IDIOMS AS MULTI-WORD EXPRESSIONS IN TURKISH

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

BY

ARZU BURCU GÜVEN

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

OCTOBER 2020

Approval of the thesis:

IDIOMS AS MULTI-WORD EXPRESSIONS IN TURKISH

submitted by **ARZU BURCU GÜVEN** in partial fulfillment of the requirements for the degree of **Master of Science in Cognitive Science Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** _____

Prof. Dr. Cem Bozşahin
Head of Department, **Cognitive Science** _____

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

Examining Committee Members:

Assist. Prof. Dr. Umut Özge
Cognitive Science Dept., METU _____

Prof. Dr. Cem Bozşahin
Cognitive Science Dept., METU _____

Assist. Prof. Dr. Burcu Can
RGLC, University of Wolverhampton _____

Date:

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: Arzu Burcu Güven

Signature :

ABSTRACT

IDIOMS AS MULTI-WORD EXPRESSIONS IN TURKISH

Güven, Arzu Burcu
M.S., Department of Cognitive Science
Supervisor: Prof. Dr. Cem Bozşahin

October 2020, 46 pages

Idioms constitute several challenges for both Natural Language Processing (NLP) and linguistic analysis. A better understanding of idioms will yield valuable insights about natural language as well as the way it is processed. The relevance of idioms, along with the fact that Turkish is a rather unexplored language from this perspective, motivates us to work on Turkish idioms. Here, we aim to demonstrate a grammatical study on Turkish idioms that were selected in accordance with distributional models.

Keywords: MWE, idiom, CCG, distributional semantics, word embedding

ÖZ

TÜRKÇEDE ÇOK SÖZCÜKLÜ İFADELER OLARAK DEYİMLER

Güven, Arzu Burcu
Yüksek Lisans, Bilişsel Bilimler Bölümü
Tez Yöneticisi: Prof. Dr. Cem Bozşahin

Ekim 2020 , 46 sayfa

Deyimler, hem Doğal Dil İşleme hem de dilbilimsel analiz için önemli bir problem teşkil eder. Deyimlerin daha iyi anlaşılması, doğal dilin hem kendisi hem de işlenmesi açısından değerli içgörüler elde etmemizi sağlayacaktır. Türkçe deyimler üzerine çalışmamızın arkasında, deyimlerin güncel öneminin yanı sıra Türkçe için bu açıdan bakıldığında henüz keşfedilmemiş alanlar olması gibi motivasyonlar var. Bu çalışmada, dağılımsal anlambilimsel modellere göre belirlenen Türkçe deyimler üzerine yapılan bir gramer çalışması sunmayı hedefliyoruz.

Anahtar Kelimeler: ÇSİ, deyim, eylem, türkçe

To my family

ACKNOWLEDGMENTS

I wish to express my gratitude, first and foremost, to Prof. Dr. Cem Bozşahin, who oversaw my transformation from a bachelor of psychology and literature to a fiercely motivated student of linguistics with dedication and diligence. I feel greatly indebted for being allowed a cordial access to his sagacity, profound insights and prolific mind.

I would also like to thank Assist. Prof. Umut Özge, who was previously my main thesis advisor and an admired instructor of a number of courses I undertook. He has also been a regular at the bi-weekly thesis/project meetings we have held with Prof. Dr. Bozşahin and Dr. Özkan Aslan. His brilliant and sophisticated mind along with his inexorable yearning for excellence has guided me in the right direction.

Special thanks are due to Dr. Özkan Aslan and Prof. Dr. Burcu Can. They are among the people who have made this thesis possible thanks to their fascinating observations and insights in the MWE project.

I thank my significant other Apo, who has been there for me from the beginning, in the form of unconditional emotional, intellectual and logistic support.

I also thank my younger sister Berna, who has taken over the role of a compassionate older sister who is always loving and supportive, along with my father Adil who has empowered and encouraged me throughout my endeavors as a scholar in training.

It would be unfair for me to not mention my dear friends and colleagues Emre Erçin and Ahmet Üstün for their invaluable insights and support throughout the process.

I would like to thank all the students and colleagues I have crossed paths with throughout my years at the institute, those who have contributed a great deal to my social and intellectual evolution.

Finally, I wish to thank my co-workers who had to endure the consequences of my sporadic absence from work and have been very understanding and supportive regardless.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

Cognitive science is tasked with examining the most prominent and complex characteristics of the human kind. Language and adaptability are two of the numerous characteristics that have rendered our species perhaps the most complicated, sophisticated and intriguing 'things' in the entire universe. Moreover, apart from its role in interacting with other characteristics and facilitating their development; language is an essential, if not the essential tool for us to conjure up new strategies for reinforcing our ability to adapt to different environments and conditions. Likewise, we also like to bend and change language to adapt it to whatever needs and whims we may have. We can expand our lexicon, add new meanings to a lexicon entry, exchange words between different lexicons, etc. The list of adaptations we can and do apply to language goes on. One rather idiosyncratic adaptation is to combine two or more entries but give them a new meaning that would not be available from the constituents. We call the object of this practice *idiom*. As they stand on the intersection between language and adaptability, idioms provide a multi-faceted examination topic for cognitive science. These qualities attracted the interest of different domains of study including psychology, linguistics and computer science. In this thesis, we are only able to examine limited aspects of this phenomenon. Before that however, a review of the how various domains approach idioms is necessary due to controversial levels of confusion regarding their classification.

From the Natural Language Processing (NLP) perspective, idioms are generally assigned into a group called Multi-word Expressions (MWEs) which is a class that includes many different types of collocations such as multi-word terms (e.g., *corpus callosum*) or phrasal verbs (e.g., *pick the book up*). Further complicating the unwelcome diversity, idioms are probably the worst kind of MWEs. There are multiple bases for this allegation, most prominent of which is that they are not static like most MWEs. This prevents us from adding them to the lexicon directly as a longer word, nor can we merely use statistics to write them off (Sag et al., 2002) They require full attention of various methods such as lexicons and statistics, as well as grammatical and semantic techniques (Sag et al., 2002).

Psychological research on idioms has focused on idiom comprehension. While early approaches considered idioms as non-compositional units stored in lexicon like words, i.e words with spaces (Bobrow and Bell, 1973; Swinney and Cutler, 1979), this view have since been challenged by later studies. Idioms themselves are argued to be a heterogeneous class with respect to their compositionality (Gibbs et al., 1989). Various theories of representation challenged the idea of non-compositional, words-with-spaces representation of idioms, pointing out that both compositional and non-

compositional strategies are used in comprehension (Cacciari and Tabossi, 1988; Caillies and Butcher, 2007). Empirical results from neuropsychological and psycholinguistic studies support the idea that idiom comprehension is linked to a complex mechanism (Cacciari and Papagno, 2012), challenging the idea that their only difference from regular lexicon entries is their length. Overall, it has been demonstrated that idioms, standing in between word comprehension and sentence comprehension, constitute an interesting niche that would help us understand language comprehension better.

From the perspective of a linguist, idioms represent a puzzling plane of language. Idioms obscure the definitions of words. They may look similar to words in terms of semantics as they both map to a single referent but they are also dissimilar as far as syntax and lexicography go (Baldwin and Kim, 2010). Earlier approaches to idioms were reluctant to incorporate them into their syntactic theory, and proposed to solve this puzzle through post-processing methods such as idiom lists. These lists would contain idiom elements with restrictions on their syntactic productivity (Weinreich, 1969; Fraser, 1970). Although it was followed by many different approaches since, the idea of considering idioms as both syntactically and semantically productive phenomena first came into fruition with Nunberg et al.'s (1994) seminal paper. They argued that different types of idioms displayed different syntactic and semantic behavior, and that it should be possible to explain these observations with grammar.

Nunberg et al. (1994) argues that the syntactic behavior of idioms reflect their semantics in a systematic way. In the case of syntactically active idioms, idiomatic heads can select for a specific component to realize. Following this insight, Bozşahin and Güven (2018) provided a framework to capture the behavior of idioms using Categorical Combinatory Grammar (CCG). This thesis aims to put this framework into practice by assigning CCG categories to Turkish verb-noun idioms. Our task consists of three steps; (i) explaining the type of idioms that are targeted, (ii) specifying the method for acquiring the said idioms, and lastly (iii) explaining the idiomatic categories.

1.1 The Outline of the Thesis

This thesis consists of three main chapters, first of which starts off with an overview of the challenges that prohibit a unanimous agreement on the definition of idioms as a concept. The elusive nature of idioms makes it difficult to distinguish them from MWEs and this impedes the task of isolating idioms that will form our dataset. In fact, instead of being divided by a clear boundary, idioms and MWEs appear to be distributed along a spectrum. Among others, the methods developed through distributional approaches stand out as viable options for extracting and identifying idioms. Specifically, we make use of the method worked out by Gong et al. (2017), which considers compositionality as a geometric relationship between a target phrase and its context. Having secured a mechanism for constructing our dataset, we move onto their syntactic and semantic representation. For this, we turn to Combinatory Categorial Grammar (CCG). Bozşahin and Güven's (2018) model, which utilizes CCG to arrive at a unified representation of syntax and semantics of idioms, is briefly introduced.

Third chapter reports the data selection process. First, we establish that metaphors are similar to literal phrases in terms of compositionality. Second, we demonstrate that idioms have lower compositionality when compared to both metaphors and literal phrases. This gap allows us to draw a line for isolating the idioms that ended up in our dataset.

Fourth chapter details a grammatical study with the idioms we selected. We demonstrate an application of Bozşahin and Güven's (2018) model to Turkish idioms, and then further address language specific complications. We conclude with a discussion about idiom semantics. Finally, in the last chapter, we conclude with an overview of our study and a discussion about future studies.

CHAPTER 2

BACKGROUND

2.1 Idioms

Idioms are a notoriously ill-defined group. They have been associated with two general categories that span other non-compositional language uses: figuratives and MWEs.[1] Figuratives include phenomena such as metonymy, metaphors, proverbs, hyperbole, irony and so on. MWEs, on the other hand, are used to refer to constructions such as multi-word terms, institutionalized phrases, phrasal verbs, light verbs and so on. While the former category is more concentrated on semantics, the latter is interested in word-like items that cross word boundaries. Idioms do relate to both due to their non-literal semantics and multi-word structure. However, in order for us to fully comprehend the unique characteristics of idioms, these categories themselves must be further deconstructed.

Figurativity has long been cited as a feature of idioms (Nunberg et al., 1994) and idioms are generally included with other figurative uses as non-literal language tools. Among the figurative uses, metaphors came to be especially confounding for idioms. One approach is to analyse idioms under the framework of conceptual metaphors (Lakoff and Johnson, 1980). Conceptual metaphors provide an overarching relationship between ideas and objects (eg. THE MIND IS A CONTAINER), and common metaphors can be found to be compliant with them as in the case of referring to ones' mind as *too full* or *empty*. Idioms also can be subjected to conceptual analysis. For example the idiom *spill the beans* is argued to be the result of the conceptual metaphors THE MIND IS A CONTAINER and IDEAS ARE ENTITIES (Gibbs et al., 1989; Gibbs, 1992); and such an analysis gives rise to the conclusion that idioms are metaphorical constructs. Although these metaphors may have played a role in the origination of idioms, it is an extrapolation to argue for metaphoricity of idioms in their current state. Unlike metaphors which are flexible and dependent on context, idioms are conventionalized and have fixed meanings. Putting aside the arguments regarding differences in behavior; bundling idioms with figuratives was also disruptive from an empirical perspective. Such misclassifications can cause problems for experimental stimuli as these distinct expression types were shown to be associated with different cognitive mechanisms (Cacciari and Papagno, 2012).

[1] Although idioms are generally subsumed under the categories of figuratives and MWEs, the term *idiom* can also be used to refer to these categories. Moon et al. (1998) outlines two versions of the term *idiom*, one sense is narrower and refers to compositionally opaque and conventionalized phrases while the other sense is broader and can be used to refer many different figuratives and MWEs.

Our work on idioms does not cover all possible types. We are only focused on idioms that consist of noun-verb pairs and also show low levels of compositionality. Thanks to this second criterion, it is possible to separate such idioms from metaphors. However, this is a topic we will address in section 3.1.3. For now, given that we focus on seemingly non-compositional idioms, we eschew the issue of metaphors and conceptual analysis for idioms. More transparent idioms -or conventionalized metaphors- can be better targets for this issue as there are many examples of phrases that are conventionalized like idioms but also involve visible metaphoric influences. For example, *gönül* 'heart' is used in many conventionalized phrases:

(1) a. Gönül ister ki yaz hiç bitmesin.
heart wants ki (that) summer never end-NEG.
'I wish that the summer never ends.'

b. Gönlümden geçeni söylemek istedim.
heart-POSS-LOC passing-ACC tell want-PAST
'I wanted tell what my heart wishes.'

Although these phrases are very conventialized, we see the repeating pattern of *gönül* representing human desire. Similarly, the words *ray* 'railroad' and *yol* 'road' represents *the proper/desirable course* in many conventionalized phrases.

(2) a. Hayatı raydan çıktı.
life-POSS railroad-ABL exit-PAST.
'Their life was in derailed.'

b. İlişkileri yolunda gitmiyordu.
relationship-POSS road-POSS-LOC go-NEG-PROG-PAST
'Their relationship was not going well.'

Different studies consider similar examples as idioms (Nunberg et al., 1994; Owens, 2016), yet whether they are actually idioms or simply conventionalized metaphors are up for debate[2]. We deliberately and with a systematic approach exclude such examples, and are therefore exempt from the metaphoricity debate.

The other group the idioms are associated with, the MWEs, are also difficult to pinpoint to an exhaustive definition. It is generally stated that they cross regular word boundaries and have somewhat word-like qualities. The cardinal problem with providing a definition for MWEs is that they do not represent a homogeneous category. MWEs include a variety of constructions with distinct syntactic and semantic features.

There is a bag of features commonly attributed to MWEs, namely collocation, word-like behavior, irregular compositionality and syntactic rigidity (Sag et al., 2002; Baldwin and Kim, 2010; Constant et al., 2017). These characteristics can be found in various types of MWEs, but with varying degrees. Whereas some MWE subtypes are

[2] Both Nunberg et al. (1994) and Owens (2016) provide examples of idiom families, where the same word participates in many different collocations. Although idiom families do not necessitate semantic similarity among their members, it is not uncommon.

quite non-compositional, others are closer to compositional than non-compositional. These different MWE subtypes are grouped together as MWEs not because they all share a specific set of features, but because they show varying degrees of the said features. In other words, they are distributed along various spectra of the aforementioned features (Moon et al., 1998), as opposed to exhibiting a binary relationship with the associated features.

To further explain this point, we can start with the feature called collocation. Collocation requires that for a phrase to be classified as a MWE, its component words must come together in a frequency that is statistically significant. There is a binary component to collocation, it is either significantly there or not. However, this alone cannot be used to justify MWE status for a phrase. Looking only at the co-occurrence frequency can yield useless pairs like *of the* or *in the*, but such examples are excluded because they neither behave like a word nor have an interesting compositionality pattern (Gries, 2008). Collocationality may not provide a strong claim for MWE status, but it provides a starting point for outlining the phenomenon.

The second feature associated with MWEs is word-like behavior. It is not as straightforward as collocation and the reason is that regular words themselves present a fuzzy category just like MWEs. Moreover, some of this fuzziness stems from the prevalence of MWEs itself (Baldwin and Kim, 2010). For simplicity's sake, we consider word like behavior of MWEs as denoting a single entity or event.

For MWEs referring to entities such as a multi-word term-borrowing the terminology of Constant et al. (2017)-(e.g., corpus callosum) or a multi-word named entity (e.g., Liquid Crystal Display (LCD)), it is readily apparent that these phrases denote single entities, just as regular words do. Furthermore, composition of their constituent words is not in the foreground; LCDs may have the properties that are explained by the words liquid, crystal and display, but the term bypasses these properties and refers directly to the entity itself.

For more idiomatic MWEs, we can generally observe them denoting a single event, but there are also many exceptions. When idioms are compliant with the word-like behavior argument (e.g., kick the bucket), we can observe that they can be substituted with a single word (e.g., die). Although having an alternate is not equivalent to having a synonym, it is a good indicator that such idioms denote a single event. We can also borrow such alternates from other languages. For example, the Turkish light verb construction *satın almak* can be translated to English as *to buy*, but there is no single word representation for it in Turkish. Word-like behavior does not apply to an entire class of idioms and they are called 'idiomatically combining expressions'. They will be discussed later.

Features of collocation and word like behavior help us distinguish MWEs from regular phrases. Compositionality and syntactic flexibility help us to do this as well, but they also play an important role in distinguishing different types of MWEs from one another. Compositionality is a property that requires the parts of a phrase to contribute to the general meaning of it. This property is generally fully realized in regular phrases but it is not always the case for MWEs. Different MWEs show different levels of compositionality. Institutionalized phrases such as *phone booth* are highly compositional but idioms such as *kick the bucket* are almost completely the

opposite. Apart from these examples that are almost polar opposites of each other; there are many MWEs that lay in between. Therefore, compositionality corresponds to a spectrum, albeit not a straightforward one.

Generally, it is assumed that if we can deduce the whole meaning of a phrase by composing its constituents, the phrase is compositional or transparent. If the constituents of the phrase do not yield the general meaning through composition, the phrase is assumed to be non-compositional or opaque. Compositionality as a continuum is also explored in the literature. According to several studies, human judgements as well as NLP methods provided results in favour of this perspective (Baldwin et al., 2003; McCarthy et al., 2003). Similarly, syntactic variability is the measure of how much syntactic modification the parts of an MWE can undergo, and how discontinuous an MWE can be.

Until now, we focused on properties of MWEs and the same properties apply to idioms. It is fair to say that idioms are more conservative compared to other MWEs with respect to compositionality and syntactic variability. However, idioms are not completely non-compositional or syntactically fixed, and they show variation among themselves.

Nunberg et al. (1994) divide idioms into two groups according to their compositionality and syntactic variability: idiomatically combining expressions (ICE) and idiomatic phrases (IP). The former group is described as compositional, and the latter as non-compositional. Compositionality of ICEs is still dissimilar to regular phrases, in fact, this attribution stems from Nunberg et al.'s (1994) special use of the term. With the term compositionality, they refer to the distribution of idiomatic meaning over parts of an idiom. In idiomatically combining expressions, parts of the expression correspond to parts of idiomatic meaning. A canonical example for this types of expressions is *spill the beans*. Here, *spill* corresponds to *reveal* and *beans* correspond to *secret*. We also see such expressions in Turkish; *çamur atmak* 'to throw mud' is an idiomatically combining expression where *mud* corresponds to *slander* while *to throw* can correspond to *to utter*.

Nunberg et al. (1994) emphasizes that syntactic variability correlates with compositionality. They argue that syntactically less rigid ICEs' behavior does not constitute an exception for idioms, and that it is rather a result of their semantics. We also observe more syntactic variability in Turkish ICEs; example (3) shows *çamur atmak* in a relative construction:

(3) Onun bana attığı bu çamur
His me throw-OP-POSS this mud
'The defamation he made against me'

So far, it has been argued that neither MWEs nor idioms are unified, homogenous groups. We support this argument by observing that they exhibit varying degrees of several features that are attributed to them. However, we can also consider the lack of clear-cut boundaries for these groups as a direct result of the nature of the phenomenon itself. According to Bolinger in (as cited in Moon et al. (1998), p.6):

> There is no clear boundary between an idiom and a collocation or between a collocation and a freely generated phrase–only a continuum with greater density at one end and greater diffusion at the other, as would be expected of a system where at least some of the parts are acquired by the later analysis of earlier wholes.

Similar to Bolinger, we conclude that MWEs cannot be made into specific sets of expressions that create strict groups. Even if such classifications are imposed, a chunk of lexical elements with valuable information would be under the risk of being left out in grey areas. As a result, we willingly refrain from advancing a definition for idioms and instead, hypothesize a compositionality and syntactic variability plane for MWEs:

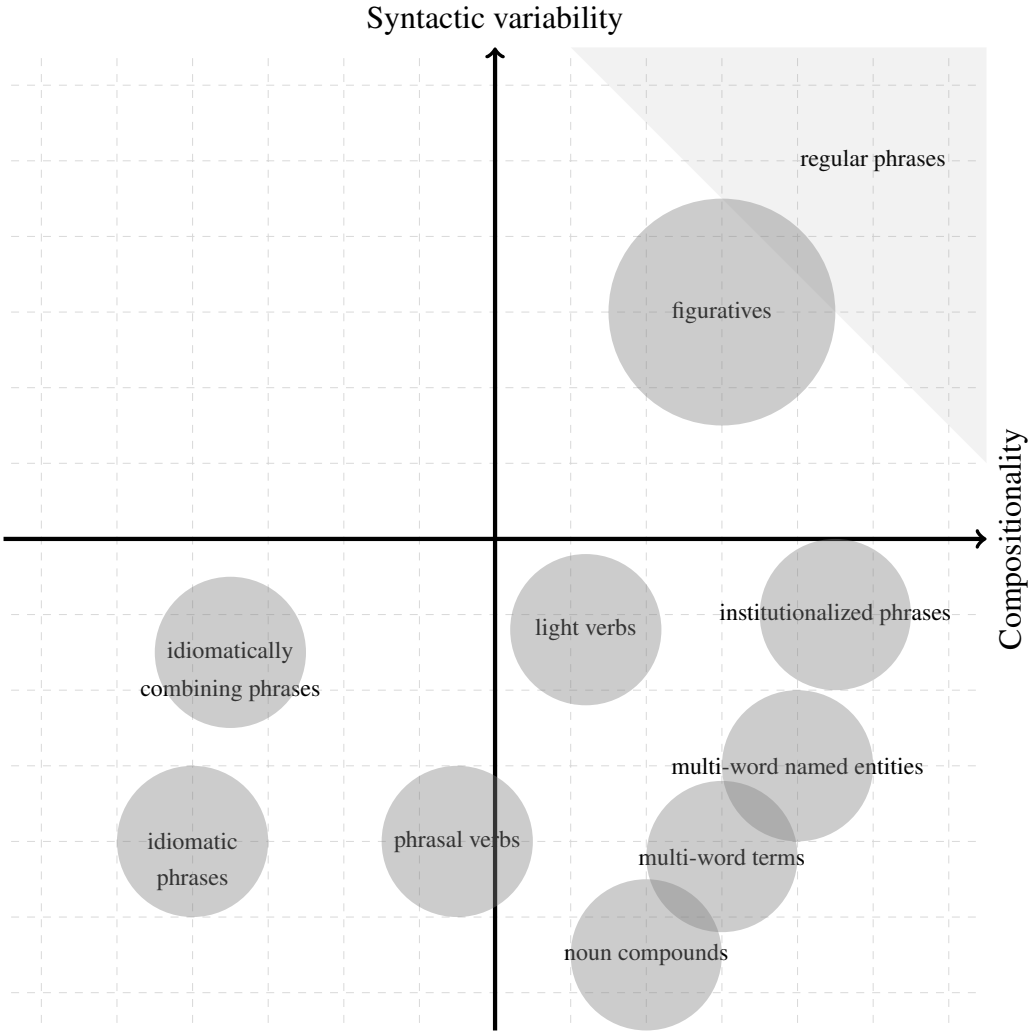


Figure 2.1: Approximate visualization of MWEs on the compositionality and syntactic variability axes.

These placements are very rough, however, fine-tuning and testing the accuracy of all

the hypothetical placements would be a task that goes far beyond the scope of this work. The aim here is to represent that idioms are outliers in terms of compositionality and support this claim through an analysis utilizing word embeddings. This is followed up with a grammatical study of idioms which aims to capture their so-called peculiar syntactic and semantic properties. Our focus group for this purpose, lies in the rather less compositional and more syntactically rigid area of this plane:

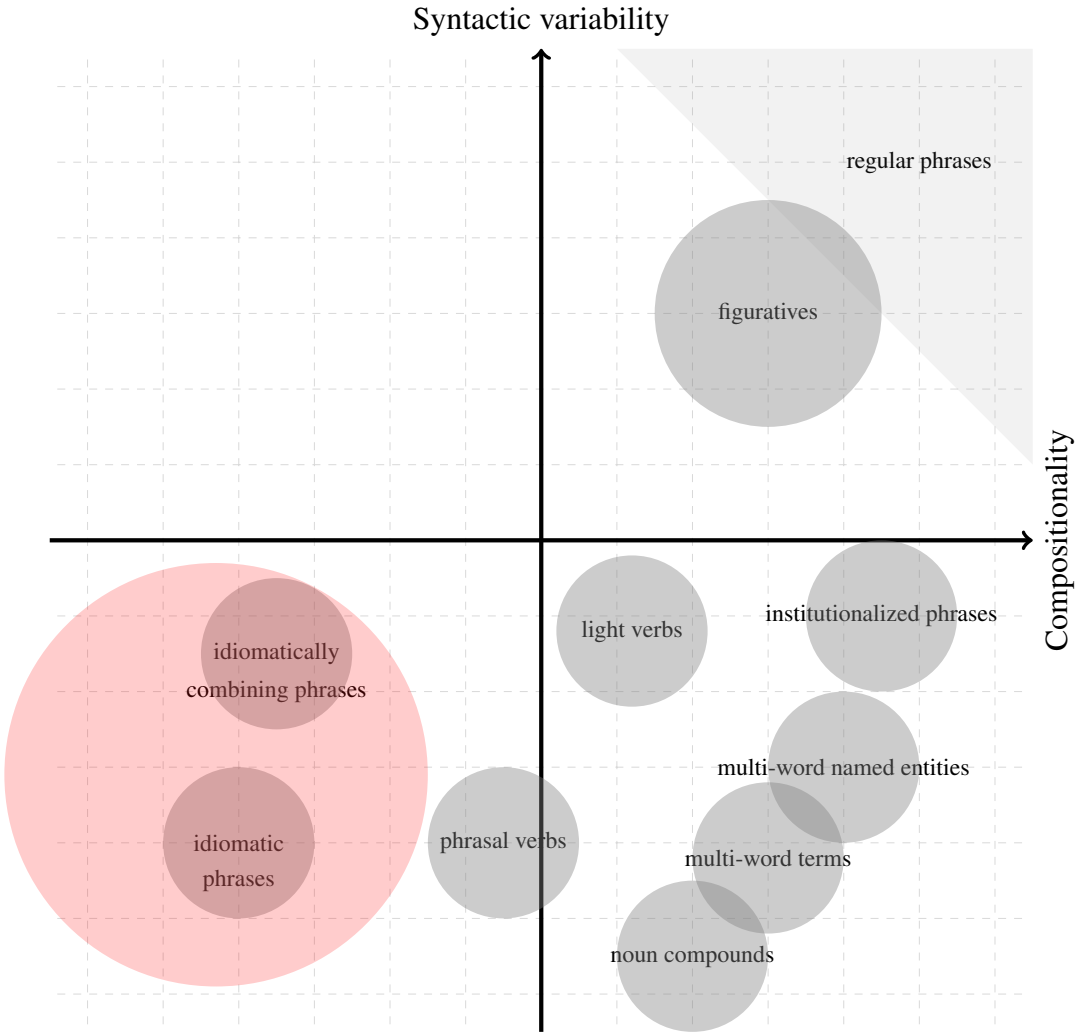


Figure 2.2: Approximate visualization of our focus area on the compositionality and syntactic variability axes.

2.1.1 Turkish Idioms and MWEs in Literature

Turkish idioms have been studied from different perspectives such as linguistics, psycholinguistics or NLP (Uzun, 1991; Adalı et al., 2016; Aydin et al., 2017; Berk et al., 2018). Although such studies are somewhat sparse in either domain, we will recount the most relevant and prominent ones in this section.

One of the earlier studies is Uzun's (1991) work on idiom forming and idiomaticity degrees with Turkish idioms. For the former, they argue that different types of figuratives such as metaphors or metonymy take part in idiom forming. They also provide three classes for idiomaticity, in which completely opaque idioms occupy the most idiomatic class, while the rest are assigned to second and third classes depending on how many of their components are able to preserve their literal or extended senses. Other studies have also looked at conceptual metaphor analysis with Turkish idioms (Ruhi and Işık-Güler, 2007).

For the NLP vein, there are a few studies that take on MWE extraction (Oflazer et al., 2004; Metin and Karaoğlan, 2010), and also a resource building effort for Turkish MWEs (Berk et al., 2018; Adalı et al., 2016). However, the way they approach idioms is quite different from ours, in the sense that we focus on compositionality, while their idiom selection criteria is more concerned with syntactic rigidity and continuous form.

2.2 Word Embedding

Although estimates regarding the exact age vary, it is well understood that children learn idioms at a later age than other words (Lodge and Leach, 1975; Kalandadze et al., 2018). Contextual effects on learning idioms, as well as new words, has been extensively documented (Levorato and Cacciari, 1992). The key is, the contextual clues that help children deduct the meaning of a word act in a different way when it comes to idioms. The context inevitably does give clues which help children deduct the meaning of an idiom, however it does not do so without cancelling the literal strategy first (Levorato and Cacciari, 1992).

To illustrate the process described in Levorato and Cacciari (1992), let us consider a sentence like *Break a leg Susan, I'm sure you'll give an outstanding performance* and assume that it is someone's first exposure to the phrase *break a leg*. In that instance, the part of the sentence that is about conveying good wishes seems to be at odds with the literal meaning of this phrase, which suggests physical violence. This mismatch between the context and the phrase signals to the child to stop using the literal strategy and try to interpret these words, although familiar, in a novel context with different meanings instead.

It is well established in psycholinguistic literature that contextual information affect not only the acquisition but also the comprehension of idioms (Caillies and Butcher, 2007). Given that both learning and comprehension of idioms are related to context, we can assume that context carries valuable information about them. A parallel insight is also utilized in studies that use distributional semantic models to extract, identify or disambiguate idioms and MWEs (Baldwin et al., 2003; McCarthy et al., 2003; Reddy et al., 2011; Peng and Feldman, 2015; Salehi et al., 2015; Gong et al., 2017). Such models exploit context to acquire semantic information about a phrase or word. There are also other studies that directly compare idiomatic and literal phrases to their local context and find differences in their relationship to their local context (Peng and Feldman, 2015; Gong et al., 2017).

One of these distributional semantic models, (neural) word embedding, is among the most popular methods (Lenci, 2018), and of particular importance to this study.

Word embedding is a method that involves mapping words into vectors and relies on neural networks. Although distributional models have a long history, prior to the recent surge of interest in the word embedding method, studies covering them had been relatively dormant (Mikolov et al., 2013). There were several reasons for this. To start with, NLP methods that preferred simple word representations as in the case of a single number representing a particular word have outperformed the then state-of-the-art distributional methods. Despite being able to deliver relatively successful models such as latent semantic analysis (LSA), distributional approaches have been overshadowed by the unprecedented rise in the popularity of word embeddings (Lenci, 2018).

Although the merits of Mikolov et al.'s (2013) so-called breakthrough model has been called into question (Lenci, 2018), this seminal work nevertheless facilitated the word embedding techniques' rise to stardom as far as semantic analysis methods go. Essentially, these methods focus on the context of a target word or the words surrounding it, and consider their relationship with the target word in order to create embeddings. Through the integration of neural networks, Mikolov et al. (2013) provides a faster way to create word vectors.

The Word2vec model treats every word as a token and analyses them through a modifiable window size. This can be better explained through an example of a possible model. Let us assume that only the sentence *Break a leg Susan, I'm sure you'll give an outstanding performance* was provided as the training data and window size was set to five, and, to keep things simple, spaces are used as the only signal of tokenization, Such a model would complete the training in seven iterations. Iterations of the complete model would look something like this:

(4) [1] Break a leg$_{\text{[target word]}}$ Susan, I'm

[2] a leg Susan$_{\text{[target word]}}$, I'm sure

[3] leg Susan I'm$_{\text{[target word]}}$ sure you'll

[4] Susan I'm sure$_{\text{[target word]}}$ you'll give

[5] I'm sure you'll$_{\text{[target word]}}$ give an

[6] sure you'll give$_{\text{[target word]}}$ an outstanding

[7] you'll give an$_{\text{[target word]}}$ outstanding performance

For training, the word2vec model can make use of several learning algorithms; two of the most common ones are the Continuous Bag of Words (CBOW) and Skip-Gram approaches. These two approaches operate in a similar but opposite way. Simply put, the CBOW approach uses the surrounding words to predict the target word. Skip-gram approach, on the other hand, uses the target word to predict the surrounding words. Mikolov et al. (2013) showed that the skip-gram method performs better than the CBOW method.

In summary, the parallelism between psycholinguistic and NLP studies emphasizes the importance of context for idioms (Levorato and Cacciari, 1992; Peng and Feldman, 2015; Gong et al., 2017). This convergence encouraged us to utilize word embeddings in our study of idioms as well. The next section includes a review of other studies that tackle MWEs with distributional models, and in the third chapter the method adapted in this study is discussed.

2.2.1 Related Work

Before the recent surge in the popularity of word embedding, distributional approaches were also used for extracting and classifying MWEs. There were a number of studies that used different distributional models, with some later studies making use of the word embedding techniques. Baldwin et al. (2003) and Schone and Jurafsky (2001) use latent semantic analysis (LSA) to disambiguate MWE types and MWE extraction, respectively. In this semantic analysis technique, words are represented as vectors. Baldwin et al. (2003), on the other hand, in an effort to distinguish MWEs according to their degree of compositionality adapted Nunberg et al.'s (1994) framework that divides idioms into groups of ICE and IP. In Baldwin et al.'s (2003) version, MWEs are divided into three groups based on their degree of compositionality: non-decomposable MWEs, idiosyncratically decomposable MWEs and simple decomposable MWEs.

Non-decomposable MWEs roughly correspond to IPs, idiosyncratically decomposable category to ICEs, and the last category corresponds to institutionalized phrases. They focus on noun compounds and verb particle constructions, and aim to find differences among the groups described above. Lastly, they compare their results with WordNet similarity results, which yields moderate correlations. Using WordNet similarity results in order to evaluate results obtained from distributional methods is common Baldwin et al. (2003), yet there are also studies that create their own gold standard test sets (McCarthy et al., 2003; Reddy et al., 2011).

Among the more recent studies, Salehi et al. (2015) uses the word2vec model to estimate the compositionality of MWEs (specifically, noun compounds and phrasal verbs from English and German). They look for any similarities between the MWE vector and the vectors of the words that constitute it, as well as the constituent words' average vector.

The Word2Vec model is also useful for disambiguating different uses of MWEs. Peng and Feldman (2015) takes advantage of word embeddings to disambiguate between literal and idiomatic instances of a pre-selected list of idioms from a corpus. Relying on the assumption that these instances would occur in different contexts, they employ word embeddings to capture the differences between each instance. There are also studies directly interested in MWE extraction that take advantage of distributional approaches (Schone and Jurafsky, 2001).

2.3 Combinatory Categorial Grammar (CCG)

Combinatory Categorial Grammar (CCG) is a radically lexicalized grammar formalism that is governed by combinatory rules and principles (Steedman, 2000; Steedman and Baldridge, 2011; Bozşahin, 2012). Radical lexicalization refers to the idea that lexical items include syntactic categories which would roughly correspond to CFG's rewrite rules. Handling the language specific information in lexical categories enables CCG to be a universally applicable mechanism.

(5) called := *(S\NP)/NP:* $\lambda x \lambda y.called'xy$

(6) S $\rightarrow$ NP VP
VP $\rightarrow$ TV NP
TV $\rightarrow$ *called*

In this sense, CCG is similar to a CFG grammar written in accepting (as opposed to producing) notation (Steedman and Baldridge, 2011). Indeed, pure CG grammars are examples of this. CCG represents the phonological, syntactic and semantic information of a lexical item with a single category.

CCG categories can be divided into two types; primitive categories and function categories (Steedman, 2000; Steedman and Baldridge, 2011). Primitive categories correspond to nouns, noun phrases, prepositions or sentences. They do not include information regarding categories other than themselves. Function categories, on the other hand, determine the type of their arguments and results, as well as the subcategorization order. They can correspond to structurally demanding lexical items such as verbs or relativizers.

(7) Primitive categories:

Mary := *NP: mary'*

β := *B:* β*'*

(8) Complex categories:

called := *(S\NP)/NP:* $\lambda x \lambda y.called'xy$

α := *A\B:* $\lambda x.\alpha' x$

One of the few combinatory rules CCG employs is Function Application (FA), which is also the only combinatory rule of pure CG grammars. It provides the basic mechanism with which we can apply arguments to functions:

(9) Forward Application (>)

$X/Y : f \quad Y : a \quad \Rightarrow \quad X : fa$

(10) Backward Application ($<$)

$Y : a \quad X\backslash Y : f \quad \Rightarrow \quad X : fa$

In the examples above, 'X' represents the range and 'Y' represents the domain. Both of these can be arguments or function categories. Apart from Function Application, CCG employs more combinators to capture the full complexity of natural languages. Combinator B is used to compose functions before their arguments are applied.

(11) Forward Composition ($>$B)

$X/Y : f \quad Y/Z : g \quad \Rightarrow \quad X/Z : \lambda x.f(gx)$

(12) Backward Composition ($< B$)

$Y\backslash Z : g \quad X\backslash Y : f \quad \Rightarrow \quad X\backslash Z : \lambda x.f(gx)$

Combinator T is used for type-raising i.e.turning arguments into functions over functions that subcategorize for them.

(13) Forward Type Raising ($>$T)

$X : a \Rightarrow \quad T/(T\backslash X) : \lambda f.f(a)$

(14) Backward Type Raising ($< T$)

$X : a \Rightarrow \quad T\backslash(T/X) : \lambda f.f(a)$

We can observe these rules in a derivation with lexical categories:

(15)
$$\dfrac{\dfrac{\dfrac{\text{John}}{S/(S\backslash NP) : \lambda p.pjohn'} \quad \dfrac{\text{called}}{(S\backslash NP)/NP : \lambda x\lambda y.called'xy}}{S/NP : \lambda x.called'xjohn'}\,{>}\mathbf{B} \quad \dfrac{\text{Mary}}{S\backslash(S/NP) : \lambda p.pmary'}}{S : called'mary'john'}\,{>}$$

All the combinatory rules are subject to the *principle of type transparency* which states that semantic composition works in parallel with syntactic composition.

2.3.1 MWEs with CCG

Nunberg et al. (1994) provides several brief proposals for grammarians looking to study idioms. Most of these were formulated as a response to the predominant transformational paradigm at the time. One of them, namely their approach to exceptional German idioms that syntactically behave like ICEs but semantically behave like IPs is of special interest to us. One of the examples they consider is shown below:

(16) beisst ins Gras
bites into the grass
'dies'

They observe that these examples can undergo some syntactic modification, and can be found in discontinuous form even though they are semantically categorized as IPs. As can be seen in example 16, *beisst ins Gras* does not distribute a semantic meaning onto its components. In response to this asymmetrical behavior, Nunberg et al. (1994) suggest that "*beisst ins Gras* could be treated as a lexical combination of an idiomatic *beisst* that selected for *ins Gras* as its complement." Moreover, they argue that the strict constraints regarding word order in the English language leads to IPs having a fixed or very minimally modifiable form. For less constrained languages such as German, however, IPs are allowed more freedom in terms of syntactic variability.

Nunberg et al.'s (1994) analysis provides a basic insight for grammatical representation of idioms. Bozşahin and Güven's (2018) work on idiomatic categories improves this basic insight in the sense that idiomatic heads similar to *beisst* are utilized to carry the idiomatic meaning and subcategorize for the complements. Example (17) shows the category for literal *beisst*, while (18) exemplifies the idiomatic one.

(17) beisst := *(S\NP)/NP:* $\lambda x \lambda y.beisst'xy$

(18) beistt := *(S\NP)/"ins Grass":* $\lambda x \lambda y.die_x'y$

The idiomatic category follows the analysis of Bozşahin and Güven (2018) They state that IPs, ICEs and phrasal verbs can be represented by lexical categories without disturbing the overall CCG rules and principles. Because CCG is radically lexicalized and it employs transparent derivation, using lexical categories to accommodate idiomatic expressions without disturbing the overall mechanism is permitted. Essentially, this is done by two modifications in lexical categories; (i) introducing singleton types and head dependencies, (ii) providing a way for them to express their idiosyncratic semantics, also called paracompositionality.

Singletons are a special type that can only be substituted by one value. They effectively represent the syntactic rigidity of phrasal verbs and phrasal idioms. They are application only and thus bear the $\backslash_\star$ star from the modal notation of Baldridge and Kruijff (2003). Function categories used for deriving phrasal verbs such as *pick up* and phrasal idioms such as *kick the bucket* contain singletons as arguments.

(19) kicked := *(S\NP)/$_\star$"the bucket": $\lambda x \lambda y.die'_x y$*

(20) pick := *(S\NP)/"up"/NP$_{heavy}$: $\lambda x \lambda y \lambda z.cause'(init'(hold'_x yz))z$*

Head dependencies can be substituted by a set of phrases that share the same head. They can be used to capture the restricted syntactic flexibility of idiomatically combining phrases. In the case of singletons, these are included in function categories used for deriving idiomatically combining phrases.

(21) spill := *(S\NP)/"the beans": $\lambda x \lambda y.reveal'_x secret'y$*

In special cases, both singletons and head dependent arguments can be used in a category:

(22) twiddled := *(S\NP$_{agr}$)/"thumbs"/NP$_{-lexc,+poss,agr}$: $\lambda x \lambda y \lambda z.pass'_y time'_{(self'_z)}z$ $\wedge$ inalien'(xyz)*

Paracompositionality is the term coined by Bozşahin and Güven (2018) to describe the idiosyncratic behavior of singletons and head dependencies in the LF of idiomatic categories. In idiomatic categories, syntactic derivation proceeds as usual except for the fact that singletons or head-marked arguments occupy the syntactic category instead of their polyvalent counterparts. However, their semantics are paracompositional i.e. idiomatic types do not take part in the argument structure like their polyvalent counterpart and instead, they modify the event modality of the predicate representation of the idiomatic phrase. In other words, singletons and head marked categories relate to the extension of these predicates. In the case of *kick the bucket*, *the bucket*'s role in LF is to signify the historicity of the phrase and that sets *kick the bucket* apart from a regular predicate like *die*. Similarly for *up* in *I picked the book up*; it adds a motion and culmination reading to the predicate as represented in the LF of example (20) (Bozşahin, 2020b).

Singletons and head dependencies may, without complicating the CCG formalism itself, extend the elements lexical categories work with. They conform to the *the principle of type transparency*. Derivation with these types works in the same way as polyvalent types, and since they have LF representations, the syntax-semantics parallelism stays intact regardless of their paracompositionality.

Singletons are compliant with the type substitution property of CCG formalism. Polyvalent types such as *NP* or *S\NP* have large sets of values that they can be substituted with. This set becomes smaller for head dependencies, and each singleton's set consists of only one value. However, since it is a modification regarding categories, the size of these sets does not complicate the substitution mechanism of CCG.

CHAPTER 3

DATA SELECTION PROCESS

3.1 Introduction

As explained in section 2.1, idioms does not refer to a well defined category. And the discussion of which types of non-literal phrases can be considered idioms does not seem to be settling soon. This means that the task of selecting idiomatic data requires special attention. As stated before, we are interested in idioms that exhibit low levels of compositionality. Given the diversity of opinion regarding classification and our requirements for specific types of idioms; an off-the-shelf dictionary or list of idioms would be unsuitable. Moreover, linguistic resources geared towards Turkish are rather limited. Even if we had no qualms about the risk of overlooking ongoing debates, using a linguistic resource such as wordnet or a dictionary of idioms for Turkish is simply impractical.

With these issues in mind, we took advantage of distributional methods to select our idiom dataset. And the distributional model we followed due to its success in idiom extraction and identification tasks is discussed in the section below.

3.1.1 Measuring the Compositionality

This section lays out the specifics of the method preferred in this study for the selection of idiom data. The algorithm used for determining the compositionality of idioms was adapted from Gong et al. (2017).[1] Their study aims to compute the compositionality of a word or phrase in a given context. They attempt to achieve this through word embeddings and principal component analysis (PCA). The algorithm is shown to be a language agnostic model through English, German and Chinese.

Gong et al. (2017) developed their method to disambiguate compositional instances from non-compositional instances in a variety of contexts. In some of these contexts the use of a target word or phrase can be literal whereas in others, it could be of a non-literal type such as idiomatic, metaphoric or sarcastic. Phrases that can alternate between literal and non-literal uses are known to cause a lot of headache for NLP researchers working on machine translation and information retrieval. This is the problem they set out to help overcome by coming up with a mechanism that can differentiate between literal and non-literal uses.

[1] https://github.com/HongyuGong/Geometry-of-Compositionality

In order to do this, Gong et al. (2017) exploits the *local linguistic context* of a target word (i.e., its surrounding words) to measure compositionality. They use word embeddings to extract vectors of target words and context words. The vectors of context words provide a *low dimensional linear subspace* and the compositional target words are geometrically related to this subspace while non-compositional ones are not.

To compute this subspace, first context words' vectors are represented by a (d×n) matrix *X* where 'd' is the number of embedding dimensions and 'n' is the number of vectors or words in the context. From this matrix *X*, PCA returns a (d×m) *X'*, where m is lower than n, thus PCA is used to lower the dimensionality of the first matrix. Column span of this new matrix returned by PCA is the subspace created by context word vectors. Compositionality is determined by projecting target phrase's embedding vector v_p onto the this subspace. This operation provides the orthogonal projection vector v'_p. Compositionality score is the cosine distance between v_p and v'_p. The size of this score is correlated positively with compositionality. High compositionality corresponds to literal use and low compositionality corresponds to idiomatic, metaphoric and sarcastic use depending on the data.

Compositional	napoleon stood with his marshals around him it was quite light above him was a clear blue sky and the sun vast orb quivered like a huge hollow crimson float on the surface of that milky sea
Non-compositional	unrealistic or impractical the author shows what is testable physics, what is blue sky nonsense, not limited by conventional notions of what is practical or feasible and what is philosophy domain

Table 3.1: Compositional and non-compositional contexts for the phrase "blue sky".

Before moving any further, one issue must be clarified. The adoption of Gong et al.'s (2017) algorithm, which was originally intended for distinguishing literal uses from idiomatic, metaphoric and sarcastic ones, should not imply that we also share their perspective on the compositionality of these phenomena. On the contrary, as discussed in section 2.1, we argue that idioms are less compositional than other figuratives and MWE types. Notwithstanding its shortcomings as an attempt to precisely and accurately distinguish literal language from non-literal language, a task that still looks insurmountable, their framework promises to be a useful tool for roughly mapping out a spectrum that spans the divide between literal and nonliteral language.

In conclusion, when selecting the idioms to write categories for, collocations that can be explained by metaphorical sense extension should be avoided. Therefore, we need to show that the selected idioms have low compositionality, while at the same time demonstrating that conventionalized metaphors do not. We aim to find quantifiable differences after comparing idioms to conventionalized metaphors in the sense that while the former need a constrained syntactic type and a special representation for their semantics, the latter do not.

3.1.2 Word Embeddings and Datasets

We used pre-trained word embedding libraries for English and Turkish from the Polyglot project Al-Rfou et al. (2013). They use wikipedia articles as training data for their models and provide word embeddings for 137 languages.

For English, we only run the algorithm on Gong et al.'s (2017) bi-context dataset to compare its compositionality scores with the results from our data. For Turkish, we manually constructed two custom bi-context datasets. One of the datasets includes conventionalized metaphors with their literal counterparts and the other one includes possible idiomatic constructions with their literal counterparts. Both idiomatic and metaphorical phrases were selected from a dictionary of idioms and proverbs (TDK, 2020).

3.1.3 Results

Metaphors

We assume that conventionalized metaphors are more compositional than idioms. This requires the conventionalized metaphors to have higher compositional scores than idioms and be closer to their literal counterparts. To test this, we created a bi-context dataset composed of 23 conventionalized metaphors from Turkish and their literal counterparts. The following table lists a summary of the compositionality scores for this data:

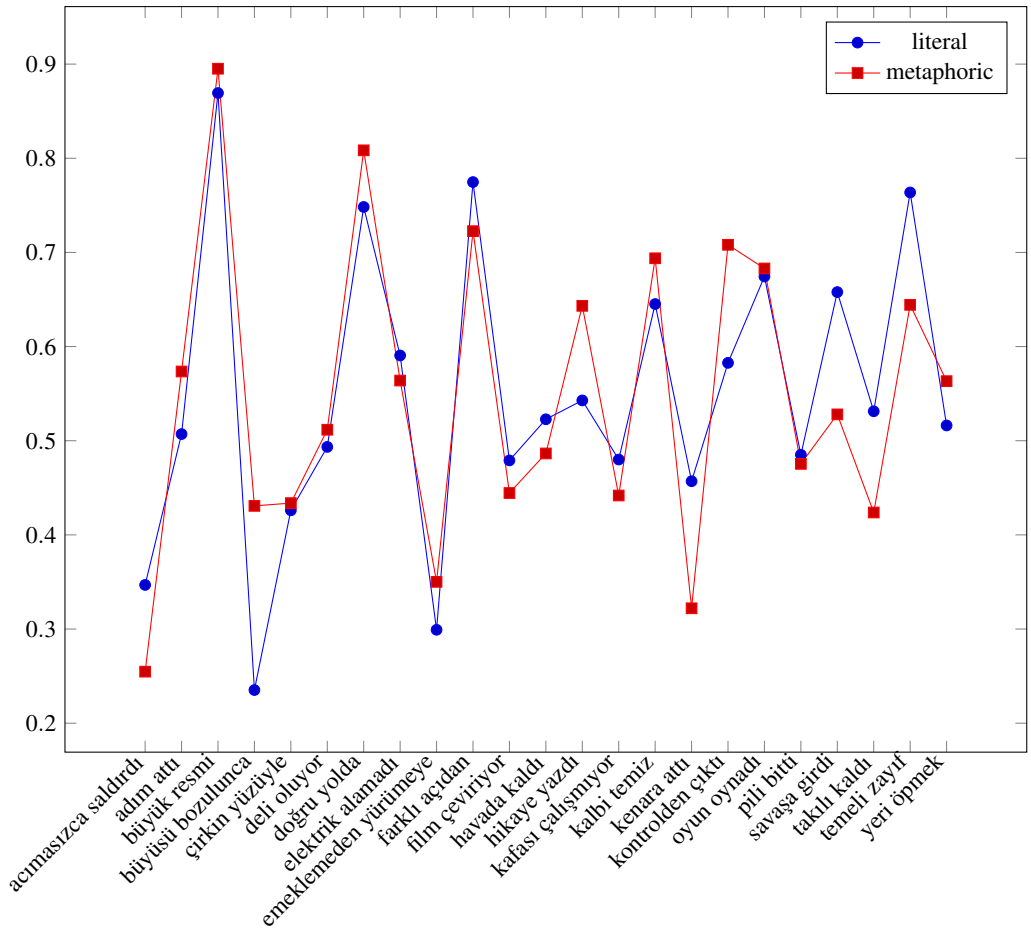


Figure 3.1: Compositionality scores of Turkish metaphor-literal pairs

We observe that metaphoric and literal uses have very similar compositionality scores with a mean difference of *0.0011* points. Moreover, we also observe that compositionality scores of phrases from either type of context rarely drops below *0.4*. Overall, these results suggest that metaphorical uses are similar to literal uses with respect to the degree of relation they have with their contexts.

Idioms

The second step of our data selection process focuses on idiomatic phrases. We created a bi-contextual dataset with *54* verb-noun idiomatic phrases to select the final set of idioms from. The following table shows the summary of compositionality scores:

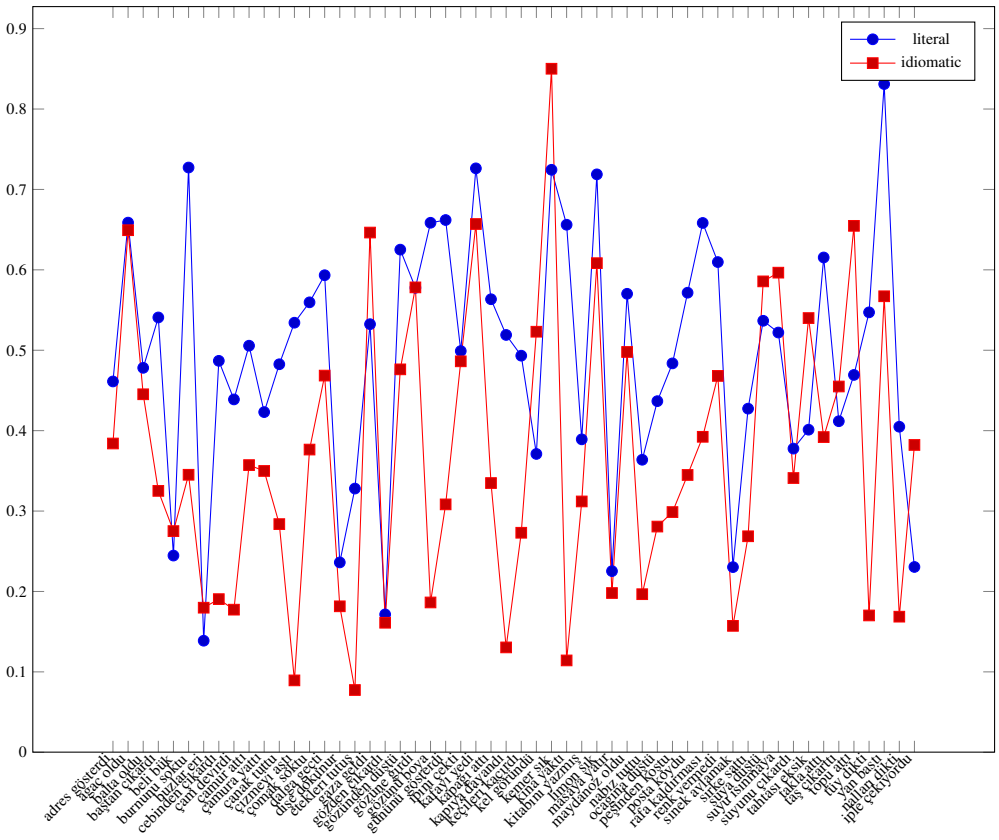


Figure 3.2: Compositionality scores of Turkish idiom-literal pairs.

These results show that idiomatic and literal uses display a larger difference in compositionality scores with a mean difference of *0.1276*. Moreover, we observe that the mean of the overall idiom scores is *0.3658*, which is lower than most (*79.7%*) of the individual compositionality scores of literal pairs from this dataset and also the metaphorical-literal pairs (*89.3%*). We also compare our results to Gong et al.'s (2017) original study of English idioms. Our results show a greater difference of *0.17* points between the means of idiomatic and literal compositionality scores, as well as greater means for both contexts. Here, we note that because The Polyglot word embeddings for English is acquired from a much larger corpus compared to the Turkish one, a direct comparison of the scores would be misleading. Other than that, it can be concluded that the pattern of compositionality scores being lower for idiomatic pairs is repeated in our dataset. No such difference is observed for metaphorical ones.

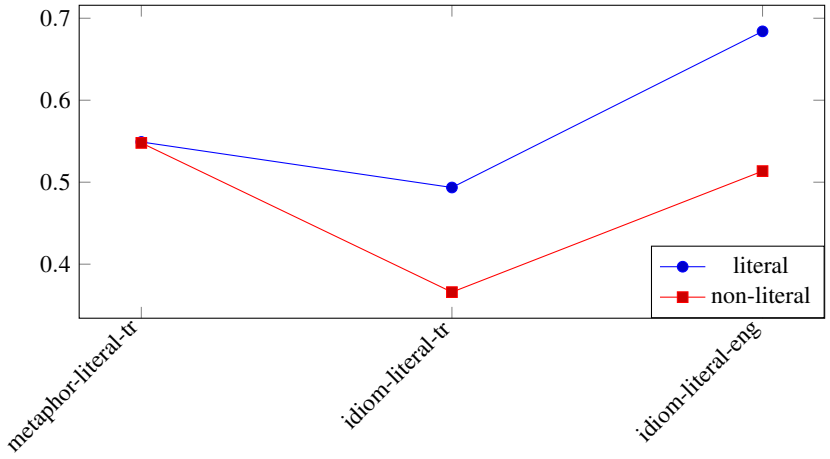


Figure 3.3: Comparision of compositionality accross the datasets.

So far, we have established that the algorithm is able to differentiate between literal and idiomatic uses when provided with Turkish datasets. We have also shown that collocationalized metaphors are comparable to literal phrases in compositionality. Having done that, we move onto the main task, which is selecting idioms that differ in semantics from their surface predicates. A closer look at compositionality scores from idiomatic contexts shows that idioms that are most disconnected to their surface predicates tend to have scores below *0.4*, which is also important because we rarely see scores lower than *0.4* from literal and metaphorical contexts. Considering this, we decided to set a threshold value of 0.4 for idiomatic compositionality, and select pairs with literal scores above and idiomatic scores below this threshold. We discarded *25* candidates that did not meet this requirement. For the remaining *29* pairs, the criteria were met with a couple of exceptions, which will be explained in the next section.

3.1.4 Discussion

We observe that all the phrases that had idiomatic scores below and literal scores above the threshold are desirable candidates for the second part of our study. They are compliant with our low compositionality criterion, which requires that their idiomatic interpretation does not occur in different contexts, and that it is not attached to either of the constutients when used alone or in a different collocation. Therefore, no exceptions were made for that group, which all ended up being included in the second part of our study. Still, it is not safe to extrapolate this pattern into the opposite direction and state that none of the discarded candidates fit our criteria. This could also mean that our cut-off point is more conservative than inclusive.

We mentioned that there were some exceptions. Four of the idiomatic phrases we decided to include despite failing to meet our criteria had compositionality scores below the threshold in both contexts. There are two possible explanations for these

low compositionality scores: either the idioms were unambiguous and did not fit a literal context or word embeddings were unable to provide a good representation of those phrases. Considering the actual pairs, we are more inclined to the conclusion that the second reason is the culprit. The example below shows shortened literal and idiomatic contexts for one of the exceptions:

(23) a. geniş elbisesinin kazayla etekleri tutuştu.
wide dress-POSS-GEN accident-INS skirt-PLU-ACC catch fire-PAST
'The skirts of her wide dress caught fire by accident'

b. sınava geciktiği için etekleri tutuştu.
exam-DAT late-OP.AGR because skirt-PLU-ACC catch fire-PAST
'She panicked because she was late for the exam'

There are several possible reasons for word embeddings to fall adequately representing some words or phrases.

Low frequency words, whose semantic representation can be skewed due to limited context, can pose a problem for distributional models. Moreover, the Turkish corpus used for training these embeddings was also smaller in size when compared to well studied languages like English or German (Al-Rfou et al., 2013).

Another reason for making these exceptions can be cited here. Since a linguistic property that is observed in the shifting semantics of idioms and discussed at length in section 4.3 was applicable to some of our exceptions, we were more inclined to include them.

Limitations

We have two limitations with respect to the scope of our data. First, because we compare literal and non-literal instances of a phrase to decide its compositionality, we cannot include unambiguous idioms.[2] Unambiguous idioms are those that are syntactically malformed or have a literal meaning that does not comply with natural language use. Following is an example of an unambiguous idiom from Turkish:

(24) Onun yaptıklarının ceremesini çekiyorum.
Him doings-POSS-GEN ?-POSS-ACC endure-TENSE
'I am paying the price for his wrong doings'

Example (24) shows an unambiguous idiom, and its constituent word *cereme* is an archaic word. Originally referring to the monetary penalty handed down to criminals, it is no longer used in contemporary Turkish outside of the idiomatic phrase. This and other unambiguous idioms cannot be included in our study because acquiring a

[2] The terms "unambiguous idiom" and "ambiguous idiom" are borrowed from psycholinguistics literature. They do not refer to the phrase's idiom status but rather the possibility of their literal counterparts. Ambiguity or unambiguity of idioms have interesting effects on their comprehension, as reviewed by Cacciari and Papagno (2012).

reliable compositionality score for them is next to impossible. Therefore, we only focus on ambiguous idioms which have viable literal interpretations. Another limitation stems from the coverage area of word embeddings. Some possible candidates were not even included in the bi-context dataset because the word embedding model did not have representations for some of their constituents.

CHAPTER 4

IDIOMATIC CATEGORIES

This chapter provides a grammatical analysis with Turkish idioms. In the following sections, we first replicate Bozşahin and Güven's (2018) singleton and head-marked category analysis with Turkish idioms. This is followed by the section that focuses on linguistic characteristics that are specific to the Turkish language, and their implications for studying idioms. Finally, a discussion on the semantic representations of idioms wraps up the second part.

4.1 Replication

As mentioned in section 2.3.1, Bozşahin and Güven's (2018) categories carry over the IP-ICE distinction in Nunberg et al. (1994). Moreover, as explained in section 2.1, Turkish idioms also comply with this distinction. To demonstrate this, we will take a closer look at the syntactic behavior of these types.

IPs are relatively fixed expressions, they do not appear in discontinuous forms and they do not permit operations like adjectival modification and relativization. In the examples below, the Turkish idiom *nalları dikmek*, meaning *to die*, is shown to behave like an IP[1]

(25) a. Adam nalları dikti.
man horseshoe-PLU-ACC stick up-PAST
'The man died'

b. #Adam büyük nalları dikti.
man big horseshoe-PLU-ACC stick up-PAST
'The man stuck up a big horseshoe'

c. #Adamın diktiği nallar
man-GEN stick up-OP-POSS horseshoe-PLU
'the horseshoes that the man stuck up'

Example (25b) only introduces an adjectival modification to the idiomatic object and example (25c) shows the idiom in relative clause construction. Both operations cause

[1] Following Bozşahin (2020b) we use # when the idiomatic reading of the phrase is not available

the idiomatic reading to disappear for IP types. In contrast, ICEs tend to preserve idiomatic meaning under the same operations. Following examples for the ICE *çam devirmek* demonstrate that the idiomatic meaning is still accessible with adjectival modification and relative clause construction.

(26) a. Adam çam devirdi.
man pine-NOM roll-PAST
'The man made a blunder'

b. Adam büyük çamı devirdi.
man big pine-ACC roll-PAST
'The man made a big blunder'

c. Adamın devirdiği çam
man-GEN roll-OP-POSS pine
'the blunder that the man made'

For the IP class, Bozşahin and Güven (2018) propose to use singleton types. Following is an example of a singleton derivation with the Turkish IP *nalları dikti*:

(27)

adam	nal-lar-ı	dik-ti
man	horseshoe-PLU-ACC	stick up-TENSE
$S/(S\backslash NP_{\text{nom}})$	$NP_{\text{nalları}}$	$(S\backslash NP_{\text{nom}})\backslash_\star "\text{nalları}"$
$: \lambda p.p\,man'$	$: nalları'$	$: \lambda x\lambda y.die_x' y$
	$S\backslash NP : \lambda y.die'_{nalları} y$ (<)	
$S : die'_{nalları} man'$ (>)		

A singleton *"B"* in a category $A\backslash$ *"B"* can only substitute for the string *"B"*. Adverbial modification on the string *"B"* makes it inaccessible to the singleton, and this is shown in the example (28a). Since singletons cannot participate in composition, hence the star at its slash, they prohibit the idiomatic category from forming a relative clause[2] (28b). In the following examples, derivations that are not possible have the '*' mark on their right:

(28) a.

#adam	büyük	nal-lar	-ı	dik-ti
man	big	horseshoe-PLU	-ACC	stick up-TENSE
$S/(S\backslash NP_{\text{nom}})$	N/N	N_{nallar}	$NP_{\text{acc}}\backslash\backslash N$	$(S\backslash NP_{\text{nom}})\backslash_\star "\text{nalları}"$
		N_{nallar} (>)		
		$NP_{nallar,acc}$ (<LEX)		
				(*<)

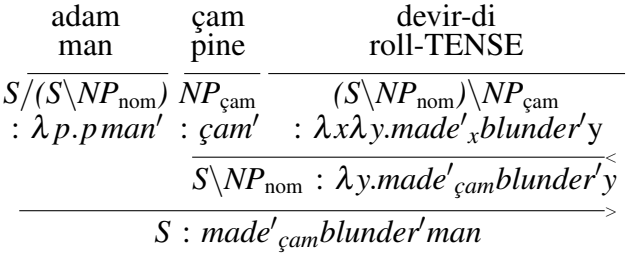
b.

#adamın	dik	-tiği	nallar
man-GEN	stick up	-OP.AGR	horseshoe-PLU
$S/(S\backslash NP_{\text{agr}})$	$(S\backslash NP)\backslash_\star "\text{nalları}"$	$(N_h/N_h)\backslash(S\backslash NP)$	N_{nallar}
$S\backslash_\star "\text{nalları}"$ (*>B)			

[2] Relative clause derivation follows Bozşahin (2002).

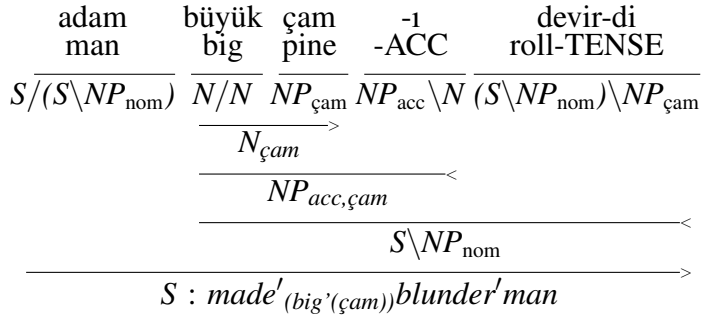
For ICEs, head-word subcategorization is used. In addition to representing the syntactic variability of this class better, it also incorporates the semantic distribution of the idiomatic meaning by mapping idiomatic representations onto surface forms in LF. The derivational process for a head-marked category is shown below:

(29)

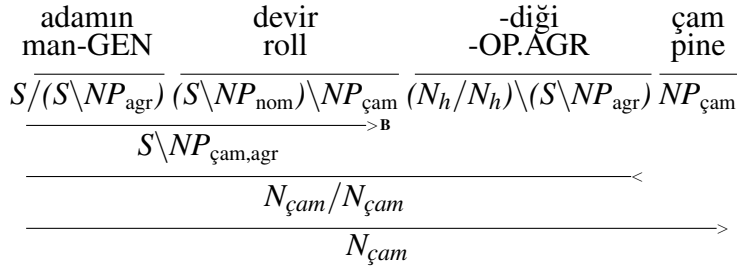


Head-marked categories can permit adverbial modification because they only select for the head. Also, since there is no restriction on their slashes, which means composition with these types are possible, they can participate in relative clauses. Example (30a) shows adverbial modification and (30b) shows relative clause derivations with head-marked types.

(30) a.

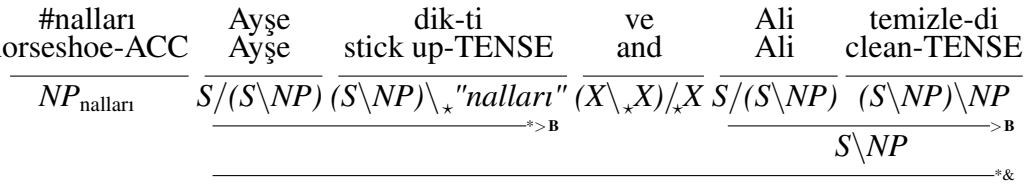


b.



So far, idiomatic categories have been shown to represent the differences between IPs and ICEs in terms of syntactical behavior. However, neither of them should be able to involve constructions like (31a) and (31b). Since singletons cannot involve composition, derivation with the idiomatic category is not possible in (31a). Head-marked category also cannot derive in (31b), because head-marking causes it to be different from the other coordinand's type.

(31) a.



b. #çamı Ayşe kes-ti ve Ali devir-di
pine-ACC Ayşe cut-TENSE and Ali roll-TENSE

$NP_{çam}$ $S/(S\backslash NP)$ $(S\backslash NP)\backslash NP$ $(X\backslash_\star X)/_\star X$ $S/(S\backslash NP)$ $(S\backslash NP)\backslash NP_{çam}$

$S\backslash NP$ (>B)

$S\backslash NP_{çam}$ (>B)

(*&)

4.2 Language Specific Properties

In the section 4.1, we looked at the structures idiomatic categories can and cannot be derived, following Bozşahin and Güven's (2018) demonstration. Given that we are dealing with Turkish examples, language-specific properties of Turkish, the so-called flexible word order and morphology, needs to be accounted for here.

4.2.1 Scrambling

As an agglutinating language, Turkish is prone to scrambling. This study follows Bozşahin's (2014) claim that Turkish is a SOV language, and develops idiomatic categories accordingly. Other word orders for Turkish is possible and shown to be reliant on phonological information (Özge and Bozşahin, 2010). Going over all possible word orders with idiomatic categories would exceed the scope of this study, but we would like to look at OSV possibilities for Turkish ICEs.

OSV word order is generally present as a secondary option for SOV languages such as Korean, Japanese, and Turkish (Bozşahin, 2014). It can be handled by type-raising and composition in CCG. The fact that singleton categories cannot compose rules OSV out as an option for them. However, because idiomatic meaning is not present in such constructions, this issue does not constitute a challenge.

(32) #Nalları adam dikti.
horseshoe-PLU-ACC man stick up-PAST
'the man stick up the horseshoes'

ICE examples for Turkish are somewhat sparse, yet the ones included in our limited list seem to preserve idiomatic meaning through local scrambling. Their derivations are possible and exemplified in (33) below:

(33) çamı ben devir-di-m
pine-ACC I roll-TENSE-AGR

$NP_{çam}$: $çam'$ NP_{nom} : I' $(S\backslash NP_{nom})\backslash NP_{çam}$: $\lambda x\lambda y.made'_x blunder'$y

$S/(S\backslash NP_{nom})$: $\lambda p.pI'$ (>T)

$S\backslash NP_{çam}$: $\lambda x.made'_x blunder'$I (>B)

S : $made'_{çam} blunder'I$ (<)

4.2.2 Morphology

Syntactic projection is considered to work without being aware of morphology (Bozşahin, 2002). For this reason, morphological operations have been generally ignored until now. However, there are some idioms that can assume different forms depending on morphology.

(34) a. Ayşe'nin etekler -i tutuştu.
Ayşe-GEN skirts -POSS$_{3s}$ catch fire-PAST
'Ayşe is in panic'

b. Benim etekler -im tutuştu.
My skirts -POSS$_{1s}$ catch fire-PAST
'I am in panic'

c. Ayşe Ahmet'in ocak -ı -na düştü
Ayşe Ahmet-GEN house -POSS$_{3s}$ -DAT fell-PAST
'Ayşe is at the mercy of Ahmet'

d. Ayşe sonunda ocak -ı -ma düştü
Ayşe finally house -POSS$_{1s}$ -DAT fell-PAST
'Ayşe is finally at my mercy'

These examples pose no problem for head-marked categories. However, because singletons can only substitute for a specific string, they are unable to capture such variations. There are two possible ways to work around this problem. One option is to provide different categories for morphological variation on the corresponding singleton arguments:

(35) a. tutuştu$_{1s}$:= $(S\backslash NP_{gen})\backslash_\star$*"eteklerim"*

b. tutuştu$_{2s}$:= $(S\backslash NP_{gen})\backslash_\star$*"eteklerin"*

c. tutuştu$_{3}$:= $(S\backslash NP_{gen})\backslash_\star$*"etekleri"*

d. tutuştu$_{1p}$:= $(S\backslash NP_{gen})\backslash_\star$*"eteklerimiz"*

e. tutuştu$_{2p}$:= $(S\backslash NP_{gen})\backslash_\star$*"etekleriniz"*

Another option is to address morphological variance of singletons by incorporating morphological functions to the singleton. Morphology differs from syntactic categories and it requires a special format. Morphological functors are represented in the $A\backslash\backslash B$ format, where the domain is required to be a lexical item, and a double slash means that it can only be used for application (Bozşahin, 2020a).

(36)
$$\dfrac{\dfrac{\text{Ayşe}}{N_{\text{nom}} : Ayşe'} \quad \dfrac{\text{-yi}}{\text{-ACC}}}{N_{\text{acc}} : Ayşe'}\ {}_{<\mathbf{LEX}} \qquad \dfrac{}{N_{\text{acc}} \backslash\backslash N : \lambda f.f'}$$

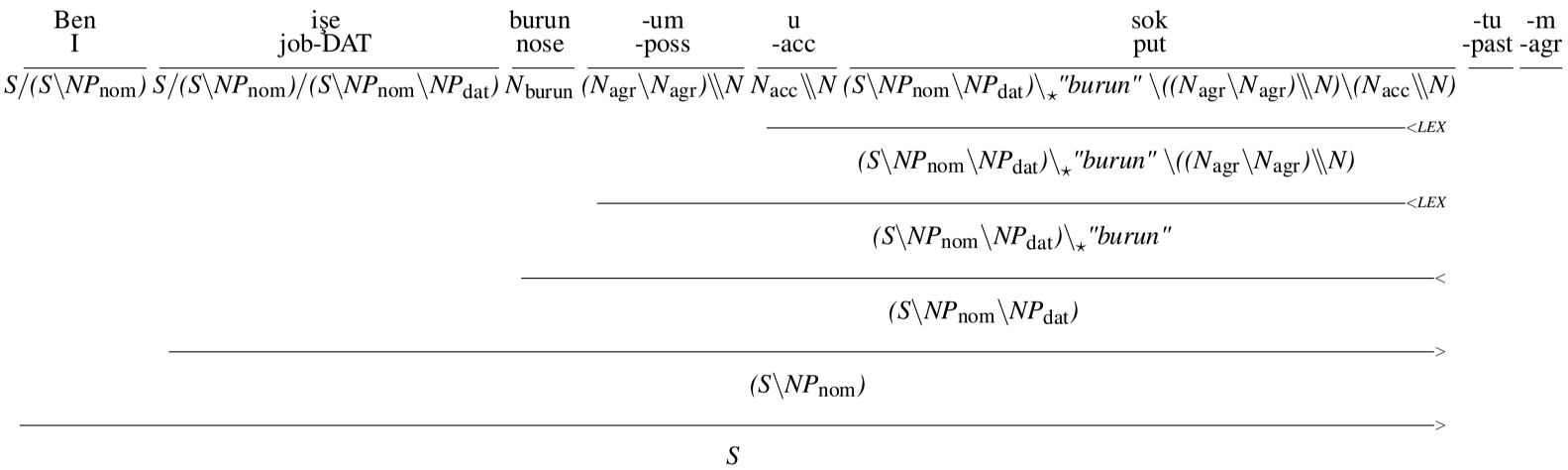
To account for the morphological variance, we propose to type raise the specific singleton categories with respect to the morphological functors they can take. This has two motivations; firstly, since singleton categories are realized in event modality instead of the argument structure; their possessive specification is semantically vacuous, and thus subcategorizing for it is not necessarily harmful. Secondly, as a result of the first argument, the possessive constructions such as *Ayşe'nin etekleri* in idiomatic categories do not actually get realized and genitive has no scope over the possessed argument. As such, we treat these morphological functions as remnants of the literal derivation. The example below shows a standard singleton category in comparison to its counterpart with morphological functions, and (38) exemplifies the latter's derivation.

(37) a. soktu := $(S\backslash NP_{agr})\backslash NP_{dat})\backslash_{\star}$"burun" : $\lambda x\lambda y\lambda z.intrude_{x}'yz$

b. soktu := $(S\backslash NP_{agr})\backslash NP_{dat})\backslash_{\star}$ ("burun" $\backslash(N_{agr}\backslash\backslash N_{agr}\backslash\backslash N)\backslash(N_{acc}\backslash\backslash N)$: $\lambda a\lambda b\lambda x\lambda y\lambda z.intrude_{xab}'yz$

The singleton category in 37b can also be likened to a polyvalent category with specified agreement and case features $NP_{1s,acc}$, except for the fact that one is already derived as such. Also, since type raising is done with morphological functors, it does not violate the *singletons can only be arguments and arguments of arguments* assumption as it concerns the syntactical domain.

(38)


Ben
I
S/(S\NP_nom)
işe
job-DAT
S/(S\NP_nom)/(S\NP_nom\NP_dat)
burun
nose
N_burun
-um
-poss
(N_agr\N_agr)\\N
u
-acc
N_acc\\N
sok
put
(S\NP_nom\NP_dat)*"burun" \((N_agr\N_agr)\\N)\(N_acc\\N)
-tu
-past
-m
-agr
<LEX
(S\NP_nom\NP_dat)*"burun" \((N_agr\N_agr)\\N)
<LEX
(S\NP_nom\NP_dat)*"burun"
<
(S\NP_nom\NP_dat)
>
(S\NP_nom)
>
S

Tense

CCG treats verbs as functions that do the structural work in verb phrases and simple sentences. Following suit, this study also considers them the linchpin of idiomatic representations. Given their importance, it is imperative to look at how common morphological functions on verbs such as tenses work with the idiomatic categories.

We chose tensed verbs instead of untensed verbs to show the idiomatic categories for convenience's sake, as it is conventionally done. However, in order to show morphological derivations, we now switch to infinitive forms.

(39) a. *dik =: IV\"nalları": λxλy.die'$_x$y*

b. *-di =: (S$_{nom}$)\\IV : λpλy.p$_{past'}$y*

Untensed verbs' lexical representation is shown in (39a) and a tense suffix is shown in (39b). Tense category is represented in the fashion of Bozşahin (2019). The dollar mark on *IV* is used to include all possible lexical variations of *IV* such as *VP\NP* and *(VP\NP)\NP*.

Untensed verbs can combine with tense, aspect, modality suffixes and personal markers offline. Without delving much into the verbal morphology of Turkish, we show a derivation with the past tense in (40).

(40)

adam	çam	devir	-di
man	pine	roll	-past
$S/(S\backslash NP_{\text{nom}})$	$NP_{\text{çam}}$	$IV\backslash NP_{\text{çam}}$	$(S\backslash NP_{\text{nom}})\backslash\backslash \$IV\$$
$: \lambda p.pman'$	$: çam'$	$: \lambda x\lambda y.make'_x blunder'y$	$: \lambda p\lambda y.p_{past'}y$
		$(S\backslash NP_{\text{nom}})\backslash NP_{\text{çam}} : \lambda x\lambda y.make'_{(past')(x)} blunder'y$ (<LEX)	
		$(S\backslash NP_{\text{nom}}) : \lambda y.make'_{(past')(çam)} blunder'y$ (<)	
		$S : make'_{(past')(çam)} blunder'man'$ (>)	

4.3 Semantics

So far, we have only focused on syntactic behavior of Turkish idioms. However, the actually challenging part when writing categories, even for a small set of idioms, is finding an appropriate semantic representation for them. One of the challenges stem from a phenomenon Nunberg et al. (1994) also observes: most idioms use concrete entities and events to represent more abstract concepts. Our data too follows this pattern without an exception.

Bozşahin (2020b) states that idioms are never equivalent to a predicate that is already in the lexicon because they always introduce an extra mannerism. In addition to

this, they also rarely correspond to simple predicates such as *die'*, but this is expected because they mostly relate to abstract concepts. Some idioms also tend to require very specific presuppositions. Let us consider the idiom *mumla aratmak*, which means *cause to search with a candle* if translated literally. As an idiom however, it refers to the situation in which 'some new person or situation turns out to be worse than the one before which might have also been bad'. Although most is not as convoluted, pinning down abstract and extraordinary meanings, such as the ones idioms have, seems to us is a task that heavily relies on native speaker intuition.

This intimidating barrier does not deter some approaches from claiming to discover systematicity in idiom semantics, however. Conceptual analysis is already discussed in section 2.1, and concluded to be inapplicable to our data. Another approach claims to have found systematicity in aspectual types of idioms. It has been argued that aspectual types of verbal idiomatic phrases comply with their literal counterparts' aspectual types (McGinnis, 2002, 2005).

This argument stems from the Distributed Morphology (DM) framework's post-syntactic approach to idiom semantics. Aspectual classes are considered to be the types of structural meaning that syntax would take into account. However, since idiom semantics are derived post-syntactically in DM, their aspectual types have to be decided ahead of that. This leads to the conclusion that the aspectual type of a phrase cannot be changed by idiomatic use. While McGinnis (2002, 2005) provides data in support of this claim, there are also others who have pointed out examples that contradict it (Glasbey, 2003, 2007; Leivada, 2017). Moreover, it has been argued that IPs do not necessarily follow the aspectual structure of their literal counterparts, unlike ICEs which tend to be compliant with them.

Although we are unable to offer a lengthy discussion about aspectual types here, it is worth noting we too observe aspectual type mismatches in our data. For example, idioms *nalları dikmek* and *etekleri tutuşmak* exhibit different aspectual types in literal and idiomatic contexts. *Nalları dikmek* has *process* or *culminated process* type when used literally, as opposed to *culmination* when used idiomatically. Similarly, *etekleri tutuşmak* has *culmination* type in literal contexts, in contrast to its *process* or *state* type in idiomatic contexts. We support these observations by Taylan's (2001) aspectual type tests.[3]

Adverbials derived with *-cE* such as *saatlerce* are only compatible with process, state and point types:

(41) a. ?saatlerce nalları dikti.
hour-PLU-ADV horseshoe-PLU-ACC stick up-PAST
'?Died for hours'

b. Saatlerce domino taşlarını dikti.
hour-PLU-ADV domino-PLU-ACC stick up-PAST
'stick up the dominos for hours'

[3] Though the studies in discussion refer to Vendlerian classes, we preferred Moens and Steedman's (1988) terminology.

The adverb *artık* only disagrees with negative culminations:

(42) a. ?artık nalları dikmedi/dikmiyor.
anymore horseshoe-PLU-ACC stick up-NEG-PAST/stick up-NEG-IMPERF
'?Did not die/is not dying anymore'

b. Artık domino taşlarını dikmedi/dikmiyor.
anymore domino-PLU-ACC stick up-NEG-PAST/stick up-NEG-IMPERF
'did not stick up/is not sticking up the dominos anymore'

Adverbials derived with *-lIğInE* or formed with *için* 'for' are only compatible with process and state types:

(43) a. Bir saatliğine/saat için etekleri tutuştu.
One hour-ADV/hour for skirt-POSS catch fire-PAST
'Panicked for an hour'

b. ?Bir saatliğine/saat için çatısı tutuştu.
One hour-ADV/hour for roof-POSS catch fire-PAST
'?Their roof caught fire for an hour'

Epeydir 'for a while' is not compatible with culmination:

(44) a. Epeydir etekleri tutuşuyor.
for a while skirt-POSS catch fire-IMPERF
'They have been panicking for some time'

b. ?Epeydir çatısı tutuşuyor.
for a while roof-POSS catch fire-IMPERF
'?Their roof have been catching fire for some time'

Example (41a) shows that the idiomatic reading of *nalları dikmek* is not compatible with *-cE* adverbials. However, the literal meaning of the same phrase is indeed compatible with them. This is further exemplified in (41b), which is achieved by changing the object of the verbal phrase and thus eliminating the idiomatic reading. These examples show that the aspectual type of *nalları dikmek* should be *[+consequent]*, while its literal counterpart have a *[-consequent]* type. Example (42a) narrows down the type of the idiom *nalları dikmek* to culmination and (42b) shows that the literal phrase cannot be a culmination. Similarly examples (43a) and (44a) show that *etekleri tutuşmak* has a *[-consequent]* aspectual type as an idiom whereas (43b) and (44b) show that its literal counterpart is of culmination type.[4] Overall, these examples illustrate that the aspectual type of a phrase can undergo a change when idiomatic reading is involved.

[4] Malformedness of example (44b) can be disputed because of its iterative interpretation. However, the selective quality of *epeydir* relies on whether the event has a temporal extension that the adverb can modify. With iteration even canonical culmination or semelfactive examples can be made compatible with *epeydir* as in the sentences *Epeydir öksürüyor* 'They have been coughing for a while' or *Epeydir hatamı buluyor* 'They have been finding my mistakes for a while'.

Aspectual type mismatches further support our analysis of idioms, which treats the semantics of individual idioms, especially IPs, as independent occurrences that do not rely on literal meanings of their components. These mismatches further corroborate the understanding that they are truly independent. As a final note, aspectual types of idioms may also provide some clues for their possible LF representations. The aspectual types utilized in idiomatic reading help us narrow down the possible predicates that can be used to represent their meaning.

CHAPTER 5

CONCLUSION

This thesis can be considered as an attempt to survey idioms from multiple perspectives. The primary aim was to apply the grammatical framework proposed by Bozşahin and Güven (2018) to Turkish idioms. This would be a straightforward task if the definition or the connotations of the term "idiom" were clear. The ambiguity of what this term denotes is further complicated by the fact that idioms are tangled up with other concepts like MWEs, which are not less challenging themselves. Thus, it became necessary that the first step is to disentangle and understand these concepts before we could start fleshing out the main object of our study.

The discussion aiming to clear the air about the specific types of idioms we want to study is followed by the solution to a more practical problem; how to select the idioms that will be included in the scope? For this purpose, in light of the insights derived from converging evidence that points to the importance of context for idioms, we settled upon word embeddings along with Gong et al.'s (2017) algorithm that takes a closer look at the relationship between phrases and their context. Accordingly, a list of idioms with the least compositional semantics was put together.

The last chapter finally focuses on the primary goal cited above and provides a case study for Turkish idioms. For this, we first replicate the framework of Bozşahin and Güven (2018) to assess the applicability of their claims in this context. This is followed by a note on the peculiarities that arise from the language-specific properties of Turkish. The chapter ends with a brief discussion regarding the semantic properties of idioms.

The idiomatic categories provided here for Turkish could serve several purposes. To start with, this process allows us to observe idiomatic behavior closely and determine how categories handle it. These observations are presented in the last chapter. Furthermore, as these categories are provided in a parsable format, they can come handy in future studies. Although this is at best a modest effort with a small set of idioms, it can still prove useful as a starting point for improving this approach.

The scope of this study can be extended in several ways. One of them involves that we exclusively use verb-noun constructions as our data. Bozşahin and Güven's (2018) framework requires that it can only be applied to idioms that are represented by a predicate. However, predicate representation is not exclusive to verbal phrases and future studies can extend their data to include other constructions. Another issue that leaves much room for improvement concerns the complexity of components. While the idioms selected here have only two components, future studies can include idioms

with more of them. As a third venue of further investigation, aspectual types can be elaborated and further examined for a better understanding of the semantic behavior of idioms.

Bibliography

Adalı, K., Dinç, T., Gokirmak, M., and Eryiğit, G. (2016). Comprehensive annotation of multiword expressions for turkish. In *TurCLing 2016, The First International Conference on Turkic Computational Linguistics at CICLING 2016*.

Al-Rfou, R., Perozzi, B., and Skiena, S. (2013). Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*.

Aydin, B., Barin, M., and Yagiz, O. (2017). Comprehension of idioms in turkish aphasic participants. *Journal of Psycholinguistic Research*, 46(6):1485–1507.

Baldridge, J. and Kruijff, G.-J. M. (2003). Multi-modal combinatory categorial grammar. In *10th Conference of the European Chapter of the Association for Computational Linguistics*.

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An empirical model of mwe decomposability. In *proceedings of the MWE workshop. ACL*, volume 10.

Baldwin, T. and Kim, S. N. (2010). Multiword expressions. *Handbook of natural language processing*, 2:267–292.

Berk, G., Erden, B., and Güngör, T. (2018). Turkish verbal multiword expressions corpus. In *2018 26th Signal Processing and Communications Applications Conference (SIU)*, pages 1–4. IEEE.

Bobrow, S. A. and Bell, S. M. (1973). On catching on to idiomatic expressions. *Memory & cognition*, 1(3):343–346.

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

Bozşahin, C. (2012). *Combinatory linguistics*. De Gruyter.

Bozşahin, C. (2014). Word order as projection. *Research in Linguistics [Dilbilim Araştırmaları]*, 22:1–23.

Bozşahin, C. (2019). Command and order by type substitution: Another way to look at word order. In *Word Order in Turkish*, pages 179–216. Springer.

Bozşahin, C. (2020a). Ccglab manual. `https://bozsahin.github.io/ccglab/CCGlab-manual.pd`.

Bozşahin, C. (2020b). Multiword expressions meet Bolinger. Submitted.

Bozşahin, C. and Güven, A. B. (2018). Paracompositionality, mwes and argument substitution. In *International Conference on Formal Grammar*, pages 16–36. Springer.

Cacciari, C. and Papagno, C. (2012). Neuropsychological and neurophysiological correlates of idiom understanding: How many hemispheres are involved. *The handbook of the neuropsychology of language*, pages 368–385.

Cacciari, C. and Tabossi, P. (1988). The comprehension of idioms. *Journal of memory and language*, 27(6):668–683.

Caillies, S. and Butcher, K. (2007). Processing of idiomatic expressions: Evidence for a new hybrid view. *Metaphor and Symbol*, 22(1):79–108.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M., and Todirascu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, 43(4):837–892.

Fraser, B. (1970). Idioms within a transformational grammar. *Foundations of language*, pages 22–42.

Gibbs, R. W. (1992). Categorization and metaphor understanding.

Gibbs, R. W., Nayak, N. P., and Cutting, C. (1989). How to kick the bucket and not decompose: Analyzability and idiom processing. *Journal of memory and language*, 28(5):576.

Glasbey, S. (2007). Aspectual composition in idioms. *Recent advances in the syntax and semantics of tense, aspect and modality*, pages 1–15.

Glasbey, S. R. (2003). Let's paint the town red for a few hours: Composition of aspect in idioms. In *Proceedings of the ACL 2003 Workshop on the Lexicon and Figurative Language*, pages 43–49.

Gong, H., Bhat, S., and Viswanath, P. (2017). Geometry of compositionality. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Gries, S. T. (2008). Phraseology and linguistic theory: A brief survey. *Phraseology: An interdisciplinary perspective*, pages 3–25.

Kalandadze, T., Norbury, C., Nærland, T., and Næss, K.-A. B. (2018). Figurative language comprehension in individuals with autism spectrum disorder: A meta-analytic review. *Autism*, 22(2):99–117.

Lakoff, G. and Johnson, M. (1980). Metaphors we live by. *Chicago: The Univ. of Chicago Press*, 242:242.

Leivada, E. (2017). The primitives of the lexicon: Insights from aspect in idioms.

Lenci, A. (2018). Distributional models of word meaning. *Annual review of Linguistics*, 4:151–171.

Levorato, M. C. and Cacciari, C. (1992). Children's comprehension and production of idioms: the role of context and familiarity. *Journal of child language*, 19(2):415–433.

Lodge, D. N. and Leach, E. A. (1975). Children's acquisition of idioms in the english language. *Journal of Speech and Hearing Research*, 18(3):521–529.

McCarthy, D., Keller, B., and Carroll, J. (2003). Detecting a continuum of compositionality in phrasal verbs. pages 73–80.

McGinnis, M. (2002). On the systematic aspect of idioms. *Linguistic Inquiry*, 33(4):665–672.

McGinnis, M. (2005). Painting the wall red for a few hours: A reply to glasbey (2003).

Metin, S. K. and Karaoğlan, B. (2010). Collocation extraction in turkish texts using statistical methods. In *International Conference on Natural Language Processing*, pages 238–249. Springer.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Moens, M. and Steedman, M. (1988). Temporal ontology and temporal reference. *Computational linguistics*, 14(2):15–28.

Moon, R. et al. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press.

Nunberg, G., Sag, I. A., and Wasow, T. (1994). Idioms. *Language*, 70(3):491–538.

Oflazer, K., Çetinoğlu, Ö., and Say, B. (2004). Integrating morphology with multiword expression processing in turkish. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing*, pages 64–71.

Owens, J. (2016). The lexical nature of idioms. *Language sciences*, 57:49–69.

Özge, U. and Bozşahin, C. (2010). Intonation in the grammar of Turkish. *Lingua*, 120(1):132–175.

Peng, J. and Feldman, A. (2015). Automatic idiom recognition with word embeddings. 656:17–29.

Reddy, S., McCarthy, D., and Manandhar, S. (2011). An empirical study on compositionality in compound nouns. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 210–218.

Ruhi, Ş. and Işık-Güler, H. (2007). Conceptualizing face and relational work in (im) politeness: Revelations from politeness lexemes and idioms in turkish. *Journal of Pragmatics*, 39(4):681–711.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. 2276:1–15.

Salehi, B., Cook, P., and Baldwin, T. (2015). A word embedding approach to predicting the compositionality of multiword expressions. pages 977–983.

Schone, P. and Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*.

Steedman, M. (2000). *The syntactic process*, volume 24. MIT press Cambridge, MA.

Steedman, M. and Baldridge, J. (2011). Combinatory categorial grammar. *Non-Transformational Syntax: Formal and explicit models of grammar*, pages 181–224.

Swinney, D. A. and Cutler, A. (1979). The access and processing of idiomatic expressions. *Journal of verbal learning and verbal behavior*, 18(5):523–534.

Taylan, E. E. (2001). On the relation between temporal/aspectual adverbs and the verb form in Turkish. *The verb in Turkish*, pages 97–129.

TDK (2020). Atasözleri ve deyimler sözlüğü. `hhttps://sozluk.gov.tr/`.

Uzun, L. S. (1991). Deyimleşme ve türkçede deyimleşme dereceleri. *Dilbilim araştırmaları dergisi*, 2:29–39.

Weinreich, U. (1969). Problems in the analysis of idioms. *Substance and structure of language*, 23:81.

APPENDIX A

CATEGORIES

Here we present the grammatical representations for our selected idioms:[1]

aştı v := (s\np[case=nom])*"çizmeyi" : \x\y.!judged !presumptuously _ x y;

attı v := (s\np[case=nom]\np[case=dat])\np[h=çamur] : \x\y\z.!made _ x !slander y z;

attı v := (s\np[case=nom]\np[case=dat])*"kapağı" : \x\y\z.!took !refuge !in _ x y z;

attı v := (s\np[case=nom])*"takla" : \x\y.!rejoiced _ x y;

attı v := (s\np[case=nom])*"topu" : \x\y.!gave !up _ x y;

boyadı v := (s\np[case=nom]\np[case=gen])*"gözünü" : \x\y\z.!misled _ x y z ;

çıkardı v := (s\np[case=nom]\np[case=acc])*"baştan" : \x\y\z.!seduced _ x y z ;

çıkardı v := (s\np[case=nom]\np[case=acc])*"cebinden" : \x\y\z.!superior to _ x y z ;

çıkardı v := (s\np[case=nom]\np[case=acc])*"gözden" : \x\y\z.!may !sacrifice _ x y z ;

çıktı v := (s\np[case=gen])*"suyu" : \x\y.!lost !meaning !by !overuse _ x y;

dayandı v := (s\np[case=nom])*"kapıya" : \x\y.!deadline !arrived _ x y;

dikti v := (s\np[case=nom]\)*"nalları" : \x\y\z.!died _ x y z;

dikti v := (s\np[case=nom]\np[case=dat])*"tüy" : \x\y\z.!worsened _ x y z;

devirdi v := (s\np[case=nom])\np[h=çam] : \x\y.!made _ x !blunder y;

düştü v := (s\np[case=nom])*"suya" : \x\y.!got !cancelled _ x y;

düştü v := (s\np[case=nom]\np[case=gen])\"ocağına": \x\y.!is !at !mercy _ x y z;

gösterdi v := (s\np[case=nom]\np[case=acc])*"adres : \x\y\z.! targeted _ x y z ;

gösterdi v := (s\np[case=nom]\np[case=dat])*"gününü" : \x\y\z.!punished _ x y z;

kaçırdı v := (s\np[case=nom])*"keçileri" : \x\y.!become !insane _ x y;

[1] Categories can also be found in https://github.com/arzuburcuguven/idiom-grammar

kaldırdı v := (s\np[case=nom]\np[case=acc])*"rafa" : \x\y\z.!put !on !hiatus _ x y z ;

koydu v := (s\np[case=nom])*"posta" : \x\y.!threatened _ x y;

koştu v := (s\np[case=nom]\np[case=gen])\"peşinden": \x\y\z.!pursued _ x y z;

sattı v := (s\np[case=nom])*"sirke" : \x\y.!sulk _ x y;

soktu v := (s\np[case=nom]\np[case=dat])*"burnunu" : \x\y\z.!intrude _ x y z;

soktu v := (s\np[case=nom]\np[case=dat])*"çomak" : \x\y\z.!prevented _ x y z;

tuttu v := (s\np[case=nom]\np[case=dat])*"çanak" : \x\y\z.!supported !maliciously _ x y z;

tutuştu v := (s\np[case=gen])\"etekleri" : \x\y.!panic _ x y;

vermedi v := (s\np[case=nom]\np[case=dat])*"renk" : \x\y\z.!concealed !emotions _ x y z;

yaktı v := (s\np[case=nom])*"kına" : \x\y.!gloat _ x y;

yatırdı v := (s\np[case=nom]\np[case=acc])*"masaya" : \x\y\z.!discussed !in !detail _ x y z;

yattı v := (s\np[case=nom])*"çamura" : \x\y.!broke !promise _ x y;

yazmış v := (s\np[case=nom]\np[case=nom])*"kitabını" : \x\y.!is !expert !at _ x y z;