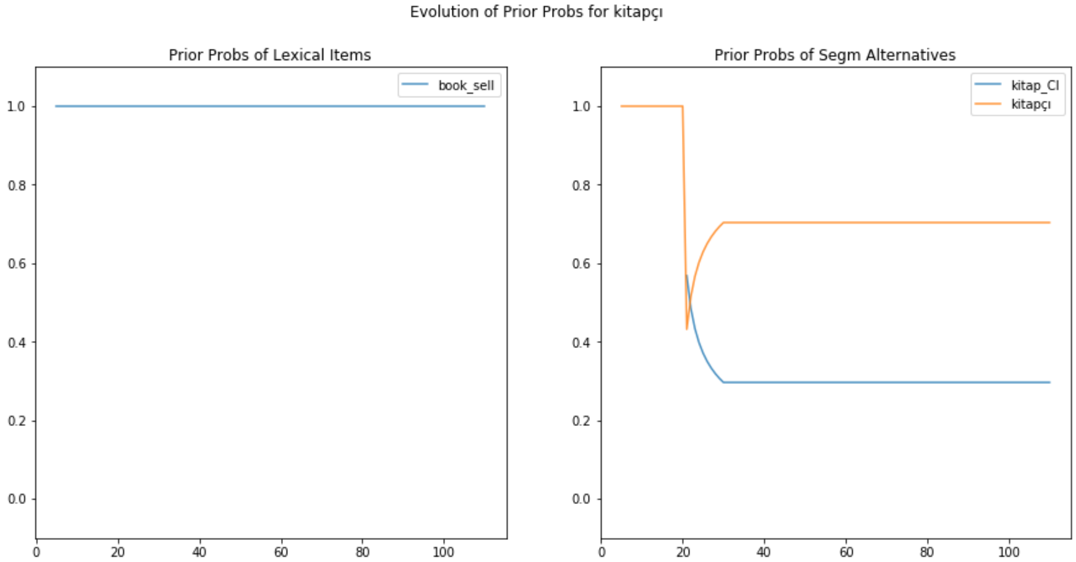


Figure 32: CT4 lexical and segmentation probabilities for *kitapçı*

In most cases, such as *kitapçı* and *Aristocu*, the initial prior probability of decomposition is higher, due to the small IOC of retrieval. With repeated observations, retrieval quickly gains prominence and converges to an asymptote. In other cases, such as *gözlükçü* and *Atatürkçü*, prior probabilities of segmentation alternatives start the same, but quickly converge in a similar fashion. These examples demonstrate that the order of exposure does not make a lasting difference; prior probabilities quickly reach a new balance after only 10 repetitions.

There are cases where not only the affix, but also the stem has homonyms and cases where the derived form itself has homonyms. Whatever the case, BOR favors the less flexible alternative. The less flexible alternative is generally expected to be retrieval, because homonymy with complex forms is less frequent.

Graphs of *kamyoncu* and *servisçi* are good examples that illustrate how the effect of BOR is present, but not as strong. Due to retrieval also producing multiple interpretations, the preference towards retrieval does not develop as quickly.

When polysemy and homonymy relations exist for both morphemes, even two-morpheme words are represented by a complex network.

(161) Different interpretations of *servisçi*

a. *servisçi* $\vdash$ N : λx1λx2.and (be shuttle x2) (drive x2 x1)

b. *servisçi* $\vdash$ N : λx1λx2.and (be cutlery x2) (sell x2 x1)

When it comes to human processing, There is significant evidence that parallels this preference towards retrieval of complex words. Bertram et al. (2000) observes that whole-word representations are

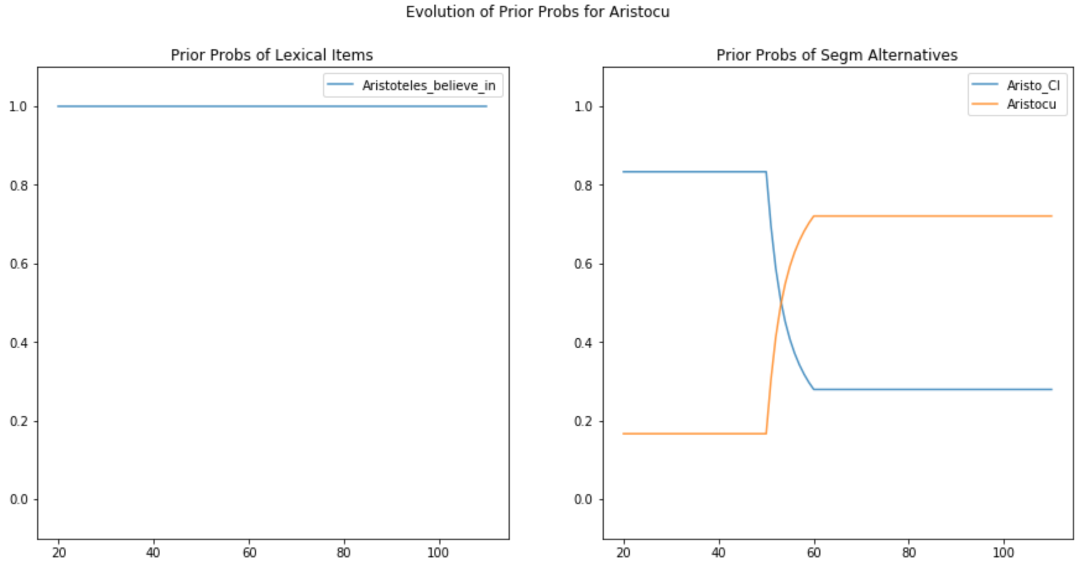


Figure 33: CT4 lexical and segmentation probabilities for *Aristocu*

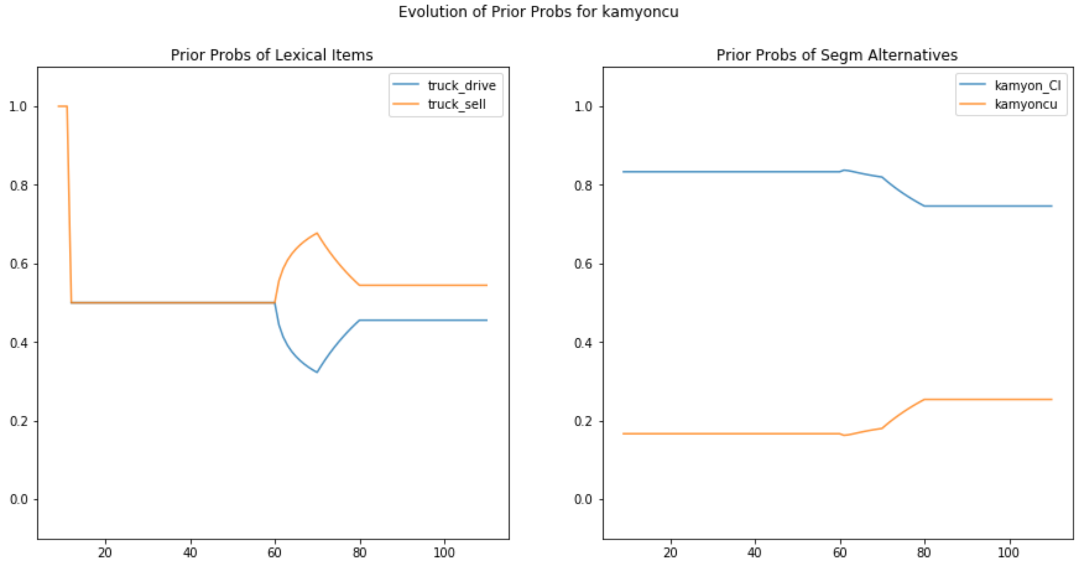


Figure 34: CT4 lexical and segmentation probabilities for *kamyoncu*

much more prominent when accessing complex words with high-frequency. He adds that several other authors have come to the same conclusion.

If their observations are correct, the mechanism behind this phenomenon might be closely represented by BOR. As they suggest, we find that complex words are more likely to be interpreted by retrieval. Again as they suggest, we find that the preference towards retrieval builds gradually; therefore, the effect is expected to be greater with high-frequency words. Additionally, we find that this effect only

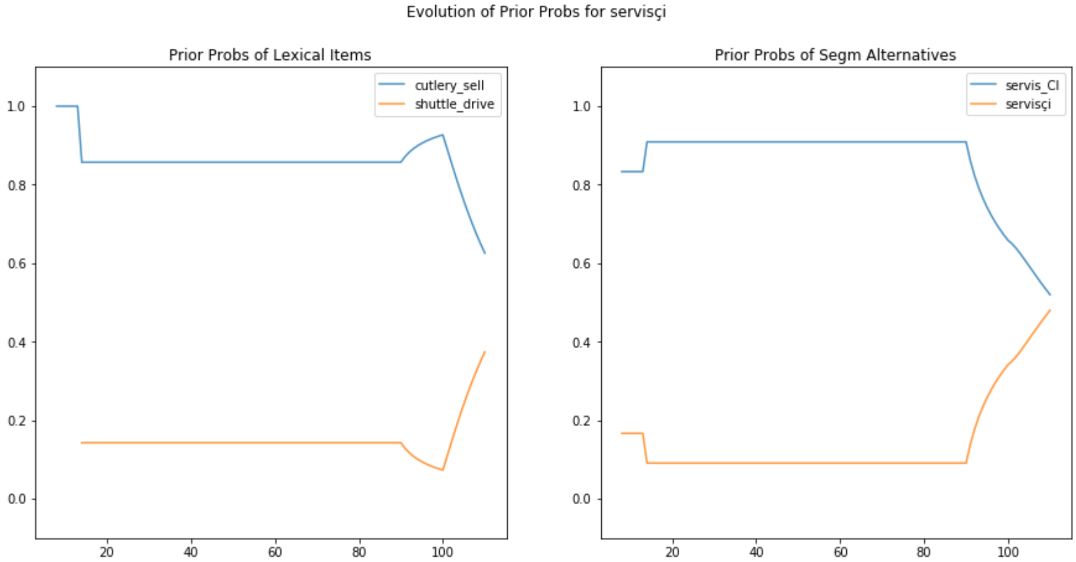


Figure 35: CT4 lexical and segmentation probabilities for *servisçi*

emerges in the presence of some level of lexical ambiguity. Neither Bertram et al. (2000) nor others make this distinction.

Before concluding this section, we must also take a look at the metrics on compact representation. For the first four trials, the number of independent parameters to represent the lexicon is at a minimum. This is due to a lack of polysemy and homonymy. In each case, we only work with a small number of segmentation alternatives. Adopting CdS does not change the number of independent parameters.

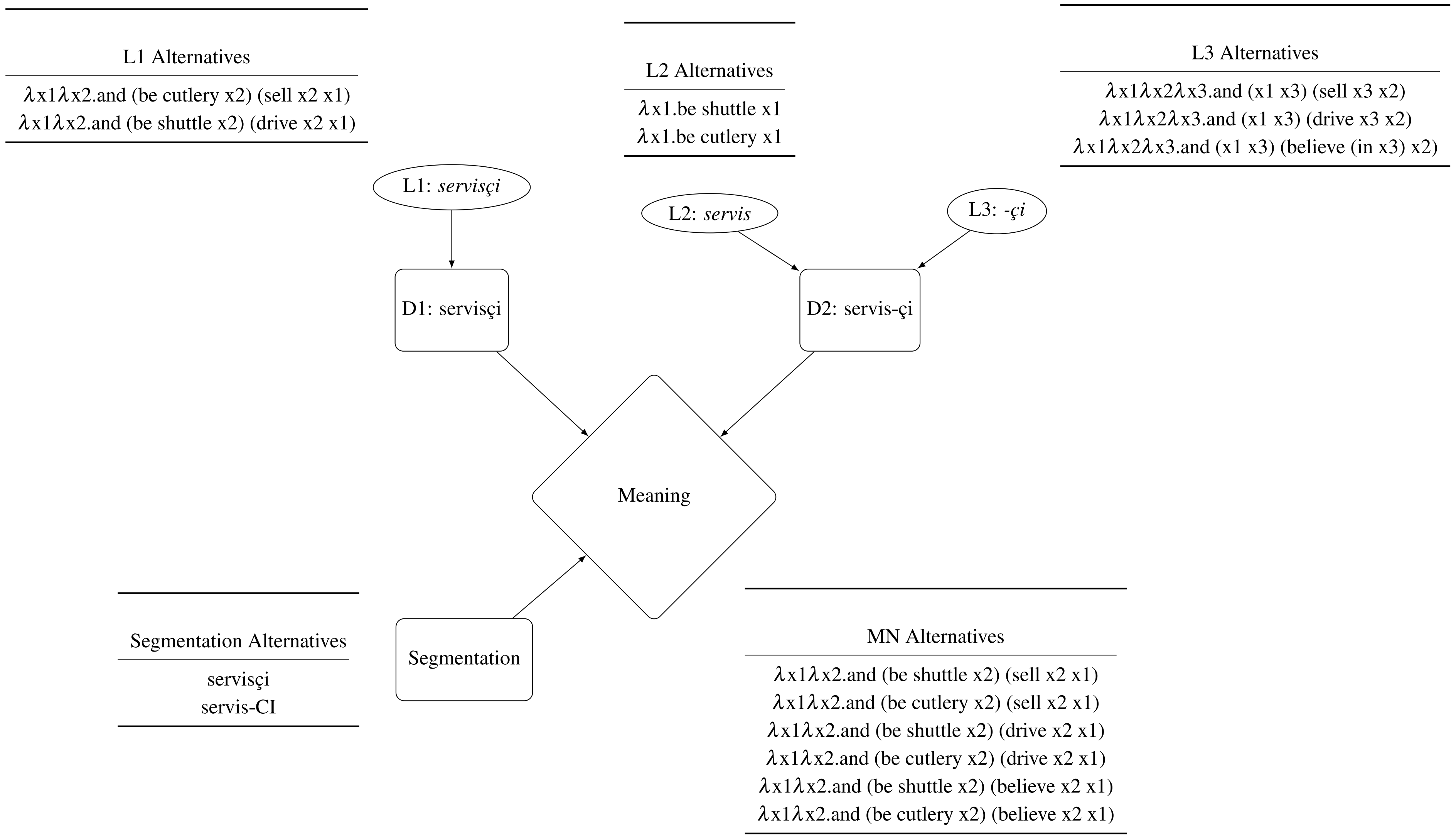
With the introduction of homonymy and polysemy, the number of interpretations for an average observation increases significantly. In CT4, we need to choose from multiple interpretations for many of the segmentation alternatives. This results in a large number of independent parameters to build an unstructured representation. On the other hand, the factorized representation of CdS decouples the segmentation layer from the lexical selection layer, and requires much fewer independent parameters. Even with the small observation list of CT4, containing only 21 distinct observations and little polysemy, we can observe a significant difference between $C_U = 1.21$ and $C_F = 0.46$.

5.3.2 Pilot Trials

In the second group of trials, we examine the representation of DM, IM and syntax. As mentioned earlier, CdS is especially suitable for representing DM. In principle, the structure and the model should also be applicable on IM and syntax in general. They do not require any assumptions specific to DM.

An overview of the pilot trials is given in Table 34. Allomorphy is always assumed from now on. In order to avoid clutter, we will often ignore explicit representation of tense.

CT5 includes inflected and derived forms, as well as combinations of DM and IM. A typical derived form contains 1-2 derivational affixes. 3-4 inflectional affixes are routinely applied on verbs and



L1 Alternatives
λx1λx2.and (be cutlery x2) (sell x2 x1)
λx1λx2.and (be shuttle x2) (drive x2 x1)

L2 Alternatives
λx1.be shuttle x1
λx1.be cutlery x1

L3 Alternatives
λx1λx2λx3.and (x1 x3) (sell x3 x2)
λx1λx2λx3.and (x1 x3) (drive x3 x2)
λx1λx2λx3.and (x1 x3) (believe (in x3) x2)

Segmentation Alternatives
servisçi
servis-CI

MN Alternatives
λx1λx2.and (be shuttle x2) (sell x2 x1)
λx1λx2.and (be cutlery x2) (sell x2 x1)
λx1λx2.and (be shuttle x2) (drive x2 x1)
λx1λx2.and (be cutlery x2) (drive x2 x1)
λx1λx2.and (be shuttle x2) (believe x2 x1)
λx1λx2.and (be cutlery x2) (believe x2 x1)

Figure 36: A toy BBN to represent the derivational structure of *servisçi*

Table 34: Summary of pilot trials

Trial	$\|\Theta\|$	$\|\theta\|$	$\|\Lambda_{t+1}\|$	C_U	C_F	Theme
CT5	71	71	96	2.51	0.97	DM, IM
CT6	86	86	116	1.46	0.83	DM, IM, MWE, Full Sentences
CT7-1/2/3/4/5	5080	510	619	5.87	1.17	Sampling with replacement

sometimes on nouns. Regarding verbal inflection, we simulate processing on TAM, copula and person marker positions. Regarding nominal inflection, we simulate processing of plural, case and possessive.

When the number of affixes on a stem increases, the number of segmentation alternatives also increases. Occasionally, incorrect segments can be recognized as pseudo-affixes, if common segments can be found in multiple lemmas. This is especially true with semantically empty affixes such as case markers. Normally, LF templates check the candidate affix to ensure that it qualifies as a valid morpheme in terms of both semantics and surface form. However, the LF of semantically empty affixes cannot be checked against the observed semantics, resulting in a flexibility not available for other affixes.

(162) Incorrect affix discovery on nominal inflection

a. *ev* $\vdash$ N : λx1.be home x1

b. *evi* $\vdash$ NACC : λx1.be home x1

c. *-I* $\vdash$ NACC\N : λx1λx2.x1 x2 (Correct recognition)

d. *evim* $\vdash$ N : λx1.and (own x1 speaker) (be home x1)

e. *-Im* $\vdash$ N\N : λx1λx2.and (own x2 speaker) (x1 x2)

f. *evimi* $\vdash$ NACC: λx1.and (own x1 speaker) (be home x1)

g. *-ImI* $\vdash$ NACC\N : λx1λx2.and (own x2 speaker) (x1 x2) (Complex affix recognition)

h. *-mI* $\vdash$ NACC\NACC : λx1λx2.and (own x2 speaker) (x1 x2) (Incorrect recognition)

The composite affix *-ImI* may be considered an acceptable, yet redundant discovery. On the other hand, the discovery of an affix *-mI* is incorrect. These incorrect discoveries should be allowed for three reasons. First, preventing such discoveries requires introducing new assumptions and rules into the otherwise simple recognition process. We prefer a simpler and more flexible process. Second, incorrect discoveries are possible also for humans. We expect human speakers to make mistakes, but to develop a preference towards correct interpretations. Third, the hypothesis selection process is already in effect and incorrect hypotheses are expected to be crowded out gradually.

Composite segments are frequently discovered on verbal inflection, too. However, incorrect discoveries do not occur, as verbal inflectional affixes are not semantically empty.

Table 35: Segmentation alternatives for *servisçileri* ‘the shuttle drivers (ACC)’

Segmentation	Categories of Segments
servisçiler + I	N + NACC\N
servisçi + lArI	N + NACC\N
servisçi + lAr + I	N + N\N + NACC\N
servis + CIlArI	N + NACC\N
servis + CIlAr + I	N + N\N + NACC\N
servis + CI + lArI	N + N\N + NACC\N
servis + CI + lAr + I	N + N\N + N\N + NACC\N

Table 36: Segmentation alternatives for *geldiydim* ‘I had come’

Segmentation	Categories of Segments
geldiydi + m	S/N + S\(S/N)
geldi + ydIm	S/N + S\(S/N)
geldi + ydI + m	S/N + S/N\(S/N) + S\(S/N)
gel + DIydIm	V + S\V
gel + DIydI + m	V + S/N\V + S\(S/N)
gel + DI + ydIm	V + S/N\V + S\(S/N)
gel + DI + ydI + m	V + S/N\V + S/N\(S/N) + S\(S/N)

(163) Composite affix discovery on verbal inflection

a. *gel* ⊢ V : λx1λt.(come x1 t)

b. *geldi* ⊢ S/N : λx1λt.and (t < tref) (come x1 t)

c. *-DI* ⊢ S/N\V : λx1λx2λx3.and (x3 < tref) (x1 x2 x3)

d. *geldim* ⊢ S : λaλt.and (be speaker a) (and (t < tref) (come a t))

e. *-DIm* ⊢ S\V : λx1λx2λx3.and (be speaker x2) (and (x3 < tref) (x1 x2 x3))

f. *-m* ⊢ S\(S/N) : λx1λx2λx3.and (be speaker x2) (x1 x2 x3)

g. *geldiydi* ⊢ S/N : λaλt.and (tref < t0) (and (t < tref) (come a t))

h. *-DIydI* ⊢ S/N\V : λx1λx2λx3.and (tref < t0) (and (x3 < tref) (x1 x2 x3))

i. *-ydI* ⊢ S/N\(S/N) : λx1λx2λx3.and (tref < t0) (x1 x2 x3)

Derivation and inflection can be applied on the same root without any inconsistency. This is true for both nominal and verbal examples. Segmentation alternatives and consequent derivations of these items reflect the full range of interpretations.

When the number of morphemes increases, categorial compatibility between adjacent segments becomes critical in keeping the number of segmentation alternatives at a tractable level. The size of the BBN can still become enormous. For instance, the text file containing BBN for the maximum set of segmentation alternatives for *servisçileri* takes up 1.5 GB of disk space.

There are also constructions that combine morphological and syntactic elements. Syntactic constructions and idiomatic expressions such as *duymazdan gelmek* 'to pretend not to hear', *görmezden gelmek* 'pretend not to see', *evsiz kalmak* 'become homeless' or *işsiz kalmak* 'become jobless' should be discoverable by the hearer using the same algorithm. Allomorphy rules should still be applicable, but just to the bound part of the construction.

CT6 contains such cases. For instance, the following sample directly leads to the discovery of the construction. No intermediate steps are needed; the construction is recognized not as a collection of morphemes, but as a whole.

(164) Sample observation list for learning constructions

a. *ev* ⊢ N : λx1.be home x1

b. *iş* ⊢ N : λx1.be job x1

c. *evsiz kal* ⊢ V : λx1λx2.become (without x2) x1

d. *işsiz kal* ⊢ V : λx1λx2.become (without x2) x1

e. *-sIz kal* ⊢ V : λx1λx2λx3.and (x1 x3) (become (without x3) x2)

The common segment between the last two observations is *-siz kalmak*. If we apply allomorphy, it affects only the bound part *-siz*.

Word order is important on the sentence level. Instances of the same sentence with different word order lead to different discoveries.

(165) Effect of word order on lexical discovery (SV)

a. *defter* ⊢ N : λx1.be notebook x1

b. *defter geldi* ⊢ S : λx1.and (be notebook x1) (came x1)

c. *geldi* ⊢ S\N : λx1λx2.and (x1 x2) (came x2)

(166) Effect of word order on lexical discovery - VS

a. *geldi* ⊢ S\N: λx1.came x1

b. *geldi defter* ⊢ S : λx1.and (be notebook x1) (came x1)

c. *defter* ⊢ S\(S\N) : λx1λx2.and (be notebook x2) (x1 x2)

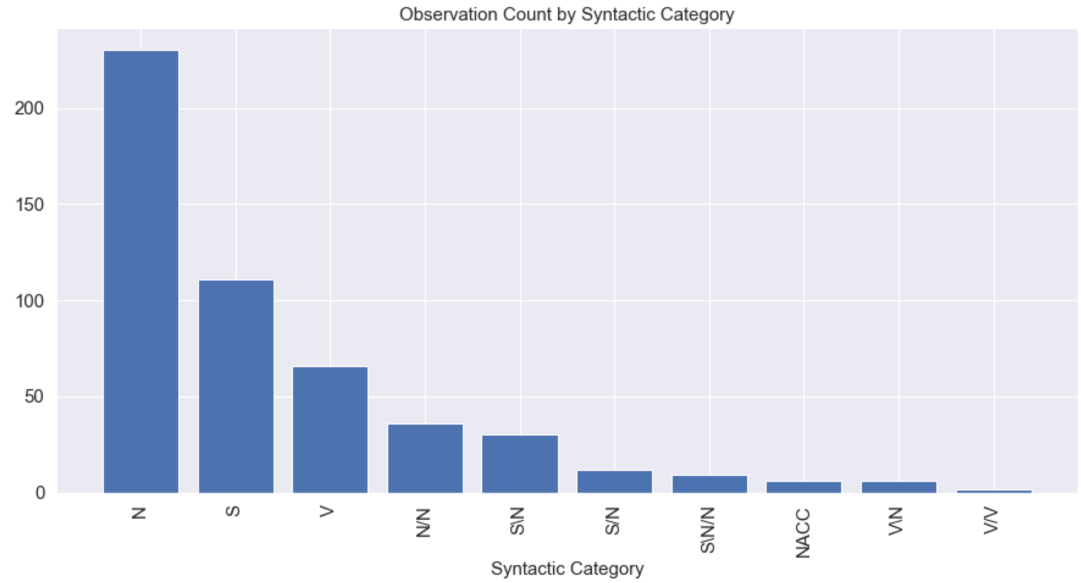


Figure 37: CT7 observations by syntactic category.

With the SV word order, if we provide the subject and the sentence, we obtain a verb with the S\N category. With the VS word order, if we provide the verb and the sentence, we obtain a subject with the S\(S\N) category. As an algorithmic choice, we carry out segmentation from the right; therefore the algorithm discovers constituents on the right. Provided that both word orders are encountered in the long run, this algorithmic choice does not put an arbitrary limit on learning word order.

The observation lists in CT5 and CT6 are still short, but we maximize learning by carefully ordering observations. The same result could be obtained with an unordered list if we sampled it a large number of times. To demonstrate this, trials in group CT7 are carried out based on large observation lists produced by sampling with replacement from a list of 508 observations. Comparing between the results of these five random trials demonstrate that order of exposure has only a temporary effect; all trials quickly converge to the steady-state lexicon.

The base observation list of CT7 contains observations of several different categories. It includes simple forms and derived forms; nouns and adjectives; transitive and intransitive verbs; as well as full sentences. Diversity in this observation list aims to reflect the general applicability of the basic principles of the proposed structure and model. As Figure 37 shows, most observations are of type N, S and V.

After the sampled observation lists (containing 5080 observations each) are processed by the algorithm, the final lexicon contains a wide variety of syntactic categories (shown in Figure 38). During affix recognition, we need to work with many different combinations of stem-lemma categories. As a result, affixes of appropriate categorization need to be created.

The final lexicon contains fairly complex forms. 125 distinct free forms out of 510 are composed of 3 or 4 segments. The distribution of lexical items according to their number of segments is given in Figure 39. This level of complexity, combined with the variety of syntactic categories, enables a comprehensive evaluation of the model's capabilities.

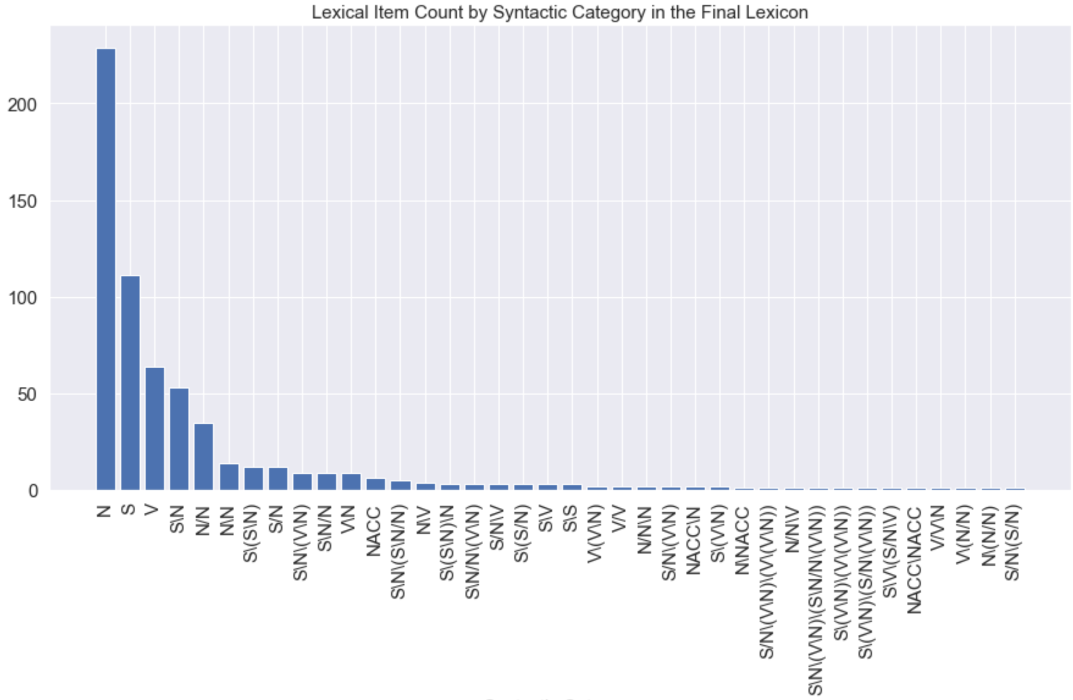


Figure 38: CT7 items in the final lexicon by syntactic category.

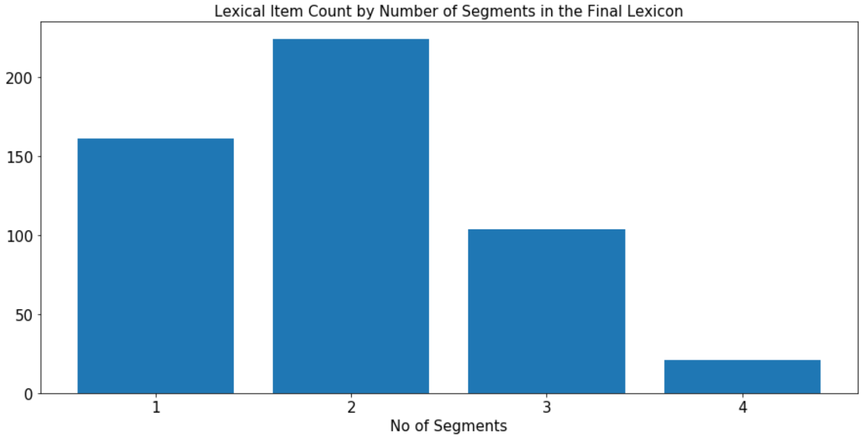


Figure 39: CT7 items in the final lexicon by number of segments.

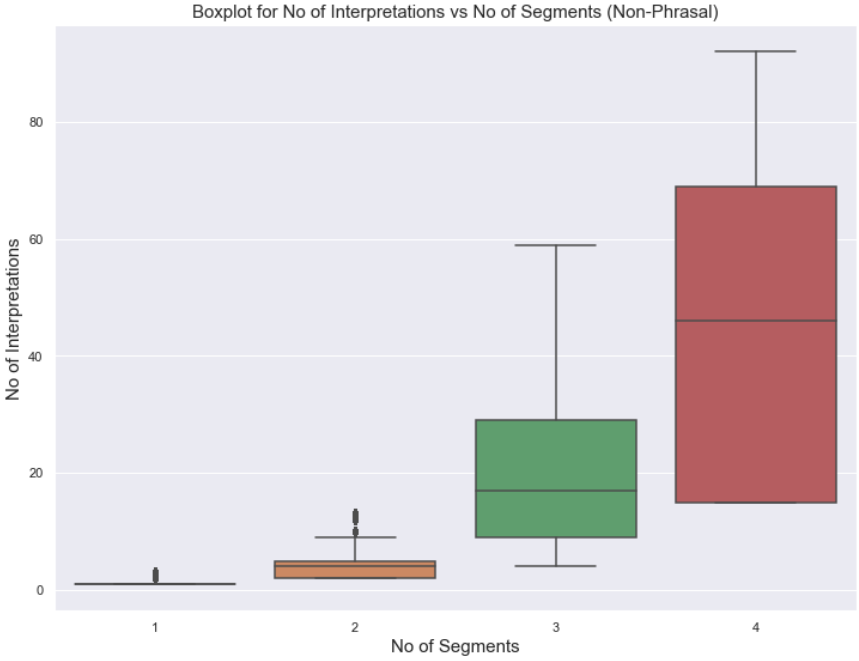


Figure 40: CT7 boxplot of the number of interpretations by the number of segments for non-phrasal items.

We calculate the number of interpretations for each lexical item. As described in Section 5.1.5, this is the number of ways an observation can be interpreted (not the number of distinct interpretations). Each segmentation alternative and lexical alternatives for each segment are taken into account to enumerate the different paths that arrive at a valid interpretation.

The number of interpretations is naturally expected to increase with the number of segments. A larger number of segments ensures not only a larger number of segmentation alternatives, but also leaves more room for polysemy and homonymy on the lexical selection layer. Figure 40 demonstrates the relationship between the number of segments and the number of interpretations. Only non-phrasal lexical items are considered in this figure, due to our focus on morphology.

Number of segments and number of interpretations are only calculated for free forms, as the hearer is not expected to encounter bound forms in isolation. We believe morphological knowledge is implicit in the multitude of ways free forms can be interpreted. Since an unstructured representation of morphological knowledge must track the salience of all these interpretations, it is the main determinant for the C_U metric.

Size of the lexicon grows with novel observations and affix recognition. This is reflected by three metrics: number of lexical items, number of unique surface forms in the lexicon and the number of unique free forms in the lexicon. The first metric gives us the size of the lexicon. The second one gives us the number of homonymy and polysemy relations. The last one gives us the number of items that can be encountered in isolation. In Figure 41, we demonstrate the evolution of these three metrics over the long observation list of CT7. We observe that results from five randomized trials converge around the 2500th observation. (Final lexicon contains 619 items, 568 distinct forms and 510 distinct

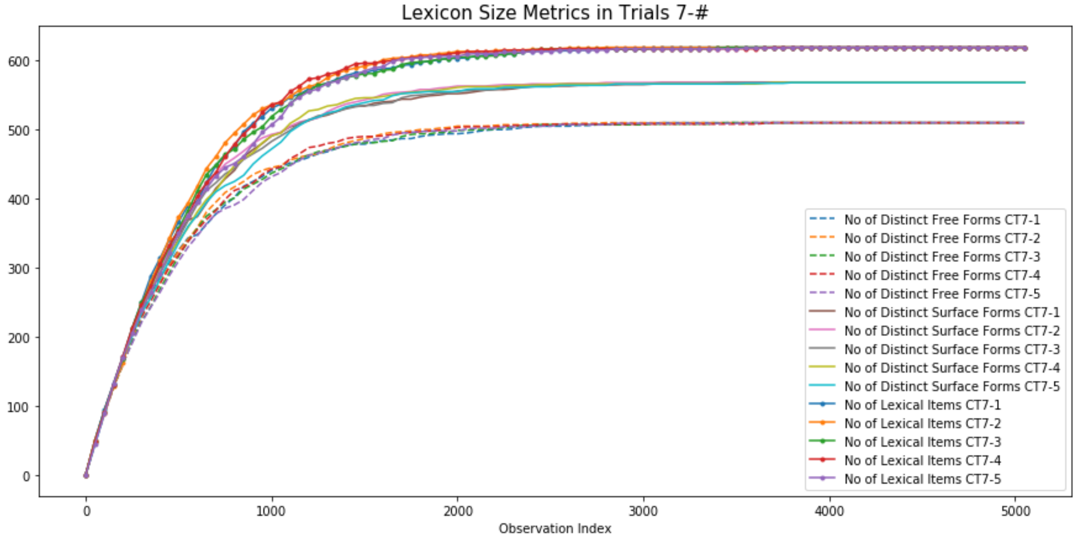


Figure 41: CT7 evolution of metrics regarding lexicon size.

free forms.) As expected, order of exposure does not have a lasting effect on the lexicon, provided that there is a little repetition (on average 5 repetition per observation in this case).

Another set of metrics is calculated based on the lexicon: the total number of interpretations and the total number of segmentation alternatives. The number of interpretations given in Figure 40 were calculated based on the final lexicon (after the last observation) and reported on a per-lexical-item basis. This time, we calculate the total number of interpretations after every 50th observation and plot its evolution in Figure 42. The total number of segmentation alternatives is calculated and visualized in the same way.

We observe that the total number of interpretations grows slowly at first, before affix recognition gains pace. This initial stage is followed by a period of rapid growth in the number of interpretations, which ends when all possible affixes have been recognized. The total number of segmentation alternatives does not increase as much, because it is not affected by the existence of lexical alternatives. This time, results from randomized trials converge around the 3500th observation for both metrics. (The total number of interpretations reaches 4144, while the total number of segmentation alternatives reaches 1185.)

Once we have the above metrics on hand, it is quite simple to calculate C_U and C_F. We calculate these two metrics after every 50th observation and plot the results. Figure 43 is very similar to Figure 42 due to the dominance of the number of interpretations and the number of segmentation alternatives in the calculations. We observe that an unstructured representation requires a much larger number of independent parameters per lexical item, compared to the factorized representation with CdS. Results from randomized trials again converge around the 3500th observation for both metrics.

Another important metric is the average probability of using the retrieval path for interpreting a lexical item. Retrieval and decomposition are the two main paths through the segmentation layer. As the lexicon grows, the average number of segmentation alternatives per observation increases, but the number of ways retrieval can be carried out remains constant at 1. As a result, we expect decomposition alter-

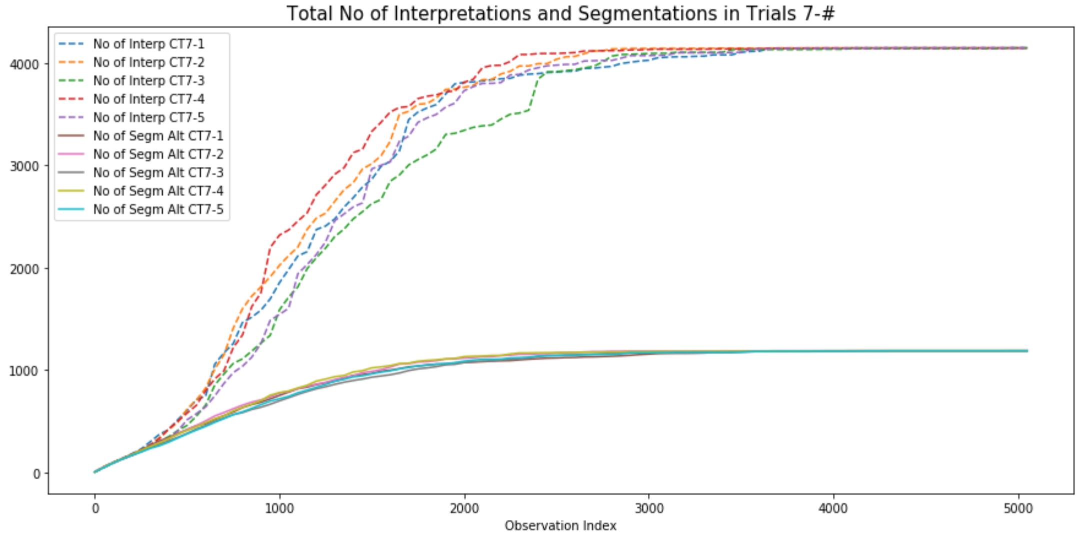


Figure 42: CT7 evolution of the total number of interpretations vs the total number of segmentations.

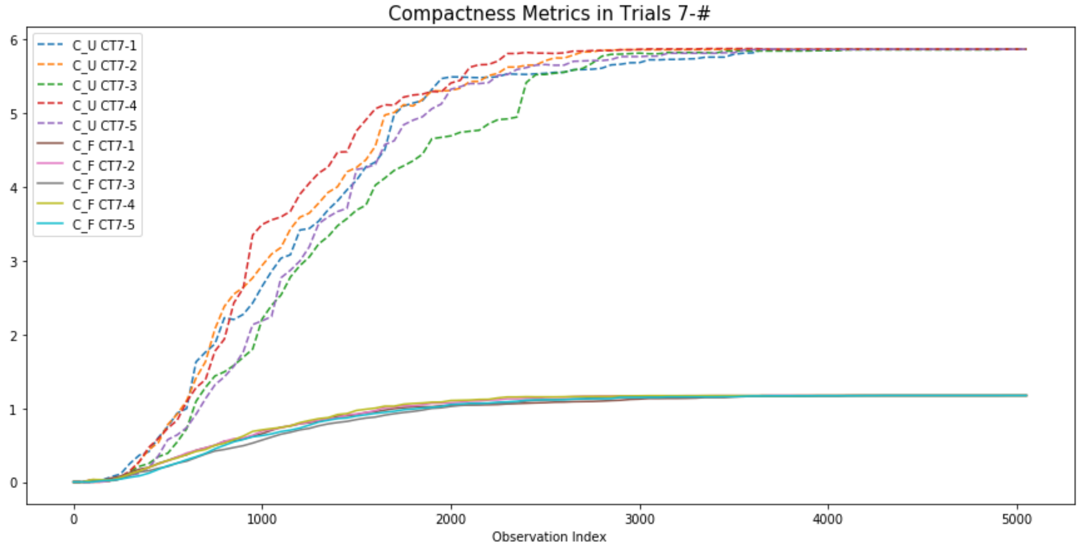


Figure 43: CT7 evolution of compactness metrics.

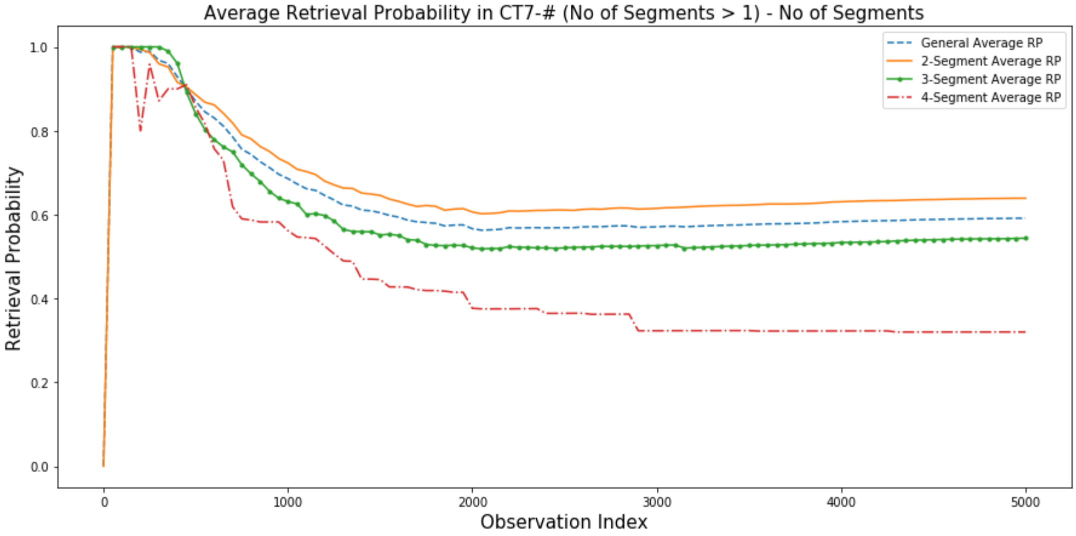


Figure 44: CT7 evolution of average retrieval probabilities grouped by number of segments.

natives to chip away at the retrieval probability. Figure 44 shows that the average retrieval probability evolves in line with this prediction. We only consider lexical items with multiple segments, because the retrieval probability is automatically 1 for single-segment items.

However, around the 2500th observation, after which virtually no new lexical items are learned, this trend reverses. In the second half of the graph, we observe a slow but steady increase in the average retrieval probability. After the effect of discovering new segmentation alternatives vanishes, BOR takes effect and gradually develops a preference towards the less flexible alternative. In most cases, retrieval is the less flexible alternative.

The effect of BOR is present in all stages of the trial, but it remains hidden until the lexicon matures. First, BOR is weaker when there are few segmentation alternatives. Second, the increasing number of segmentation alternatives in the growth phase strongly drives down the retrieval probability. After convergence, the gradual increase due to BOR is slow, because only 5 additional repetitions are made per the average observation.

We observe that BOR's preference towards retrieval is more pronounced for lexical items with fewer segments. Also, retrieval is especially hard for 4-segment items. The difference is quite large, even compared to 3-segment items.

In Figure 45, we look at the relationship between the final retrieval probability and the number of interpretations for each lexical item. As expected, a larger number of interpretations predict a lower retrieval probability, because the former implies a larger number of segmentation and lexical alternatives.

When we break the data down in terms of phrasal vs. non-phrasal distinction, we observe that for non-phrasal items retrieval probability decreases less quickly with the number of interpretations. This might be because for non-phrasal items, the number of lexical alternatives is higher (due to affix

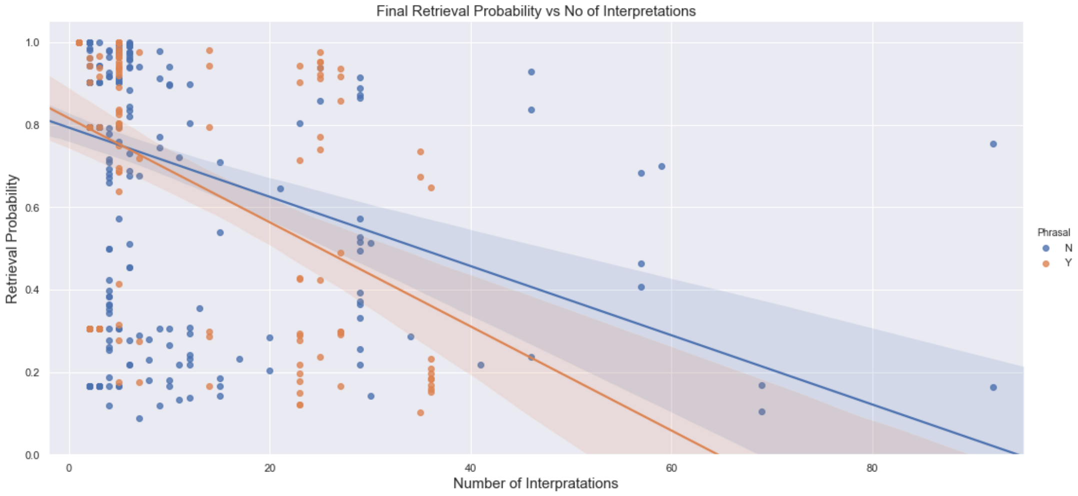


Figure 45: CT7 final retrieval probabilities by the number of interpretations, grouped by items' being phrasal/non-phrasal.

homonymy and polysemy) compared to phrasal items. As a result, BOR might have a stronger effect on non-phrasal items, pushing their retrieval probability a little higher.

5.4 Discussion

Having reviewed the psycholinguistics literature in Chapter 2, we have argued that the traditional approaches to understanding morphology processing is not sufficient. Traditional approaches mostly deal with lexical selection and derivation, while the "correct" segmentation is assumed to be an input of the model. We proposed a new structure, CdS, to relax this assumption and introduce a segmentation layer to the morphological structure. In this chapter, we devised a BBN representation for CdS and presented an implementation. We demonstrated the adequacy of this framework to the task at hand.

We also argued in Chapter 2 that learning must be driven by the recognition of smaller segments inside larger expressions. This is not only an alternative approach to the data, there are cases where learning individual morphemes simply is not enough for understanding an expression. Constructions are not simply the sum of their constituent morphemes. They may come with very specific selection criteria. The trials presented in this chapter constitute a proof-of-concept for our claims.

Core trials demonstrate that for complex words, there is a larger tendency towards the retrieval path. This is due to BOR developing a preference towards less flexible hypotheses. This observation coincides with similar observations in the psycholinguistics literature.

Pilot trials demonstrate that the problem of constructing a BBN for a long sequence of morphemes, as well as carrying out Bayesian inference remains tractable even with a lexicon of non-trivial size. The effect of polysemy and synonymy is countered by categorial restrictions during the segmentation stage.

Overall, our investigations based on a BBN implementation of CdS can be considered a good start for a novel approach towards morphological processing.

CHAPTER 6

DISCUSSION AND FUTURE RESEARCH

In this thesis, our aim has been to put forward a comprehensive examination of Turkish derivational morphology. We have drawn knowledge and insights from many different areas of research including psycholinguistics, computational linguistics, theoretical linguistics, diachronic linguistics, distributional semantics, categorial grammar and Bayesian statistics. We believe the nature of cognitive science forces one to cover so many different areas. We also believe that we succeeded in covering the ground necessary for truly understanding DM, at least in the case of Turkish.

In each chapter, we picked an area of research, reviewed the literature, analyzed the data, and tried to come up with insights regarding Turkish DM. These insights provided the basis for further investigation in other areas, which revealed further insights. Collection of these insights culminated in five distinct contributions.

First, we have demonstrated that the a simple morphological structure is not sufficient to explain morphological processing. Lexical ambiguity and derivation ambiguity are not the only kinds of ambiguity present during processing. Segmentation ambiguity must also be taken into account. This is corroborated by evidence from psycholinguistics. We proposed a new structure for morphological processing and called this the Conventionalized Structure (CdS).

Second, we reviewed descriptive grammars of Turkish and found that current classifications of morphemes make it virtually impossible to pursue a rule-based investigation of Turkish DM. This is despite the fact that speakers, including children, can easily be shown to possess an awareness and active use of DM. With insights from Old Turkic, and using semantic (polysemy, synonymy etc.) and thematic relations (agent, patient etc.), we obtained a more organized classification for Turkish DM. This new classification has been the basis for our efforts on semantic representation.

Third, we looked into the distributional semantics of derivational morphemes. The literature on distributional semantics is primarily built on word embeddings, although there are many alternative approaches. Believing that these alternatives fail to do justice to the semantic relations between stems, derived forms and affixes; we set out to find a way to estimate affix embeddings. Using simple vector arithmetic, we demonstrated that affix embeddings can be reliably estimated. We used these estimations during affix recognition, as semantic similarity restrictions on stem-lemma pairs.

Fourth, we developed a comprehensive rule set that ensures the consistent semantic representation of a wide variety of lexical items. We followed the Categorial Grammar framework, with a lot of inspiration from Combinatory Categorial Grammar. We discovered an asymmetry between representations of syntactic and morphological operations, and managed to reflect it in a simple way inside logical forms. We compared our position regarding this asymmetry with Sezer (1991), and showed that our conclusions largely coincide with and constitute an interpretation of his claims.

In the final Chapter, we proposed a model for the Conventionalized Structure, based on the findings in previous chapters. Recognizing the statistical relations between different components of the structure, we adopted the Bayesian Belief Networks (BBN) framework. We demonstrated the adequacy of this framework to the current task and examined how it represents the trade-offs arising from the Conventionalized Structure. We fully implemented the model, along with custom subroutines for the CYK algorithm, segmentation and learning of semantics by latent syntax. Trials with several variation sets demonstrated that the model behaves as expected; it develops preference towards retrieval and decomposition in a similar way to the observations presented in the psycholinguistics literature. We explained how these trial results, while based on small observation lists, relate to real-life outcomes during processing.

This thesis can be considered as a proof-of-concept for a novel approach to morphological processing. The structure that we propose does not replace the traditional morphological structure, but extends it. By integrating a segmentation layer into the picture, CdS promises to shed new light on the decades-old debate on retrieval vs. decomposition. As discussed in Chapter 2, we believe that traditional approaches to morphology fell short in adequately representing the trade-offs of morphology. CdS is a more appropriate structure to study morphology.

The hierarchical structure we adopt in CdS makes it possible to represent even complex morphology in a compact way. By decoupling segmentation and lexical selection layers from each other, we ensure that the number of interactions do not grow exponentially with the number of constituents. CdS works with compact probability tables to encode interactions between alternatives.

We model the competition between segmentation and lexical alternatives statistically. Psycholinguistics literature provides ample evidence to suggest that the competition between linguistic alternatives is indeed statistical. Prominence of an alternative depends on its frequency of use compared to others.

The hypothesis selection scheme we employ based on this representation, Bayesian Occam's Razor (BOR), is also suitable to the task. BOR's predictions such as longer words' being more likely to be retrieved are corroborated by evidence from psycholinguistics.

Future studies may find other ways to extend the morphological structure, or prefer to carry on the effort based on CdS. Our BBN-based model may serve as a concrete basis for further investigation and communication on this subject. Furthermore, predictions of the model may inspire new experimental studies that specifically focus on the interaction between segmentation and lexical layers. In turn, results from these experiments may validate or falsify the predictions of the model, laying the groundwork for new paths of investigation.

We do not make any language-specific assumptions. We limited ourselves to analyzing segmental morphology and adopted an Item-and-Arrangement perspective for simplicity. Other kinds of morphology (and other methods of representation) can be implemented without getting in conflict with the core assumptions of the model.

Several implications follow from the claims made in this thesis. First, we endorse the parallel non-interactive processing point of view in the ancient parallel vs serial processing debate of psycholinguistics. We claim, and CdS requires, that (a) all segmentation and lexical alternatives are processed in parallel (b) segmentation, lexical and derivation layers come one after the other in a non-interactive manner.

Second, we create lexical items for all segments identified on observations. This results in a “maximal” lexicon, containing both free forms and bound forms, as well as entire phrases. We assume that there are only two kinds of linguistic input for the hearer: the sound sequence (which we approximate with written form) and the context. Constituents cannot be directly observed, they must be discovered. We assume common segments to be the basis for this discovery.

In this view, language acquisition is not bottom-up, it is top-down. The child learns whole expressions first, not their constituents. If the sound sequence represents a word, it is associated with an object or a property. If it represents a sentence, it is associated with an event. Constituents in previously observed sound sequences need to be discovered by the child based on common segments.

Third, the “maximal” lexicon includes both regular and irregular forms. In this view, the lexicon is not just a list of irregularities. As a result, we expect lexicalization to precede irregularity. For a derived form to gain an irregular meaning, it must already be frozen in the lexicon.

(167) a. Ordinary meaning: *haçlı* ‘Someone with a Christian cross’

b. Irregular meaning: *haçlı* ‘Crusader’

c. Frozen lexical item: *haçlı* ‘Crusader’

(168) a. Ordinary meaning: *haçlı* ‘Someone with a Christian cross’

b. Lexical item: *haçlı* ‘Someone with a Christian cross’

c. Irregular meaning: *haçlı* ‘Crusader’

d. Lexical item: *haçlı* ‘Crusader’

Both alternative meanings must be stored in the lexicon. In the long-run, one of them gains prominence at the expense of the other, based on new observations. The irregular item becomes frozen after winning in its competition against the regular item.

Fourth, we expect that the more polysemy relations exist for the constituents, the more the retrieval route gains prominence. This is a consequence of the hypothesis selection scheme embedded in BBN, Bayesian Occam’s Razor (BOR), When multiple lexical alternatives exist for the same surface form, the number of ways to interpret the observation increases. BOR prefers simpler, inflexible hypotheses, and punishes polysemy. This finding is compatible with psycholinguists’ observation that complex words are more likely to be retrieved.

Fifth, we expect the tendency towards retrieval to be stronger for DM, compared to IM. We have shown that polysemy is much more prevalent in DM. This is also compatible with empirical facts from psycholinguistics.

The proposed approach comes with several limitations. Some of these limitations are due to the scope of this thesis and could be overcome with the help of extensions to our model. Some others are due to the assumptions and claims we make along the way; therefore, it may be tricky to find ways around them.

First, we explicitly restrict ourselves to segmental morphology. With segmentation taking such a central stage, it may not be trivial to extend the model to cover non-segmental morphology. Relegating non-segmental morphology entirely to the lexicon is one alternative way forward, but this approach would need to be carefully justified. Otherwise, we must find ways to represent non-segmental morphological processes algorithmically and incorporate these processes into the segmentation layer. Until such an extension, CdS and the accompanying model should be applicable for languages which lend themselves well to an Item-and-Arrangement analysis.

Second, we only study morphological processing from a comprehension point of view. With comprehension in mind, we take surface form and sequence as given, and carry out segmentation to identify constituents. With production in mind, one would have to generate surface form and sequence from the constituents. While similar structures can be expected to underlie both processes, an entirely different mechanism would be necessary to simulate production. It is not straightforward to modify the current model to make predictions regarding production. However, our assumptions hold in a production setting. For instance, we can easily explain how a speaker may invent novel forms derived from existing morphemes. Even if an affix can be observed on just a few distinct forms (perhaps LT is lower for some speakers), speakers may pick up the common meaning and use the affix productively. *çaysamak* 'to crave tea' is a nice example of this. Such cases demonstrate that derivation is regular to a large extent, also from the perspective of production.

Third, we ignore the interaction between different layers of processing. We adopt a holistic view of CdS during hypothesis selection, but the processing mechanism takes place strictly following a sequence from segmentation to lexical selection to derivation. If the interaction between different layers is not negligible, this methodology would fall short in explaining the whole picture. CdS can be extended by the introduction of new dependence relations between pairs of nodes that we assumed independent. The BBN framework allows this.

Acknowledging these limitations, we can identify several paths for future research. First, we can look for ways to modify the segmentation layer to accommodate non-segmental morphology. This could be made possible by representing non-segmental processes as a functions over stems. In that case the adjacency assumption would have to be abandoned. Derivation and affix recognition mechanisms would need to be modified as well.

Second, we could adopt a production point of view to complement the current analysis. Such a study would still use CdS and the accompanying model, but run the network in reverse. In other words, instead of receiving form as input, the algorithm would receive a logical form. The segmentation problem would not be an issue in production, the sequencing problem would take its place.

Third, we could use contextual embeddings for our investigation of distributional semantics. Contextual embeddings produce more accurate vectors, since they are able to differentiate between lexical items with homonymy and polysemy relations. We did not attempt this due to time limitation.

Bibliography

O. Abend, T. Kwiatkowski, N. J. Smith, S. Goldwater, and M. Steedman. Bootstrapping language acquisition. *Cognition*, 164:116–143, 2017.

F. Akkuş. Suspended affixation with derivational suffixes and lexical integrity. In *Mediterranean Morphology Meetings*, volume 10, pages 1–15, 2016.

A. Aksu-Koç and D. I. Slobin. The acquisition of Turkish. In *The crosslinguistic study of language acquisition*, pages 839–878. Psychology Press, 2017.

A. Aksu-Koç, F. N. Ketrez, K. Laalo, and B. Pfeiler. Agglutinating languages: Turkish, Finnish, and Yucatec Maya. In *Typological perspectives on the acquisition of noun and verb morphology*, pages 47–58. Antwerp: Antwerp University, 2007.

R. Al-Rfou, B. Perozzi, and S. Skiena. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*, 2013.

S. Alibekiroğlu. Türkiye Türkçesinde ikili biçim birimler [binary format units in Turkish]. *TÜRÜK Uluslararası Dil, Edebiyat ve Halkbilimi Araştırmaları Dergisi*, 7(17):164–176, 2019.

J. M. Anderson. The non-autonomy of syntax. *Folia Linguistica*, 39(3-4):223–250, 2006. ISSN 0165-4004.

S. Andrews. Morphological influences on lexical access: Lexical or nonlexical effects? *Journal of Memory and Language*, 25(6):726–740, 1986.

J. M. Anglin, G. A. Miller, and P. C. Wakefield. Vocabulary development: A morphological analysis. *Monographs of the society for research in child development*, pages i–186, 1993.

M. Aronoff. *Morphology by itself: Stems and inflectional classes*. MIT press, 1994.

M. Aronoff and K. Fudeman. *What is morphology?* John Wiley & Sons, 2022.

S. Arora, A. May, J. Zhang, and C. Ré. Contextual embeddings: When are they worth it? *arXiv preprint arXiv:2005.09117*, 2020.

Ö. Aslan, S. Günal, and B. T. Dinçer. A computational morphological lexicon for Turkish: Trlex. *Lingua*, 206:21–34, 2018.

E. Avcu. Nouns-first, verbs-first and computationally-easier first: A preliminary design to test the order of acquisition. Master's thesis, Middle East Technical University, 2014.

E. Aydın. *Yenisey Yazıtları*. Kömen Yayınları, 2015.

R. H. Baayen. *Word frequency distributions*, volume 18. Springer Science & Business Media, 2001.

R. H. Baayen and R. Schreuder. Towards a psycholinguistics computational model for morphological parsing. *Philosophical Transactions of the Royal Society A*, 358:1–13, 2000.

R. H. Baayen, P. Milin, D. F. Đurđević, P. Hendrix, and M. Marelli. An amorphous model for morphological processing in visual comprehension based on naive discriminative learning. *Psychological review*, 118(3):438, 2011.

M. Baker. *Incorporation: A Theory of Grammatical Function Changing*. Chicago University Press, 1988.

M. Baker. On agreement and its relationship to case: Some generative ideas and results. *Lingua*, 130: 14–32, 2013.

R. Bamler and S. Mandt. Dynamic word embeddings. In *International Conference on Machine Learning*, pages 380–389. PMLR, 2017.

C. Başdaş. Türkçede üçüncü şahıs iyelik eki ve zamir N'si [Third person possessive and pronominal N in Turkish]. *The Journal of Academic Social Science Studies*, 30:147–161, 2014.

R. Bertram, M. Laine, and M. M. Virkkala. The role of derivational morphology in vocabulary acquisition: Get by with a little help from my morpheme friends. *Scandinavian Journal of Psychology*, 41(4):287–296, 2000.

T. Blanchard, T. Lombrozo, and S. Nichols. Bayesian Occam's razor is a razor of the people. *Cognitive Science*, 42(4):1345–1359, 2018.

J. P. Blevins. *Word and paradigm morphology*. Oxford University Press, 2016.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

G. Booij. Inflection and derivation. *Morphology*, 17:360–369, 2000.

G. Booij. Lexical integrity as a formal universal: A constructionist view. In *Universals of language today*, pages 83–100. Springer, 2009.

G. Booij. *Construction Morphology*. Oxford University Press, 2010.

J. Botha and P. Blunsom. Compositional morphology for word representations and language modelling. In *International Conference on Machine Learning*, pages 1899–1907. PMLR, 2014.

S. R. Bowman, J. Gauthier, A. Rastogi, R. Gupta, C. D. Manning, and C. Potts. A fast unified model for parsing and sentence understanding. *arXiv preprint arXiv:1603.06021*, 2016.

C. Bozşahin. Lexical integrity and type dependence of language. Technical report, METU, 2007.

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

C. Bozşahin. *Combinatory Linguistics*. de Gruyter Mouton, 2012.

C. Bozşahin. *CCGlab Manual*, 2017. URL `http://bozsahin.github.io/ccglab`.

C. Bozşahin. Some binary concepts in Turkish morphology. Cem Bozşahin, 2018.

A. Buran. Türkçede kelimelerin ekleşmesi ve eklerin kökeni. In *3. Uluslararası Türk Dili Kurultayı Bildirileri*, pages 207–214, 1996.

C. Burani and A. Caramazza. Representation and processing of derived words. *Language and cognitive processes*, 2(3-4):217–227, 1987.

J. L. Bybee. *Morphology: A study of the relation between meaning and form*, volume 9. John Benjamins Publishing, 1985.

B. Can and S. Manandhar. Tree structured dirichlet processes for hierarchical morphological segmentation. *Computational Linguistics*, 44(2):349–374, 2018.

B. Can, H. Aleçakır, S. Manandhar, and C. Bozşahin. Joint learning of morphology and syntax with cross-level contextual information flow. *Natural Language Engineering*, 28(6):763–795, 2022.

K. Cao and M. Rei. A joint model for word embedding and word morphology. *arXiv preprint arXiv:1606.02601*, 2016.

A. Caramazza, A. Laudanna, and C. Romani. Lexical access and inflectional morphology. *Cognition*, 28(3):297–332, 1988.

A. Carstairs-McCarthy. *The Evolution of Morphology*. Oxford University Press, 2010.

N. Chater, J. B. Tenenbaum, and A. Yuille. Probabilistic models of cognition: Conceptual foundations. *Trends in cognitive sciences*, 10(7):287–291, 2006.

S. Choi, T. Kim, J. Seol, and S. goo Lee. A syllable-based technique for word embeddings of Korean words. *arXiv preprint arXiv:1708.01766*, 2017.

N. Chomsky. *Essays on Form and Interpretation*. Amsterdam: North-Holland, 1977.

N. Chomsky. *Lectures on government and binding: The Pisa lectures*. Number 9 in Studies in generative grammar. Walter de Gruyter, 1993.

E. V. Clark. Acquisition of derivational morphology. In R. Lieber and P. Štekauer, editors, *The Oxford Handbook of Derivational Morphology*, page 424–439. Oxford University Press Oxford, 2014.

E. V. Clark. Morphology in language acquisition. *The Handbook of Morphology*, pages 374–389, 2017.

S. Clark and J. R. Curran. Log-linear models for wide-coverage CCG parsing. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 97–104, 2003.

S. Clark and J. R. Curran. Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552, 2007.

R. Cotterell and H. Schütze. Morphological word embeddings. *arXiv preprint arXiv:1907.02423*, 2019.

E. A. Cowper. *A concise introduction to syntactic theory: The government-binding approach*. University of Chicago Press, 1992.

M. Creutz and K. Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(1):1–34, 2007.

Q. Cui, B. Gao, J. Bian, S. Qiu, and T.-Y. Liu. Knet: A general framework for learning word embedding using morphological knowledge. *ACM Transactions on Information Systems (TOIS)*, 34(1):1–25, 2015.

A. Cutler and D. Norris. The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14(1):113, 1988.

K. Demirci. Dilbilgiselleşme üzerine bir inceleme. *bilig*, 45:131–146, 2008.

D. Dowty. *Word Meaning and Montague Grammar*. Dordrecht: Reidel, 1979.

D. Dowty. Quantification and the lexicon: A reply to Fodor and Fodor. *The scope of lexical rules*, pages 79–106, 1981.

D. Dowty. Thematic proto-roles and argument selection. *Language*, 67(3):547–619, 1991.

W. U. Dressler. Prototypical differences between inflection and derivation. *STUF - Language Typology and Universals*, 42(1):3–10, 1989.

L. G. Duncan, S. Casalis, and P. Colé. Early metalinguistic awareness of derivational morphology: Observations from a comparison of english and french. *Applied Psycholinguistics*, 30(3):405–440, 2009.

O. Durmuş. -(y)arak zarf-fiil ekinin kökeni üzerine. *Türkbilig*, 23:19–60, 2012.

Ö. Ekmekçi. Creativity in the language acquisition process. *Studies on Modern Turkish*, pages 203–210, 1987.

M. Enç. Topic switching and pronominal subjects in Turkish. In *Studies in Turkish Linguistics*. John Benjamins, 1986.

M. Erdal. *Old Turkic Word Formation*. Turcologica 9. Wiesbaden: Harrassowitz, 1991.

M. Erdal. Clitics in Turkish. In A. Göksel and C. Kerslake, editors, *Studies on Turkish and Turkic Languages*, 2000.

M. Erdal. *A Grammar of Old Turkic*, volume 3. Leiden Brill, 2004.

M. Ergin. *Türk Dil Bilgisi*. Bayrak, 2009.

E. Erguvanlı-Taylan, editor. *The Verb in Turkish*. Amsterdam: John Benjamins, 2001.

J. A. Erickson. On the origin of the directive case in Turkic. *Acta Orientalia*, 55(4):403–411, 2002.

İ. Z. Eyüboğlu. *Türk Dilinin Etimoloji Sözlüğü*. Say Yayınları, 2017.

M. C. Frank, N. D. Goodman, and J. B. Tenenbaum. Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, 20(5):578–585, 2009.

U. H. Frauenfelder and R. Schreuder. Constraining psycholinguistic models of morphological processing and representation: The role of productivity. In *Yearbook of Morphology 1991*, pages 165–183. Springer, 1992.

K. Fukushima. Compositionality, lexical integrity, and agglutinative morphology. *Language Sciences*, 51:67–85, 2015.

É. Gaussier. Unsupervised learning of derivational morphology from inflectional lexicons. In *Unsupervised Learning in Natural Language Processing*, 1999.

L. George and J. Kornfilt. Finiteness and boundedness in Turkish. *Binding and Filtering*, 105:127, 1981.

T. Givon and D. Slobin. Function, structure and language acquisition. *The Crosslinguistic Study of Language Acquisition*, 2:1005–1028, 1985.

A. Gladkova, A. Drozd, and S. Matsuoka. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the NAACL Student Research Workshop*, pages 8–15, San Diego, California, June 2016. Association for Computational Linguistics.

H. Gökdayı and T. Sebzecioğlu. Türkiye Türkçesinde {-CA} biçimbiriminin türleri [Types of {-CA} morpheme in Modern Turkish]. *bilig*, 58(1):147–172, 2011.

A. Göksel and C. Kerslake. *Turkish: A comprehensive grammar*. London: Routledge, 2005.

J. Goldsmith. Unsupervised learning of the morphology of a natural language. *Computational linguistics*, 27(2):153–198, 2001.

S. Goldwater and T. Griffiths. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 744–751, 2007.

L. M. Gonnerman. *Morphology and the lexicon: Exploring the semantics-phonology interface*. University of Southern California, 1999.

J. Good and C. L. Alan. Morphosyntax of two Turkish subject pronominal paradigms. In *Proceedings of the North East Linguistic Society 30*, volume 2, 2005.

A. Gopnik and L. Schulz. Mechanisms of theory formation in young children. *Trends in cognitive sciences*, 8(8):371–377, 2004.

A. Gopnik, C. Glymour, D. M. Sobel, L. E. Schulz, T. Kushnir, and D. Danks. A theory of causal learning in children: causal maps and Bayes nets. *Psychological review*, 111(1):3, 2004.

S. Gouws and A. Søgaard. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, Denver, Colorado, May–June 2015. Association for Computational Linguistics.

E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. *arXiv preprint arXiv:1802.06893*, 2018.

J. Grimshaw. *Argument Structure*. MIT Press, 1990.

S. Gündoğdu. Il/-In morphology in Turkish: Implications for U-syncretism. *Proceedings of PICGL4*, pages 85–103, 2016.

M. Güven. Türkiye Türkçesinde iyelik gruplarının dil bilgisel düzeni ve iyelik gruplarındaki dil bilgisel aykırılıklar. *International Journal of Languages' Education and Teaching*, 1(4):658–670, 2021.

K. Hale and S. J. Keyser. *Prolegomenon to a Theory of Argument Structure*. MIT Press, 2002.

M. Halle and A. Marantz. Some key features of distributed morphology. *MIT working papers in linguistics*, 21(275):88, 1994.

M. Halle, A. Marantz, K. Hale, and S. J. Keyser. Distributed morphology and the pieces of inflection. *1993*, pages 111–176, 1993.

W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv:1605.09096*, 2016.

H. Hammarström and L. Borin. Unsupervised learning of morphology. *Computational Linguistics*, 37 (2):309–350, 2011.

J. Hankamer. Morphological parsing and the lexicon. *Lexical Representation and Process*, pages 392–408, 1989.

Z. S. Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

M. Haspelmath. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia linguistica*, 45(1):31–80, 2011.

D. Heckerman. A tutorial on learning with Bayesian networks. In D. E. Holmes and L. C. Jain, editors, *Innovations in Bayesian Networks: Theory and Applications*, pages 33–82. Springer Berlin Heidelberg, 2008.

C. F. Hockett. Two models of grammatical description. *Word*, 10(2-3):210–234, 1954.

Y. Ichisugi and N. Takahashi. A formal model of the mechanism of semantic analysis in the brain. In *Biologically Inspired Cognitive Architectures Meeting*, pages 128–137. Springer, 2018.

N. İlhan and M. G. Öz. Türkçede kelimelerin ekleşmesiyle ortaya çıkan ekler. *The Journal of Academic Social Science Studies*, 75:149–162, 2019.

G. Jurdzinski. Word embeddings for morphologically complex languages. *Schedae Informaticae*, 25, 2017.

B. Kabak. Turkish suspended affixation. *Linguistics*, 45(2):311–347, 2007.

M. Kaşgarlı. *Divanü Lugati't-Türk*. Kabalci Yayınevi, 2005.

C. Kennedy. Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, 30:1–45, 2007.

C. Kennedy and L. McNally. Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81(2):345–381, 2005.

F. N. Ketrez. Early verbs and the acquisition of Turkish argument structure. *MA Thesis, Boğaziçi University, İstanbul*, 1999.

F. N. Ketrez and A. Aksu-Koç. The (scarcity of) diminutives in Turkish child language. *Language Acquisition and Language Disorders*, 2007.

Ö. Kılıç and C. Bozşahin. Selection of linker type in emphatic reduplication: Speaker's intuition meets corpus statistics. In *35th Annual Meeting of Cognitive Science Society*, volume 35, 2013.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Z. Korkmaz. Türkçede –acak/-ecek gelecek zaman (futurum) ekinin yapısı üzerine [On the structure of -acak/-ecek future tense affix in Turkish]. *Ankara Üniversitesi Dil ve Tarih Coğrafya Fakültesi Dergisi*, 17(1-2):159–168, 1959.

J. Kornfilt. On copular clitic forms in Turkish. *ZAS Papers in Linguistics*, 6:96–114, 1996.

J. Kornfilt. On the syntax and morphology of relative clauses in Turkish. *Dilbilim Araştırmaları Dergisi*, 8:24–51, 1997.

J. Kornfilt. Revisiting "suspended affixation" and other coordinate mysteries. In L. Brugè, A. Cardinaletti, G. Giusti, N. Munaro, and C. Poletto, editors, *Functional Heads: The Cartography of Syntactic Structures*, volume 7, page 181–196. Oxford University Press, 2012.

M. Koç. Eski Anadolu Türkçesinde /duk/ ekli geçmiş zaman çekimi [perfective inflection by /duk/ in Old Anatolian Turkish]. *Türkbilig*, 13(23):11–18, 2012.

G.-J. M. Kruijff and J. Baldridge. Generalizing dimensionality in combinatory categorial grammar. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 191–197, 2004.

A. Küntay and D. I. Slobin. The acquisition of Turkish as a native language. a research review. *Turkic languages*, 3(2):151–188, 1999.

A. Küntay and D. I. Slobin. Listening to a Turkish mother: Some puzzles for acquisition. In *Social interaction, social context, and language*, pages 283–304. Psychology Press, 2014.

U. C. Kunter and C. Bozşahin. CCG for Turkish finite verb inflection. ISBCS 2018 Presentation, 2018.

U. C. Kunter, G. N. Özdemir, and C. Bozş. Distributional and lexical exploration of semantics of derivational morphology. ISBCS 2020 Presentation, 2020. URL `https://www.youtube.com/watch?v=zziRn2PR4ok`.

M. Kürüm. Türkçede iyelik eklerinin ele alınışı ve çağdaş lehçelerde kullanılışı [Turkish possessive affixes and their usage in modern dialects]. *Mavi Atlas*, pages 76–95, 2015.

P. I. Kuznetsov. Türkiye Türkçesinin morfoetimolojisine dair [On the morphoetymology of Turkish]. *Türk Dili Araştırmaları Yıllığı - Belleten 1995*, pages 193–262, 1997.

T. Kwiatkowksi, L. Zettlemoyer, S. Goldwater, and M. Steedman. Inducing probabilistic CCG grammars from logical form with higher-order unification. In *Proceedings of the 2010 conference on empirical methods in natural language processing*, pages 1223–1233, 2010.

T. Kwiatkowski, L. Zettlemoyer, S. Goldwater, and M. Steedman. Lexical generalization in CCG grammar induction for semantic parsing. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1512–1523, 2011.

B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

A. Laudanna, W. Badecker, and A. Caramazza. Processing inflectional and derivational morphology. *Journal of Memory and Language*, 31(3):333–348, 1992.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 50(2):157–194, 1988.

A. Lazaridou, M. Marelli, R. Zamparelli, and M. Baroni. Compositional-ly derived representations of morphologically complex words in distributional semantics. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1517–1526, 2013.

R. Lieber. *Deconstructing morphology: word formation in syntactic theory*. University of Chicago Press, 1992.

R. Lieber and S. Scalise. The lexical integrity hypothesis in a new theoretical universe. *Lingue e linguaggio*, 6(1):7–32, 2006.

Q. Liu, M. J. Kusner, and P. Blunsom. A survey on contextual embeddings. *arXiv preprint arXiv:2003.07278*, 2020.

S. Liu, P. Bremer, J. J. Thiagarajan, V. Srikumar, B. Wang, Y. Livnat, and V. Pascucci. Visual exploration of semantic relationships in neural word embeddings. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):553–562, 2018.

J. Lyons. *Language and Linguistics*. Cambridge University Press, 1981.

L. Manelis and D. A. Tharp. The processing of affixed words. *Memory & Cognition*, 5(6):690–695, 1977.

A. Marantz. No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon. *University of Pennsylvania working papers in linguistics*, 4(2):14, 1997.

A. Marantz. No escape from morphemes in morphological processing. *Language and cognitive processes*, 28(7):905–916, 2013.

W. Marslen-Wilson, L. K. Tyler, R. Waksler, and L. Older. Morphology and meaning in the English mental lexicon. *Psychological review*, 101(1):3, 1994.

S. Masliyah. Four Turkish suffixes in Iraqi Arabic. *Journal of Semitic Studies*, XLI(2):291–300, 1996.

B. Mayo. *A computational model of derivational morphology*. PhD thesis, Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky, 1999.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

P. H. Miller. Post-lexical cliticization vs. affixation: Coordination criteria. In *Proceedings of the 28th Meeting of the Chicago Linguistic Society*, pages 382–396. CLS Chicago, 1992.

R. Mitkov. *The Oxford handbook of computational linguistics*. Oxford University Press, 2022.

M. Moens and M. Steedman. Temporal ontology and temporal reference. *Computational Linguistics*, 14:15–28, 1988.

F. Moghimifar, A. Rahimi, M. Baktashmotlagh, and X. Li. Learning causal Bayesian networks from text. *arXiv preprint arXiv:2011.13115*, 2020.

T. Musil, J. Vidra, and D. Mareček. Derivational morphological relations in word embeddings. *arXiv preprint arXiv:1906.02510*, 2019.

W. E. Nagy, I.-A. N. Diakidoy, and R. C. Anderson. The acquisition of morphology: Learning the contribution of suffixes to the meanings of derivatives. *Journal of reading Behavior*, 25(2):155–170, 1993.

M. V. Nalbant. -DUK eki ve Divanü Lugati't Türk'te -DUK ekli görülen geçmiş zaman çekimi. *Türkoloji Dergisi*, XV(1):193–203, 2002.

J. W. Ney. The non-existence of autonomous syntax. *Language Sciences*, 4(1):35–54, 1982.

I. Nikolaeva. Altaic. In R. Lieber and P. Štekauer, editors, *The Oxford handbook of derivational morphology*, pages 493–508. Oxford University Press Oxford, 2014.

S. Nişanyan. *Nişanyan Sözlük-Çağdaş Türkçenin etimolojisi [Nişanyan Sözlük-The Etymology of Modern Turkish]*. Liberus, 2021.

O. Nizam. Eski Türkçede çokluk ve topluluk yapıları. Master's thesis, Eskişehir Osmangazi Üniversitesi, 2017.

K. Oflazer, E. Göçmen, and C. Bozşahin. An outline of Turkish morphology. Technical report, Middle East Technical University and Bilkent University, 1995. Re-issued in 2014.

S. Olsen. Delineating derivation and compounding. In R. Lieber and P. Štekauer, editors, *The Oxford handbook of derivational morphology*, pages 26–49. Oxford University Press Oxford, 2014.

M. Öner. Türkçede -prAk zarf fiili. *Modern Türklük Araştırmaları Dergisi*, 4(3):68–73, Sept. 2007.

M. Özmen. *Türkçede -Ken Zarf-Fiili*. TDK Yayınları, 2014.

B. Öztürk. Turkish as a non-pro-drop language. In E. E. Taylan, editor, *The Verb in Turkish*, pages 239–259. John Benjamins Publishing Company, 2001.

B. Öztürk and E. Taylan. Possessive constructions in Turkish. *Lingua*, 182:88–108, 2016.

F. Paoli. Comparative logic as an approach to comparison in natural language. *Journal of Semantics*, 16:67–96, 1999.

C. Paradis. Adjectives and boundedness. *Cognitive Linguistics*, 12(1):47–65, 2001.

T. Parsons. Thematic relations and arguments. *Linguistic Inquiry*, pages 635–662, 1995.

M. Paster. Allomorphy. In R. Lieber and P. Štekauer, editors, *The Oxford handbook of derivational morphology*, pages 219–234. Oxford University Press Oxford, 2014.

J. Pearl. *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann, 1988.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passosand, D. Cournapeauand, M. Brucher, M. Perrot, and E. Duchesnay. Scikitlearn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

A. Perfors, J. B. Tenenbaum, and T. Regier. The learnability of abstract syntactic principles. *Cognition*, 118(3):306–338, 2011.

A. M. Peters. Language segmentation: Operating principles for the perception and analysis of language. In *The crosslinguistic study of language acquisition*, pages 81–116. Psychology Press, 2013.

M. E. Peters, M. Neumann, L. Zettlemoyer, and W. tau Yih. Dissecting contextual word embeddings: Architecture and representation. *arXiv preprint arXiv:1808.08949*, 2018.

S. T. Piantadosi. *Learning and the Language of Thought*. PhD thesis, Massachusetts Institute of Technology, 2011.

S. T. Piantadosi, N. D. Goodman, B. A. Ellis, and J. Tenenbaum. A Bayesian model of the acquisition of compositional semantics. In *Proceedings of the 30th annual conference of the cognitive science society*, 2008.

W. V. O. Quine. Word and object. new edition. *Cambridge, Massachusetts: The*, 1960.

F. Rainer. Polysemy in derivation. In R. Lieber and P. Štekauer, editors, *The Oxford handbook of derivational morphology*, pages 338–353. Oxford University Press Oxford, 2014.

S. Ravi and K. Knight. Minimized models for unsupervised part-of-speech tagging. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 504–512, 2009.

B. Roark and R. Sproat. *Computational Approaches to Morphology and Syntax*. Oxford University Press, 2007.

R. Rosa and Z. Žabokrtský. Attempting to separate inflection and derivation using vector space representations. In *Proceedings of the 2nd Int. Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019)*, page 61–70, 2019.

T. Ruokolainen, O. Kohonen, K. Sirts, S.-A. Grönroos, M. Kurimo, and S. Virpioja. A comparative study of minimally supervised morphological segmentation. *Computational Linguistics*, 42(1):91–120, 2016.

J. R. Saffran, R. N. Aslin, and E. L. Newport. Statistical learning by 8-month-old infants. *Science*, 274 (5294):1926–1928, 1996.

G. S. Sagidolda. Declension system of the Turkic languages: Historical development of case endings. *Russian Federation Bulletin of the Kalmyk Institute for Humanities of the Russian Academy of Sciences*, 23(1):166–173, 2016.

G. G. Şahin. *Building of Turkish PropBank and Semantic Role Labeling of Turkish*. PhD thesis, Istanbul Technical University, 2018.

G. G. Şahin and M. Steedman. Character-level models versus morphology in semantic role labeling. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 386–396, 2018.

S. Scalise. Inflection and derivation. *Linguistics*, 26(4):561–582, 1988.

S. Scalise. Generative morphology. In *Generative Morphology*. De Gruyter Mouton, 2011.

A. M. Şçerbak. Türkçe morfoloji tarihini inceleme meselesi uzerine. *Türk Dili Araştırmaları Yıllığı - Belleten*, 37:317–321, 1989.

J. Schreiber. Pomegranate: Fast and flexible probabilistic modeling in Python. *Journal of Machine Learning Research*, 18(164):1–6, 2018.

R. Schreuder and R. H. Baayen. Modeling morphological processing. *Morphological aspects of language processing*, 2:257–294, 1995.

M. S. Seidenberg and L. M. Gonnerman. Explaining derivational morphology as the convergence of codes. *Trends in cognitive sciences*, 4(9):353–361, 2000.

F. E. Sezer. *Issues in Turkish Syntax*. PhD thesis, Harvard University, 1991.

D. Siegel. *Topics in English Morphology*. PhD thesis, MIT, 1974.

D. I. Slobin. Universal and particular in the acquisition of language. *Language acquisition: The state of the art*, 57, 1982.

P. Smolensky. Grammar-based connectionist approaches to language. *Cognitive Science*, 23(4):589–613, 1999.

H. Sofu. Acquisition of reduplication in Turkish. *Studies on Reduplication*, page 493, 2005.

M. Steedman. *Surface Structure and Interpretation*. MIT Press, Cambridge, MA, 1996.

M. Steedman. Information structure and the syntax-phonology interface. *Linguistic Inquiry*, 31:649–689, 2000.

M. Steedman and J. Baldridge. Combinatory categorial grammar. *Non-Transformational Syntax: Formal and Explicit Models of Grammar. Wiley-Blackwell*, pages 181–224, 2011.

P. Štekauer. Derivational paradigms. In R. Lieber and P. Štekauer, editors, *The Oxford handbook of derivational morphology*, pages 354–369. Oxford University Press Oxford, 2014.

R. Stockwell and D. Minkova. *English words: History and structure*. Cambridge University Press, 2001.

G. T. Stump. A paradigm-based theory of morphosemantic mismatches. *Language*, pages 675–725, 1991.

G. T. Stump. Position classes and morphological theory. In G. Booij and J. van Marle, editors, *Yearbook of Morphology 1992*, pages 129–180. Springer Netherlands, Dordrecht, 1993.

G. T. Stump. Inflection. In A. Spencer and A. M. Zwicky, editors, *The handbook of morphology*, pages 11–43. Wiley Online Library, 2017.

Y. Sugioka. *Interaction of Derivational Morphology and Syntax in Japanese and English*. Routledge, 1986.

M. Swadesh. Nootka internal syntax. *International Journal of American Linguistics*, 9(2/4):77–102, 1938.

M. Taft and K. I. Forster. Lexical storage and retrieval of prefixed words. *Journal of verbal learning and verbal behavior*, 14(6):638–647, 1975.

M. Taft and X. Zhu. The representation of bound morphemes in the lexicon: A Chinese study. *Morphological aspects of language processing*, pages 293–316, 1995.

N. Takahashi and Y. Ichisugi. Restricted quasi Bayesian networks as a prototyping tool for computational models of individual cortical areas. In *Advanced Methodologies for Bayesian Networks*, pages 188–199. PMLR, 2017.

E. E. Taylan. Pronominal versus zero representation of anaphora in Turkish. In *Studies in Turkish linguistics*, page 209. John Benjamins, 1986.

TDK. Güncel Türkçe Sözlük, 2019. URL `https://sozluk.gov.tr/`.

T. Tekin. *A Grammar of Orkhon Turkic*. Research Center for the Language Sciences, Indiana University, 1968.

T. Tekin. On the Old Turkic verbal noun suffix -dOk. *Türk Dilleri Araştırmaları*, 7:5–12, 1997.

T. Tekin. Türkçe ilgi hali ekinin kökeni üzerine [On the etymology of the Turkish genitive]. *Dil Araştırmaları*, 13:157–162, 2013.

T. Tekin. *Orhon Türkçesi Grameri [A Grammar of Orkhon Turkic]*. Türk Dil Kurumu Yayınları, 2016.

P. Ten Hacken. Delineating derivation and inflection. In R. Lieber and P. Štekauer, editors, *The Oxford handbook of derivational morphology*, pages 10–25. Oxford University Press Oxford, 2014.

J. B. Tenenbaum. *A Bayesian Framework for Concept Learning*. PhD thesis, Massachusetts Institute of Technology, 1999.

J. B. Tenenbaum, C. Kemp, T. L. Griffiths, and N. D. Goodman. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285, 2011.

A. Tyler and W. Nagy. The acquisition of English derivational morphology. *Journal of memory and language*, 28(6):649–667, 1989.

O. Ünal. Türkçe üçüncü tekil şahıs iyelik ekinin ve pronominal N sesinin kokeni üzerine [On the etymology of Turkish third person singular possessive affix and pronominal N]. *Türkiyat Araştırmaları Enstitüsü Dergisi*, 64:47–69, 2019.

A. Üstün. Probabilistic learning of Turkish morphosemantics by latent syntax. Master's thesis, Middle East Technical University, 2017.

A. Üstün and B. Can. Unsupervised morphological segmentation using neural word embeddings. In *International conference on statistical language and speech processing*, pages 43–53. Springer, 2016.

A. Üstün, M. Kurfalı, and B. Can. Characters or morphemes: How to represent words? In *Proceedings of the 3rd Workshop on Representation Learning for NLP*, page 144–153, 2018.

L. Van der Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

A. Wang, T. Kwiatkowski, and L. Zettlemoyer. Morpho-syntactic lexical generalization for CCG semantic parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1284–1295, 2014.

A. Wedel. Turkish emphatic reduplication. Technical report, UC Santa Cruz: Working Papers, 1999.

E. Williams. Dumping lexicalism. *The Oxford handbook of linguistic interfaces*, pages 353–382, 2007.

F. Xu and J. B. Tenenbaum. Word learning as Bayesian inference. *Psychological review*, 114(2):245, 2007.

P. Yavuzarslan. Türk dilinde kişi eklerinin tarihsel gelişimi ve değişimi. *International Congress of Asian and North African Studies*, 38:1953–1966, 2011.

M. L. Yener. Türkçede ek eylemin işlevi: Ad tümcelerini yeniden düşünmek. *RumeliDE Dil ve Edebiyat Araştırmaları Dergisi*, pages 134–152, 2018.

B. Yet. Personal communication, 2021.

G. Yıldırım. Türkçede benzerlik, eşitlik ifade eden isimden isim yapma ekleri. Master's thesis, Gazi Üniversitesi, 2011.

M. Yüceol Özezen. Türkiye Türkçesinde +A ve -A ekli zarflar üzerine. *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 17(2):329–344i, 2008.

M. Yüceol Özezen. Türkçede zarf-fiiller ve zarf-fiillerde yapılaşma süreçleri. *Çukurova Üniversitesi Sosyal Bilimler Enstitüsü Dergisi*, 24(1):79–97, 2018.

D. Yüret and F. Türe. Learning morphological disambiguation rules for Turkish. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 328–334, 2006.

H. Zargayouna, I. Tellier, D. Buscaldi, and T. Charnois. Exploring vector spaces for semantic relations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1814–1823, 2017.

L. Zettlemoyer and M. Collins. Online learning of relaxed CCG grammars for parsing to logical form. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 678–687, 2007.

L. S. Zettlemoyer. *Learning to map sentences to logical form.* PhD thesis, Massachusetts Institute of Technology, 2009.

L. S. Zettlemoyer and M. Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv:1207.1420*, 2012.

Appendix A

CLASSIFICATION OF TURKISH AFFIXES

A few cells are marked to indicate that additional clarifications are available regarding our choices. We list those clarifications below:

(169) Notes on the classification matrix

[1] Slot V-3.3 is occupied by the complex verb.

[2] sA is unique in its ability to occur after one of the other copular markers have already occurred.

[3] Auxiliary is a defective verb that may only receive a restricted set of TAM markers.

[4] We mark recursivity Yes, only if we observed cases of recursion.

[5] Number of polysemous uses reflect the uses we could find so far.

[6] Zeroth position applies directly on the base.

Table 37: Derivational affixes and the number of lemmas they appear in the reannotated TrLex dataset

Affix	Type Count	Affix	Type Count	Affix	Type Count	Affix	Type Count
AAD_CA+	9	JND_IZ	4	NND_CIK	150	NVD_MA	7949
AAD_CAK	2	JND_KI+	3	NND_CIL	61	NVD_MAC+	31
AAD_DAN+	7	JND_LI	2290	NND_CIN+	12	NVD_MAN	23
AAD_KI+	4	JND_LIK	76	NND_DA+	1	NVD_MIK+	14
AAD_LIK+	1	JND_MAN+	1	NND_DAM+	3	NVD_MIS+	4
AAD_SA+	6	JND_MIS+	1	NND_DAS	54	NVD_MUR+	1
AJD_CA	308	JND_SAL	192	NND_DIRIK+	8	NVD_S+	2
AJD_CASINA+	12	JND_SER	22	NND_GA+	1	NVD_SI+	3
AJD_DAN+	13	JND_SI	135	NND_GEN	11	NVD_TI	176
AJD_SA+	1	JND_SIL+	9	NND_GILLER+	182	NVD_V+	3
AND_ADAK+	23	JND_SIZ	1476	NND_IR+	12	NVD_YIS	1149
AND_ADAN+	2	JVD_A+	1	NND_LAR+	149	VJD_AL	41
AND_ADANAK+	1	JVD_ACAN+	4	NND_LI+	181	VJD_AR	12
AND_ARI+	4	JVD_AGAN	13	NND_LIK	2634	VJD_IK+	6
AND_CA	102	JVD_AK	69	NND_MAN+	1	VJD_IMSE	12
AND_CAK	3	JVD_AL+	8	NND_RA+	2	VJD_LA+	31
AND_DA+	34	JVD_AN	15	NND_SAK+	7	VJD_LAS	627
AND_DAN+	69	JVD_ARI+	2	NND_TI+	6	VJD_LAT+	7
AND_DIR+	1	JVD_ASI	3	NVD_A+	6	VJD_SA	16
AND_IN+	9	JVD_DAK+	3	NVD_ACAK	22	VND_A	12
AND_LA	58	JVD_DIK+	8	NVD_AK	72	VND_AR+	7
AND_LARI+	6	JVD_GIN	100	NVD_ALAK+	3	VND_DA	87
AND_LEYIN+	6	JVD_I	31	NVD_ALGA+	6	VND_ET	2
AND_LIK+	1	JVD_ICI	73	NVD_AM+	6	VND_IR+	16
AND_RA+	2	JVD_IK	10	NVD_AMAK	3	VND_KIR+	11
AND_SA+	1	JVD_IR	66	NVD_AN	34	VND_LA	1120
AVD_A+	2	JVD_MA	4	NVD_ANAK	23	VND_LAN	541
AVD_ARAK	3	JVD_MAN+	5	NVD_ASI+	1	VND_LAS+	310
JAD_KI+	12	JVD_MAZ	50	NVD_AY+	7	VND_LAT+	6
JAD_LI+	1	JVD_MIS	29	NVD_C+	20	VND_SA	18
JJD_CA	62	JVD_S+	2	NVD_CA+	16	VND_SIN+	5
JJD_CAK+	2	NAD_LIK+	13	NVD_CAK+	3	VVD_A+	1
JJD_CIK	16	NJD_CA	2	NVD_GA	29	VVD_AKLA	12
JJD_IMSI	29	NJD_KI+	6	NVD_GAC	71	VVD_ALA	20
JJD_LIK+	2	NJD_LIK	2759	NVD_GAN	97	VVD_GA+	2
JJD_MAN+	7	NJD_MAN+	3	NVD_GI	115	VVD_I+	2
JJD_MTRAK	15	NND_A+	2	NVD_GIC	14	VVD_IMSE+	3
JJD_RAK+	10	NND_AC+	11	NVD_I	119	VVD_P+	2
JJD_SIN+	4	NND_AK+	8	NVD_ICI	306	VVD_USTUR	31
JND_CA+	13	NND_ALAK+	3	NVD_IK	218		
JND_CIK	1	NND_AY+	7	NVD_IM	335		
JND_DA+	1	NND_CA	150	NVD_IN+	4		
JND_IL+	47	NND_CAGZ	8	NVD_INC	13		
JND_IMSI	76	NND_CAK	10	NVD_INTI	86		
JND_INCI	22	NND_CI	1832	NVD_IT	6		

Table 38: Classification of VVD and VVI affixes

Group	[4]Recursivity	[5]Polysemy	Deriv. Allomorphy	Semantic Selection	Dictionary entries	Change in Base Arg. Str.	Change in Base POS	Phrasal Scope	[6]Order of Application	Inv. of Order of Appl.	Member of a Paradigm	Suspended Affixation	Required by Syntax	Affix Class	Verdict
VVD AklA / IklA / AlA	No	3+	3	Extensive	12 / 21	No	No	No	V-0	No	No	No	No	Suffix	Definitely DM
VVD A / I	No	None	2	Extensive	1 / 2	No	No	No	V-0	No	No	No	No	Suffix	Definitely DM
VVD p	No	None	None	Extensive	2	No	No	No	V-0	No	No	No	No	Suffix	Definitely DM
VVD ImsA	No	None	None	Extensive	3	No	No	No	V-0	No	No	No	No	Suffix	Definitely DM
VVD IştIr	No	None	None	Extensive	31	No	No	No	V-0	No	No	No	No	Suffix	Definitely DM
Voice (Causative)	Yes	None	3+	Exceptional	1573	Yes	No	No	V-1	No	No	No	No	Suffix	Definitely DM
Voice (Passive / Reflexive)	No	None	2	Exceptional	1389 / 22	Yes	No	No	V-1	Yes	No	No	No	Suffix	Leans DM
Voice (Reciprocal)	No	None	None	Exceptional	168	Yes	No	No	V-1	Yes	No	No	No	Suffix	Leans DM
Negation	No	None	None	None	N/A	No	No	No	V-2 / V-3.4	Yes	No	No	No	Suffix	Leans IM
Gerundium (y)A	No	3+	None	None	N/A	No	Yes	Yes	[1]V-3.1	Yes	No	No	No	Suffix	Leans DM
TAM K-paradigm	No	None	None	None	N/A	No	Yes	Yes	V-4	Yes	Yes	No	Yes (Any V-4)	Suffix	Definitely IM
TAM Z-par. (mAlI)	No	None	None	None	N/A	No	Yes	Yes	V-4	Yes	Yes	No	Yes (Any V-4)	Suffix	Definitely IM
TAM Z-par. (Other)	No	None	None	None	N/A	No	Yes	Yes	V-4	Yes	Yes	No	Yes (Any V-4)	Suffix	Definitely IM
TAM Imp. / Vol.	No	None	None	None	N/A	No	Yes	Yes	V-4	Yes	Yes	No	Yes (Any V-4)	Suffix	Definitely IM
Question M.	No	None	None	None	N/A	N/A	N/A	Yes	V-5 / V-9 / N-7	No	No	Yes	No	Clitic (Free)	Clitic
[3]**Auxiliary**	Yes	None	Yes	None	N/A	N/A	Yes	Yes	V-6.1 / N-6.1	No	No	Yes	No	Clitic (Bound / Free)	Clitic
Copula DI / mIş	No	None	None	None	N/A	No	Yes	Yes	V-6.2 / N-6.2	Yes	Yes	No	Yes (Any V-6.2)	Suffix	Definitely IM
Copula sA	No	None	None	None	N/A	No	Yes	Yes	[2]V-6.2 / N-6.2	Yes	Yes	No	Yes (Any V-6.2)	Suffix	Definitely IM
Agr. K-paradigm	No	None	None	None	N/A	N/A	Yes	Yes	V-7 / N-7	Yes	Yes	No	Yes (Any V-7)	Suffix	Definitely IM
Agr. Z-paradigm	No	None	None	None	N/A	N/A	Yes	Yes	V-7 / N-7	Yes	Yes	Yes	Yes (Any V-7)	Clitic (Bound)	Clitic
Epistem. Cop.	No	None	None	Exceptional	N/A	N/A	No	Yes	V-8 / N-8	No	No	Yes	No	Clitic (Bound)	Clitic

Table 39: Affixes producing deverbal verbs

Function	DIr / t / Ir	Il	In	Iş	mA	AklA / IklA / AlA	IştIr	A / I	p	ImsA
causative	X									
passive		X	X							
reflexive		X	X							
reciprocal				X						
negative					X					
repetition						dürt-ükle-, it-ekle-, dürt-ele-, it-ele-, ov-ala-, serp-ele-, eş-ele-	ara-ştır-, kırp-ıştır-, serp-iştir-			
diminutive						dur-akla-, it-ekle-, uyu-kla-, dur-ala-, it-ele-, şaş-ala-		tık-a-, kaz-ı-, sür-ü-	kır-p-, ser-p-	an-ımsa-, gül-ümse-, duy-umsa-

Table 40: Classification of NVD affixes

Group	[4]Recursivity	[5]Polysemy	Deriv. Allomorphy	Semantic Selection	Dictionary entries	Change in Base Arg. Str.	Change in Base POS	Phrasal Scope	[6]Order of Application	Inv. of Order of Appl.	Member of a Paradigm	Suspended Affixation	Required by Syntax	Affix Class	Verdict
NVD mAK	No	2+	None	None	N/A	Yes	Yes	Yes	V-4	Yes	No	No	No	Suffix	Leans IM
NVD (y)Iş	No	4+	None	None	298	Yes	Yes	Yes	V-4	Yes	No	No	No	Suffix	Leans IM
NVD GA / GI / AlgA	No	9+	3	Extensive	22 / 89	Yes	Yes	Yes	V-4	Yes	No	No	No	Suffix	Leans DM
NVD I / IK	No	10+	2	Extensive	117 / 139	Yes	Yes	Yes	V-4	Yes	No	No	No	Suffix	Leans DM
NVD (y)An / AGAn / GAn / GHn	No	7+	4	Extensive	44 / 6 / 58 / 89	Yes	Yes	Yes	V-4	Yes	No	No	No	Suffix	Leans DM
NVD (G)AÇ / GIÇ / (y)IcI	No	8+	3	Extensive	41 / 5	Yes	Yes	Yes	V-4	Yes	No	No	No	Suffix	Leans DM
NVD Ar / Ir / mAz	No	4+	2	Extensive	66 / 50	Yes	Yes	Yes	V-4	Yes	No	No	No	Suffix	Leans DM
NVD AK / AnAK	No	6+	2	Extensive	84 / 23	Yes	Yes	Yes	V-4	Yes	No	No	No	Suffix	Leans DM
NVD Am / (y)Im	No	4+	2	Extensive	6 / 265	Yes	Yes	Yes	V-4	Yes	No	No	No	Suffix	Leans DM
NVD (In)Ç	No	3+	None	Extensive	33	Yes	Yes	No	V-4	Yes	No	No	No	Suffix	Leans DM
NVD mA / mAÇ	No	5+	2	None	2538 / 31	Yes	Yes	Yes	V-4	Yes	No	No	No	Suffix	Leans DM
NVD DI / DIK / TI / IntI	No	3+	4	Extensive	- / 8 / 176 / 86	Yes	Yes	Yes	V-4	Yes	No	No	No	Suffix	Leans DM
NVD (y)AcAK	No	2+	None	Extensive	22	Yes	Yes	Yes	V-4	Yes	No	No	No	Suffix	Leans DM
NVD mIş	No	2+	None	Extensive	34	Yes	Yes	Yes	V-4	Yes	No	No	No	Suffix	Leans DM
AVD Ip	No	None	None	None	N/A	No	Yes	Yes	V-4	Yes	Yes	No	Yes (Any V-4)	Suffix	Definitely IM
AVD (y)ArAK	No	None	None	None	N/A	No	Yes	Yes	V-4	Yes	Yes	No	Yes (Any V-4)	Suffix	Definitely IM
AVD mAdAn	No	None	None	None	N/A	No	Yes	Yes	V-4	Yes	Yes	No	Yes (Any V-4)	Suffix	Definitely IM
AVD ken	No	None	None	None	N/A	No	Yes	Yes	V-6.2 / N-6.2	Yes	No	No	No	Suffix	Leans IM

Table 41: Affixes producing deverbal nouns

Thematic Role	mAK	(y)Iş	GA / GI / AlgA	I / IK	(y)An / AGAn / GAn / GIn	(G)AÇ / GIÇ / (y)IcI	Ar / Ir / mAz	AK / AnAK	Am / (y)Im	(In)Ç	mA / mAç	DI / DIK / TI / IntI	(y)AcAK	mIş
action	koş-mak	ulusa seslen-iş, güne bak-ış, dal-ış, diren-iş, yürüy-üş	öv-gü	koş-u, gör-ü, öngör-ü, içgör-ü	bas-kın, sür-gün				kullan-ım, dol-um, etkileş-im, evr-im, iletiş-im, ak-ım, bunal-ım, ayr-ım, seç-im, deney-im, geril-im, devin-im	bas-ınç, diren-ç	ağaçlandır-ma, hışırda-ma, değer ver-me, yürüt-me, meydan oku-ma, gönül al-ma, pusulayı şaşır-ma			
1-stimulus (anim.)														
2-agent			bil-ge		sürün-gen, konuş-kan, et-ken, sıç-an, bak-an	bil-giç, dal-gıç, yar-gıç, öğren-ci, dilen-ci	kes-er, yara-r, gel-ir, oku-r, yaz-ar, değ-er							
3-? (anim. cause)														
4-stimulus (inanim.)			sez-gi, duy-gu, coş-ku, sev-gi	kok-u, kork-u, sez-i, duy-u										
5-force / nat. cause			et-ki, dal-ga			yaz-ıcı, yatıştır-ıcı, uyuştur-ucu								
6-cause														
7-instrument (anim.)														
8-? (non-part. İnstr.)														
9-instrument (inanim.)			del-gi, göster-ge, süpür-ge, çiz-elge, sil-gi, sür-gü	kay-ık		bağla-ç, dokun-aç, büyüt-eç, del-geç		ele-k, kay-ak						
10-manner		düş-üş, ak-ış, gör-üş, bak-ış, yaradıl-ış							yaklaş-ım, ver-im					
11-experiencer			bit-ki, çeliş-ki (?)											er-miş
12-beneficiary														
13-theme		bul-uş, göster-iş	bildir-ge, öner-ge, gör-gü, ver-gi, iç-ki, bil-gi	art-ı, yet-i, duyur-u, ver-i, kon-uk	gez-egen, kay-gan	sür-eç	aç-maz	ada-k, gör-enek	bil-im, anla-m, dön-em, dön-üm, tut-am, tut-um, varsay-ım, kur-am, böl-üm			uy-du, tanı-dık, bil-dik, söylen-ti, alın-tı	gel-ecek	geç-miş, dol-muş
14-patient	ek-mek, çak-mak, ye-mek		çiz-ge, bas-kı, der-gi, sömür-ge, diz-ge	yaz-ı, yar-ı				tut-anak, öde-nek	yatır-ım, biç-em, biç-im, kur-am, kur-um		bas-ma, kıy-ma, in-me, yırt-maç, bula-maç	alın-dı, çık-tı, bağır-tı	yiy-ecek, giy-ecek, iç-ecek	
15-purpose														
16-source / origin (anim.)														
17-source / origin (inanim.)						başlan-gıç								
18-? (anim. location)														
19-location		gir-iş	yerleş-ke	bat-ı				dur-ak, kon-ak, sığın-ak, tapın-ak, yayla-(k), kışla-(k)						
20-recipient														
21-direction / goal												doğrul-tu		
22-? (source time)														
23-time														
24-? (goal time)														

Table 42: Affixes producing deverbal adjectives

Thematic Role	mAK	(y)Iş	GA / GI / AlgA	I / IK	(y)An / AGAn / GAn / GIn	(G)AÇ / GIÇ / (y)IcI	Ar / Ir / mAz	AK / AnAK	Am / (y)Im	(In)Ç	mA / mAç	DI / DIK / TI / IntI	(y)AcAK	mIş
1-stimulus (anim.)										iğren-ç, gülün-ç				
2-agent				içe dön-ük	uç-an (daire), düşün-en (adam), danış-an, kır-an, tamla-yan, çalış-kan, ısır-gan, gez-gin, et-kin	yap-ıcı	çal-ar, okuryaz-ar (?)				tepeden in-me, sonradan gör-me			
3-? (anim. cause)														
4-stimulus (inanim.)						üz-ücü, yor-ucu				iğren-ç, gülün-ç				
5-force / nat. cause														
6-cause														
7-instrument (anim.)														
8-? (non-part. İnstr.)														
9-instrument (inanim.)														
10-manner														
11-experiencer				soğu-k	bit-kin, yor-gun, kır-gın, dur-gun, kır-gın	utan-gaç		ürk-ek, kork-ak						
12-beneficiary														
13-theme				büyü-k, ışı-k, dol-u, dur-u, sol-uk	dur-ağan, ol-ağan, dur-gun		tüken-mez, bit-mez, anlaşıl-maz	kur-ak			kulaktan dol-ma			
14-patient				asıl-ı, seril-i, dikil-i, sık-ı, buna-k, esne-k, öl-ü, kır-ık	kamyon çarp-an, edil-gen						dol-ma, dök-me (demir), yekpare taştan yapıl-ma, serp-me, şiş-me, süz-me, döv-me (dondurma), bur-ma, dondur-ma, sür-me, yaz-ma			
15-purpose														
16-source / origin (anim.)														
17-source / origin (inanim.)														
18-? (anim. location)														
19-location														
20-recipient														
21-direction / goal														
22-? (source time)														
23-time														
24-? (goal time)														

Table 43: Affixes producing deverbal adverbs

Thematic Role	(y)Ip	(y)ken	(y)ArAK	mAdAn	A
1-stimulus (anim.)					
2-agent					
3-? (anim. cause)					
4-stimulus (inanim.)					
5-force / nat. cause					
6-cause					
7-instrument (anim.)					
8-? (non-part. İnstr.)					
9-instrument (inanim.)					
10-manner		hasta-yken, gelmiş-ken	koş-arak geldi	koş-madan geldi	
11-experiencer					
12-beneficiary					
13-theme					
14-patient					
15-purpose					
16-source / origin (anim.)					
17-source / origin (inanim.)					
18-? (anim. location)					
19-location					
20-recipient					
21-direction / goal					
22-? (source time)	koş-up geldi, yat-ıp uyuyacak				
23-time		hasta-yken, gelmiş-ken			
24-? (goal time)					
action					koş-a-biliyorum

Table 44: Classification of VND affixes

Group	[4]Recursivity	[5]Polysemy	Deriv. Allomorphy	Semantic Selection	Dictionary entries	Change in Base Arg. Str.	Change in Base POS	Phrasal Scope	[6]Order of Application	Inv. of Order of Appl.	Member of a Paradigm	Suspended Affixation	Required by Syntax	Affix Class	Verdict
VND DA	No	None	None	Extensive	87	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
VND KIr	No	None	None	Extensive	11	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
VND lA	No	14+	None	Extensive	1119	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
VND lAn	No	8+	None	Extensive	541	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
VND lAş	No	5+	None	Extensive	318	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
VND lAt	No	2+	None	Extensive	6	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
VND ImsA / sA	No	5+	2	Extensive	12 / 34	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
VND Ar	No	2+	None	Extensive	19	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
VND Al	No	1+	None	Extensive	39	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM

Table 45: Affixes producing denominal verbs

Thematic Relations	DA	KIr	lA	lAn	lAş	lAt	ImsA / sA	Ar	Al
onomatopoeia	şakır-da-, vızıl-da-, çatır-da-, kıpır-da-	fış-kır-, püs-kür-, hıç-kır-	gıdak-la-, miyav-la-, hav-la-, of-la- ah-la-						
result			alkış-la-, alıntı-la-, ara-la-, ayrım-la-, fotoğraf-la-, hesap-la-, ıska-la-, özet-le-, örgüt-le-, örnek-le-, panik-le-, plan-la-, sıra-la-, sınıf-la-, sorgu-la-, tanı-la-	arıza-lan-, cephe-len-, köpük-len-, nefes-len-	ahit-leş-, ant-laş-, şaka-laş-, yardım-laş-		durak-sa-, ayrım-sa-		
action			adım-la-						
1-stimulus (anim.)									
2-agent									
3-? (anim. cause)									
4-stimulus (inanim.)			kok-la-	cesaret-len-, dehşet-len-, evham-lan-, heves-len-, kuşku-lan-, öfke-len-, sevda-lan-, stres-len-, telaş-lan-					
5-force / nat. cause				böcek-len-, güve-len-, güneş-len-, rutubet-len-					
6-cause									
7-instrument (anim.)									
8-? (non-part. İnstr.)									
9-instrument (inanim.)			algı-la-, alçı-la-, asfalt-la-, ateş-le-, arşın-la-, bant-la-, cilt-le-, gaga-la-, iğne-le-, kelepçe-le-, kilit-le, kira-la-, nokta-la-, pençe-le-, poşet-le-, sap-la-, sepet-le-, taş-la-	ağaç-lan-, ayak-lan-, bilgi-len-, et-len-, ev-len-, görev-len, kir-len-	yüz-leş-, email-leş-	kir-let-	kut-sa-		
10-manner			köpek-le-	alev-len-, horoz-lan-, karınca-lan-, öbek-len-, sürat-len-, şiddet-len-	anıt-laş-, arap-laş-, bir-leş-, blok-laş-, bozkır-laş-, çete-leş-, deyim-leş-		kap-sa-, gerek-sin-, güç-sün-, yük-sün-, öz-ümse-, benim-se-		
11-experiencer									
12-beneficiary									
13-theme			hedef-le-		haber-leş-		önem-se-, su-sa-		
14-patient			av-la-	borç-lan-					
15-purpose									
16-source / origin (anim.)									
17-source / origin (inanim.)			sol-la-						
18-? (anim. location)									
19-location									
20-recipient									
21-direction / goal			aşağı-la-, arşiv-le-, ilik-le-, katalog-la-					baş-ar-, iç-er-, ön-er-	
22-? (source time)									
23-time			güz-le-, akşam-la-, kış-la-						
24-? (goal time)			sabah-la-, önce-le-						

Table 46: Affixes producing deadjectival verbs

Thematic Relations	DA	KIr	lA	lAn	lAş	lAt	ImsA / sA	Ar	Al
result									
1-stimulus (anim.)									
2-agent									
3-? (anim. cause)									
4-stimulus (inanim.)									
5-force / nat. cause									
6-cause									
7-instrument (anim.)									
8-? (non-part. İnstr.)									
9-instrument (inanim.)									
10-manner									
11-experiencer				deli-len-, rahatsız-lan-				deli-r-	
12-beneficiary									
13-theme			zayıf-la-, hafif-le-, serin-le-, şişman-la-	bilmez-len-, sakat-lan-	züppe-leş-, yuvarlak-laş-, yassı-laş-, usta-laş-, taze-leş-, suskun-laş-, sıradan-laş- katı-laş-, fena-laş-, derin-leş-, güzel-leş-, koyu-laş-, kır-laş-			ağ-ar-, bol-ar-, kara-r-, mor-ar-, sar-ar-, yaş-ar-, yeş-er-	alça-l-, az-al-, boş-al-, çoğ-al-, doğru-l-, düz-el-, kısa-l-, ince-l-, kör-el-, sivri-l-, yamu-l-
14-patient			ak-la-, kuru-la-	hoş-lan-		aydın-lat-, bol-lat-, derin-let-, keskin-let	az-ımsa-, küçü-mse-, garip-se-, hafif-se-, ırak-sa-, küçük-se-, mühim-se-		
15-purpose									
16-source / origin (anim.)									
17-source / origin (inanim.)									
18-? (anim. location)									
19-location									
20-recipient									
21-direction / goal									
22-? (source time)									
23-time									
24-? (goal time)									

Table 47: Classification of DM-leaning NND affixes

Group	[4]Recursivity	[5]Polysemy	Deriv. Allomorphy	Semantic Selection	Dictionary entries	Change in Base Arg. Str.	Change in Base POS	Phrasal Scope	[6]Order of Application	Inv. of Order of Appl.	Member of a Paradigm	Suspended Affixation	Required by Syntax	Affix Class	Verdict
NND dAş	No	1+	None	Extensive	54	N/A	No	No	N-0	No	No	No	No	Suffix	Definitely DM
NND gil / giller	No	2+	2	Extensive	182	N/A	No	No	N-0	No	No	No	No	Suffix	Leans DM
NND tI	No	1+	None	Extensive	6	N/A	No	No	N-0	No	No	No	No	Suffix	Definitely DM
NND gen	No	1+	None	Extensive	11	N/A	No	No	N-0	No	No	No	No	Suffix	Leans DM
NND lIK	Yes	14+	None	Extensive	2139	N/A	Yes	Yes	N-0 / V-5	No	No	No	No	Suffix	Definitely DM
NND CI	No	5+	None	Extensive	1832	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
NND CA / CAnA / CAsI / CAsInA	No	10+	4	Extensive	662 / - / - / 12	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
NND CAK / CIK / CAcIK / cağız	No	5+	4	Extensive	20 / 167 / - / 8	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
NND ImsI / ImtraK / rAK / sI	No	2+	4	Extensive	105 / 15 / 10 / 135	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
NND Al / Il / sAl / sIl	No	3+	4	Extensive	- / 47 / 580 / 31	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
NND Ø	Yes	2+	None	Extensive	1+	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
NND CIl	No	2+	None	Extensive	61	N/A	Yes	No	N-0	No	No	No	No	Suffix	Definitely DM
NND lI	No	3+	None	Extensive	2192	N/A	Yes	Yes	N-0	No	No	No	No	Suffix	Leans DM
NND sIz	No	1+	None	Extensive	1476	N/A	Yes	Yes	N-0	No	No	No	No	Suffix	Leans DM
NND IncI	No	1+	None	Extensive	22	N/A	Yes	No	N-0	No	No	No	No	Suffix	Leans DM
NND Iz	No	1+	None	Extensive	4	N/A	Yes	No	N-0	No	No	No	No	Suffix	Leans DM
NND şAr	No	1+	None	Extensive	22	N/A	Yes	No	N-0	No	No	No	No	Suffix	Leans DM

Table 48: Classification of IM-leaning NND affixes

Group	**[4]Recursivity**	**[5]Polysemy**	**Deriv. Allo-morphy**	**Semantic Selection**	**Dictionary entries**	**Change in Base Arg. Str.**	**Change in Base POS**	**Phrasal Scope**	**[6]Order of Application**	**Inv. of Order of Appl.**	**Member of a Paradigm**	**Suspended Affixation**	**Required by Syntax**	**Affix Class**	**Verdict**
NND lA	No	1+	None	Extensive	58	N/A	Yes	No	N-0	No	No	No	No	Suffix	Leans IM
NND DAn	No	2+	None	Extensive	69	N/A	Yes	No	N-3	No	No	No	No	Suffix	Leans IM
NND DA	No	1+	None	Extensive	34	N/A	Yes	No	N-3	No	No	No	No	Suffix	Leans IM
NND I	No	1+	None	Extensive	6	N/A	Yes	No	N-2	No	No	No	No	Suffix	Leans IM
NND ki	Yes	None	None	None	N/A	N/A	Yes	Yes	N-1 (After time adverbs) / N-4 (After GEN / LOC)	No	No	Yes?	No	Clitic?	Clitic?
Plural	No	None	None	Exceptional	N/A	N/A	No	Yes	N-1	Yes	No	Yes	Yes?	Suffix	Definitely IM
Possessive	No	None	None	None	N/A	?	No	Yes	N-2 / V-4.2	Yes	Yes	Yes	Yes	Suffix	Leans IM
Case (Instr)	No	None	None	None	N/A	N/A	No	Yes	N-3	Yes	Yes	Yes	Yes	Clitic (Bound / Free)	Clitic
Case (Other)	No	None	None	None	N/A	N/A	No	Yes	N-3	Yes	Yes	Yes	Yes	Suffix	Definitely IM
Topic M. dA	No	None	None	None	N/A	N/A	No	Yes	N-5 / V-3.2	No	No	Yes	No	Clitic (Free)	Clitic

Table 49: Category-preserving affixes producing denominal nominals

Semantics	dAş	gil / giller	tI	gen	lIK	CI	CA / CAnA / CAsI / CAsInA	CAK / CIK / CAcIK / cağız	ImsI / ImtraK / rAK / sI	Al / Il / sAl / sIl
sharing of	ad-aş, boy-daş, arka-daş, ev-deş, duygu-daş, çağ-daş, din-daş, fikir-deş, karın-daş, kök-teş, meslek-taş, pay-daş, vatan-daş, yan-daş, yurt-taş									
group / family		Ahmet-gil, amcası-gil, Ahmet-ler, annen-ler								
family / genus		ananas-giller, ayı-giller, bakla-giller, domuz-giller								
onomatopoeia			gıcır-tı, takır-tı, gürül-tü, hışır-tı, horul-tu							
number of corners				üç-gen, dört-gen, beş-gen						
apparel for					iç-lik, göz-lük, baş-lık					
characteristic of					abajurcu-luk, abla-lık, abone-lik, adanalı-lık, aday-lık, adliyeci-lik, asilzade-lik, baba-lık, donkişot-luk					
dedicated to					iftariye-lik, yol-luk, zeytin-lik, mezar-lık					
adopted family					ana-lık, evlat-lık					
banknote of					on-luk, yirmi-lik, yüz-lük					
salary every					hafta-lık, ay-lık, gün-lük					
container of					odun-luk, kitap-lık, pabuç-luk					belge-sel, kum-sal
affinitity towards						abartı-cı, acele-ci, akıl-cı, akşam-cı, şüphe-ci, yayılma-cı, içki-ci, palavra-cı, devrim-ci, geri-ci				
engaged in						akın-cı, isyan-cı, yol-cu, kapkaç-çı				
professionally in						aba-cı, adliye-ci, afiş-çi, akademi-ci, arpa-cı, avanta-cı, av-cı, bağlama-cı, balta-cı				
ideologically in						aristoteles-çi, atatürk-çü				
language of							arap-ça, alman-ca, türk-çe, kırgız-ca, osmanlı-ca			
repetition of							kovalama-ca, korkutma-ca			
diminutive								(ı)sı-cak, yavru-cak, ada-cık, ana-cık, bebe-cik, beyin-cik, bronş-çuk, göl-cük, söz-cük		
tool for								oyun-cak, salın-cak		
location										
diminutive							acı-ca, ağır-ca, ak-ça, ala-ca, çevik-çe, geniş-çe, enli-ce, iri-ce, kalın-ca, sarı-ca	büyü-cek, az-ıcık, biraz-cık, ince-cik, körpe-cik, mini-cik, sıca-cık, ufa-cık, yumuşa-cık, dar-acık	acı-msı, bordo-msu, kekre-msi, sarı-msı, tatlı-msı, acı-mtrak, beyaz-ımtrak, acı-rak, alça-rak, kısa-rak	
characteristic of										yok-sul, var-sıl
colored in										
diminutive								demin-cek, hemen-cecik, hazır-cacık, çabu-cak		

Table 50: Category-changing affixes producing denominal adjectives and deadjectival nouns

Semantics	∅	lIK	ImsI / ImtraK / rAK / sI	CI	CIl	CA / CAnA / CAsI / CAsInA	Al / Il / sAl / sIl	lI	sIz	IncI	Iz	şAr
characteristic of	aynasız-, balıkçıl-, köylü-, üniversiteli-	acayip-lik, acılı-lık, aciz-lik, adi-lik, abartıcı-lık, adaş-lık, afyonkeş-lik, ayrımcı-lık, aziz-lik, sağır-lık, iyi-lik										
dedicated to		adak-lık, atıştırma-lık, çıra-lık, erkanıharp-lik, ısırım-lık, içim-lik, idam-lık, sofra-lık, aşure-lik, dolma-lık, elbise-lik, iki kişi-lik										
recurrent every		hafta-lık, ay-lık, gün-lük										
amount of		yüz milyon *-luk, beş kilo-luk, yüz milyon lira-lık, üç saat-lik										
approx. age		doksan-lık, elli-lik										
characteristic of			abide-msi, ağac-ımsı, fikra-msı, ıspanağ-ımsı, platin-imsi, şeytan-ımsı, taş-ımsı, zımpara-msı, ağaç-sı, altın-sı, anıt-sı, aslan-sı, bakır-sı, diken-si, eylem-si, insan-sı									
affinitity towards				abla-cı	bütün-cül, ölüm-cül, ön-cül son-cul, ana-cıl, baba-cıl, barış-çıl, iyi-cil, kötü-cül, insan-cıl, ev-cil, ben-cil							
feeding on					balık-çıl, böcek-çil, çürük-çül, et-çil							
result of						düzme-ce, kesme-ce, kurma-ca, seçme-ce						
large amount						o-nca, bu-nca, onlar-ca, binler-ce, kilolar-ca, tonlar-ca, günler-ce, yıllar-ca, hektarlar-ca						
related to							anayasa-l, ard-ıl, birinci-l, dişi-l, çoğ-ul, doğa-l, finans-al, giz-il, ilk-el, açı-sal, akıl-sal, altyapı-sal, anıt-sal, anlam-sal, araç-sal, biçim-sel, bitki-sel, birey-sel, bütün-sel					
person from								adana-lı, ankara-lı, arjantin-li, belçika-lı, anadolu-lu, iskandinavya-lı, üniversite-li, köy-lü, bura-lı, nere-li				
presence of								abartı-lı, acı-lı, ağaçlık-lı, mavi-li, ne-li (dondurma), üç-*-lü, dört-*-lü, altı-*-lı, kısa saç-lı, dört çocuk-lu, mavi elbise-li, bindokuzyüz-lü (yıllar), deniz manzara(sı)-lı				
lack of									acıma-sız, acı-sız, ağrı-sız, ahlak-sız, aile-siz, akıl-sız, aralık-sız, ayıp-sız, azim-siz, baba-sız, badana-sız, para-sız			
ordinal number										bir-inci, kaç-ıncı, bilmem kaç-ıncı, son-uncu		
member of a multiple											iki-z, üç-üz, dörd-üz	
partition each												bir-er, kaç-ar, yarım-şar

Table 51: Category-changing affixes producing denominal adverbs and deadverbial nominals

Semantics	lIK	CA / CAnA / CAsI / CAsInA	CAK / CIK / CAcIK / cağız	lA	lI	DAn	DA	I	∅	ki
characteristic of	ayrıca-lık, biteviye-lik, boşuna-lık, çoğun-luk, evvel-lik, farkında-lık, göre-lik, günde-lik, kendiliğinden-lik, nite-lik, onda-lık, önce-lik, yerinde-lik									
in terms of		adet-çe, anlam-ca, beden-ce, biçim-ce, boy-ca, durum-ca, kafa-ca, hak-ça, sayı-ca								
according to		ben-ce, biz-ce								
collectively as		ev-ce	aile-cek, ev-cek, mahalle-cek							
characteristic of		adam-ca, ahbap-ça, arkadaş-ça, amir-ce, aptal-ca, asker-ce, çocuk-ça, derviş-çe, erkek-çe, iblis-çe, hak-çası, erkek-çesi		açıklık-la, afiyet-le, çoğunluk-la, ekseriyet-le, güçlük-le, güzellik-le, hayranlık-la, hayret-le, hız-la, ısrar-la, içtenlik-le, ivedilik-le, kesinlik-le, memnuniyet-le, öncelik-le, özellik-le, rahatlık-la, sabır-la, zaman-la						
conjunction					ana-lı kız-lı, kız-lı oğlan-lı, gece-li gündüz-lü					
17-source / origin (inanim.)						sıra-dan, ne-den, iç-ten, top-tan				
19-location							ayak-ta, baş-ta, dünya-da, gerçek-te, ileri-de, oracık-ta, orta-da, ortalık-ta, söz-de, yedek-te			
23-time								sabahlar-ı, önceler-i, akşamlar-ı, sonralar-ı, evveller-i		
characteristic of		ahmak-ça, açık-ça, ahlaksız-ca, alçak-ça, amansız-ca, arsız-ca, avanak-ça, bencil-ce, bilge-ce, aptal-casına, budala-casına, çılgın-casına, deli-cesine, güzel-cene, kolay-cana							parasız-, sensiz-, arabasız-	
17-source / origin (inanim.)						açık-tan, ani-den, çok-tan, eski-den, fazla-dan hafif-ten, ilk-ten, incecik-ten, ince-den, sahi-den, yakın-dan, yeni-den				
23-time										akşam-ki, bugün-kü, demin-ki, dün-kü, evvel-ki, gece-ki, gündüz-ki, önce-ki, sabah-ki, sonra-ki, şimdi-ki, yarın-ki

Appendix B

TRIALS

B.1 Core Trials

B.1.1 CT0

(170) Observation List for CT0

a. *diz* ⊢ n : λx1.(be knee x1)

b. *dirsek* ⊢ n : λx1.(be elbow x1)

c. *bilek* ⊢ n : λx1.(be wrist x1)

d. *dizlik* ⊢ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon)

e. *dirseklik* ⊢ n : λx1λx2.and (be elbow x2) (wear (on x2) x1 anon)

f. *bileklik* ⊢ n : λx1λx2.and (be wrist x2) (wear (on x2) x1 anon)

g. *dizlik* ⊢ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon) (4 times)

h. *bileklik* ⊢ n : λx1λx2.and (be wrist x2) (wear (on x2) x1 anon) (4 times)

Morphology LF TEMPLATE:
λx0λ*Optional_New_Variables* λ*Stem_Variables*.and (x0 *Stem_Variables*) *New_Term*

Syntax LF TEMPLATE:
λx0λ*Optional_New_Variables* λ*Stem_Variables*.and (x0 *Reversed_Stem_Variables*) *New_Term*

(171) Grammar Learned from CT0 with Final Prior Probabilities

a. *diz* ⊢ n : λx1.be knee x1

b. *dirsek* ⊢ n : λx1.be elbow x1

c. *bilek* ⊢ n : λx1.be wrist x1

d. *dizlik* ⊢ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon)

Table 52: Segmentation prior probabilities at the end of CT0

Surface Form	Segmentation	Prior Probability
diz	diz	1.00
dirsek	dirsek	1.00
bilek	bilek	1.00
dizlik	dizlik	0.31
dirseklik	dirseklik	1.00
-lik	lik	1.00
bileklik	bilek-lik	0.83
bileklik	bileklik	0.17
dizlik	diz-lik	0.69

e. *dirseklik* $\vdash$ n : λx1λx2.and (be elbow x2) (wear (on x2) x1 anon)

f. *-lik* $\vdash$ n\n : λx1λx2λx3.and (x1 x3) (wear (on x3) x2 anon)

g. *bileklik* $\vdash$ n : λx1λx2.and (be wrist x2) (wear (on x2) x1 anon)

B.1.2 Demonstration of the Algorithm on CT0

(172) Learning Parameters

a. Learning Threshold (LT): 2

b. Initial Observation Count (IOC): 0.2

c. Maximum Embedding Dissimilarity (MED): 3

(173) Lexicon at $t = 0$

a. Initial lexicon is empty.

(174) Segmentation Alternatives at $t = 0$

a. Initial list is empty.

(175) Observation 1

a. Observation: *diz* $\vdash$ n : λx1.(be knee x1)

b. Segmentation: No valid segmentation
diz: Not attested
di-z: Not attested

d-i-z: Not attested
d-iz: Not attested

c. BBN Construction: No BBN without a valid segmentation

d. Post BBN: No inference without BBN
Unsegmented form added to the lexicon with 0.2 IOC.
No valid segmentation

e. Affix Recognition: No affix candidates
di-: Not attested
d-: Not attested

(176) Lexicon at $t = 1$

a. *diz* $\vdash$ n : λx1.(be knee x1) (Prob: 1.00)

(177) Segmentation Alternatives at $t = 1$

a. *diz*: diz (Prob: 1.00)

(178) Observation 2

a. Observation: *dirsek* $\vdash$ n : λx1.(be elbow x1)

b. Segmentation: No valid segmentation
dirsek: Not attested
dirse-k: Not attested
dirs-e-k: Not attested
dir-s-e-k: Not attested
...

c. BBN Construction: No BBN without a valid segmentation

d. Post BBN: No inference without BBN
Unsegmented form added to the lexicon with 0.2 IOC.
No valid segmentation

e. Affix Recognition: No affix candidates
dirse-: Not attested
dirs-: Not attested
dir-: Not attested
di-: Not attested
d-: Not attested

(179) Lexicon at $t = 2$

a. *diz* $\vdash$ n : λx1.(be knee x1) (Prob: 1.00)

b. *dirsek* $\vdash$ n : λx1.(be elbow x1) (Prob: 1.00)

(180) Segmentation Alternatives at $t = 2$

a. *diz*: diz (Prob: 1.00)

b. *dirsek*: dirsek (Prob: 1.00)

(181) Observation 3

a. Observation: *bilek* $\vdash$ n : λx1.(be wrist x1)

b. Segmentation: No valid segmentation
bilek: Not attested
bile-k: Not attested
bil-e-k: Not attested
bi-l-e-k: Not attested
...

c. BBN Construction: No BBN without a valid segmentation

d. Post BBN: No inference without BBN
Unsegmented form added to the lexicon with 0.2 IOC.
No valid segmentation

e. Affix Recognition: No affix candidates
bile-: Not attested
bil-: Not attested
bi-: Not attested
b-: Not attested

(182) Lexicon at $t = 3$

a. *diz* $\vdash$ n : λx1.(be knee x1) (Prob: 1.00)

b. *dirsek* $\vdash$ n : λx1.(be elbow x1) (Prob: 1.00)

c. *bilek* $\vdash$ n : λx1.(be wrist x1) (Prob: 1.00)

(183) Segmentation Alternatives at $t = 3$

a. *diz*: diz (Prob: 1.00)

b. *dirsek*: dirsek (Prob: 1.00)

c. *bilek*: bilek (Prob: 1.00)

(184) Observation 4

a. Observation: *dizlik* $\vdash$ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon)

b. Segmentation: No valid segmentation
dizlik: Not attested
dizli-k: Not attested
dizl-i-k: Not attested
diz-l-i-k: Not attested
...

c. BBN Construction: No BBN without a valid segmentation

d. Post BBN: No inference without BBN
Unsegmented form added to the lexicon with 0.2 IOC.
No valid segmentation

e. Affix Recognition: No viable affix candidates
dizli-: Not attested
dizl-: Not attested
diz-: Attested
Affix Candidate: -lik
Stem-Lemma Candidate: *dizlik*,*diz* (Embedding Dissimilarity=2.26 < MED)
Number of Attested Stems = 1 < LT
di-: Not attested
d-: Not attested

(185) Lexicon at $t = 4$

a. *diz* $\vdash$ n : λx1.(be knee x1) (Prob: 1.00)

b. *dirsek* $\vdash$ n : λx1.(be elbow x1) (Prob: 1.00)

c. *bilek* $\vdash$ n : λx1.(be wrist x1) (Prob: 1.00)

d. *dizlik* $\vdash$ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon) (Prob: 1.00)

(186) Segmentation Alternatives at $t = 4$

a. *diz*: diz (Prob: 1.00)

b. *dirsek*: dirsek (Prob: 1.00)

c. *bilek*: bilek (Prob: 1.00)

d. *dizlik*: dizlik (Prob: 1.00)

(187) Observation 5

a. Observation: *dirseklik* ⊢ n : λx1λx2.and (be elbow x2) (wear (on x2) x1 anon)

b. Segmentation: No valid segmentation
dirseklik: Not attested
dirsekli-k: Not attested
dirsekl-i-k: Not attested
dirsek-l-i-k: Not attested
...

c. BBN Construction: No BBN without a valid segmentation

d. Post BBN: No inference without BBN
Unsegmented form added to the lexicon with 0.2 IOC.
No valid segmentation

e. Affix Recognition: 1 viable affix candidate
dirsekli-: Not attested
dirsekl-: Not attested
dirsek-: Attested
Affix Candidate: -lik
Stem-Lemma Candidate: *dizlik*,*diz* (Embedding Dissimilarity=2.26 < MED)
Stem-Lemma Candidate: *dirseklik*,*dirsek* (Embedding Dissimilarity=0.77 < MED)
Number of Attested Stem-Lemma Candidates = 2 >= LT
Common Lambda Terms in the LFs of Candidate Lemmas: (wear (on x2) x1 anon)
Adopt Morphology Template: λx0λx1 λx2.and (x0 x2) (wear (on x2) x1 anon)
Renumber Candidate LF Variables: λx1λx2 λx3.and (x1 x3) (wear (on x3) x2 anon)
LFs obtained by deriving stem candidates match LFs of lemma candidates.
Add affix to the lexicon with 1 IOC.
dirse-: Not attested
dirs-: Not attested
...

(188) Lexicon at $t = 5$

a. *diz* ⊢ n : λx1.(be knee x1) (Prob: 1.00)

b. *dirsek* ⊢ n : λx1.(be elbow x1) (Prob: 1.00)

c. *bilek* ⊢ n : λx1.(be wrist x1) (Prob: 1.00)

d. *dizlik* ⊢ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon) (Prob: 1.00)

e. *dirseklik* ⊢ n : λx1λx2.and (be elbow x2) (wear (on x2) x1 anon) (Prob: 1.00)

f. *-lik* ⊢ n\n : λx1λx2λx3.and (x1 x3) (wear (on x3) x2 anon) (Prob: 1.00)

(189) Segmentation Alternatives at $t = 5$

a. *diz*: diz (Prob: 1.00)

b. *dirsek*: dirsek (Prob: 1.00)

c. *bilek*: bilek (Prob: 1.00)

d. *dizlik*: dizlik (Prob: 1.00)

e. *dirseklik*: dirseklik (Prob: 1.00)

(190) Observation 6

a. Observation: *bileklik* ⊢ n : λx1λx2.and (be wrist x2) (wear (on x2) x1 anon)

b. Segmentation: 1 valid segmentation
bileklik: Not attested
bilekli-k: Not attested
bilekl-i-k: Not attested
bilek-lik: Attested
...

c. BBN Construction
SN: bilek-lik
LN1: *bilek*
bilek ⊢ n : λx1.(be wrist x1)
LN2: *-lik*
-lik ⊢ n\n : λx1λx2λx3.and (x1 x3) (wear (on x3) x2 anon)
DN1: bilek-lik
bileklik ⊢ n : λx1λx2.and (be wrist x2) (wear (on x2) x1 anon)
MN
bileklik ⊢ n : λx1λx2.and (be wrist x2) (wear (on x2) x1 anon)

d. Post BBN: Bayesian inference trivial due to MN containing a single interpretation
Unsegmented form added to the lexicon with 0.2 IOC.
Segmentation alternative added to the lexicon with 0.2 IOC.

e. Affix Recognition: 1 viable affix candidate
bilekli-: Not attested
bilekl-: Not attested
bilek-: Attested
Affix Candidate: -lik
Stem-Lemma Candidate: *dizlik*,*diz* (Embedding Dissimilarity=2.26 < MED)
Stem-Lemma Candidate: *dirseklik*,*dirsek* (Embedding Dissimilarity=0.77 < MED)
Stem-Lemma Candidate: *bileklik*,*bilek* (Embedding Dissimilarity=1.03 < MED)
Number of Attested Stems = 3 >= LT
Common Lambda Terms in the LFs of Candidate Lemma: (wear (on x2) x1 anon)

Adopt Morphology Template: λx0λx1 λx2.and (x0 x2) (wear (on x2) x1 anon)

Renumber Candidate LF Variables: λx1λx2 λx3.and (x1 x3) (wear (on x3) x2 anon)

LFs obtained by deriving stem candidates match LFs of lemma candidates.

Candidate affix is already present in the lexicon.

bile-: Not attested
bil-: Not attested
...

(191) Lexicon at $t = 6$

a. *diz* ⊢ n : λx1.(be knee x1) (Prob: 1.00)

b. *dirsek* ⊢ n : λx1.(be elbow x1) (Prob: 1.00)

c. *bilek* ⊢ n : λx1.(be wrist x1) (Prob: 1.00)

d. *dizlik* ⊢ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon) (Prob: 1.00)

e. *dirseklik* ⊢ n : λx1λx2.and (be elbow x2) (wear (on x2) x1 anon) (Prob: 1.00)

f. *-lik* ⊢ n\n : λx1λx2λx3.and (x1 x3) (wear (on x3) x2 anon) (Prob: 1.00)

g. *bileklik* ⊢ n : λx1λx2.and (be wrist x2) (wear (on x2) x1 anon) (Prob: 1.00)

(192) Segmentation Alternatives at $t = 6$

a. *diz*: diz (Prob: 1.00)

b. *dirsek*: dirsek (Prob: 1.00)

c. *bilek*: bilek (Prob: 1.00)

d. *dizlik*: dizlik (Prob: 1.00)

e. *dirseklik*: dirseklik (Prob: 1.00)

f. *bileklik*: bilek-lik (Prob: 0.83)

g. *bileklik*: bileklik (Prob: 0.17)

(193) Observation 7

a. Observation: *dizlik* ⊢ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon)

b. Segmentation: 2 valid segmentations
dizlik: Attested
dizli-k: Not attested
dizl-i-k: Not attested

diz-lik: Attested

...

c. BBN Construction

SN: dizlik,diz-lik

LN1: *dizlik*

dizlik $\vdash$ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon)

LN2: *diz*

diz $\vdash$ n : λx1.(be knee x1)

LN3: *-lik*

-lik $\vdash$ n\n : λx1λx2λx3.and (x1 x3) (wear (on x3) x2 anon)

DN1: dizlik

dizlik $\vdash$ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon)

DN2: diz-lik

dizlik $\vdash$ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon)

MN

dizlik $\vdash$ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon)

d. Post BBN: Bayesian inference is carried out.

Unsegmented form is already in the lexicon.

New segmentation alternative diz-lik added to the lexicon with 0.2 IOC.

e. Affix Recognition: 1 viable affix candidate

dizli-: Not attested

dizl-: Not attested

diz-: Attested

Affix Candidate: -lik

Stem-Lemma Candidate: *dizlik*,*diz* (Embedding Dissimilarity=2.26 < MED)

Stem-Lemma Candidate: *dirseklik*,*dirsek* (Embedding Dissimilarity=0.77 < MED)

Stem-Lemma Candidate: *bileklik*,*bilek* (Embedding Dissimilarity=1.03 < MED)

Number of Attested Stems = 3 >= LT

Common Lambda Terms in the LFs of Candidate Lemma: (wear (on x2) x1 anon)

Adopt Morphology Template: λx0λx1 λx2.and (x0 x2) (wear (on x2) x1 anon)

Renumber Candidate LF Variables: λx1λx2 λx3.and (x1 x3) (wear (on x3) x2 anon)

LFs obtained by deriving stem candidates match LFs of lemma candidates.

Candidate affix is already present in the lexicon.

di-: Not attested

d-: Not attested

(194) Lexicon at $t = 7$

a. *diz* $\vdash$ n : λx1.(be knee x1) (Prob: 1.00)

b. *dirsek* $\vdash$ n : λx1.(be elbow x1) (Prob: 1.00)

c. *bilek* ⊢ n : λx1.(be wrist x1) (Prob: 1.00)

d. *dizlik* ⊢ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon) (Prob: 1.00) (Prob: 1.00)

e. *dirseklik* ⊢ n : λx1λx2.and (be elbow x2) (wear (on x2) x1 anon) (Prob: 1.00)

f. *-lik* ⊢ n\n : λx1λx2λx3.and (x1 x3) (wear (on x3) x2 anon) (Prob: 1.00)

g. *bileklik* ⊢ n : λx1λx2.and (be wrist x2) (wear (on x2) x1 anon) (Prob: 1.00)

(195) Segmentation Alternatives at $t = 7$

a. *diz*: diz (Prob: 1.00)

b. *dirsek*: dirsek (Prob: 1.00)

c. *bilek*: bilek (Prob: 1.00)

d. *dizlik*: dizlik (Prob: 0.31)

e. *dirseklik*: dirseklik (Prob: 1.00)

f. *bileklik*: bilek-lik (Prob: 0.83)

g. *bileklik*: bileklik (Prob: 0.17)

h. *dizlik*: diz-lik (Prob: 0.69)

(196) Observation 8

a. ...

B.1.3 CT1

(197) Observation List for CT1

a. *dizlik* ⊢ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon)

b. *dirseklik* ⊢ n : λx1λx2.and (be elbow x2) (wear (on x2) x1 anon)

c. *bileklik* ⊢ n : λx1λx2.and (be wrist x2) (wear (on x2) x1 anon)

d. *diz* ⊢ n : λx1.(be knee x1)

e. *dirsek* ⊢ n : λx1.(be elbow x1)

f. *bilek* ⊢ n : λx1.(be wrist x1)

g. *dizlik* ⊢ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon) (4 times)

Table 53: Segmentation prior probabilities at the end of CT1

Surface Form	Segmentation	Prior Probability
dizlik	dizlik	0.79
dirseklik	dirseklik	1.00
bileklik	bileklik	0.31
diz	diz	1.00
dirsek	dirsek	1.00
bilek	bilek	1.00
-lik	lik	1.00
dizlik	diz-lik	0.21
bileklik	bilek-lik	0.69

h. *bileklik* ⊢ n : λx1λx2.and (be wrist x2) (wear (on x2) x1 anon) (4 times)

(198) Grammar Learned from CT1 with Final Prior Probabilities

a. *dizlik* ⊢ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon) (Prob: 1.00)

b. *dirseklik* ⊢ n : λx1λx2.and (be elbow x2) (wear (on x2) x1 anon) (Prob: 1.00)

c. *bileklik* ⊢ n : λx1λx2.and (be wrist x2) (wear (on x2) x1 anon) (Prob: 1.00)

d. *diz* ⊢ n : λx1.be knee x1 (Prob: 1.00)

e. *dirsek* ⊢ n : λx1.be elbow x1 (Prob: 1.00)

f. *bilek* ⊢ n : λx1.be wrist x1 (Prob: 1.00)

g. *-lik* ⊢ n\n : λx1λx2λx3.and (x1 x3) (wear (on x3) x2 anon) (Prob: 1.00)

B.1.4 CT2

(199) Observation List for CT2A

a. *diz* ⊢ n : λa.(be knee a)

b. *dirsek* ⊢ n : λa.(be elbow a)

c. *bilek* ⊢ n : λa.(be wrist a)

d. *dizlik* ⊢ n : λaλb.and (be knee b) (wear (on b) a anon)

e. *dirseklik* ⊢ n : λaλb.and (be elbow b) (wear (on b) a anon)

f. *bileklik* ⊢ n : λaλb.and (be wrist b) (wear (on b) a anon)

g. *dizlikçi* ⊢ n : λaλbλc.and (and (be knee c) (wear (on c) b anon)) (sell b a)

h. *dirseklikçi* ⊢ n : λaλbλc.and (and (be elbow c) (wear (on c) b anon)) (sell b a)

i. *bileklikçi* ⊢ n : λaλbλc.and (and (be wrist c) (wear (on c) b anon)) (sell b a)

j. *dizlik* ⊢ n : λaλb.and (be knee b) (wear (on b) a anon) (4 times)

k. *dizlikçi* ⊢ n : λaλbλc.and (and (be knee c) (wear (on c) b anon)) (sell b a) (4 times)

l. *bileklik* ⊢ n : λaλb.and (be wrist b) (wear (on b) a anon) (4 times)

m. *bileklikçi* ⊢ n : λaλbλc.and (and (be wrist c) (wear (on c) b anon)) (sell b a) (4 times)

(200) Grammar Learned from CT2A with Final Prior Probabilities

a. *diz* ⊢ n : λx1.be knee x1 (Prob: 1.00)

b. *dirsek* ⊢ n : λx1.be elbow x1 (Prob: 1.00)

c. *bilek* ⊢ n : λx1.be wrist x1 (Prob: 1.00)

d. *dizlik* ⊢ n : λx1λx2.and (be knee x2) (wear (on x2) x1 anon) (Prob: 1.00)

e. *dirseklik* ⊢ n : λx1λx2.and (be elbow x2) (wear (on x2) x1 anon) (Prob: 1.00)

f. *-lik* ⊢ n\n : λx1λx2λx3.and (x1 x3) (wear (on x3) x2 anon) (Prob: 1.00)

g. *bileklik* ⊢ n : λx1λx2.and (be wrist x2) (wear (on x2) x1 anon) (Prob: 1.00)

h. *dizlikçi* ⊢ n : λaλbλc.and (and (be knee c) (wear (on c) b anon)) (sell b a) (Prob: 1.00)

i. *dirseklikçi* ⊢ n : λaλbλc.and (and (be elbow c) (wear (on c) b anon)) (sell b a) (Prob: 1.00)

j. *-likçi* ⊢ n\n : λx1λx2λx3λx4.and (and (x1 x4) (wear (on x4) x3 anon)) (sell x3 x2) (Prob: 1.00)

k. *-çi* ⊢ n\n : λx1λx2λx3λx4.and (x1 x3 x4) (sell x3 x2) (Prob: 1.00)

l. *bileklikçi* ⊢ n : λaλbλc.and (and (be wrist c) (wear (on c) b anon)) (sell b a) (Prob: 1.00)

(201) Observation List for CT2B

a. *et* ⊢ n : λa.(be meat a)

b. *çiçek* ⊢ n : λa.(be flower a)

c. *bisiklet* ⊢ n : λa.(be bicycle a)

d. *etçi* ⊢ n : λaλb.and (be meat b) (sell b a)

e. *çiçekçi* ⊢ n : λaλb.and (be flower b) (sell b a)

Table 54: Segmentation prior probabilities at the end of CT2A

Surface Form	Segmentation	Prior Probability
diz	diz	1.00
dirsek	dirsek	1.00
bilek	bilek	1.00
dizlik	dizlik	0.31
dirseklik	dirseklik	1.00
-lik	lik	1.00
bileklik	bilek-lik	0.83
bileklik	bileklik	0.17
dizlikçi	dizlikçi	0.22
dirseklikçi	dirseklikçi	1.00
-likçi	likçi	1.00
-çi	çi	1.00
bileklikçi	bilek-lik-çi	0.28
bileklikçi	bilek-likçi	0.28
bileklikçi	bileklik-çi	0.28
bileklikçi	bileklikçi	0.17
dizlik	diz-lik	0.69
dizlikçi	diz-lik-çi	0.26
dizlikçi	diz-likçi	0.26
dizlikçi	dizlik-çi	0.26

Table 55: Segmentation prior probabilities at the end of CT2B

Surface Form	Segmentation	Prior Probability
et	et	1.00
çiçek	çiçek	1.00
bisiklet	bisiklet	1.00
etçi	etçi	0.31
çiçekçi	çiçekçi	1.00
-çi	çi	1.00
bisikletçi	bisiklet-çi	0.83
bisikletçi	bisikletçi	0.17
etçi	et-çi	0.69

f. *bisikletçi* ⊢ n : λaλb.and (be bicycle b) (sell b a)

g. *etçi* ⊢ n : λaλb.and (be meat b) (sell b a) (4 times)

h. *bisikletçi* ⊢ n : λaλb.and (be bicycle b) (sell b a) (4 times)

(202) Grammar Learned from CT2B with Final Prior Probabilities

a. *et* ⊢ n : λa.(be meat a) (Prob: 1.00)

b. *çiçek* ⊢ n : λa.(be flower a) (Prob: 1.00)

c. *bisiklet* ⊢ n : λa.(be bicycle a) (Prob: 1.00)

d. *etçi* ⊢ n : λaλb.and (be meat b) (sell b a) (Prob: 1.00)

e. *çiçekçi* ⊢ n : λaλb.and (be flower b) (sell b a) (Prob: 1.00)

f. *-çi* ⊢ n\n : λx1λx2λx3.and (x1 x3) (sell x3 x2) (Prob: 1.00)

g. *bisikletçi* ⊢ n : λaλb.and (be bicycle b) (sell b a) (Prob: 1.00)

B.1.5 CT3

(203) Observation List for CT3A and CT3B

a. *Amerika* ⊢ n : λa.(be USA a)

b. *Ankara* ⊢ n : λa.(be Ankara a)

c. *Antalya* ⊢ n : λa.(be Antalya a)

d. *Mısır* ⊢ n : λa.(be Egypt a)

e. *Muğla* $\vdash$ n : λa.(be Muğla a)

f. *Türkiye* $\vdash$ n : λa.(be Türkiye a)

g. *Senegal* $\vdash$ n : λa.(be Senegal a)

h. *Yemen* $\vdash$ n : λa.(be Yemen a)

i. *Bolu* $\vdash$ n : λa.(be Bolu a)

j. *İstanbul* $\vdash$ n : λa.(be Istanbul a)

k. *Amerikalı* $\vdash$ n : λaλb.and (be USA b) (be (from b) a)

l. *Ankaralı* $\vdash$ n : λaλb.and (be Ankara b) (be (from b) a)

m. *Antalyalı* $\vdash$ n : λaλb.and (be Antalya b) (be (from b) a)

n. *Mısırlı* $\vdash$ n : λaλb.and (be Egypt b) (be (from b) a)

o. *Muğlalı* $\vdash$ n : λaλb.and (be Muğla b) (be (from b) a)

p. *Senegalli* $\vdash$ n : λaλb.and (be Senegal b) (be (from b) a)

q. *Türkiyeli* $\vdash$ n : λaλb.and (be Türkiye b) (be (from b) a)

r. *Yemenli* $\vdash$ n : λaλb.and (be Yemen b) (be (from b) a)

s. *Bolulu* $\vdash$ n : λaλb.and (be Bolu b) (be (from b) a)

t. *İstanbullu* $\vdash$ n : λaλb.and (be Istanbul b) (be (from b) a)

u. *Amerikalı* $\vdash$ n : λaλb.and (be USA b) (be (from b) a) (4 times)

v. *Muğlalı* $\vdash$ n : λaλb.and (be Muğla b) (be (from b) a) (4 times)

(204) Observation List for CT3C

a. Simple Forms in the Observation List for CT3A and CT3B (without allomorphy)

b. Complex Forms in the Observation List for CT3A and CT3B (without allomorphy)

c. Repeated Trials in the Observation List for CT3A and CT3B (without allomorphy)

d. Complex Forms in the Observation List for CT3A and CT3B (with allomorphy)

e. Repeated Trials in the Observation List for CT3A and CT3B (with allomorphy)

f. Muğlalı $\vdash$ n : λaλb.and (be Muğla b) (be (from b) a) (40 times) (with allomorphy)

(205) Affixes Recognized from CT3A with Final Prior Probabilities

a. *-lı* ⊢ n\n : λx1λx2λx3.and (x1 x3) (be (from x3) x2) (Prob: 1.00)

b. *-li* ⊢ n\n : λx1λx2λx3.and (x1 x3) (be (from x3) x2) (Prob: 1.00)

c. *-lu* ⊢ n\n : λx1λx2λx3.and (x1 x3) (be (from x3) x2) (Prob: 1.00)

(206) Affixes Recognized from CT3B with Final Prior Probabilities

a. *-lI* ⊢ n\n : λx1λx2λx3.and (x1 x3) (be (from x3) x2) (Prob: 1.00)

(207) Affixes Recognized from CT3C with Final Prior Probabilities

a. *-lı* ⊢ n\n : λx1λx2λx3.and (x1 x3) (be (from x3) x2) (Prob: 1.00)

b. *-li* ⊢ n\n : λx1λx2λx3.and (x1 x3) (be (from x3) x2) (Prob: 1.00)

c. *-lu* ⊢ n\n : λx1λx2λx3.and (x1 x3) (be (from x3) x2) (Prob: 1.00)

d. *-lI* ⊢ n\n : λx1λx2λx3.and (x1 x3) (be (from x3) x2) (Prob: 1.00)

B.1.6 CT4

(208) Observation List for CT4

a. *kitap* ⊢ n : λa.(be book a)

b. *mobilya* ⊢ n : λa.(be furniture a)

c. *gözlük* ⊢ n : λaλb.and (be eye b) (wear (on b) a anon)

d. *servis* ⊢ n : λa.(be cutlery a)

e. *kamyon* ⊢ n : λa.(be truck a)

f. *kitapçı* ⊢ n : λaλb.and (be book b) (sell b a)

g. *mobilyacı* ⊢ n : λaλb.and (be furniture b) (sell b a)

h. *gözlükçü* ⊢ n : λaλbλc.and (and (be eye c) (wear (on c) b anon)) (sell b a)

i. *servisçi* ⊢ n : λaλb.and (be cutlery b) (sell b a)

j. *kamyoncu* ⊢ n : λaλb.and (be truck b) (sell b a)

k. *taksi* ⊢ n : λa.(be taxi a)

l. *servis* ⊢ n : λa.(be shuttle a)

m. *kamyoncu* ⊢ n : λaλb.and (be truck b) (drive b a)

n. *taksici* ⊢ n : λaλb.and (be taxi b) (drive b a)

o. *servisçi* ⊢ n : λaλb.and (be shuttle b) (drive b a)

p. *Atatürk* ⊢ n : λa.(be Atatürk a)

q. *Epikür* ⊢ n : λa.(be Epicurus a)

r. *Aristo* ⊢ n : λa.(be Aristoteles a)

s. *Atatürkçü* ⊢ n : λaλb.and (be Atatürk b) (believe (in b) a)

t. *Epikürcü* ⊢ n : λaλb.and (be Epicurus b) (believe (in b) a)

u. *Aristocu* ⊢ n : λaλb.and (be Aristoteles b) (believe (in b) a)

v. *kitapçı* ⊢ n : λaλb.and (be book b) (sell b a) (10 times)

w. *gözlükçü* ⊢ n : λaλbλc.and (and (be eye c) (wear (on c) b anon)) (sell b a) (10 times)

x. *Atatürkçü* ⊢ n : λaλb.and (be Atatürk b) (believe (in b) a) (10 times)

y. *Aristocu* ⊢ n : λaλb.and (be Aristoteles b) (believe (in b) a) (10 times)

z. *kamyoncu* ⊢ n : λaλb.and (be truck b) (sell b a) (10 times)

{. *kamyoncu* ⊢ n : λaλb.and (be truck b) (drive b a) (10 times)

|. *taksici* ⊢ n : λaλb.and (be taxi b) (drive b a) (10 times)

}. *servisçi* ⊢ n : λaλb.and (be cutlery b) (sell b a) (10 times)

~. *servisçi* ⊢ n : λaλb.and (be shuttle b) (drive b a) (10 times)

(209) Affixes Recognized from CT4 with Final Prior Probabilities

a. *-CI* ⊢ n\n : λx1λx2λx3.and (x1 x3) (sell x3 x2) (Prob: 0.55)

b. *-CI* ⊢ n\n : λx1λx2λx3.and (x1 x3) (drive x3 x2) (Prob: 0.27)

c. *-CI* ⊢ n\n : λx1λx2λx3.and (x1 x3) (believe (in x3) x2) (Prob: 0.18)

(210) Homonyms from CT4 with Final Prior Probabilities

a. *servis* ⊢ n : λx1.be cutlery x1 (Prob: 0.66)

b. *servis* ⊢ n : λx1.be shuttle x1 (Prob: 0.34)

c. *servisçi* ⊢ n : λx1λx2.and (be cutlery x2) (sell x2 x1) (Prob: 0.63)

Table 56: Segmentation prior probabilities at the end of CT4

Surface Form	Segmentation	Prior Probability
kitap	kitap	1.00
mobilya	mobilya	1.00
gözlük	gözlük	1.00
servis	servis	1.00
kamyon	kamyon	1.00
kitapçı	kitapçı	0.70
mobilyacı	mobilyacı	1.00
-CI	CI	1.00
gözlükçü	gözlük_CI	0.02
gözlükçü	gözlükçü	0.98
servisçi	servis_CI	0.52
servisçi	servisçi	0.48
kamyoncu	kamyon_CI	0.75
kamyoncu	kamyoncu	0.25
taksi	taksi	1.00
taksici	taksi_CI	0.08
taksici	taksici	0.92
Atatürk	Atatürk	1.00
Epikür	Epikür	1.00
Aristo	Aristo	1.00
Atatürkçü	Atatürk_CI	0.06
Atatürkçü	Atatürkçü	0.94
Epikürcü	Epikür_CI	0.50
Epikürcü	Epikürcü	0.50
Aristocu	Aristo_CI	0.28
Aristocu	Aristocu	0.72
kitapçı	kitap_CI	0.30

d. *servisçi* ⊢ n : λx1λx2.and (be shuttle x2) (drive x2 x1) (Prob: 0.37)

e. *kamyoncu* ⊢ n : λx1λx2.and (be truck x2) (sell x2 x1) (Prob: 0.54)

f. *kamyoncu* ⊢ n : λx1λx2.and (be truck x2) (drive x2 x1) (Prob: 0.46)

B.2 Pilot Trials and Post-Evaluation

B.2.1 CT5

(211) Observation List for CT5

a. *gel* ⊢ v : λaλt.(come a t)

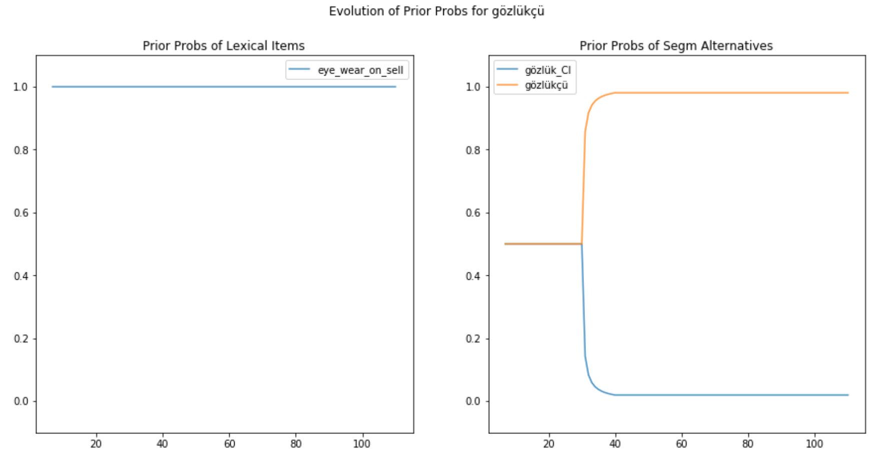


Figure 46: CT4 lexical and segmentation probabilities for *gözlükçü*

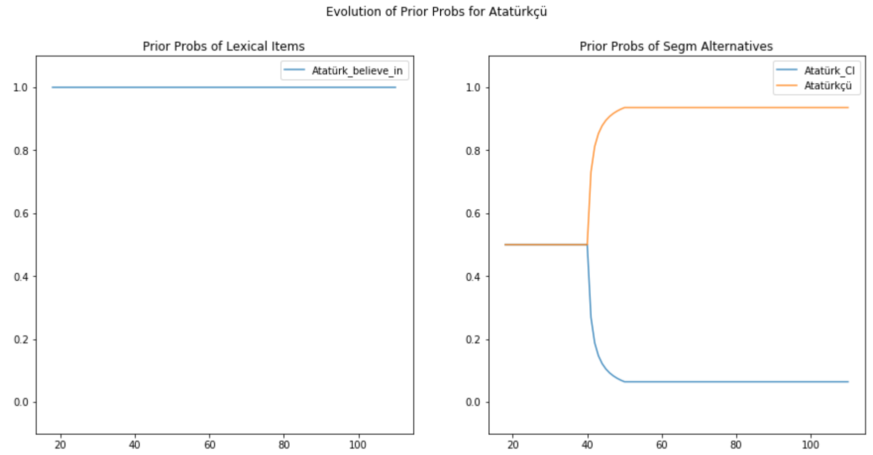


Figure 47: CT4 lexical and segmentation probabilities for *Atatürkçü*

b. *koş* ⊢ v : $\lambda a \lambda t$.(run a t)

c. *düş* ⊢ v : $\lambda a \lambda t$.(fall a t)

d. *geldi* ⊢ s/n : $\lambda a \lambda t$.and (t < tref) (come a t)

e. *koştu* ⊢ s/n : $\lambda a \lambda t$.and (t < tref) (run a t)

f. *düştü* ⊢ s/n : $\lambda a \lambda t$.and (t < tref) (fall a t)

g. *geldim* ⊢ s/n : $\lambda a \lambda t$.and (be speaker a) (and (t < tref) (come a t))

h. *koştum* ⊢ s/n : $\lambda a \lambda t$.and (be speaker a) (and (t < tref) (run a t))

i. *düştüm* ⊢ s/n : $\lambda a \lambda t$.and (be speaker a) (and (t < tref) (fall a t))

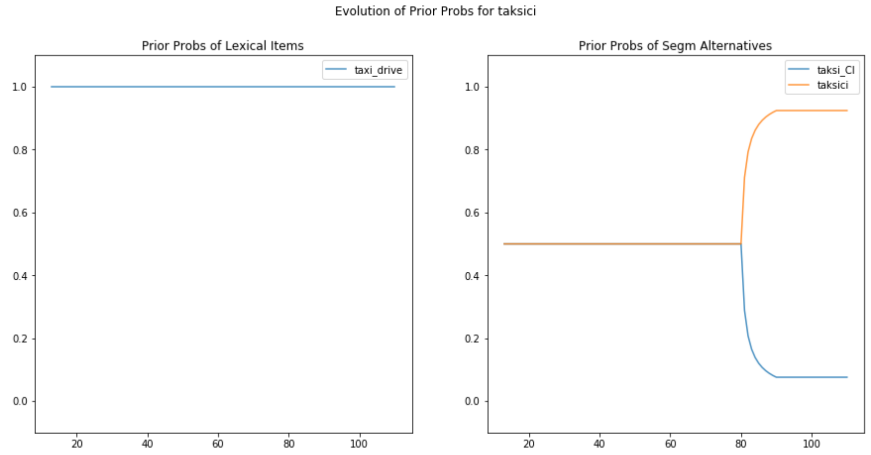


Figure 48: CT4 lexical and segmentation probabilities for *taksici*

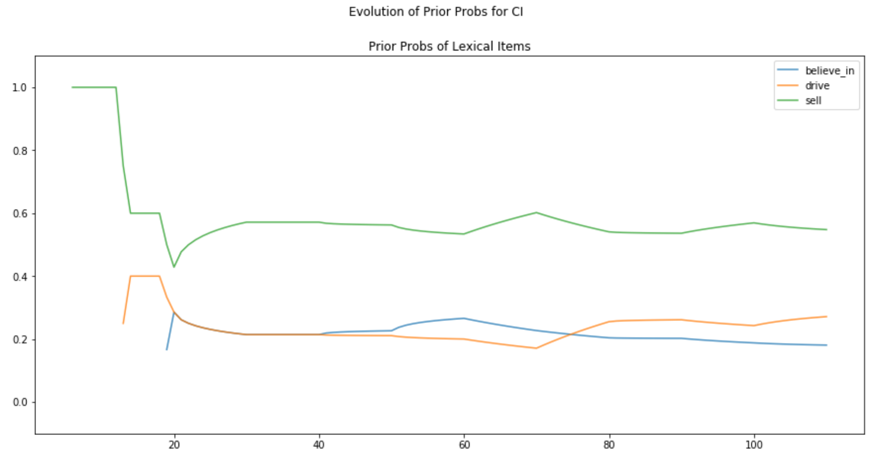


Figure 49: CT4 lexical and segmentation probabilities for -*CI*

j. *geldiydi* ⊢ s/n : λaλt.and (tref < t0) (and (t < tref) (come a t))

k. *koştuydu* ⊢ s/n : λaλt.and (tref < t0) (and (t < tref) (run a t))

l. *düştüydü* ⊢ s/n : λaλt.and (tref < t0) (and (t < tref) (fall a t))

m. *geldiydim* ⊢ s : λaλt.and (be speaker a) (and (tref < t0) (and (t < tref) (come a t)))

n. *koştuydum* ⊢ s : λaλt.and (be speaker a) (and (tref < t0) (and (t < tref) (run a t)))

o. *düştüydüm* ⊢ s : λaλt.and (be speaker a) (and (tref < t0) (and (t < tref) (fall a t)))

p. *gömlek* ⊢ n : λa.(be shirt a)

q. *kitap* ⊢ n : λa.(be book a)

r. *defter* ⊢ n : λa.(be notebook a)

s. *gömlekler* ⊢ n : λa.and (be shirt a) (be plural a)

t. *kitaplar* ⊢ n : λa.and (be book a) (be plural a)

u. *defterler* ⊢ n : λa.and (be notebook a) (be plural a)

v. *ev* ⊢ n : λa.(be home a)

w.*okul* ⊢ n : λa.(be school a)

x. *kalem* ⊢ n : λa.(be pen a)

y. *evi* ⊢ nACC : λa.(be home a)

z. *okulu* ⊢ nACC : λa.(be school a)

{. *kalemi* ⊢ nACC : λa.(be pen a)

|. *evim* ⊢ n : λa.and (own a speaker) (be home a)

}. *okulum* ⊢ n : λa.and (own a speaker) (be school a)

~. *kalemim* ⊢ n : λa.and (own a speaker) (be pen a)

-. *evlerim* ⊢ n : λa.and (own a speaker) (and (be home a) (be plural a))

Ă.*okullarım* ⊢ n : λa.and (own a speaker) (and (be school a) (be plural a))

Ą.*kalemlerim* ⊢ n : λa.and (own a speaker) (and (be pen a) (be plural a))

Ć.*evimi* ⊢ nACC : λa.and (own a speaker) (be home a)

Č.*okulumu* ⊢ nACC : λa.and (own a speaker) (be school a)

Ď.*kalemimi* ⊢ nACC : λa.and (own a speaker) (be pen a)

Ě.*mobilya* ⊢ n : λa.(be furniture a)

Ę.*servis* ⊢ n : λa.(be cutlery a)

Ğ.*kamyon* ⊢ n : λa.(be truck a)

Ĺ.*mobilyacı* ⊢ n : λa.and (be furniture a) (sell b a)

Ľ.*servisçi* ⊢ n : λa.and (be cutlery a) (sell b a)

Ł.*kamyoncu* ⊢ n : λa.and (be truck a) (sell b a)

Ń.*servis* ⊢ n : λa.(be shuttle a)

Ň.*taksi* ⊢ n : λa.(be taxi a)

■. *taksici* ⊢ n : λa.and (be taxi a) (drive b a)

Ő.*servisçi* ⊢ n : λa.and (be shuttle a) (drive b a)

Ŕ.*kamyoncu* ⊢ n : λa.and (be truck a) (drive b a)

Ř.*taksiciler* ⊢ n : λa.and (and (be taxi a) (drive b a)) (be plural a)

Ś.*servisçiler* ⊢ n : λa.and (and (be shuttle a) (drive b a)) (be plural a)

Š.*kamyoncular* ⊢ n : λa.and (and (be truck a) (drive b a)) (be plural a)

Ş.*taksicileri* ⊢ nACC : λa.and (and (be taxi a) (drive b a)) (be plural a)

Ť.*servisçileri* ⊢ nACC : λa.and (and (be shuttle a) (drive b a)) (be plural a)

Ţ.*Atatürk* ⊢ n : λa.(be Atatürk a)

Ű.*Epikür* ⊢ n : λa.(be Epicurus a)

Ů.*Aristo* ⊢ n : λa.(be Aristoteles a)

Ÿ.*Atatürkçü* ⊢ n : λa.and (be Atatürk a) (believe (in b) a)

Ź.*Epikürcü* ⊢ n : λa.and (be Epicurus a) (believe (in b) a)

Ž.*Aristocu* ⊢ n : λa.and (be Aristoteles a) (believe (in b) a)

Ż.*el* ⊢ n : λa.(be hand a)

IJ.*baş* ⊢ n : λa.(be head a)

İ. *elle* ⊢ v\n : λx1λx2λx3.and (be hand x3) (do (with x3) sth x1 x2)

đ. *başla* ⊢ v\n : λx1λx2λx3.and (be head x3) (do (with x3) sth x1 x2)

§. *miyav* ⊢ n : λa.(be meow_sound a)

ă. *hav* ⊢ n : λa.(be bark_sound a)

ą. *miyavla* ⊢ v : λx1λx2λx3.and (be meow_sound x3) (sound (like x3) x1 x2)

ć. *havla* ⊢ v : λx1λx2λx3.and (be bark_sound x3) (sound (like x3) x1 x2)

č. *temiz* ⊢ n/n : λx1λx2.and (x1 x2) (be clean x2)

ď.*hazır* ⊢ n/n : λx1λx2.and (x1 x2) (be ready x2)

ě. *temizle* ⊢ v\n: λx1λx2λx3λx4λx5.and (and (x4 x5) (be clean x5)) (make (like x4) x1 x2 x3)

ę. *hazırla* ⊢ v\n: λx1λx2λx3λx4λx5.and (and (x4 x5) (be ready x5)) (make (like x4) x1 x2 x3)

ğ. *miyavladım* ⊢ s : λx1λx2λx3.and (be speaker x1) (and (x2 < tref) (and (be meow_sound x3) (sound (like x3) x1 x2)))

Í. *havladım* ⊢ s : λx1λx2λx3.and (be speaker x1) (and (x2 < tref) (and (be bark_sound x3) (sound (like x3) x1 x2)))

B.2.2 CT6

(212) Observation List for CT6

a. *defter* ⊢ n : λa.(be notebook a)

b. *gömlek* ⊢ n : λa.(be shirt a)

c. *kitap* ⊢ n : λa.(be book a)

d. *su* ⊢ n : λa.(be water a)

e. *odun* ⊢ n : λa.(be wood a)

f. *defterci* ⊢ n : λaλb.and (be notebook b) (sell b a)

g. *gömlekçi* ⊢ n : λaλb.and (be shirt b) (sell b a)

h. *kitapçı* ⊢ n : λaλb.and (be book b) (sell b a)

i. *defterler* ⊢ n : λa.and (be notebook a) (be plural a)

j. *gömlekler* ⊢ n : λa.and (be shirt a) (be plural a)

k. *kitaplar* ⊢ n : λa.and (be book a) (be plural a)

l. *suluk* ⊢ n : λaλb.and (be water b) (contain b a)

m. *kitaplık* ⊢ n : λaλb.and (be book b) (contain b a)

n. *odunluk* ⊢ n : λaλb.and (be wood b) (contain b a)

o. *defter geldi* ⊢ s : λa.and (be notebook a) (came a)

p. *gömlek geldi* ⊢ s : λa.and (be shirt a) (came a)

q. *kitap geldi* ⊢ s : λa.and (be book a) (came a)

r. *defterci geldi* ⊢ s : λaλb.and (and (be notebook b) (sell b a)) (came a)

s. *gömlekçi geldi* ⊢ s : λaλb.and (and (be shirt b) (sell b a)) (came a)

t. *kitapçı geldi* ⊢ s : λaλb.and (and (be book b) (sell b a)) (came a)

u. *defterler geldi* ⊢ s : λa.and (and (be notebook a) (be plural a)) (came a)

v. *gömlekler geldi* ⊢ s : λa.and (and (be shirt a) (be plural a)) (came a) w.*kitaplar geldi* ⊢ s : λa.and (and (be book a) (be plural a)) (came a)

x. *suluk geldi* ⊢ s : λaλb.and (and (be water b) (contain b a)) (came a)

y. *kitaplık geldi* ⊢ s : λaλb.and (and (be book b) (contain b a)) (came a)

z. *odunluk geldi* ⊢ s : λaλb.and (and (be wood b) (contain b a)) (came a)

{. *geldi* ⊢ s\n : λa.(came a)

|. *gitti* ⊢ s\n : λa.(went_away a)

}. *düştü* ⊢ s\n : λa.(fell a)

~. *geldi defter* ⊢ s : λa.and (be notebook a) (came a)

-. *gitti defter* ⊢ s : λa.and (be notebook a) (went_away a)

Ă.*düştü defter* ⊢ s : λa.and (be notebook a) (fell a)

Ą.*geldi defterci* ⊢ s : λaλb.and (and (be notebook b) (sell b a)) (came a)

Ć.*gitti defterci* ⊢ s : λaλb.and (and (be notebook b) (sell b a)) (went_away a)

Č.*düştü defterci* ⊢ s : λaλb.and (and (be notebook b) (sell b a)) (fell a)

Ď.*geldi defterler* ⊢ s : λa.and (and (be notebook a) (be plural a)) (came a)

Ě.*gitti defterler* ⊢ s : λa.and (and (be notebook a) (be plural a)) (went_away a)

Ę.*düştü defterler* ⊢ s : λa.and (and (be notebook a) (be plural a)) (fell a)

Ğ.*geldi kitaplık* ⊢ s : λaλb.and (and (be book b) (contain b a)) (came a)

Ĺ.*gitti kitaplık* ⊢ s : λaλb.and (and (be book b) (contain b a)) (went_away a)

Ľ.*düştü kitaplık* ⊢ s : λaλb.and (and (be book b) (contain b a)) (fell a)

Ł.*defter yok* ⊢ s : λa.and (be notebook a) (not_exist a)

Ń.*gömlek yok* ⊢ s : λa.and (be shirt a) (not_exist a)

Ň.*kitap yok* ⊢ s : λa.and (be book a) (not_exist a)

■. *doğum yerim* ⊢ n : λa.(be my_birth_place a)

Ő.*memleketim* ⊢ n : λa.(be my_hometown a)

Ŕ.*nüfus kaydım* ⊢ n : λa.(be my_civil_registry a)

Ř.*doğum yerim Cape Town* ⊢ s : λa.and (be my_birth_place a) (be Cape_Town a)

Ś.*memleketim Cape Town* ⊢ s : λa.and (be my_hometown a) (be Cape_Town a)

Š. *nüfus kaydım Cape Town* ⊢ s : λa.and (be my_civil_registry a) (be Cape_Town a)

Ş. *defterler geldi yan yana* ⊢ s : λa.and (and (and (be notebook a) (be plural a)) (came a)) (side_by_side a)

Ť. *gömlekler geldi yan yana* ⊢ s : λa.and (and (and (be shirt a) (be plural a)) (came a)) (side_by_side a)

Ţ. *kitaplar geldi yan yana* ⊢ s : λa.and (and (and (be book a) (be plural a)) (came a)) (side_by_side a)

Ű. *ev* ⊢ n : λa.be home a

Ů. *okul* ⊢ n : λa.be school a

Ÿ. *kalem* ⊢ n : λa.be pen a

Ź. *evi* ⊢ nACC : λa.be home a

Ž. *okulu* ⊢ nACC : λa.be school a

Ż. *kalemi* ⊢ nACC : λa.be pen a

Ĳ. *evim* ⊢ n : λa.and (own a speaker) (be home a)

İ. *okulum* ⊢ n : λa.and (own a speaker) (be school a)

đ. *kalemim* ⊢ n : λa.and (own a speaker) (be pen a)

§. *evimi* ⊢ nACC : λa.and (own a speaker) (be home a)

ă. *okulumu* ⊢ nACC : λa.and (own a speaker) (be school a)

ą. *kalemimi* ⊢ nACC : λa.and (own a speaker) (be pen a)

ć. *iş* ⊢ n : λa.(be job a)

č. *evsiz kal* ⊢ v : λaλb.and (be home b) (become (without b) a)

ď. *işsiz kal* ⊢ v : λaλb.and (be job b) (become (without b) a)

ě. *bul* ⊢ v\n : λaλbλt.find a b t

ę. *ye* ⊢ v\n : λaλbλt.eat a b t

ğ. *gör* ⊢ v\n : λaλbλt.see a b t

Ĺ. *evi bul* ⊢ v : λaλbλt.and (be home a) (find a b t)

ľ. *evi ye* ⊢ v : λaλbλt.and (be home a) (eat a b t)

ł. *evi gör* ⊢ v : λaλbλt.and (be home a) (see a b t) ń. *evi buldu* ⊢ s/n : λaλbλt.and (t < tref) (and (be home a) (find a b t))

ň. *evi yedi* ⊢ s/n : λaλbλt.and (t < tref) (and (be home a) (eat a b t))

■. *evi gördü* ⊢ s/n : λaλbλt.and (t < tref) (and (be home a) (see a b t))

ő. *evi buldum* ⊢ s : λaλbλt.and (be speaker b) (and (t < tref) (and (be home a) (find a b t)))

ŕ. *evi yedim* ⊢ s : λaλbλt.and (be speaker b) (and (t < tref) (and (be home a) (eat a b t)))

ř. *evi gördüm* ⊢ s : λaλbλt.and (be speaker b) (and (t < tref) (and (be home a) (see a b t)))

ś. *evi bulduydu* ⊢ s/n : λaλbλt.and (tref < t0) (and (t < tref) (and (be home a) (find a b t)))

š. *evi yediydi* ⊢ s/n : λaλbλt.and (tref < t0) (and (t < tref) (and (be home a) (eat a b t)))

ş. *evi gördüydü* ⊢ s/n : λaλbλt.and (tref < t0) (and (t < tref) (and (be home a) (see a b t)))

t'. *evi bulduydum* ⊢ s : λaλbλt.and (be speaker b) (and (tref < t0) (and (t < tref) (and (be home a) (find a b t))))

ţ. *evi yediydim* ⊢ s : λaλbλt.and (be speaker b) (and (tref < t0) (and (t < tref) (and (be home a) (eat a b t))))

ű. *evi gördüydüm* ⊢ s : λaλbλt.and (be speaker b) (and (tref < t0) (and (t < tref) (and (be home a) (see a b t))))

B.2.3 CT7

The base observation list for CT7 includes the union of observation lists of all other trials. The base list contains 508 items. Observation lists for 5 randomized trials are generated by sampling with replacement from the base list 5080 times.

CURRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: Kunter, Utku Can

EDUCATION

Degree	Institution	Year of Graduation
M.S.	METU Department of Industrial Engineering	M.S. 2015
B.S.	METU Department of Industrial Engineering	B.S. 2013
High School	Ankara Atatürk Anadolu Lisesi	2008

PROFESSIONAL EXPERIENCE

Year	Place	Enrollment
2021-	Roketsan A.Ş.	HRM Solutions Manager
2016-2021	Roketsan A.Ş.	ERP & Data & System Analyst
2015-2016	University of Chicago Booth School of Business	PhD Student, Management Sc.
2013-2015	METU Department of Industrial Engineering	Research Assistant

PUBLICATIONS

International Conference Publications

U. C. Kunter, G. N. Özdemir, and C. Bozşahin. *Distributional and lexical exploration of semantics of derivational morphology*. ISBCS, Online, May 31 2020.

U. C. Kunter, and C. Bozşahin. *CCG for Turkish finite verb inflection*. ISBCS, Istanbul, Turkey, May 6 2018.

U. C. Kunter. *City logistics system design under cost uncertainty*. MS Thesis, Middle East Technical University, 2015.

U. C. Kunter, C. İyigün, and H. Süral. *City logistics system design under cost uncertainty*. INFORMS Annual Meeting 2015, Philadelphia, USA, November 1-4 2015.

U. C. Kunter, C. İyigün, and H. Süral. *City logistics system design under cost uncertainty*. 2015 TSL Workshop, Berlin, Germany, July 6-8 2015.

U. C. Kunter, C. İyigün, and H. Süral. *Handling travel time uncertainty in city logistics*. INFORMS Annual Meeting 2014, San Francisco, USA, November 9-12 2014.