Formulaicity of affixes in Turkish

by

Heikal Badrulhisham

B.A., University of Wisconsin-Madison, 2016

Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Arts

in the
Department of Linguistics
Faculty of Arts and Social Sciences

SIMON FRASER UNIVERSITY
Fall 2019

Approval

Name:	**Heikal Badrulhisham**
Degree:	**Master of Arts**
Title:	**Formulaicity of affixes in Turkish**
Examining Committee:	**Chair:** Suzanne Hilgendorf Associate Professor
	John Alderete Senior Supervisor Professor .....
	Maite Taboada Supervisor Professor .....
	Paul Tupper Supervisor Professor Department of Mathematics .....
	Julian Brooke External Examiner Instructor Department of Linguistics University of British Columbia .....
Date Defended/Approved:	November 27, 2019

Abstract

This study examines whether suffix sequences in a Turkish corpus distribute as units (formulas). Most research on formulaicity focused on word-level formulas. As for affix-level formulas, most evidence for them comes from psycholinguistic studies, whereas there is less evidence from corpus data. This study examines the pattern of cooccurrence of suffixes on verbs in the Turkish National Corpus. To capture formulaicity between suffixes, this study uses a measurement called risk ratio, which is a novel way to measure collocation. The analysis of the risk ratio data suggests that 1) affix formulaicity likely does exist in the corpus, 2) affix formulaicity is a gradient rather than discrete phenomenon, and 3) formulaicity also holds between affixes and stems. The existence of affix formulas suggests that some polymorphemic sequences are stored as wholes in the mental lexicon, despite their apparent decompositionality. Theoretically, the results support psycholinguistic models of morphological processing with both analytic and holistic processing.

Keywords: formulaicity; affixes; Turkish; corpus; risk ratio; psycholinguistic morphological processing

Table of Contents

List of Tables

List of Figures

Chapter 1. Introduction

1.1. Overview

Formulaicity is the notion that some analyzable sequences of items (e.g., words) may function as a single entity psycholinguistically (Wray 2002). In addition to the separate lexical entry of each member item therein, such sequences may be stored as single wholes in the mental lexicon and are retrieved as wholes in language processing. Such sequences are called *formulas* or *formulaic sequences*. Examples of formulaic sequences are not limited to word-level sequences such as idioms (e.g., *the jury is still out*) and recurrent phrases (e.g., *and she was like*), but they may also include sequences of affixes or polymorphemic words (Wray 2002). However, most formulaicity research has focused on formulas that are sequences of words. In contrast, there has been relatively less attention on formulaicity among affixes.

Nevertheless, there has been research on affix formulaicity using both psycholinguistic approaches and corpus-based approaches. From psycholinguistic approaches, the key findings are based on the idea that certain affix sequences or polymorphemic words are processed as a unit. The first form of psycholinguistic evidence for affix formulaicity is the processing advantage of affix formulas. Using a lexical decision task on plural nouns, Sereno & Jongman (1997) found that responses were faster to plural nouns whose plural form is more frequent than the singular. The results were taken to imply that such high-frequency affixed words are stored as wholes, which enables faster processing because retrieval then involves a single item. A similar relationship between

affixed form frequency and processing speed has been found for other languages and other classes of affixes (e.g., Bertram et al. (2000), Niemi et al. 1994, Lehtonen et al. 2007, Soveri et al. 2007 for Finnish, Bertram et al. (2009) for Dutch).

The second form of psycholinguistic evidence for affix formulaicity concerns the perception of compositionality of words. Using an affixedness rating task, Wurm (1997) and Hay (2001) found that subjects were able to judge different degrees of affixedness for similarly affixed words, and decide which of two complex words is more complex than the other (e.g., *settlement* was judged to be 'more affixed' than *government*). This implies that some complex words were perceived to be more preassembled.

In addition to psycholinguistic approaches, affix formulaicity has also been studied using corpus-based approaches, albeit to a lesser extent. One such study is Durrant (2013), which analyzed the distribution of suffixes on verbs in a corpus of Turkish. Looking at the most frequent suffixes, the study found that they occurred most frequently only with specific suffixes. The study also found that there are longer suffix combinations that were so frequent that they may be behaving as units. Finally, the study found that some of these combinations are biased to occur with certain verb stems over others. These observations were taken to imply that sequences of multiple morphemes can become lexicalized.

From the above review of the two methodologies, it is apparent that there is a relative scarcity of evidence for affix formulaicity from corpus data. As such, this study is intended to contribute to the body of evidence for affix formulaicity. Specifically, what is proposed is a replication of the Durrant (2013) study. A replication is proposed because the study has several design flaws that may restrict its generalizability. The study in this thesis will feature design improvements, most importantly in terms of corpus data and a more robust quantitative measurement to capture affix formulaicity (explained more in Sections 3.1 and 3.3). It is important that there is more corpus-based evidence for affix formulaicity. This is because corpus data represent language in natural use, rather than a product of experimental conditions.

The general objective of this study is to investigate affix formulaicity in a corpus of Turkish. The first step in this investigation is establishing the existence of affix formulaicity in Turkish. If affix formulaicity does exists, the next question is whether the observed formulaicity is a discrete or gradient phenomenon. Finally, in addition to investigating formulaicity of affixes, the scope of the study is then widened to include formulaicity involving stems and affixes. In addition to contributing to the body of evidence for affix formulaicity with the first step, the points of focus here are also intended to contribute to the consideration of psycholinguistic models of morphological processing. This is particularly relevant to the aspect of gradience, which was not an explicit focus in Durrant (2013), or in most previous formulaicity studies for that matter. However, this aspect of this study may further refine said discussion by posing additional challenges for the processing models to be considered.

1.2. Motivation

The theoretical motivation for this study is to contribute to the theory of morphological processing. The possible existence of affix formulas implies the possibility of preassembled polymorphemic sequences. This bears on whether there is a strict division between lexicon and grammar, because there may exist in storage items that can alternatively be combined grammatically. Also relevant is the extent to which morphological processing is analytic or holistic. This question can be construed as a comparison between different psycholinguistic models of language comprehension. The extreme positions are the purely analytic models (e.g., Taft & Forster 1975) and full listing models (e.g. Manelis & Tharp 1977). In purely analytic models polymorphemic words are decomposed in processing. In contrast, in full listing models words are only accessed by complete forms. If affix formulas are observed in the data, this would be a challenge for purely analytic models. If formulaicity does not appear to apply to all affix sequences, this would be a challenge for full listing models.

However, it is likelier that the formulaicity data constitute an intermediate between the two extremes. Thus, there may be support for models that have both analytic and holistic processing. Such models are:

1) Hybrid models (e.g., Marslen-Wilson et al. 1994, Caramazza et al. 1988)

2) Interactive activation models (Taft 1994)

3) Distributed connectionist models (Seidenberg & McClelland 1989)

First, in hybrid models words go through one of two routes of processing, either holistic processing or decomposition. Second, in interactive activation models, there are nodes corresponding to various units (e.g., syllables, morphemes, words) that are hierarchically connected and are activated in processing. Third, the distributed connectionist models consist of a network of weighted connections between neuron-like processing units, which learns to map from one domain to another (e.g., sound to meaning). Various aspects of the affix formulaicity data in this study may support one of these models over the others. These models are explained in more detail in Section 5.2. Figure 1 provides graphical representations of these models.

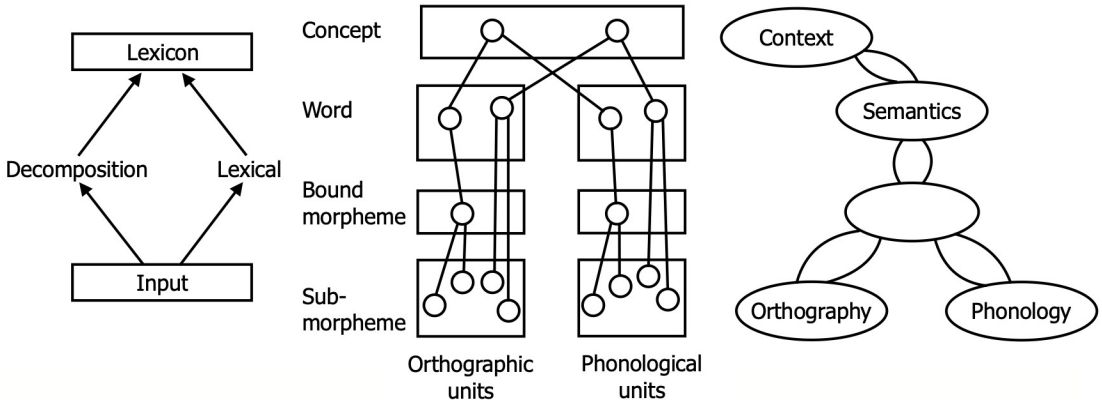


Figure 1: Models of morphological processing (based on Seidenberg & Gonnerman 2000).

From left: 1) Hybrid models, 2) Interactive activation model, 3) Distributed connectionist models.

One aspect of affix formulaicity related to deciding between these models is whether affix formulaicity is a gradient or discrete phenomenon. From the literature, it appears that there are at least two senses in which formulaicity may be gradient. The first sense can be inferred from studies on the processing advantages of affix formulas. This is gradience in that the magnitude of affix formulaicity comes in a continuum. The magnitude of formulaicity could be construed in reference to variables such as increase in processing speed in psycholinguistic studies or to the values of collocation statistics in corpus studies. The second sense can be inferred from Hay & Baayen's (2005) discussion of gradient structure in morphology. This is gradience in the sense of the extent to which affix formulas consist of discrete constituents. The first type of gradience may bear on the question of whether affix formulaicity and morphology are probabilistic. The second type of gradience may bear on the status of morphemes as discrete units. More formally, gradience may support one of the aforementioned models over the others. This is because not all of the models are probabilistic and some of the models assume the morpheme as a discrete entity.

As for methodological motivation, this thesis is intended to contribute evidence for affix formulaicity from a different data type. Most research on affix formulaicity uses experimental/psycholinguistic approaches. It is important however to find evidence from corpus studies as well; this parallels what has been done in word-level formulaicity research, in which there are numerous corpus studies complementing psycholinguistic findings (see Biber (2009) for examples). Although psycholinguistic approaches can reveal behaviors corresponding to affix formulaicity, they do so in terms of specific experimental tasks. Thus, there is a risk that a piece of experimental evidence is an artifact of laboratory conditions. In contrast, corpus data represent language in natural use. Additionally, corpus data may be a reflection of psycholinguistic reality, thereby supporting results from experimental approaches. Thus, this replication study is intended to address the need for more corpus findings in conjunction with existing psycholinguistic findings.

Another methodological contribution of this thesis is in the usage of quantitative measurements to capture formulaicity. One often-used measurement for capturing collocations and formulaicity is raw frequency (as used in Durrant (2013)). However, raw frequency alone may not be reliable in identifying and capturing formulaicity (Wray 2002). Furthermore, which statistic of collocation is the most appropriate for a study is an unsettled issue (Evert 2005). This study contrasts previous ones by using a measurement called risk ratio (Agresti 2019), which is a novel way of measuring collocation in language data. However, risk ratio avoids the disadvantages of other measurements. The usage of risk ratio in this study may contribute to the discussion in quantitative linguistics and natural language processing on the matter of measuring collocation.

1.3. Thesis outline

The following is the outline of this thesis:

2. Background

2.1. Introduction to formulaicity: a definition of formulaicity and a survey of research on formulaicity.

2.2. Affix formulaicity: a description of research on affix formulaicity, including Durrant (2013) and a critique of it.

2.3. Turkish morphology overview: an overview of suffixal morphology and morphosyntax in Turkish.

3. Methods

3.1. Corpus data source: a description of the Turkish National Corpus, its content and how it was used.

3.2. Procedures: a description of the steps taken in processing the corpus data.

3.3. Measurements of association: a comparison of multiple measurements of associations and a justification for the choice of risk ratio for this study.

4. Results and analysis

4.1. Establishing affix formulaicity: an analysis of the risk ratio data to uncover patterns implicating the existence of affix formulaicity.

4.2. Sequences that are formulaic: an examination of what formulaic suffix sequences there are and linguistic reasons for their formulaic status.

4.3. Gradience in affix formulaicity: an analysis of the risk ratio data to determine if affix formulaicity is discrete or gradient.

4.4. Affix formulaicity and the lexicon: an analysis of formulaicity between suffixes and stems.

5. Discussion: a discussion of the implications of affix formulaicity for language and an exemplification of potential application of the results in another domain of language research.

Chapter 2. Background

This chapter provides background information relevant to the thesis. Section 2.1 provides an existing definition of formulaicity and an introduction to formulaicity research. Wray (2002, ch. 1) is recommended for a more extensive introduction to the topic. Next, Section 2.2 discusses the findings of the Durrant (2013) study and critiques it. Based on the critique, a replication study will be proposed with an explanation of methodological changes over the previous study. Finally, Section 2.3 is an overview of Turkish morphology to give the reader an impression of what patterns can be expected in the suffix data.

2.1. Introduction to formulaicity

2.1.1. What is formulaicity

A variety of terminologies have been used to describe different subphenomena of formulaicity. In face of this fragmentation, the definition adopted for this thesis is the one by Wray (2002), which is a prominent attempt at a unified description of formulaicity. In formulaicity, the object of interest are formulaic sequences or formulas. A formulaic sequence is defined as follows:

> a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (Wray 2002)

This definition is meant to be neutral and inclusive. It allows for the possibility of formulaic sequences consisting of units other than words, such as affixes. It also allows for formulas whose fixed elements are not necessarily next to each other and formulas with slots of variable elements (e.g., *<NP> set +<tense> <pronoun> sights on <NP>/<VP>*: *They are setting their sights on colonizing Mars*). Also, it is important that a formulaic sequence is fully analyzable; every unit within the sequence has its distinctive shape and function or meaning. This distinguishes formulaic sequences from cases of many-to-one mapping between meanings and an exponent (Caballero & Harris 2012). Thus, constituents in a formulaic sequence function as a unit even though each individual can occur independently of each other.

Examples of formulaic sequences include idioms (e.g., *water off a duck's back*), conventionalized phrases (e.g., *How's it going*?), phrasal verbs (e.g., *turn up*), fixed binomials (e.g., *nuts and bolts*) and recurring word sequences (e.g., *in the middle of, the fact that*). A sequence can be considered a formula due to several properties. Some of these include semantic opacity (such as with idioms), conventionalization (such as with greetings), fixedness or resistance to modification (such as with idioms and fixed binomials) and highly recurrent usage.

2.1.2. Occurrence of formulaicity

This section reviews some of the research that provides evidence for the existence of formulaicity in language. The two main research areas concerned are psycholinguistics and corpus linguistics.

In psycholinguistics, it has been argued that formulaicity is related to economy of effort in processing language (Perkins 1999). Formulaic sequences function as frameworks for building expressions such that an entire string need not be built from scratch in every instance of language use or perception (Becker 1975). Effectively, using formulas allows the language user to overcome the limitations of working memory, and devote more mental resources to other concurrent tasks (Conklin & Schmitt 2012). Concurrently, exploiting formulas enables more fluent speech (Kuiper 1996).

As such, studies using various experimental designs point to the privileged status of formulas in psycholinguistic processing. For example, experiments using reading tasks have found that formulaic sequences were read more quickly than non-formulaic ones (Conklin & Schmitt 2008; Tremblay, et al. 2011). In the same vein, Underwood et al. (2004) found that when subjects read formulaic sequences, fixations on terminal words were shorter and less frequent than when the same sequences were used nonformulaically. Conversely, a deviation in a formulaic sequence, such as substitution (*cheap cost* as opposed to *low cost*) (Millar 2011) or reversal in a fixed binomial (*groom and bride* instead of *bride and groom*) (Siyanova-Chanturia et al. 2011) could slow down reading speed. Formulaicity also helps with memory as Tremblay et al. (2011) found that formulaic sequences were more likely to be correctly recalled. Jiang & Nekrasova (2007) found that in a grammaticality judgment task subjects were faster and more accurate when the stimulus was formulaic than when it was not. Van Lancker et al. (1981) compared formulaic sequences in both idiomatic and literal contexts and found that the former were articulated faster, while the latter were articulate slower with longer and more frequent pauses, changes in pitch, and less precision of pronunciation. Using an observational approach, Kuiper (1996) examined the language of sports commentators and found that commentaries in faster-paced horse races contained more formulaic sequences than commentaries in slower-paced cricket games.

These studies concluded that such results imply that formulaic sequences are holistically processed, as opposed to processed constituent by constituent. The premise for this conclusion is that the elements in a formula came preassembled, and accessing a single

preassembled string rather than building it online reduces processing burden and leads to faster language processing.

Another body of evidence implying the existence of formulaicity came from corpus linguistics. Studies have shown that the distribution of words in corpora display 'unrandomness' that is unexpected based on grammatical rules alone (Sinclair 1995). One manifestation of this unrandomness is that recurring strings of words account for a substantial portion of a corpus (Biber et al. 1999). Corpus studies also show the unit-like behavior of formulas. For example, when formulas are repeated they tend to be repeated as wholes (e.g., *I think that I think that DNA is a very good example* ...), and when pauses occur they occur at the boundary of formulas, rather than interrupting them (e.g., *I mean they fought valiantly for peace but I, I think that erm <pause> the maternity bill*) (ibid.; emphasis added). Another evidence is the existence of preferred strings, that is there is a preferred way of expressing a meaning or function over other grammatically valid alternatives. For example, Biber et al. (1998) found in a corpus that *large number* is five times more common than *great number*, even though both are grammatically valid way of expressing the same meaning. Formulas can also exhibit preferences more abstractly. Stefanowitsch & Gries (2003) examined the construction <*Noun*> *waiting to happen* and found that words that are disproportionately attracted to the <*Noun*> slot are words of negative connotations such as *accident* or *disaster*. Given findings such as these, connections have been made between corpus data and psychological reality. It has been suggested that recurring strings are so frequently relied upon that they may be treated as a unit of their own (Sinclair 1995, Ellis 1996, Biber et al. 1999). In support of this conclusion, there are psycholinguistic studies showing a processing advantage of formulas extracted from corpora (e.g., Underwoord et al. 2004, Jiang & Nekrasova 2007, Siyanova-Chanturia et al. 2011).

The body of evidence reviewed above implies the existence of formulaic sequences in language, which challenges certain models of grammar. At a more abstract level, formulaicity is problematic for the Chomskian (1965) generative view of language. A model which posits combinatorial rules as the basis for longer strings is intolerant of internally

complex units. This is because such a model assumes a strict division between grammar and the lexicon. Formulas contradict this assumption because they are theoretically analyzable and yet they are mentally stored as wholes. Although a purely analytical grammar can explain speakers' capacity for producing and understanding completely novel utterances, the significance of this capability has been overstated; according to Pawley & Syder (1983), not all grammatically possible sequences are equally likely to occur or judged to be equally native-like or natural. More technically, formulaicity challenges psycholinguistic models based only on rules computed in the brain (e.g., Taft & Forster 1975, Pinker 1997, Ullman 2001). Instead, the findings reviewed in this section support models that have both analytical processing and holistic processing (e.g., Marslen-Wilson et al. 1994, Caramazza et al. 1988, Taft 1994, Seidenberg & McClelland 1989).

2.1.3. Applications of formulaicity

Following the previous review of research establishing formulaicity, this section reviews several fields that engage formulaic sequence data. The purpose of this review is to highlight the further importance of formulaicity in linguistics.

First language learning

Formulaic sequences have been found to play a role in child first language acquisition. For a child, there are two types of formulaic sequences: sequences that a child has created and stored as a whole, and sequences that adult speakers understand to have more complex structure than the child does (Peters 1983). An example of this underanalysis is a child uttering *time for a cup of coffee* when requesting a biscuit, presumably after associating the acquisition of the target object with such utterance from an adult. Peters (1977) hypothesized that in some stage of child language acquisition children take a combination of analytical and holistic approaches to the input language. According to

Nelson (1973), for a child, formulaic sequences are more likely in expressive language (language related to interactional functions (e.g., requests)) than in referential language (language for labeling entities in the world). As for utility, using formulaic sequences assists acquisition by requiring less processing attention. Using formulaic sequences enables a child to produce utterances more completely than they could from scratch (Nelson 1973). This enables them to converse with adults more before they are able to create novel utterances more frequently. Formulaicity in child language also has a methodological consequence. In acquiring the target language, the learner acts upon 'units' in the input, but these units may or may not correspond to individual words or morphemes. Given this, in assessing acquisition it may be more informative to count units rather than words (Peters 1983). A further complication is that different children or the same children at different stages may define a unit differently. Because of this variation, some children may appear more advanced than others, while still subject to the same cognitive limitations as other children.

Second language learning/teaching

Formulaic sequences also play a role in second language acquisition and teaching. The reason for this is that only a small set of grammatically possible strings sound natural or native-like. Thus, a language learner may still fail to approximate native capability despite having mastered the target language grammar. As such, one strand of formulaicity research in second language learning/teaching is identifying strategies for learning or teaching formulaic sequences in the target language and testing the effectiveness of such measures (e.g., Wray & Fitzpatrick 2008, Erman, 2009). Other studies examine how learners acquire formulas. Some of the earliest phrases learners acquire in a second language are often formulas such as greetings. Second language learners also consciously learn formulas such as by memorizing phrases (Stevick 1989). However, it is not easy for learners to identify what sequences are formulaic by themselves, thus learners tend to acquire some formulas in the target language but not others (Pawley & Syder 1983). This is also due to that instructional speech tends to be less idiomatic than speech between native speakers (Wray 2002).

However, for second language learners there is also a conflict between formulaicity and creativity in language use. On one hand, learners can also overuse the formulas they do know to the effect of sounding unnatural (Jaworski 1990). On the other hand, speakers can take risks in constructing novel constructions rather than relying on formulas (e.g., relying on item-by-item translation)(Biskup 1992). This can happen when there is sufficient similarity between the native and target languages. In this case, the similarity may lead a learner to conjecture that knowledge of the first language provides correct intuition about the target language.

Aphasia research

Aphasia research has uncovered properties of formulaic sequences that would not be possible to discover with non-aphasic speakers. It has been found that formulaic language tends to persist even as other parts of the language faculty have deteriorated (Wray 2002). One line of research in aphasia and formulaicity is whether formulaic sequences are psycholinguistically equivalent to words, that is whether they are stored as units. If formulaic sequences are word-like, then it is expected aphasics can produce them fluently. If formulaic sequences are like phrases and clauses, then disfluency is expected in aphasic production. Various studies do suggest that formulaic sequences in aphasic language behave like words psycholinguistically. Similar to words, formulaic sequences are internally resilient, resistant to omissions or substitutions in aphasic speech (Benton & Joynt 1960). Related to this is that items that persist within formulaic sequences cannot be used creatively by aphasics (Critchley 1970) (e.g., being able to utter the phrase *son of a bitch* but not *son* for its individual meaning (Van Lancker 1988)). Although aphasics can be fluent with fixed sequences, they are less so with semi-fixed sequences. This is possibly due to difficulties in retrieving words to fill variable slots in semi-fixed sequences (Gardner 1985). Other parallels between formulaic sequences and words are that they are subject to word-retrieval difficulty (anomia), and they are subject to substitution (paraphasia), often using sequences that are semantically or structurally similar (Van Lancker 1988, Semenza et al.

1997). Besides production, aphasia also affects language comprehension. It has been found that prompted with an idiom, aphasics select a non-literal interpretation over a literal interpretation of the idiom (Van Lancker & Kempler 1987). However, the aforementioned effects affect different sequences differently. Thus the notion that they are equivalent to words may be too simple; there may be different classes of formulaic sequences (Wray 2002). It is clear that aphasics are able to use formulaic sequences, and their usage enables aphasics to recover some degree of fluency, appearing less impaired than they actually are. Thus, in assessing the extent of an impairment it may be necessary to examine patients beyond just fluency and output structure (ibid.).

Language change

It has been hypothesized that formulaicity drives language change (Booij 2010; Bybee 2003; Bybee & Cacoullos 2009). The proposal is that content words located in high-frequency sequences can undergo phonological reduction and drift from its independent meaning (Wray 2012). Over time, such words become more associated with the meaning of the overall sequence, thus they are perceived to have a grammatical role (ibid.). One example of this is the sequence *be going to*, which mainly entailed physical movement a few centuries ago, with the meaning of future intention only implied (Beckner et al. 2009). Eventually, it lost its movement-related meaning and took on the intentional function accompanied by phonological reduction to *(be) gonna* (ibid.). Adding to this, Bybee (2003) argued that frequency or repetition is a major factor of grammaticalization. Repeated usage of the same phrase leads to semantic bleaching due to habituation, and phonological reduction (ibid.). The components in the sequence disassociate from their independent meanings and the sequence becomes semantically opaque, enabling it to appear in more contexts and gain new pragmatic associations (ibid.). As the sequence becomes more autonomous and entrenched in the language, it preserves obsolete morphosyntactic structures (ibid.). Not only that, Bybee & Cocoullos (2009) argued that grammaticalization in high-frequency sequences can spread elsewhere in language. Focusing on constructions, they found that tokens of the

construction with the highest frequency were the earliest and fastest to grammaticalize by semantically bleaching and showing sings of 'unithood'. The high-frequency tokens of the construction attracted newer types of lexical items and increasing the construction's productivity. For example, Bybee & Cocoullos (2009) examined the *can* construction in English, originating from Old English *cunnan* 'to know', and how it acquired the meaning of 'being able to'. They determined that tokens of the construction, *kan/can seye* (Modern English *can say*) and *kan/can telle* (Modern English *can tell*), as formulaic tokens of the construction, based on frequency and their roles as rhetorical devices. They found that *kan/can seye* and *kan/can telle* occur more frequently to convey ability rather than knowledge compared to other tokens of *kan/can*. They take this as indicating that *kan/can seye* and *kan/can telle* has lead the grammaticalization of the *can* construction as a whole towards the ability function.

Natural language processing

Formulaic sequences pose a challenge for various tasks in natural language processing (NLP), where they are commonly referred to as 'multiword expressions'. General methods in various tasks face an overgeneration problem (producing grammatical, compositional output that are unnatural-sounding) and an idiomaticity problem ('knowing' that a sequence's meaning is not perfectly predictable from its constituents) (Sag et al. 2002). As such, there is research on different NLP tasks devoted to handling multiword expressions to improve existing methods. In machine translation, some of the approaches that have been proposed are as follows: 1) using a lexicon containing multiword expressions (Koehn et al. 2003), 2) aligning multiword sequences in a source corpus to the equivalent in a target corpus (Bouamor et al. 2012), and 3) treating multiword sequences as a single token before training an alignment model (Lambert & Banchs 2005). Next, another area concerned with multiword sequences is information retrieval. Accounting for multiword sequences is useful in this area; multiword sequences have higher information content and specificity than single terms. Because of this, multiword expressions more accurately represent a text

content. Thus they are better for making queries and for ranking documents (Vechtomova 2005). Yet another area concerned with multiword sequences is word sense disambiguation, which is the assignment to each word in a text the appropriate entry from a sense inventory. A word sense disambiguation algorithm that cannot correctly detect the multiword expressions that are listed in its sense inventory will not only miss those sense assignments, it will also incorrectly assign senses to constituents that themselves have sense entries (Finlayson & Kulkarni 2011). Also important for this task is detecting multiword expressions, because they are less polysemous than single words, which can reduce the number of possible senses for a string containing multiword expressions (ibid.)

2.2. Affix formulaicity

To bring this introductory section closer to the theme of this study, this section reviews some research on affix formulaicity. The definition of formulaicity by Wray (2002) presented in Section 2.1.1 can accommodate sequences of affixes or morphologically complex words. Although most formulaicity research is on formulas of words, there is research on affix formulaicity to a lesser extent.

The research on affix formulaicity encompasses psycholinguistics and corpus linguistics. In psycholinguistics, it has been accepted that individual morphemes are recognized in processing (Taft & Forster 1975). This means that in language processing a multimorphemic word is decomposed into its individual morphemes. This is in contrast to full-listing processing where polymorphemic words are processed as wholes (Manelis & Tharp 1977). However, some language processing models have been proposed that have both types of processing (Gonnerman et al. 2007).

Experimental studies have shown that some words that are composed of distinct morphemes can nevertheless be processed as wholes. Evidence for this comes from findings of two types: difference in processing loads of polymorphemic words and difference in the

perception of compositionality of polymorphemic words. Some of the findings of the first type of evidence are as follows. Using a lexical decision task on nouns in English, Sereno & Jongman (1997) compared nouns that are more common in the singular than in the plural (high-base, low-plural) and vice versa (low-base, high-plural). When stimuli were presented in the singular, reaction times were faster for high-base, low-plural nouns. Conversely, when stimuli were presented in the plural, reaction times were faster for low-base, high-plural nouns. The result was taken to imply that such high-frequency affixed words are stored as wholes. A similar relationship between affixed form frequency and processing speed has been found for other languages and other classes of affixes (e.g., New et al. (2004) for French, Soveri et al. (2007) for Finnish, Bertram et al. (2009) for Dutch).

Next, some of the findings of the second type of evidence concerning compositionality are as follows. Wurm (1997) and Hay (2001) used an affixedness rating task and found that subjects could judge different degrees affixedness for similarly affixed words, and decide which of two complex words is more complex than the other (e.g., *settlement* was reported as 'more affixed' than *government*). This implies that some affixed words were perceived to be more fused with their affixes. According to Hay (2001), derived forms that were judged to be less complex are the ones that are more frequent than their bases. In addition to that, Marslen-Wilson et al. (1994) compared pairs containing stems and derived forms that are either semantically transparent (e.g., *happiness ~ happy*) or semantically opaque (e.g., *apartment ~ apart*). If accessing a morphologically complex word involves accessing the stem, this should affect response to the stem alone. Using a lexical decision task, they found that semantically opaque word pairs did not show a priming effect, while semantically transparent pairs did. The interpretation is that semantically opaque polymorphemic words are stored as morphologically simple lexical entries despite their apparent morphological compositionality.

There have also been some corpus studies on affix formulaicity. For example, Hefferman & Sato (2017) examined the usage of *mitai-na* 'similar to (similar-attributive)' in a corpus of Japanese. They found that the full form is highly frequent compared to the

constituents overall and that *mitai* is more likely to occur with *-na* than other adjectives than can take the affix. The authors concluded that this indicates that *mitai-na* is behaving as a single unit.

Another corpus study on affix formulaicity is Durrant (2013), which involved a larger set of affixes (all verbal suffixes) using a corpus of Turkish. That study found the following. First, suffixes were strongly collocated with a limited set of other suffixes but not others. Second, the study found that there were some high-frequency three-morpheme sequences that occurred across a wide range of verb roots. These three-morpheme combinations were interpreted to be behaving as a unit. Third, verb stems tended to attract certain morpheme bundles and repel others, and vice versa.

However, the study has several design drawbacks, which weaken its generalizability. The main issues are the corpus data and the measurement used to capture affix formulaicity. First, the corpus data used by that study came from a collection of newspaper articles (news and opinion pieces) personally collected by the author in the course of personal news consumption. In terms of size, the dataset consisted of 765 texts and 375,000 words. Thus, the corpus was small, genre-unbalanced and unlikely to be representative of language use in general. However, Durrant (2013) justified this on the grounds that the corpus should represent the language experience of a single actual language user. However, even on that basis, the more appropriate data would need to be output data rather than input data, because input data are more likely representing the language behavior of individuals other than the subject. As an improvement, this study is designed to capture the language experience of Turkish users in general. Also related to data source is that the study only considered affixes attached to the 20 highest-frequency verbs in the corpus. The focus on verbs alone is justified; verbal morphology is the most productive and abundant in the language, with each verb being able to host numerous suffixes. However, it is clear that the amount of data covered can be increased substantially. Second, the study used frequency (raw frequency and percentage of occurrence) as a measurement of formulaicity between affixes. For example, two suffixes were considered to form a formula if the other suffix was

the most frequent collocate for the first suffix (percentage of occurrence). Another example, a three-morpheme bundle was considered a formula if it is one of the most frequent three-morpheme bundles in the corpus (raw frequency). However, frequency as a measure of formulaicity can lead to misleading results (see Section 3.3).

In response to this, what is proposed is a replication study of Durrant (2013). The main improvements are in terms of corpus data and measurement of formulaicity. First, the source of data for this study is the Turkish National Corpus. Using a more balanced, representative and larger corpus should lead to more generalizable results. As for data coverage, the suffixes to be observed are from over 700 highest-frequency verbs in the corpus. The second type of design improvement is in terms of the method of capturing affix formulaicity. For this study, what is proposed is that statistical association is used for this purpose. Thus, it is assumed that association is an operationalization of formulaicity. More specifically, the measurement of association to be used is risk ratio. Measurements of association go beyond simple frequency and is more robust as they take into account the frequency of each item in a sequence. The corpus and the choice of measurement of association are discussed in Sections 3.1 and 3.3 respectively.

2.3. Turkish morphology overview

An overview of Turkish morphology is provided in this section. This is to aid the reader in contextualizing the suffix cooccurrence data in this study. The review here is based on the Turkish grammar by Göksel and Kerslake (2005), which is an often-cited descriptive grammar of Turkish. The scope of this overview is as follows. Turkish has an entirely suffixing morphology. The only exceptions are compounding, and partial and full reduplications, the latter two of which are restricted to a small set of words. Although Turkish has both derivational and inflectional morphemes, this overview focuses on inflectional ones because they are the scope of this study. This focus is possible because derivational suffixes attach to stems before any inflectional suffixes. Although this study

focuses on verbal suffixes, nominal morphology will be described as well. This is because in subordinate clauses the subordinate verb is nominalized in that it can host nominal suffixes.

Nominal morphology

The suffixes that attach to nouns are the plural, possessive and case markers. There is a possessive morpheme for each person. The case suffixes are the dative, accusative, ablative, locative, instrumental/comitative and genitive. However, there is no explicit marker for the nominative case. Table 1 lists these nominal suffixes. (Capital letters in suffixes represent vowels that alternate in phonological processes such as vowel harmony or devoicing. Letters in parentheses indicate segments that appear in epenthesis).

Table 1: Nominal inflectional suffixes

Function/meaning	Suffix	Function/meaning	Suffix
Plural	-lAr	Case	
Possessive		• Dative	-(n)A
• Third person singular	-(s)I	• Accusative	-(n)I
• Third person plural	-(s)I	• Ablative	-(n)DAn
• Second person singular	-(I)n	• Locative	-(n)DA
• Second person plural	-(I)nIz	• Instrumental/comitative	-(y)lA
• First person singular	-(I)m	• Genitive	-(n)In
• First person plural	-(I)mIz	Relative	-ki

When multiple suffixes are attached to a noun, the order follows the following template:

	Noun base	+ Plural	+ Possessive	+ Case	+ Relative
e.g:	sokak	-lar	-ımız	-da	-ki
	street	-PL	-3.PL.POSS	-LOC	-REL
	‘the ones on our streets’				

Verbal morphology

There are numerous categories of suffixes that attach to verbs, and verbs can host numerous suffixes at once. In compound verbs, all inflection is on the main verb. Table 2 lists these suffix categories and the specific morphemes under them.

Table 2: Inflectional verb suffixes

Function/meaning	Suffix	Function/meaning	Suffix
Voice		Imperfective	-mAktA
• Reflexive	-(I)n	Necessitative	-mAlI
• Reciprocal	-Iş	Optative	-A
• Causative	-DIr, -t, -(A/I)r	Copular	
• Passive	-Il, -(I)n	• Past	-(y)DI
Negation	-mA	• Evidential	-(y)mIş
Modality		• Conditional	-(y)sA
• Possibility	-Abil	Assertion	-DIr
• Non-possibility	-AmA	Subordinate	
• Non-premeditative	-Iver	• Subjunctive	-mA
Aspect		• Past indicative	-DIk
• Progressive	-(I)yor	• Future indicative	-AcAk
• Future	-AcAK	Adverbial	
• Aorist	-(I/A)r	• 'while doing'	-(y)kAn
Tense		• 'by doing so'	-ArAk
• Past	-DI	• 'after doing so'	-Ip
• Evidential	-mIş	• 'without doing'	-mEdEn
• Conditional	-sA		

Some suffixes may appear to be in incorrect categories (e.g., future in aspect, conditional in tense). Such category memberships are due to that the suffix patterns like the other suffixes in the category.

After the suffixes in Table 2, person suffixes are attached last in most cases. The person suffixes consist of several paradigms. The *-k* paradigm is used when the preceding suffix is the evidential *-mIş* or the past tense *-DI*. In other cases the *-z* paradigm is used[1]. The *-z* paradigm is also used in copular sentences, where the personal ending is attached to the predicate (e.g., a noun or an adjective). There are also separate paradigms for verbs in the optative and the imperative. The paradigms of person suffixes are shown in Table 3.

Table 3: Person suffixes

Person	1st		2nd		3rd	
	Singular	Plural	Singular	Plural	Singular	Plural
Nominative pronouns	ben	biz	sen	siz	o	onlar
Possessive	-(I)m	-(I)mIz	-(I)n	-(I)nIz	-(s)I	-(s)I
Verbal endings						
• *z* paradigm	-Im	-Iz	-sIn	-sInIz	ϕ	*-lAr
• *k* paradigm	-m	-k	-n	-nIz	ϕ	-*lAr
• Optative	-(y)AyIm	-(y)AlIm	-(y)EsIn	-(y)EsInIz	-(y)A	-(y)ElEr
• Imperative	N/A	N/A	-*In	-(y)In(Iz)	-sIn	-sInlAr

Asterisks indicate optional suffixes.

Inflection on verbs follows several different 'paths' specifying possible sequencing of suffixes. These paths are as follows:

[1] The names of these two paradigms are based on the contrast within the first person plural.

Base + Voice + Negation + Modality

→ Aspect + Tense + Copular + Person + Assertion

→ Imperfective/Necessitative/Optative + Copular + Person+Assertion

→ Subordinate + Possessive + Case

→ Infinitive + Case

→ Adverbial

→ Imperative

Morphosyntax

A description of Turkish morphosyntax is also provided here, because some inflectional morphology is driven by syntactic relationships between words in a sentence, such as agreement. The default word order in the language is (S)OV. Deviations from this do occur but mainly for deemphasis, by which the deemphasized element is placed after the verb. In main clauses, the main verb or the copular predicate must agree with the subject, even if the subject is dropped. Agreement is achieved by attaching the personal verbal suffix corresponding to the subject. Also, agreement is required in copular sentences, where the personal verbal suffix is attached to the predicate:

Non-copular:	(Ben/Sen/O)	bir	kedi	bul-du-m/n/ϕ
	(1.SG/2.SG/3.SG)	one	cat	find-PAST-1.SG/2.SG/3.SG
	"(I/You/It) found a cat"			

Copular:	(Ben/Sen/O)	hasta-yım/sın/ϕ
	(1.SG/2.SG/3.SG)	sick-1.SG/2.SG/3.SG
	"(I/You/It) am/are/is sick"	

Note that the person suffixes in the non-copular and copular sentences are from the *-k* and *-z* paradigms respectively.

Conversely, there is no subject agreement for verbs that are not the primary verb in the clause. This would the case when the verb ends with the infinitive or adverbial suffixes:

Biz	yür-erek	gel-di-k
1.PL	walk-'by doing so'	come-PST-1.PL
'We came by walking'		

Ye-mek	ist-iyor-uz
eat-INF	want-PROG-1.PL
'We want to eat'	

Another aspect where agreement is required is in the possessive construction. In this construction, the possessed noun has to agree with the possessor, even if the possessor is dropped. When the possessor is a pronoun, a set of suppletive genitive pronouns is used. If the possessor is a noun and not a pronoun, the genitive suffix is attached to the possessor.

(Onun) ev-i	3.SG.GEN house-3.SG.POSS	'Its house'
(Levent-in) ev-i	Levent-GEN house-3.SG.POSS	'Levent's house'
(Onların) ev-i	3.PL.GEN house-3.PL.POSS	'Their house'
(Benim) ev-im	1.SG.GEN house-1.SG.POSS	'My house'
(Bizim) ev-imiz	1.PL.GEN house-1.PL.POSS	'Our house'
(Senin) ev-in	2.SG.GEN house-2.SG.POSS	'Your house'
(Sizin) ev-iniz	2.PL.GEN house-2.PL.POSS	'Your (pl.) house'

Unlike in the possessive construction, the genitive is omitted in noun-noun compounds where the two nouns are not in a possessor-possessed relationship. Nevertheless, the third person possessive is still attached to the head noun:

	köpek	balığ-ı
	dog	fish-3.SG.POSS
	'Shark (lit. dog fish)'	

Compare to:	köpeğ-in	balığ-ı
	dog-GEN	fish-3.SG.POSS
	'Dog's fish'	

So far, subject-verb agreement in main clauses has been explained, but not subject-verb agreement in embedded clauses. Agreement in embedded clauses follows the genitive-possessive pattern in possessive constructions. First, verbs in embedded clauses are either attached with the subjunctive subordinate suffix *-mA* or with indicative subordinate suffixes *-AcAk*/*-DIk* (SUB.FUT/SUB.PAST). Next, the embedded subject is in the genitive, and the embedded main verb agrees with the embedded subject using personal suffixes from the possessive paradigm (examples 1, 2 below). Intuitively speaking, the embedded subject 'owns' the embedded verb. Additionally, if the embedded phrase is a complement of a verb in the main clause, the embedded verb receives the appropriate case marker (examples 3, 4 below).

(1) [Levent'-in bul-duğ-u] baykuş
Levent-GEN find-SUB.PAST-3.SG.POSS owl
The owl that Levent found'

(2) [Levent'-in öl-me-si] lazım
Levent-GEN die-SUB.SUBJ-3.SG.POSS necessary
'That Levent dies is necessary'

(3) [Bizim zıpla-dığ-ımız-ı] biliyor
1.PL.GEN jump-SUB.PAST-1.PL.POSS-ACC he.knows
'He knows that we jumped'

(4) [Bizim zıpla-ma-mız-ı] istiyor
1.PL.GEN jump-SUBJ-1.PL.POSS-ACC he.wants
'He wants us to jump'

Chapter 3. Methods

This section details the corpus, data collection and processing procedure, and the numerical measurement of association used in this study.

3.1. Corpus data source

The corpus used in Durrant (2013) was a collection of 765 articles from 7 online newspapers collected over 6 months amounting to 375,000 words. The data were collected in the course of the author's personal news consumption.

As an improvement on this, the data source chosen for this study is the Turkish National Corpus (TNC). Its design and construction are detailed in Aksan et al. (2012). At 50 million words, the TNC is designed to be a balanced, representative corpus containing data of contemporary Turkish over a span of 20 years, from 1990 to 2009. The data consist of 98% written sources and 2% transcribed speech. Modeled on the British National Corpus, data in the TNC were sampled proportionally from a variety of types of sources.

Textual data in the corpus cover informative (i.e. nonfictional) and imaginative (i.e. fictional) domains. The textual data consist of 81% informative texts and 19% imaginative texts. The text mediums in the TNC consist of books, periodicals, published texts, unpublished texts (e.g., student essays, emails, blogs) and texts written for spoken delivery (e.g., news script, screenplays). Table 4 shows the composition of textual data in the TNC.

Table 4: Composition of the TNC

Text domain	Percentage	Text medium	Percentage
Imaginative	19	Books	58
Social science	16	Periodicals	32
Art	7	Miscellaneous published documents	5
Commerce/finance	8	Misc. unpublished documents	3
Belief and thought	4	Speech documents	2
World affairs	20		
Applied science	8		
Natural science	4		
Leisure	14		

The 2% of spoken data in the TNC are sourced from spontaneous, informal conversations and speeches in formal settings such as meetings and lectures. The spoken part of the TNC constitutes a million words, equally divided between transcriptions of informal and formal speech.

The TNC can be accessed via its website, www.tnc.org.tr, using a free account. Through its browser-based user interface users can perform search queries for corpus data based on words, lemmas or affixes. After a search is performed, the interface presents the relevant data points. Each data point consists of the search term (or word token if searching by lemma) and concordance windows (up to five words) on the left and the right. The edges of the concordance windows may or may not correspond to sentence boundaries. Register information (spoken or written) and metadata are also included in each data point. The data returned from a single search query can be downloaded as files in the format of .csv (comma-separated values) or .tsv (tab-separated values). Figure 2 shows the structure of information in a downloaded TNC data file as viewed in a spreadsheet software.

	A	B	C	D	
1	Register	Text	Left	-5;5	Right
2	w	W-GC06E1B-3080-1821	Ψ(canlı Ψ(canlı kedi)+Ψ(ölü kedi)+Ψ(ölü kedi)+Ψ(ölü	kedi)	şeklinde olmalı. "Bettlemann'ın çorapları" hikâyesinin
3	w	W-IA16B2A-2672-2487	şık arabaların durduğu sokakta, şimdi	kediler	çöp tenekelerini karıştırıyor, karşı apartmandan
4	w	W-UE36E1B-3358-26	şölene dönüşen spor festivalleri, semirmiş	kedileri,	çimenlerinde kablosuz internet erişimi, solak
5	w	W-UE36C0A-1679-885	şuursuzca, sırf eğlence olsun diye	kedi	asan, kuş kafası koparan çocuk
6	w	W-EA14B1A-1616-499	Şurda küçük bir kız bir	kedi	kurtarmış. İmza istiyor diye olay
7	w	W-UE36E1B-3358-2413	şu kediyi o pencereden atamazsınız	(kedisini	gösteriyor). Duble, yani iki ayrı
8	w	W-CA16B2A-0159-1594	Şoparlar papiklendikleri vakit, işe çıkardıkları	kedileri,	en uzak yıldızlara kadar fırlatabiliyorlardı.
9	w	W-FA16B2A-0998-155	şimdi... Hayır canım martı değil	kedileri...	Evet Norma bir süre buraları
10	w	W-CA16B3A-2682-1231	şimdi. Gülünç olursun. Biri görse...	Kedinizi	getirdim, dedi bir ses. Bir
11	w	W-GA16B1A-0732-1850	Şimdi vur da, gör gününü."	Kedi	bir pati vurdu, "ebeee!" dedi
12	w	W-GC06E1B-3080-1059	şimdi makroskobik bir sistem olan	kedinin	kaderi de anık parçacığın davranışına
13	w	W-TC29D1B-2798-998	şiddetlenen kalp atışları ve ter;	kedi	ile karşı karşıya kalan farenin
14	w	W-SI45F1D-4793-688	şeyler katmaksızın yapabildiğim şeyleri düşünüyorum.	Kedilerimi	seyretmek mesela. Bunu saatlerce bıkmadan
15	w	W-CA16B1A-0505-1630	şeyler fısıldadı. Sonra süt dökmüş	kedi	gibi girdi içeri. Semaverin altını
16	w	W-OG37E1B-2922-1727	şeyler dönmüş olmalı. Ev darmadağın.	Kedi	girmiştir herhalde, çok yüz verdim
17	w	W-DA16B2A-1325-1937	şeyler de öğrendim bu arada.	Kedileri,	sokak kedilerini, uyuz kedileri besleyen
18	w	W-NE36C3A-0079-101	şeyi özetliyordu; "Ben Sizin Sandığınız	Kedi	Değilim." Çoğu insanın aksine ve
19	w	W-EA16B2A-0448-2325	şey, bir başkasının yemeği. Yaban	kedisi	ırmaktan balık pençeler. Kara kartal,
20	w	W-TA15B2A-0545-538	şey yaptım Onu kucakladım... Bir	kedi	geçti Önümden, Önemsemedi bile Çekti
21	w	W-EA16B2A-1205-2111	şey taşıması gerekiyor. Ayrıca, kara	kediyi	uğursuz sayıyorlar. Onlara göre, sayıların

Figure 2: View of concordance lines in a TNC data file

Contents in the 'Text' column would appear as buttons in the web interface. Clicking the button therein reveals text metadata for the associated data point.

A relevant limitation of the corpus interface is that data cannot be searched for by part-of-speech. This impacts this study's procedure in that it is not possible to retrieve all strings that are verbs. However, the interface does allow searching by a lemma with a specified POS. Given this feature, the workaround to said limitation is performing multiple individual searches based on a list of verb lemmas that would collectively return a substantial amount of data. The extensiveness of the verb list and the basis for it ensure that the procedure would result in a wide coverage of verbs in the corpus. The list of verb lemmas used in this study is detailed in Section 3.2.

3.2. Procedure

This section details the procedure of processing data from the TNC. The end product of this procedure is data on the distribution of suffixes in the corpus. The data processing here was carried out by computer programs written in Python specifically for this study. The programs are stored and can be accessed at the following repository on the version control website Github: https://github.com/heikalb/thesis-scripts. The programs were written to output various data files (e.g., files of morphological parses, files of association measurement data). However, most of the data files themselves are not on the Github repository due to memory restrictions. Generating those data files require cloning the repository onto a local computer and running the programs therein. However, the files containing the data in the main analysis of this thesis, the risk ratio data, are available on Github to help with replication or scrutiny of this study or reuse for a different study.

The outline of the data processing procedure is as follows:

1. Prepare query terms
2. Perform iterative queries on the TNC interface
3. Download data file after each query
4. Spell-checking
5. Morphological parsing
6. Isolate verb parses
7. Derive suffix collocate pairs
8. Calculate values of association measures on collocate pairs
9. Repeat step 8 by register and verb stem

The first step was obtaining verbs from the TNC. In Durrant (2013), the verb types that were used were the 20 most frequent verb types in the corpus of that study. However, this study is designed to cover a wider scope of verbs. As stated in Section 3.1, the TNC web interface does not allow searches by POS alone. To overcome this, iterative searches were performed on different verb types. For this purpose, *A Frequency Dictionary of Turkish* (Aksan et al. 2017), which lists the 5000 most frequent words in the TNC, was referred to. All 732 verb types (in the form of the stem) were extracted from the dictionary. That the

frequency dictionary is based on the TNC and that the verb list is also based on frequency make it likely that the search results will be a comprehensive coverage of verbs in the corpus. All of the 20 verb types in Durrant (2013) are also in the *Frequency Dictionary* verb list. Table 5 lists the 20 most frequent verb types in the *Frequency Dictionary* and all of the 20 verb types used in Durrant (2013).

Table 5: First 20 verb types (in stem form) in this study and Durrant (2013)

In *A Frequency Dictionary of Turkish*		In Durrant (2013)	
ol	**'be'**	**ol**	**'be'**
et	**'do/make'**	**et**	**'do/make'**
yap	**'do/make'**	**yap**	**'do/make'**
al	'take'	**ver**	**'give'**
de	**'say'**	**de**	**'say'**
gel	'come'	**çık**	**'emerge'**
ver	**'give'**	**çalış**	**'work'**
gör	'see'	konuş	'talk'
çık	**'emerge'**	**geç**	**'pass'**
bul	'find'	yaşa	'live'
git	'go'	gir	'enter'
çalış	**'work'**	**bak**	**'look'**
iste	'want'	bırak	'leave'
geç	**'pass'**	**anla**	**'understand'**
bil	'know'	geliş	'develop'
anla	**'understand'**	sağla	'provide'
kal	'remain'	yarat	'create'
söyle	'say'	koru	'protect'
bak	**'look'**	paylaş	'share'
ye	'eat'	**önle**	**'prevent'**

Overlaps between the two lists are in bold.

Next, each of the 732 verb stems in the list was separately entered as a query on the TNC interface. Since words in the TNC are lemmatized, each search would 'hit' any word that is a morphological variant of the query word. After the search results were returned, the data file of the query results was downloaded. The histogram in Figure 3 shows the

number of data points (verb tokens and their concordance windows) by each verb type in the corpus, which is equivalent to the token frequency of each verb type. Collectively the queries returned 1.49 million data points.

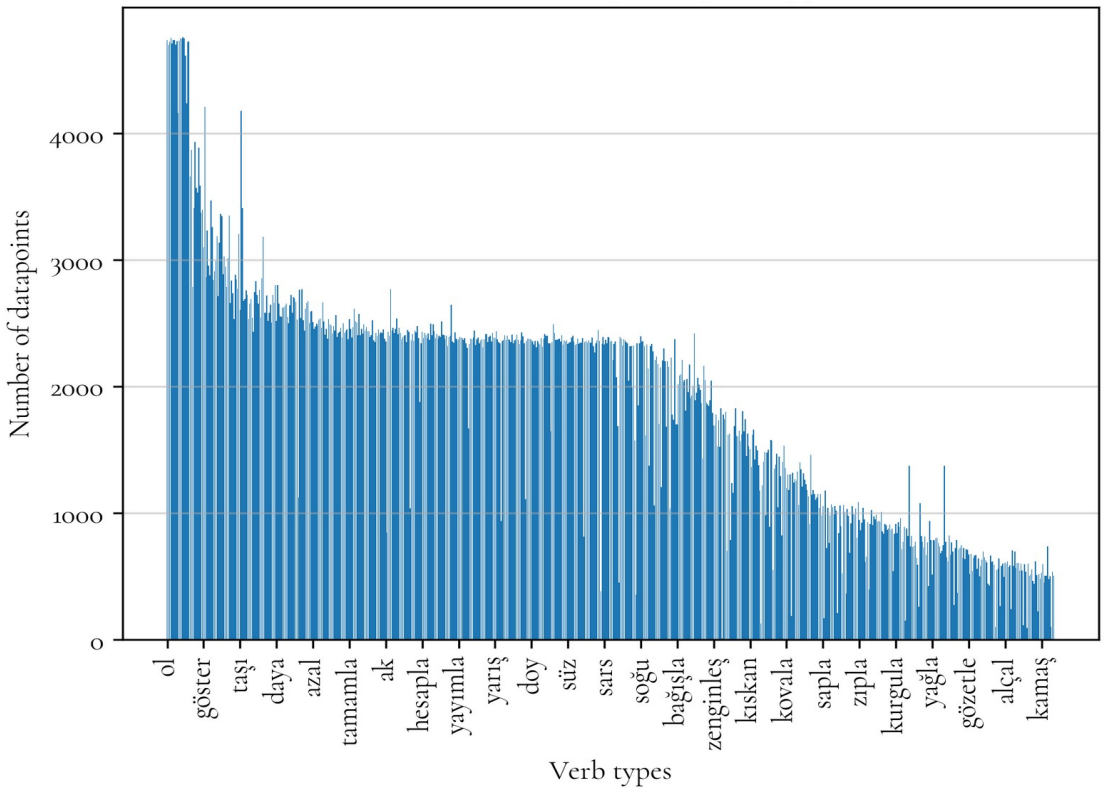


Figure 3: Distribution of data points by verb types

Verb types (not all labeled) arranged in descending frequency from left to right on the x-axis.

Next, verb tokens in all data points were spell-corrected using a set of spelling substitutions that take into account colloquial spelling, which exists in the corpus due to the range of genres of data. This was done as opposed to just correcting any spelling error found because a reliable spell correcter could not be found. Also, it appeared that colloquial

spelling accounts for the vast majority of spelling deviations in the dataset. The spell-correction step was needed because correct and formal spelling was needed for the morphological parser to operate properly.

Next, for each data point, all of the words therein were morphologically parsed. This means each word was decomposed into its constituent stems and suffixes. The morphological parser that was used comes from Zemberek-NLP (Akın & Akın 2018), a library written in Java. It can be freely accessed on the following Github repository: https://github.com/ahmetaa/zemberek-nlp. Although the focus of the study is on suffixes on verbs, all the other words in the 10-word context windows were parsed as well; parses of surrounding words enable the parser to disambiguate the parsing of the target verbs in cases where there are multiple plausible parses for the target verb. Parsing errors resulted when the parser was unable to produce a parse or when the parse labels the base word as a non-verb. There were 9438 parsing errors, reducing the number of usable verb parses from 1.49 million to 1.48 million.

Next, for every remaining verb parse, suffix pairs were created based on two-suffix combinations within the parse. To illustrate, given the parse of *stem + suffix1 + suffix2 + suffix3*, the pairs (*suffix1, suffix2*), (*suffix1, suffix3*), (*suffix2, suffix3*) are derived. These shall be referred to as collocate pairs. The collocate pairs are not restricted to bigrams (adjacent pairs); it is not presumed that association between suffixes is only local. Then these collocate pairs were tallied across the dataset. As a result, there were 1108 collocate pair types and 3.1 million collocate pair tokens. Table 6 shows summary statistics of the dataset.

Table 6: Summary statistics of the dataset

	Type frequency	Token frequency
Verb stems	732	1.48 million
Suffixes	77	3.5 million
Collocate pairs	1108	3.1 million

After deriving collocate pairs, a measurement of association was calculated for each pair. For this study that measurement is risk ratio, a choice justified in Section 3.3. Because risk ratio is not symmetric, two risk ratio values were calculated for each collocate pair. For example, given the collocate pair (*suffix1, suffix2*), two risk ratio values were calculated: one with *suffix1* as the conditioning variable and one with *suffix2* as the conditioning variable. Furthermore, separate sets of calculations were done on various divisions of the dataset: by the whole dataset, by register (written or spoken) and by verb stem. However, risk ratio was not the only measurement of association that was considered. Hence, the values of the other measurements were also recorded in the main data file.

The end product of the above procedure are .csv files containing values for every association measure for every collocate pair derived from the corpus. Association data for the whole dataset can be found on this study's Github repository at the directory path d5_statistics/association_stats/000__association_stats.csv. Association data for the entire written and spoken registers are located at d5_statistics/association_stats_written/000__association_stats_written.csv and d5_statistics/association_stats_spoken/000__association_stats_spoken.csv respectively. Association data for stems and suffix trigrams are located at d5_statistics/trigram/stem_trigram_rr.csv. The data files should appear as in Figure 4 when opened in a spreadsheet software.

	A	B	C	D	E	F	G
1	collocate_pair	risk_ratio	risk_ratio_reverse	odds_ratio	mutual_information	dice_coefficient	t_score
2	('Gen', 'Rel→Pron')	41919.4074074074	33024.25464191	47701.2567049808	7.24599448468746	0.000348887919756	1.98682560535
3	('NarrPart→Adj', 'Ness→Noun')	9044.10901162791	1812.2544395924	4528.63609898108	10.3276795041911	0.001476014760148	1.41311350989
4	('NarrPart→Adj', 'Aslf→Adj')	4522.05377906977	1512.4115646259	4535.23469387755	10.3276795041911	0.002949852507375	1.99844486512
5	('NarrPart→Adj', 'Ly→Adv')	2261.02688953488	1133.4825200291	2265.96504005827	10.3276795041911	0.00221320545924	1.73070377260
6	('With→Adj', 'Become→Verb')	1268.73793016618	446.35778349046	1834.34038554749	5.62851258700001	0.077434429196696	47.5432791957
7	('Loc', 'Rel→Adj')	1019.77069339112	1399.2638807678	2050.08553061669	7.22364022551821	0.059504958746562	26.5447225374
8	('Inf3→Noun', 'Equ')	665.876711910003	603.82843205189	708.57521634459	11.1836396285767	0.157556270096463	6.99703981028
9	('Related→Adj', 'Become→Verb')	398.246598203323	439.33139951841	572.688623459173	5.90702862977807	0.024726305378398	26.4798616309
10	('PresPart→Noun', 'Without→Adj')	353.090924647767	118.37033656589	176.555504848831	7.12521200970446	0.000160436386973	0.992837101272
11	('FeelLike→Noun', 'Equ')	262.567035748181	266.73677444237	269.987377187846	9.29762776264406	0.026845637583893	1.99682306519
12	('Agt→Noun', 'Become→Verb')	257.081227078997	341.06715075465	344.381645401382	5.68662476153508	0.001035357457162	5.37090398834
13	('Caus→Verb', 'ActOf→Noun')	237.408051753911	8.6263937882883	118.709103718469	3.30102435221643	0.000192544670364	3.48003995037
14	('FeelLike→Noun', 'P1sg')	116.534239990338	321.42318411427	323.162784588441	7.45921908675539	0.014009188822776	9.58897734448
15	('PresPart→Noun', 'Rel→Pron')	78.4646372457678	54.641603504044	78.4823161725084	8.12521200970446	0.000641642605069	1.99283810390

Figure 4: Screenshot (partially obscured) of a data file of association measurement values

3.3. Measurements of association

This section explains the numerical method used to capture affix formulaicity. What is proposed is the use of association to measure formulaicity. Specifically, for every collocate pair, a value of a measurement of association between the suffixes therein was calculated. As a justification for the choice of risk ratio for this study, this section compares the following measurements of association that can be found in the literature of corpus linguistics and natural language processing:

1. Frequency
2. Pointwise mutual information
3. Dice coefficient
4. t-score
5. Pearson's chi-squared
6. Risk ratio/odds ratio

Before proceeding with the comparison, it may be helpful to clarify the conceptual basis of the numerical approach in this study. Most importantly is the relationship between the terms that will be encountered: association, formulaicity and collocation. To begin, measures such as pointwise mutual information or Dice coefficient are different ways of expressing association. Association, in turn, is a statistical concept that concerns the relationship between variables. Studies on the distribution of items in a corpus have used association as an operationalization of collocation between items. This study assumes association as an operationalization of formulaicity; if a sequence is formulaic then the items within should be associated with each other. However, collocation and formulaicity are not equivalent. Collocation is a distributional concept; it concerns how elements are distributed in relation to each other (Clear 1993). In contrast, formulaicity is fundamentally a psychological concept; it concerns the unitary status of a sequence of elements (Wray 2002). Formulaic status could manifest as distributional behavior in a corpus. Thus, elements in a formulaic sequence may also appear as collocates.

Next, before proceeding with the comparison, a useful construct to explain in conceptualizing these measurements is that of contingency tables, which capture the relationship between categorical variables. 2x2 contingency tables show one variable as columns and the other variable as rows. The number of columns or rows depends on the number of possible values for a variable. Higher-order contingency tables with more than two variables are also possible. Each cell in a contingency table is populated with the frequency of observations of the intersecting variables.

For a concrete nonlinguistic example, consider a hypothetical study that seeks to find whether cats or dogs are more likely to be collared. In this study's sample, individuals vary on two variables: 1) pet type (cat or dog) and, 2) collared-ness (uncollared or collared). In this case, one may consider pet type as a conditioning variable and collaredness as a response variable or event. The cells in Table 7 show the frequencies of individuals who are uncollared cats, uncollared dogs, collared cats and collared dogs. In addition to these, marginal total, that is the total frequency in an entire row or column, may also be displayed at the edges; in this case, the marginal totals are the number of all cats, all dogs, all uncollared pets and all collared pets. The grand total, or the number of individuals in the table, may be displayed on the bottom right corner of the table.

Table 7: Frequencies of collar-wearing by pet type in a hypothetical sample

	Pet type		
Collaredness	Cats	Dogs	Total
Uncollared	400	300	700
Collared	2	100	102
Total	402	400	802

In the context of collocational studies, the variables in a contingency table are: 1) the occurrence/non-occurrence of an element (e.g., word, morpheme) *x*, and 2) the occurrence/nonoccurrence of another element *y*. The occurrence of *x* may be considered as the conditioning variable on the occurrence of *y*. The contingency table for this example is

shown as Table 8. Here the frequencies are represented with variable notation that will be used in the rest of this section. A cooccurrence *x*/*x*′ and *y*/*y*′ is only counted if both of them are within a predetermined domain of locality, such as arbitrary word spans, adjacency or linguistic structures (e.g., phrases) (Evert 2005). For example, if the domain of locality is a 10-word span, then *fxy* is the number of 10-word spans in the corpus in which *x* and *y* cooccur.

Table 8: Cooccurrence frequencies of elements *x* and *y* in a hypothetical corpus

	x	*x*′ (not *x*)	Total
y	*fxy*	*fx*′*y*	*fy*
y′ (not *y*)	*fxy*′	*fx*′*y*′	*fy*′
Total	*fx*	*fx*′	$N = fx + fx'$ or $N = fy + fy'$

The following is a review of the aforementioned association measurements.

Frequency

One way frequency is used in finding formulaic sequences in a corpus is by ranking sequences of *n*-items by their number of occurrences (Manning & Schütze 1999). This can also be relativized to examining collocates with respect to a target item by finding the other element that occurs most frequently with the target item (e.g., finding the adjective that is most likely to precede the word 'coffee')(ibid.). If the count of items alone is used, this is considered as raw frequency.

One problem with frequency is that it does not take into account the overall frequency of each element in the collocate. Consequently, two elements can be deemed to be associated by virtue of one or all elements in the collocate being highly frequent (Manning & Schütze 1999). Thus, although element *y* may occur the most frequently with element *x*, there may not be a meaningful association if *y* occurs frequently with many other

elements as well, that is, if *y* is frequent but not specifically with *x*. Because of this, collocations identified by frequency are often compositional and are not lexically particular (Thanapoulos 2002). As an illustration, consider a hypothetical corpus of adjective-noun sequences specified in Table 9. By frequency alone, for the word *bird* the most associated sequence is *large bird*. However, this is likely because *large* occurs so frequently with nouns in general in the corpus. Instead, it is likelier that *migratory bird* is the more associated sequence. Although *migratory bird* is less frequent than *large bird*, *bird* is more predictable if the prior is *migratory* rather than *large*; 4 out of 14 cases of *migratory* are followed by *bird*, whereas 6 out of 986 cases of *large* are followed by *bird*.

Table 9: Hypothetical adjective-noun sequence corpus

Sequence	Count
large bird	6
migratory bird	4
large <other noun>	980
migratory <other noun>	10

Frequency is also dependent on corpus size, thus operates on different scales, making it not comparable across corpora (Gablasova et al. 2017). The solution to this is normalization, such as using percentages (ibid.). In reference to Table 8, the percentage of occurrence of *y* with respect to *x* is calculated as follows:

$$Occurrence(x, y) = \frac{f_{xy}}{f_x}$$

Unlike raw frequency, normalized frequency is not dependent on corpus size, operates on a normalized scale and thus is comparable across corpora. However, normalized frequency

still faces the same problem as raw frequency in that it does not take into account the overall frequency of each element in the collocate.

Pointwise mutual information

Pointwise mutual information[2] is a metric from information theory that measures mutual dependence between two random variables; this is the information gained about one variable after observing another variable. Its use for corpus studies was first demonstrated in Church & Hanks (1990). It has since become one of the most commonly used measurement of collocational strength (Gablasova et al. 2017). In reference to Table 8, the pointwise mutual information I between elements x and y is calculated as follows:

$$I(x, y) = \log_2 \frac{P(x, y)}{P(x)P(y)} = \log_2 \frac{f_{xy}/N}{(f_x/N)(f_y/N)}$$

The probability of an element x (i.d., $P(x)$) is its frequency (i.d., fx) divided by the number of sequences in the corpus, N. The formula compares the probability of observing x and y together with the probability of observing x and y independently. If there is an association between x and y, $P(x, y)$ will be larger than $P(x)P(y)$, therefore $I(x, y) > 0$. If there is no relationship between x and y, $P(x, y)$ will be almost equal to $P(x)P(y)$, therefore $I(x, y) \approx 0$. If x and y are in complementary distribution, $P(x, y)$ will be less than $P(x)P(y)$, therefore $I(x, y) < 0$ (Church & Hanks 1990). The larger the value the more exclusively the two words are associated and the rarer the combination is. As for its scale, pointwise mutual information is a normalized score, thus it is comparable across corpora. It does not have a theoretical minimum and maximum, but contextualizing specific values is nevertheless possible because of the cutoff point of 0 (Gablasova et al. 2017).

[2] This is in contrast to *mutual information*. Pointwise mutual information applies to one pair of variable values, whereas mutual information is an average of pointwise mutual information of all possible pairs of variable values. However, in the literature *mutual information* has been used even though *pointwise mutual information* is meant.

One drawback of pointwise mutual information as a measure of association is that it favors rarer events. This can be seen with fully exclusive collocates (that is collocates where the members only occur with each other and nothing else, such as *ceteris paribus*). Given two fully exclusive collocates, the less frequent one will have a higher pointwise mutual information (Manning & Schütze 1999). This is a disadvantage because the measure rewards sequences for which there is less evidence in the corpus (Gablasova et al. 2017). Pointwise mutual information also tends to highlight rare fully exclusive collocates. Thus it strongly favors more idiosyncratic sequences such as names and specialized or technical terms (ibid.). However, the most crucial drawback is that pointwise mutual information is better as a measure of independence rather than dependence (Manning & Schütze 1999). This is explained as follows. When x and y are independent, $P(x, y)$ is equal to $P(x)P(y)$, so $I(x, y)$ reduces to 0. However, when x and y are fully dependent, $P(x, y)$ is equal to $P(x)$ and $P(y)$, so $I(x, y)$ reduces to either $\log_2 1/P(x)$ or $\log_2 1/P(y)$. Thus for dependence the measure depends on the frequency of the individual words, and collocates with less frequent elements get higher values. This is not a preferable property; ideally, more frequent dependent collocates should score higher because there is more evidence for them.

Dice coefficient

Dice coefficient was introduced to capture the association between species in ecological studies (Dice, 1945). Dice (1945) introduced this as *coincidence index*, but the term *Dice coefficient* can be found elsewhere. Dice coefficient compares (1) the cooccurrence frequency of elements x and y and (2) the individual frequencies of x and y. In reference to Table 8, the Dice coefficient between x and y is calculated as follows:

$$Dice(x, y) = \frac{2 f_{xy}}{f_x + f_y}$$

This is the harmonic mean (an average of ratios) of two proportions: (1) the association of y to x, f_{xy}/f_x, and (2) the association of x to y, f_{xy}/f_y. Each of the two proportions is unidirectional (there may dependence of one element on another but not necessarily the other way around) and each value is dependent on the variable that is the basis of comparison (the denominator variable). By taking the harmonic mean of proportions (1) and (2), Dice coefficient then is a symmetric measure (Dice 1945). Its values range from 0 to 1. 1 indicates that x and y always cooccur, while 0 indicates that x and y never cooccur.

t-score

t-score is a measurement of association that is a derivative of the Student's t-test. In reference to Table 8, the t-score between x and y is calculated as follows:

$$t-score = \frac{f_{xy} - \frac{f_x f_y}{N}}{\sqrt{f_{xy}}}$$

In the numerator, the actual frequency of the collocate, $f_{xy,}$ is subtracted with its expected frequency (i.e. frequency by chance). As with a conventional t-test, the t-score assumes a normal distribution of the data (Manning & Schütze 1999). In its usage, the t-score is not a measurement of the magnitude of association; it is conceptualized as the certainty to which there is an association of any kind (Gablasova et al. 2017). In an analogy to the t-test, the null hypothesis is that the cooccurrence frequency is equal to the level expected by chance.

The following are problems with the t-score as a measure of association. First, applying the Student's t-test to cooccurrence frequency data is mathematically dubious; this lies in the fact that in a conventional t-test the data are a sample of value observations, but for frequency data the sample would have to consist of boolean indicator/dummy variables

(Evert 2005). Because of this, it is not possible to establish a valid rejection region for the null hypothesis (Gablasova et al. 2017). Second, the t-score assumes a random distribution of language, that is the calculation involves an expected frequency, an assumption that Gablasova et al. (2017) found to be problematic for language data. Third, the assumption of approximate normality may not be valid for corpus data. Finally, there are some issues regarding data quantity. The t-score is dependent on corpus size, implying a lack of a standardized scale and thus incomparability across corpora (ibid.). Additionally, as with other significance tests of association, a high association score could be due to either the existence of a strong association of any kind, or it could be due to the availability of a large amount of evidence (Evert 2005). This measure cannot distinguish between the two effects. For this reason, the t-score also tends to highlight frequent combinations. Indeed, it has been observed that rankings based on t-score are similar to rankings based on frequency (Gablasova et al. 2017). However, even this aspect is opaque because similarly high-frequency sequences may differ in t-score rankings. This led to Gablasova et al. (2017) remarking, "While all collocations identified by the t-score are frequent, not all frequent word combinations have a high t-score."

Pearson's Chi-squared

Pearson's Chi-squared is another measurement that is in the paradigm of a significance test. In reference to Table 8, the chi-squared statistic between *x* and *y* is calculated as follows:

$$X^2 = \frac{(f_{xy} - f_x f_y)^2}{f_x f_y} + \frac{(f_{xy'} - f_x f_{y'})^2}{f_x f_{y'}} + \frac{(f_{x'y} - f_{x'} f_y)^2}{f_{x'} f_y} + \frac{(f_{x'y'} - f_{x'} f_{y'})^2}{f_{x'} f_{y'}}$$

Each term in the formula is the squared difference between expected and observed frequencies scaled by the expected frequency. As in any hypothesis test, the chi-squared statistic is calculated to determine if there is significance of any type of association, rather

than a measure of the magnitude of association. Unlike the t-score, the chi-squared test does not assume that the data are normally distributed. However, Manning & Schütze (1999) demonstrated that the two can be similar, with collocates having the highest t-scores in a corpus also having the highest chi-squared scores. Also unlike the t-score, the chi-squared test is originally designed for categorical or count data. Thus its usage for collocational studies is mathematically justified.

A drawback of chi-squared is that it gives a poor approximation when there is a low frequency in any of the cells in the contingency table (Manning & Schütze 1999, Evert 2005). Additionally, as with other significance tests of association, a high association score could be due to either the existence of a strong association, or it could be due to the availability of a large amount of evidence (Evert 2005). This measure cannot distinguish between the two effects and is thus biased towards high-frequency collocates (ibid.)

Risk ratio / odds ratio

Risk ratio (also called relative risk) is a measurement that is not commonly used in collocational studies, but it is widely used in other fields such as medicine and social science (Schmidt & Kohlmann 2008). It expresses how likely an event is given one condition compared to the likelihood of the same event given another condition (Agresti 2012). In reference to Table 8, the risk ratio between *x* and *y* is calculated as follows:

$$\textit{Risk ratio} = \frac{P(x|y)}{P(x|y')} = \frac{f_{xy}/(f_{xy}+f_{x'y})}{f_{xy'}/(f_{xy'}+f_{x'y'})}$$

Risk ratios range over non-negative numbers only with a minimum of 0 and no theoretical maximum. A risk ratio of 1 occurs when the event is equally likely under both conditions, that is, the two variables are independent. A value of more than 1 occurs when *x* is likelier due to *y* than to *y'* (association). A value of less than 1 occurs when *x* is less likely

due to y than to y' (disassociation). However, one problem with risk ratio is that the calculation is undefined if there is a zero frequency in a denominator. One way of avoiding an undefined calculation is by replacing zero frequencies with another nonnegative number; a common substitute is 0.5. This adjusted calculation has been shown to be well-behaved in various studies (Agresti 2019).

A related measurement to the risk ratio is the odds ratio. This is the ratio of the odds of an event given one condition and the odds of the event given another condition. In reference to Table 8, the odds ratio between x and y is calculated as follows:

$$Odds\ ratio = \frac{P(x|y)/(1-P(x|y))}{P(x|y')/(1-P(x|y'))}$$

Its properties are the same as that of risk ratio in terms of range, interpretation of values and a problem with zero frequencies. However, it is different in that the value does not change when the orientation of the contingency table is reversed, that is when the column variables become the row variables and vice versa (Agresti 2019). Thus it is not necessary to identify which variable is the condition or the response variable. In contrast, risk ratio does depend on the orientation of the contingency table.

Risk ratio and odds ratio are related measures and they are related as follows:

$$Odds\ ratio = Risk\ ratio \times \frac{1 - P(x \mid y')}{1 - P(x \mid y)}$$

When the frequency of the event (occurrence of x) is near zero, the two measures are similar in values (Agresti 2019). Consequently, odds ratio can be used as an estimate of risk ratio when risk ratio cannot be calculated. However, when the event is of higher probability, odds ratio is higher than risk ratio for a given incidence level, thus one cannot be used to

estimate the other (Schmidt & Kohlmann 2008). Factors for choosing one over the other include study design and ease of interpreting values. For example, only odds ratio is suitable for case-control studies and some authors consider the language of risk ratio to be more intuitively understandable (Schmidt & Kohlmann 2008).

What is common of both measurements is that for any risk ratio and odds ratio value a confidence interval of a certain confidence level can be derived. More precisely, confidence intervals can be derived for values of log risk ratio and log odds ratio. This is because risk ratios and odds ratios are not normally distributed but their logarithms are approximately so. The confidence interval is calculated as follows (Agresti 2019):

$$\log(Risk \text{ or } Odds\ ratio) \pm z_{a/2} SE$$

The SE, standard error, for risk ratio and odds ratio are $(fxy'/fxy(fxy+fxy') + fx'y'/fx'y(fx'y+fx'y'))^{1/2}$ and $(1/fxy + 1/fx'y + 1/fxy' + 1/fx'y')^{1/2}$ respectively. $za/2$ is the z-statistic, which in the case of the commonly used 95% confidence level would be 1.96. To get the confidence intervals for the risk ratio or odds ratio themselves, the confidence intervals of their logarithms need to be exponentiated. The confidence interval can be used to capture the level of certainty of a particular value of risk ratio or odds ratio, thus the certainty of the type of association.

Selecting an association measure

Following the previous review of association measures, a justification of choosing one measure over others is presented. This can be approached from an empirical or a mathematical perspective. From the empirical perspective, various studies have attempted to evaluate association measures with experiments (e.g., Evert 2005, Krenn & Evert 2001, Thanapoulos et al. 2002, Gablasova et al. 2017). However, methodology and datasets vary significantly in such studies. A common approach is to use measurements to rank collocates

extracted from a test corpus, and then the effectiveness of the measurements is gauged based on how many of the top *n* collocates overlap with a gold standard dataset. Examples of gold data include a list of figurative PP-verb combinations (Krenn & Evert 2001) and lists of named entities (Thanapoulos 2002). The gold data that have been used in such studies are unlikely to approximate the full range of collocates in language. Thus, empirical work seems far from settled to guide the selection of an association measure, especially a measurement for affixes.

Another method that is somewhat related to the empirical approach is to manually inspect a sample of collocates. This is related to the empirical approach in that the choice is influenced by the data, albeit less systematically. One way this is done in some studies is by coming up with a sample of high-ranking collocates as determined by different candidate measures. Then a candidate measure is evaluated based on whether it detects 'interesting' collocates (e.g., Church & Hanks 1990, Clear 1993, Thanapoulos 2002). Some of the criteria for such evaluation include the lexicality or compositionality of the collocate, or the semantic proximity of the member words. However, given that this study focuses on inflectional suffixes, such criteria are not usable in this context. This is because all the suffix collocates in this study will consist entirely of grammatical items.

The second approach is to choose between measurements based on their mathematical properties. For the purpose of this study the useful properties of an association measurement are the following: 1) the measurement can capture the magnitude of association, and not just determine the existence of an association, and 2) the values can be meaningfully interpreted, or more specifically there is a non-arbitrary threshold separating association and negative association. Criterion (1) is due to the interest in gradience of formulaicity in this study. Criterion (2) is due to the need to observe how many affix sequences can be considered formulaic. This is different from a common approach in other collocational studies, which often are interested in ranking collocates (for example, to find the word that is most associated with a target word).

With these criteria, a measurement can be selected via a process of elimination. Criterion (1) excludes measurements based on significance tests such as t-score and Pearson's chi-squared. The reason is that these tests only capture the certainty of there being an association of any kind. Thus, the possible outcomes are a significant association, a non-significant association, a significant non-association, a non-significant non-association. Furthermore, one test alone does not indicate the type of association and the directionality of the effect in the contingency table. For narrowing that down, additional tests (post hoc tests) are necessary. A test of significance will also not capture the different magnitudes of association. However, some studies do use the test statistic themselves as scores of associations (e.g., Krenn & Evert 2001, Thanapoulos et al. 2002). However, those test statistics are not meaningful on their own, other than for deriving a *p*-value, and different test statistic values are not meaningfully comparable.

By criterion (2), pointwise mutual information also has to be excluded. Although it is a measurement with gradient values, it is more of a test of independence rather than dependence. This is because low pointwise mutual information indicates independence but higher values may not necessarily indicate dependence (Manning & Schültze 1999). This limits its usefulness for finding associated sequences.

Next, although Dice coefficient satisfies criterion (1), it does not satisfy criterion (2). Dice coefficient ranges between 0 and 1. However, there is no non-arbitrary point in between that serves as a threshold for separating sequences that are associated and negatively associated.

Finally, the remaining measure is risk ratio/odds ratio. They satisfy criterion (1) by ranging over all nonnegative numbers, and that different values are meaningfully comparable. They also satisfy criterion (2) by having the value 1 as the threshold between association and negative association. Furthermore, the threshold of 1 is non-arbitrary because it is the point where the two conditional probabilities or the two odds are equal. Another advantage of either of these measures is that they can be used in conjunction with confidence intervals. Confidence intervals can be used to capture the degree of certainty of an association. However, one limitation of both of these measures is a possible zero frequency in a denominator leading to an undefined calculation. In response to

this, the calculations in this study shall adopt the common workaround of replacing zero frequencies with 0.5. Such situation occurs when all of *y* occurs with *x*, or in reference to Table 8, when *fxy'* is zero. This is a case in which *y* is completely associated with *x*. Replacing *fxy'* with a small value like 0.5 will result in a large value for risk ratio/odds ratio, which is nevertheless useful in identifying such extreme collocates. Regardless, it turns out that this adjustment may not be that crucial overall; in the calculation of risk ratios for all verb types, this adjustment was resorted to only for 9 collocate pairs (out of 1108).

The next step is deciding between risk ratio or odds ratio. As previously mentioned, both measurements are mathematically related. Not only that, the risk ratio and odds ratio values calculated from the corpus are strongly correlated (Pearson's $r = 0.72$ or Pearson's $r = 0.89$ for risk ratios in reverse table orientations). Indeed, all collocate pairs have the same type of association under both measures; collocate pairs with a risk ratio above 1 also have an odds ratio above 1, and collocate pairs with a risk ratio below 1 also have an odds ratio below 1. Thus, as far as identifying associated sequences, there is no loss in effectiveness in choosing one measure over the other.

However, there are some characteristics that make risk ratio more useful than odds ratio for this study. First, risk ratio is not independent of the orientation of the contingency table, while odds ratio is. This may seem like a practical advantage for odds ratio. However, the asymmetry of risk ratio may lend itself to uncover additional patterns in the collocation data. Specifically, risk ratio data could capture asymmetric association (e.g., when *y* is strongly associated with *x* but *x* is not as strongly associated with *y*). Thus, what is proposed here is that for a collocate pair such as (*suffix1, suffix2*) two risk ratio values are calculated, one with s*uffix1* as the conditioning variable and another with *suffix2* as the conditioning variable. To conclude, risk ratio shall be used to measure the association between suffixes in collocate pairs.

Before closing this discussion on association measures, it must be pointed out that this is not the only method used in studies on collocations or formulaic sequences. In addition to the metric-based approaches such as in the studies cited in this section, there are also studies using algorithmic approaches. In metric-based approaches, collocations or formulaic sequences are measured and ranked based on an association measure. In contrast, in algorithmic approaches, collocations or formulaic sequences are discovered in corpus data using a set of processes in an algorithm (e.g., Brooke et al. 2017, Schneider et al. 2014, Newman et al. 2012). Although algorithmic approaches in

this area are relatively recent, studies have found they can outperform metric-based approaches (ibid.). Nevertheless, metric-based approaches have the advantage of ease of implementation. Furthermore, that some algorithmic approaches incorporate an association measure (e.g., Brooke et al. 2017) indicates that association measures are still useful in identifying collocates or formulaic sequences.

Chapter 4. Results and analysis

The procedure detailed in Section 3.2 resulted in data files which list collocate pairs along with their values for each association measurement. This section analyzes the risk ratio data within such files. The analysis proceeds as follows: first, the data are examined for indications of the existence of affix formulaicity. The second step of the analysis concerns what sequences are formulaic and possible reasons for their formulaic status. Third, the data are examined for indications of the gradience of affix formulaicity. Finally, the data are examined for indications of associations between affixes and lexical items.

4.1. Establishing affix formulaicity

The first step of the analysis is to establish the existence of formulaicity among suffixes in the dataset and to show the extent of its occurrence. Since risk ratio is calculated to capture the association between suffixes, formulaicity here is operationalized as a collocate pair having a risk ratio greater than one. If affix formulaicity is present in the language, it is expected that there is a non-trivial proportion of collocate pairs with risk ratio above 1.

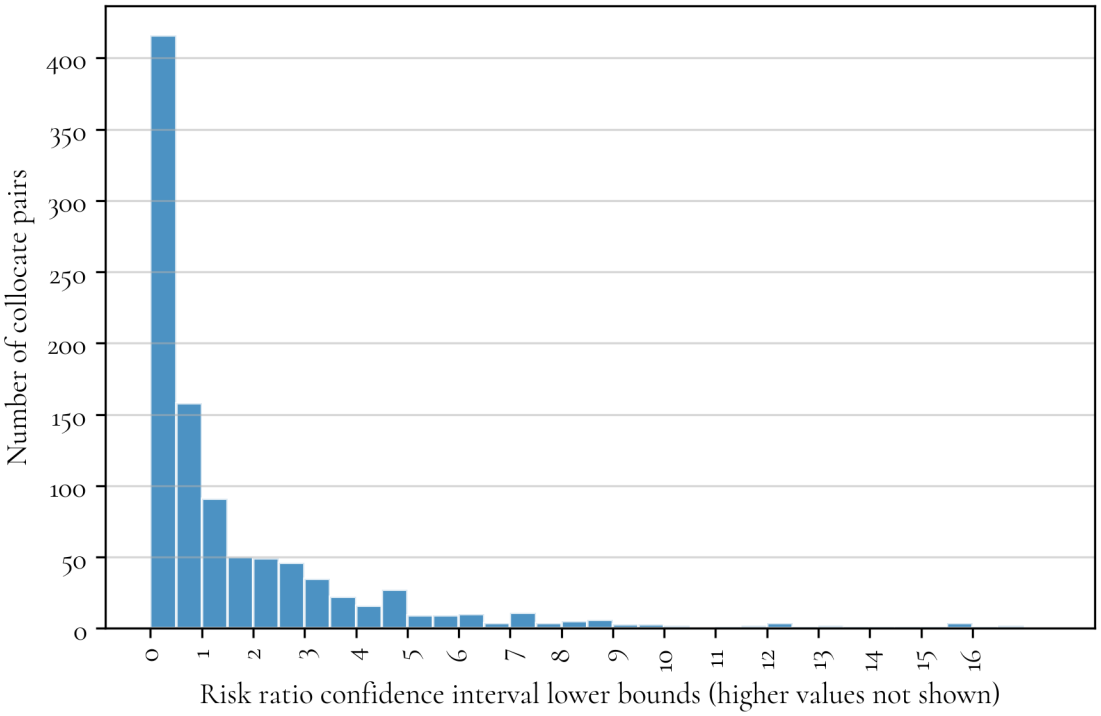
To make the examination more stringent, two requirements on the data are imposed. First, only collocate pairs where the frequency of the first and second suffixes are both at least 100 are considered. The reason for this is that if one of the suffixes is extremely rare, there is low confidence that the whole range of that suffix's distributional behavior has

been observed. Nevertheless, the exact cutoff frequency is arbitrary. As a result of this data exclusion, the number of collocate pairs considered decreases by 90 from 1108 to 1018. The second requirement is that the confidence intervals (95%) of risk ratio values are to be considered as well. Thus, for a collocate pair to be considered formulaic, its risk ratio confidence interval needs to have a lower bound greater than 1. This is done to account for the possible variability in the data. Seen in another way, the collocate pairs surpassing this requirement can be believed to be formulaic at a 95% confidence level.

Table 10 and Figure 5 show the distribution of risk ratio values in the dataset. Several presentational decisions here are to be remarked upon. First, the set of values considered are the ones in which the first suffix is the condition in the risk ratio equation rather than the reverse. The choice of one contingency table orientation or the other does not result in a loss of generalization, because when a risk ratio value is above/below 1, the reverse value is also above/below 1. Second, in Table 10 a distinction is made between risk ratio between 1 and 2, and risk ratio above 2. This is done in anticipation of a concern that a risk ratio just above 1 may not be considered a substantial association. This interval was chosen because here the probability of the event given one condition is only marginally greater than the probability of the event given the other condition.

Table 10: Distribution of risk ratio values of collocate pairs

Risk ratio range	Point values	Confidence interval lower bounds
$x \leq 1$	516 (51 %)	574 (56 %)
$1 < x < 2$	156 (15%)	142 (14 %)
$x \geq 2$ (maximum of 41919.41)	346 (34%)	302 (30 %)



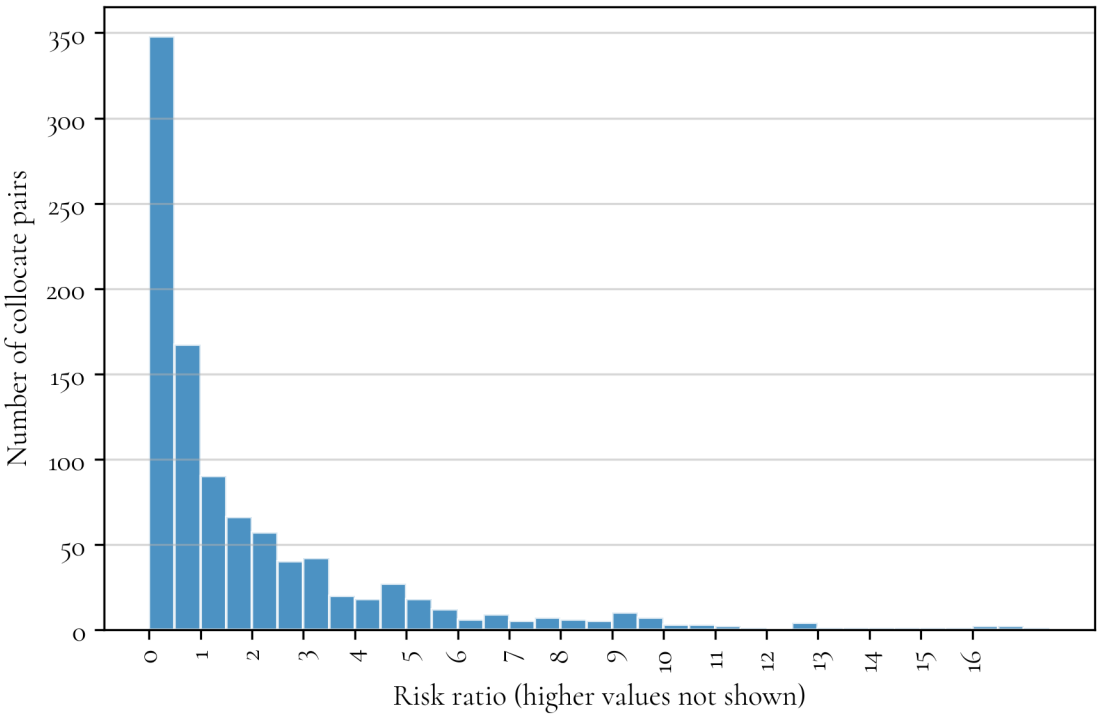


Figure 5: Distribution of risk ratio values

Next, inferences are made based on the risk ratio data as presented in Table 10 and Figure 5. The data show that around the majority of collocates are not associated, that is, have a risk ratio below 1. Nevertheless, around half of collocate pairs have risk ratio greater than one. Even after excluding risk ratio just above 1, the proportion is still nontrivial (more than a third). Furthermore, the result holds even after considering the lower bounds of the confidence intervals. On the question of whether affix formulaicity exists in the corpus, based on the number of associated collocate pairs it appears that it does. Thus, it can be inferred that formulaicity is a major part of language use in the corpus. That the majority of collocate pairs are not associated is expected; this represents more novel combinations in language use, which should consist of more item types that have fewer tokens each. That associated collocate pairs constitute a minority is also expected. For formulas to serve processing economy, it is expected that they make up a restricted set of items that are recurrently relied upon.

Although a substantial number of collocate pairs are associated, the conclusion that this implies the existence of affix formulaicity in the corpus can be undermined if these collocate pairs were in fact a part of longer formulas at the word level. If that were true, it would not be clear if formulaicity is taking place at the affix level. This possibility cannot be addressed directly because the detection of word-level formulas in Turkish is outside the scope of this study. However, the stem-wise frequencies of collocate pairs may indicate that this is not so. If a suffix sequence is formulaic by virtue of being in a word-level formula, it is expected that such sequences tend to occur with the verbs that are stored with that word-level formula. Among all of the 575 collocate pairs with risk ratio above 1, 293 (51%) occur with at least 100 verb types, and 413 (72%) occur with at least 10 verb types. Given the wide number of verb types that most collocate pairs occur with, it is unlikely that most collocate pairs are predominantly appearing in word-level formulas.

The following question is whether such data distribution does indeed imply the occurrence of affix formulaicity. Although the opposite is unlikely given this data distribution, the reason for this caution is that risk ratio is only assumed as an

operationalization of formulaicity. Thus the following additional analyses are done to supplement the risk ratio distribution data. The aspects of the data to be examined are: 1) frequency of usage of formulas, 2) distinction between registers, 3) distinction between adjacent and nonadjacent suffixes.

The first way that the risk ratio data could implicate affix formulaicity is through the frequency of usage of formulas. One expected property of formulas is the high number of contexts in which they occur (Durrant 2013). Presumably, this is related to the notion of formulas aiding lexical retrieval (Wray 2002), and occurring in the widest use cases maximizes the processual utility of formulas. Thus, high risk ratio collocate pairs are expected to have more verb stems hosting them. However, there does not appear to be a correlation between risk ratio and number of hosting verb stems (Pearson's r = 0.04). Another way of approaching this question is to focus on the most stem-wise frequent collocate pairs and observe how many of them are formulaic. Among the 24 collocate pairs that occur with more than 700 (out of 732) verb stems, 22 have risk ratio above 1 (Table 11). Thus, these highly frequent collocate pairs are more likely to be formulaic.

Table 11: Collocate pairs with more than 700 verb stem hosts

Collocate pair	Risk ratio	Collocate pair	Risk ratio
('Narr', 'A3sg')	2.18	('Prog1', 'Past')	3.83
('Past', 'A3sg')	2.63	('Inf2→Noun', 'Acc')	1.67
('Prog1', 'A3sg')	1.96	('Narr', 'Past')	3.52
('Inf2→Noun', 'Dat')	7.96	('PastPart→Noun', 'P2sg')	9.22
('Inf2→Noun', 'P3sg')	4.76	('Pass→Verb', 'A3sg')	0.61
('Aor', 'A3sg')	2.13	('P2sg', 'Acc')	10.04
('PastPart→Noun', 'P3sg')	3.03	('PastPart→Adj', 'P3sg')	7.82
('Inf1→Noun', 'A3sg')	2.99	('Inf2→Noun', 'Gen')	4.95
('PastPart→Noun', 'Acc')	7.01	('P3sg', 'Dat')	8.88
('Neg', 'A3sg')	0.92	('Narr', 'Cop')	5.28
('P3sg', 'Acc')	8.01	('Inf2→Noun', 'A3pl')	2.46
('Fut', 'A3sg')	1.96	('Inf2→Noun', 'Abl')	2.63

The second way the risk ratio data could implicate affix formulaicity is in the difference between spoken and written registers. It is expected that formulaic collocate pairs would be more prevalent in the spoken register than in the written register. This is based on the assumption that speech, being a performative act, would entail more processing pressures where formulas are needed for faster lexical retrieval (Kuiper 1996). In contrast, writing is a more planned act, which would make novel suffix combinations more likely.

Table 12 displays the number and proportion of collocate pairs that are formulaic within each register. Both type frequency and token frequency of collocate pair are considered. By both measures, formulaic collocate pair types are more common in the spoken register than in the written register (56% vs. 50% by type frequency, 82% vs. 77% by token frequency). Conversely, nonformulaic pairs are more common in the written register than in the spoken register (50% vs. 44% by type frequency, 23% vs. 18% by token frequency). This is consistent with the written register being more planned and allowing for more novel forms, while the spoken register has a greater reliance on recurring forms. Also noteworthy is that the difference between formulaic and non-formulaic collocate pairs is more pronounced when considering pair token frequency. More specifically, although formulaic pairs are the majority in both registers according to token frequency, formulaic pairs have a bigger dominance in the spoken register (difference of 64% (82%-18%) in spoken register vs. difference of 54% (77%-23%) in written register). This contrasts the relative parity between formulaic and non-formulaic pairs in both registers by type frequency. Assuming that pair token frequency represents language in use, it is expected that formulas are even more common in speech by this measure. To conclude, the difference between the written and spoken registers is as expected if there were affix formulaicity in the data.

Table 12: Type and token frequency of collocate pairs by register

Risk ratio range	Written register		Spoken register		Whole dataset	
	Pair types	Pair tokens	Pair types	Pair tokens	Pair types	Pair tokens
Up to 1	506 (50%)	667703 (23%)	278 (44%)	37369 (18%)	516 (51%)	705546 (23%)
Above 1	506 (50%)	2234703 (77%)	357 (56%)	172001 (82%)	502 (49%)	2406649 (77%)

The third way the risk ratio data could implicate affix formulaicity is the distinction between adjacent and nonadjacent suffixes. The motivation for this line of inquiry is that although the definition of formulaicity includes non-continuous sequences (Wray 2002), formulas are prototypically continuous. Furthermore, Durrant (2013) found that for a given a suffix, it is the suffixes adjacent to it that are the most predictable. Thus it is expected that association is stronger among adjacent suffixes than non-adjacent ones.

Collocate pairs are divided into three sets where the suffixes in the pair are: 1) always adjacent 2) sometimes adjacent, or 3) never adjacent. For this examination only associated collocate pairs are considered (502 out of 1108). The breakdown of the three sets are shown in Table 13. A one-way analysis of variance (ANOVA) was conducted to compare the risk ratio of collocate pairs in these sets. To fulfill the assumptions of the test, the log of risk ratios were calculated so that the distribution approximates normality and the variance of the sets are more similar. There is a significant effect of adjacency on log risk ratio ($F(2, 502) = 4.07$, $p = 0.018$). Not only do the risk ratios of the three sets differ, it can be observed from Table 13 that median risk ratio is highest in the always adjacent set and lowest in the never adjacent set. Additionally, that median risk ratios of the sometimes adjacent and never adjacent sets are nevertheless above 1 is expected. The two sets may be analogous to formulas with variable slots, with the elements interfering between the two suffixes in the collocate pairs as being the variable elements. Another noteworthy observation is that the always adjacent and sometimes adjacent pairs are far more numerous than the never adjacent set. Given that the collocate pairs considered here are the 502 that are associated, it appears that suffixes need to occur close to each other to attain formulaic status. This is expected because

previous research found that formulas are not unbounded (i.e., can be arbitrarily spaced) and instead they occur within some domain of locality (e.g., Biber 2009, Stefanowitsch & Gries 2009).

Table 13: Risk ratio of associated adjacent and non-adjacent pairs

Pair set	Number of pair types	Mean risk ratio	Median risk ratio
Always adjacent	166	26.68	5.25
Sometimes adjacent	273	4.05	2.63
Never adjacent	63	13.23	1.87

In conclusion, these three additional analyses may further support the conclusion that the risk ratio data do indeed imply the existence of affix formulaicity in the language. The next question is whether there are more patterns among suffixes in addition to formulaicity. Other possible patterns that may exist are repulsion and asymmetric association.

The first possible additional pattern is repulsion among suffixes. In addition to lack of association (for risk ratios around 1), there may be negative association or repulsion with pairs on the low end of the distribution. Figure 5 shows that the most frequent risk ratio values are below 1. The range between 0 and 1 may be too short to reveal further distinction among pairs in this range. Taking the logarithm of risk ratios can make an existing distinction more apparent. This is because at lower values logarithmic functions have higher rates of change. The distribution of log risk ratio (base 2) is shown in Figure 6. Because the critical threshold for association in risk ratio is 1, the critical threshold in log risk ratio is 0. It can be observed that below 0, there is a wide range of risk ratio values that extend far below 0. The values nearer to 0 may indicate a lack of association. The values towards the

left of the distribution may indicate repulsion between suffixes. What collocate pairs are in these ranges will be detailed in the next section.

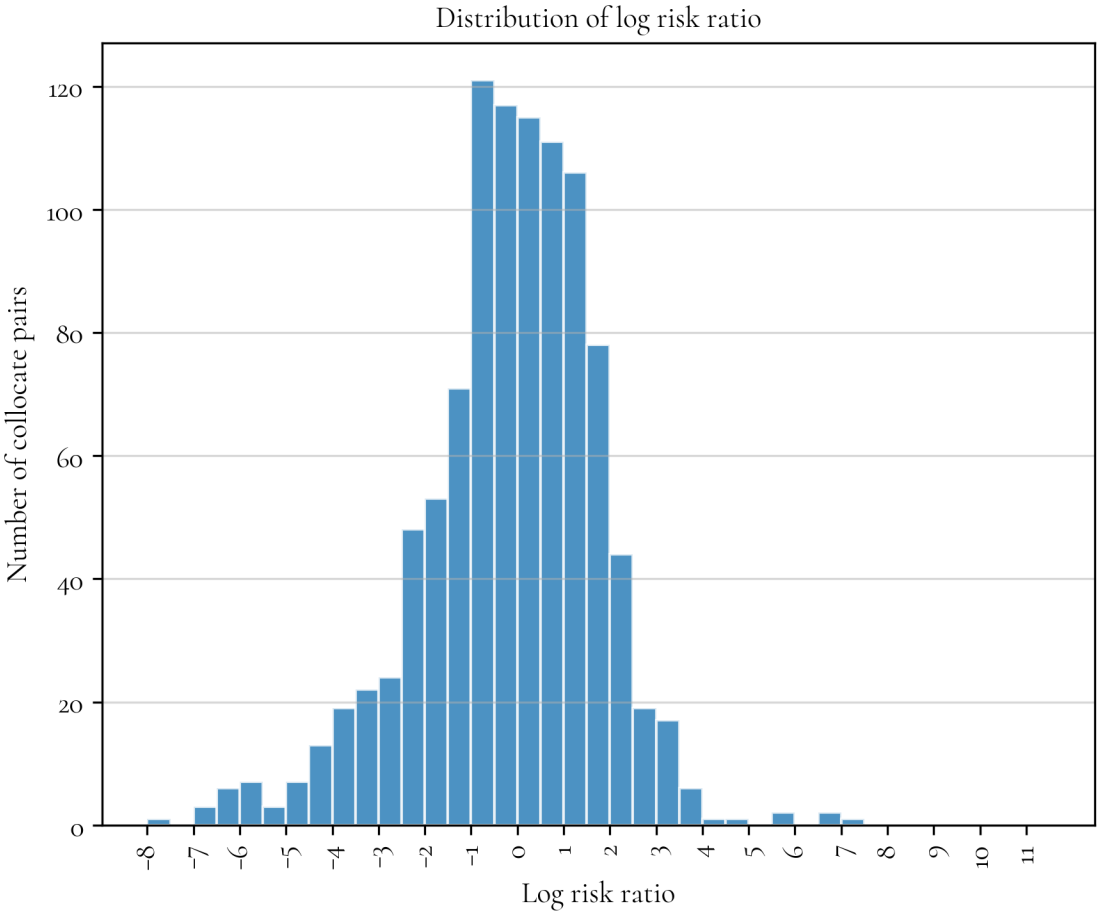


Figure 6: Distribution of log risk ratio

The second possible additional pattern is asymmetric association among suffixes. This would be a case in which one suffix is associated with another suffix more than the other way around. Technically, this would be a case in which the risk ratio value is significantly higher with one contingency table orientation than with the other orientation. To examine asymmetric association, the following measurement of symmetry was calculated:

$$Symmtery(x, y) = max\{\frac{Risk\ ratio(x, y)}{Risk\ ratio(y, x)}, \frac{Risk\ ratio(y, x)}{Risk\ ratio(x, y)}\}$$

This is a ratio of a collocate pair's risk ratio to its risk ratio in reverse. Which value is in the numerator or denominator is determined based on which configuration results in a higher value. Thus, the measurement quantifies how much one value is larger than the other. If the association between the two suffixes are more or less symmetric, symmetry is expected to be close to 1. If the association between the two suffixes are more asymmetric, symmetry is expected to be greater than 1.

The distribution of symmetry values is displayed in Table 14 for collocate pairs with risk ratio above 1. As can be seen in Table 14, for most collocate pairs the values of this ratio are close to 1. However, there is a minority of collocate pairs in which one risk ratio value is larger than the other one. Not only that, there is a large range of magnitude by which one risk ratio dwarves the other. Based on this study's assumption that if a sequence is formulaic, then the elements within should be associated with each other, collocate pairs that are more or less symmetric (bottom row of Table 14) may represent affix formulas, as they suggest a behavior as units. In contrast, more asymmetric collocate pairs (first row of Table 14) may represent a subset of affix formulas in which one suffix has a more restricted distribution.

Table 14: Symmetry of collocate pairs with risk ratios above 1

Symmetry	Number of collocate pairs	Range
≥ 2	22	2.08 – 2667.98
< 2	480	1.00 – 1.97

To conclude this section, the distribution of risk ratios indicates that there is affix formulaicity in the data. Further analyses based on frequency of usage of affix formulas, distinction between registers and distinction between adjacent and nonadjacent suffixes

make this conclusion likelier. In addition to association, the data also exhibit repulsion and asymmetric association between suffixes.

4.2. Sequences that are formulaic

After concluding that there is affix formulaicity in the dataset in the previous section, this section examines what collocates are formulaic and if there are longer sequences that may be formulaic. Manual observations may reveal linguistic reasons for some collocate pairs being formulaic or non-formulaic. For example, Table 15 shows that the aorist suffix is highly associated with suffixes such as the 'while' adverbial suffix, the abilitative and the conditional. This is possibly because such adverbial suffixes imply propositions with an indefinite time scale. The association of the pair (Narrative/Evidential, Past) may be consistent with the narrative/evidential often being used to report an event. The optative is highly associated with the first person singular and plural because the optative is often used in requesting permission or making a suggestion (similar to English *Let's*). These functions may also explain why the optative is negatively associated with the past suffix. (Abilitative, 'By doing so') and (Abilitative, 'Passive') may be negatively associated because the 'By doing so' suffix denotes a manner adverb, showing that a verb is carried out using another verb (e.g., *They went by walking*). Given this function, presumably speakers prefer to use the 'By doing so' suffix with action verbs rather than stative verbs, thus their repulsion with the abilitative. In conclusion, the association or repulsion between suffixes may be explained by semantic, grammatical or functional factors. The magnitude of repulsion of these grammatically conflicting collocate pairs is apparent when considering the log of their risk ratios, also displayed in Table 15. These collocate pairs occupy the left end of the log risk ratio distribution.

Table 15: Some highly associated, non-associated, and negatively associated collocate pairs

Collocate Pair	Risk ratio	Log risk ratio
(Aorist, 'While')	37.15	5.22
(Abilitative, Aorist)	9.09	3.18
(Aorist, Conditional)	12.93	3.69
(Narrative/Evidential, Past)	1.91	1.81
(Optative, 1.SG)	2.86	1.52
(Optative, 1.PL)	15.83	3.98
(Abilitative, 'By doing so')	0.005	-7.59
(Abilitative, 'Passive')	0.002	-9.37
(Optative, Past)	0.04	-4.56

The examination of the data so far has only considered collocates of two suffixes. It is possible that there are multiple collocate pairs that constitute units of more than two suffixes. In other words, there may be longer suffix formulas. That the suffix sequences considered so far are pairs follows from the need to calculate measures of association. To uncover these longer sequences, suffix trigrams were derived from the data. The adjacent suffixes in the trigram must occur adjacently to each other, and any two adjacent suffixes must have a risk ratio greater than 1. This procedure is based on the assumption that a chain of associated suffixes may function as a single formula. 2555 formulaic trigram types were derived by this criterion. Table 16 lists some of the most frequent of these. With three suffixes in a sequence it would be more difficult to manually discern any linguistic motivation for the three suffixes to cohere together. However, given that these trigrams are highly frequent, a plausible conjecture is that these formulaic trigrams facilitate language processing as much as possible by being available for use in the widest range of contexts.

Table 16: Ten most frequent formulaic trigrams

Trigrams	Suffix 1-2 risk ratio	Suffix 2-3 risk ratio	Frequency
(Pass→Verb, Inf2→Noun, P3sg)	3.70	4.76	40294
(Pass→Verb, Narr, A3sg)	4.69	2.18	23034
(Prog1, Past, A3sg)	3.83	2.63	22522
(Narr, Past, A3sg)	3.52	2.63	19155
(Pass→Verb, Aor, A3sg)	2.40	2.13	13567
(Inf2→Noun, P3sg, Dat)	4.76	8.88	10710
(Caus→Verb, Pass→Verb, Inf2→Noun)	20.33	3.70	10547
(Able→Verb, Aor, A3sg)	9.09	2.13	10485
(Inf2→Noun, P3sg, Acc)	4.76	8.01	10353
(Pass→Verb, Narr, Cop)	4.69	5.28	9815

At this point, a comparison can be made between the formulaic trigrams derived here and the 'morphemic bundles' found in Durrant (2013). Those bundles appear to be formulaic here as well according to the measurement and procedures of this study. Table 17 shows that the morphemic bundles/trigrams identified in Durrant (2013) have risk ratios above 1 for both adjacent suffix pairs within. One subset of the data that is of interest concerns subordinate markers. Many of the morphemic bundles identified in Durrant (2013) have subordinate markers. Based on this, Durrant posited that suffix formulas assist the processing of utterances containing multiple clauses. That finding is reflected here as well; 591 of the 2555 formulaic trigrams here contain one of the markers of subordinate clauses (i.e.: Inf2→Noun, PastPart→Noun, FutPart→Noun). In fact, the most frequent formulaic trigram contains the subordinate subjunctive Inf2/-mA morpheme (Table 16).

Table 17: Risk ratio of morphemic bundles identified in Durrant (2013)

Trigram	Risk ratio between pairs
(PastPart→Noun, P3sg, Acc)	3.03, 8.01
(Pass→Verb, Inf2→Noun, P3sg)	3.7, 4.76
(Neg, Past, A3sg)	1.64, 2.63
(FutPart→Noun, P3sg, Acc)	2.72, 8.01
(Inf2→Noun, P3sg, Acc)	4.76, 8.01
(PastPart→Noun, P3sg, Dat)	3.03, 8.87

4.3. Gradience in affix formulaicity

In Sections 4.1 and 4.2 it is apparent that there are collocate pairs that are formulaic and collocate pairs that are non-formulaic. Given this, the following question is whether affix formulaicity is a categorical system as grammar has been traditionally conceived as (Hay & Baayen 2005), or is there finer distinction among affix formulas in the data. Based on this question, this section explores whether the affix formulaicity in the dataset is a discrete or a gradient phenomenon. The exploration here shall adopt the two notions of gradience discussed in Section 1.2: gradience in magnitude of formulaicity and structural gradience of formulas.

The first question is whether the affix formulaicity in the data is gradient in magnitude. One approach to this question is to examine the risk ratio values themselves. This assumes that risk ratio values are correlated with magnitudes of formulaicity. Theoretically, whether affix formulaicity is gradient is ostensibly true. This is because risk ratio is a continuous variable ranging over positive numbers. However, this property alone may not be sufficient to conclude that formulaicity is gradient because the variation could be due to noise in the data. Instead, one approach is to observe if risk ratio values tend to concentrate around a limited number of values. Those points of concentration may

represent discrete magnitudes of formulaicity. Thus, if formulaicity is discrete, it can be expected that the distribution of risk ratio follows a bimodal or multimodal distribution. The histograms of risk ratios in Figure 5 show that this may not be the case visually; the mode of the distribution is below 1 and the frequency of risk ratio values tapers off rightwards. A similar impression is given by the distribution of log risk ratios as shown in Figure 6, which appears to approximate a normal distribution. Visually it is apparent that the data (risk ratios and log risk ratios) do not approximate a multimodal distribution either. Thus the data do not suggest discrete magnitudes of formulaicity.

At this point it is worth questioning whether gradience in risk ratios actually corresponds to gradience in psycholinguistic processing as assumed in the previous analysis. This is because risk ratio is an operationalization of formulaicity rather than a direct measurement of it. Indirectly, this assumption may be justified to some extent. This is based on findings from previous research that more frequent words are recognized faster and more easily (Tremblay et al. 2011). Relating frequency with risk ratio may be appropriate to the extent that collocates that are more frequent can have higher risk ratios, all else being equal. However, relating gradience of risk ratios to gradience in psycholinguistic processing must be treated with caution. This is because it is not clear how closely the two domains correspond. Furthermore, it is not clear what variable of psycholinguistic processing would the risk ratios correspond to. Although this limits the interpretive power of the previous analysis, this does point to the need for further psycholinguistic studies to validate any gradient pattern found in corpus data.

The second question is whether the affix formulas found in the data are gradient structures. Structural gradience is in contrast to structures being composed of discrete constituents (Hay & Baayen 2005). For the purpose of this analysis, a possible interpretation of structural gradience is related to the integrity of formulas, which can be defined as how often the constituents of the formula cooccur contiguously. For each collocate pair type, this is a subset of all of its tokens. This is because the procedure for deriving collocate pairs (Section 3.2) means that suffixes do not have to be adjacent; they just have to cooccur in the

same word. Integrity then is calculated as follows (# denotes the cardinality of a set, i.e., its number of members):

$$Integrity(x, y) = \frac{\#\{\, Collocate\ pairs\ where\ x \text{ and } y\ are\ adjacent \,\}}{f_{xy}}$$

In the context of pairs, integrity here is measured as the ratio of the following: 1) the frequency of the two affixes occurring adjacently, and 2) the frequency of cooccurrences of the two affixes. If the suffix pair is fully integral, that is the suffixes always appear adjacently, then the ratio would be equal to 1. The property captured by this measurement is the extent to which a formula behaves as a single contiguous element, instead of strictly composed of discrete constituents. If a formula is less integral and is more composed of discrete constituents, then it can be expected that its members can be separated as expected by grammar alone.

Table 18 shows the distribution of this integrity ratio for all collocate pairs, formulaic collocate pairs and nonformulaic collocate pairs. Several observations are apparent when considering just formulaic collocate pairs. A substantial proportion of formulaic collocate pairs (33%) always occur adjacently. Some of these cases could be due to morphological factors such that two suffixes must attach successively. More clearly, a plurality of formulaic collocate pairs (43%) occur adjacently at least half of the time. Collocate pairs in these two categories may represent affix formulas that are perceived more as a unit and less as a sequence of discrete morphemes. Although gradient structures can be seen in formulaic collocate pairs, it does not appear that integrity values themselves are as gradient. The distribution of integrity values as shown in Figure 7 shows that integrity values tend to concentrate at or near 1. Below that there is a range of integrity values observed, encompassing far fewer formulaic collocate pairs.

A remark can also be made about the contrast between formulaic and nonformulaic collocate pairs when it comes to integrity values. The distribution of integrity values of

formulaic collocate pairs and nonformulaic pairs are asymmetric. The integrity values of formulaic pairs are concentrated in higher ranges, while the integrity values of nonformulaic pairs are concentrated in lower ranges. In fact, the majority of nonformulaic pairs (53%) never occur adjacently. Such asymmetry suggests a hypothesis that formulaic status leads to a constraint on how far apart the constituents of a formula can be, or that recurrently adjacent suffixes enable their development into formulaic status.

Table 18: Distribution of collocate pair integrity values

Integrity (0-1)	Number of collocate pairs		
	All	Formulaic	Non-formulaic
$x = 1$	256 (25%)	166 (33%)	90 (17%)
$0.5 \leq x < 1$	299 (29%)	214 (43%)	85 (17%)
$0 < x < 0.5$	128 (13%)	59 (12%)	69 (13%)
$x = 0$	335 (33%)	63 (12%)	272 (53%)

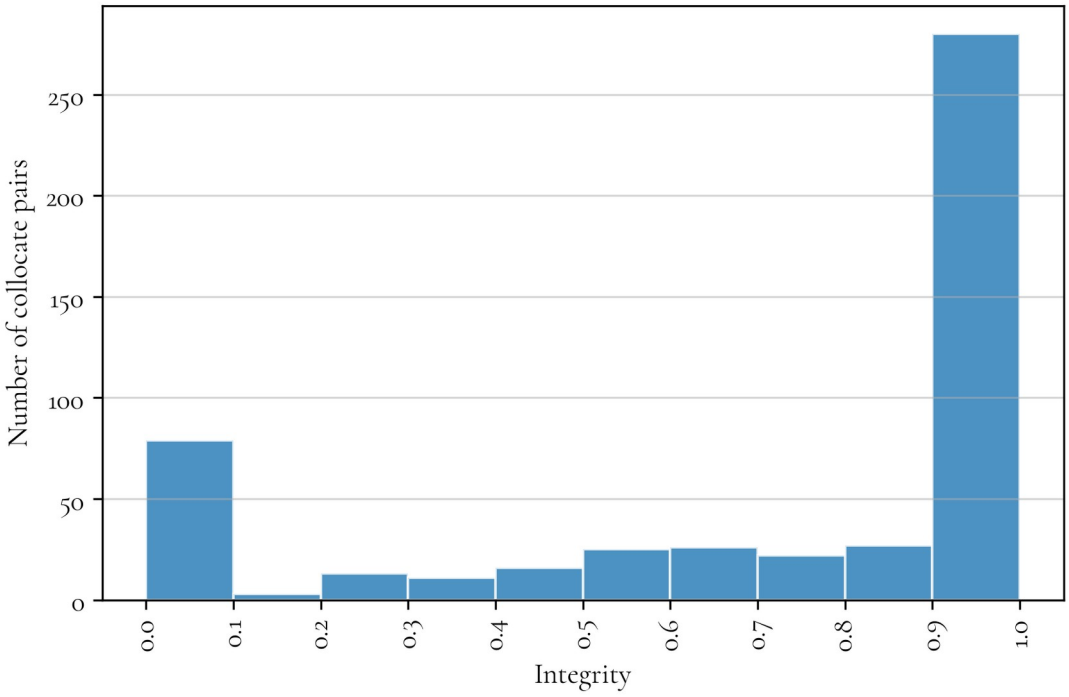


Figure 7: Frequency of collocate pair integrity of formulaic collocate pairs

Another possible interpretation of gradience in the structure of affix formulas concerns their internal cohesion. Given an affix formula of more than two suffixes, it is possible that one link (between two suffixes) may be more associated than another link within the formula. This is related to the notion that structural gradience contrasts a formula being composed of discrete constituents. If an affix formula is strictly composed of discrete constituents, then it is expected that the association between any two adjacent affix is equal, because the constituents are equal in status in the formula. If there is an inequality in association within an affix formula, this may suggest more internal structure that is unexpected from pure concatenation alone. To investigate this, the following measurement is calculated on each trigram in Section 4.1 (note that the function has only one input, which is the trigram (*x*, *y*, *z*), rather than three inputs, *x*, *y*, *z*):

$$Trigram\ link\ ratio((x, y, z)) = min\{\frac{Risk\ ratio(x, y)}{Risk\ ratio(y, z)}, \frac{Risk\ ratio(y, z)}{Risk\ ratio(x, y)}\}$$

Table 19 shows the frequency of trigram link ratio values on trigrams in which all adjacent suffix pairs within are formulaic. It can be seen that there is a minority of trigrams where the first and second suffix links are approximately equally associated. There are also numerous trigrams where one link is substantially greater than the other. Again, it can be observed that these ratios come in a range of values. These different ranges suggest that there are different types of affix formulas among these trigrams. First, the trigrams in the first row of Table 19 may be suffix sequences that constitute stored wholes; that the links in these trigrams are almost equally associated may mean that these sequences are internally cohesive structures. Second, that there are trigrams where one link is substantially more associated than the other may imply that there is a distinction between core and adjunct structures. The suffixes that are linked by the significantly stronger link would form the core structure, and the suffix that is on the weaker link would be the adjunct. This means that such a trigram may not be a single stored whole, rather it has a core structure that may be a stored whole, and the adjunct is incidentally associated either to the core or to one of the suffixes in the core.

Table 19: Distribution of trigram link ratio values

Trigram link ratio (0-1)	Frequency
$0.9 \leq x \leq 1$	114 (7%)
$0.5 \leq x < 0.9$	562 (37%)
$0.1 \leq x < 0.5$	701 (46%)
$x < 0.1$	148 (10%)

To conclude, this section adopted two notions of gradience in formulaicity: gradience in magnitude and structural gradience. For structural gradience, the analysis in this section adopted two interpretations: integrity and internal cohesion of affix formulas. These approaches to the data suggest that affix formulaicity in the corpus is a gradient

phenomenon. That affix formulaicity is likely to be gradient may figure into the comparison of psycholinguistic models of morphological processing in Section 5.2.

4.4. Affix formulas and the lexicon

The previous sections analyzed formulaicity among suffixes in the dataset. Expanding the scope of analysis, the final part of the analysis concerns the distribution of suffix formulas with respect to lexical items, or specifically, verb stems. The basis of this examination is the question of whether there is a separation between grammar and the lexicon. In this respect, Durrant (2013) found that there are attraction or repulsion relationships between verb stems and tri-morphemic bundles. This restriction in cooccurrence was interpreted as problematic for models of language that posit a separation between grammar and the lexicon.

To organize suffix sequences in the data, suffixes cooccurring in the same word were grouped into trigrams. Trigrams were selected for the following reasons. The first reason is for comparability with Durrant's (2013) examination on stem-affix formulaicity. Second, if suffixes were organized into unigrams instead, this may miss patterns involving longer sequences of suffixes. On the other hand, not using any predetermined groupings at all (that is considering all the suffixes in a word as a single sequence), would result in too many distinct suffix sequences to process in time. Thus, trigrams were deemed to be an optimal compromise. However, because trigrams are arbitrary, convenient analytic devices, it must be cautioned that any patterns between stems and suffixes found here may be incomplete.

Following the procedure with suffix collocate pairs, risk ratios were calculated for pairs of stems and suffix trigrams that cooccur in the dataset. For example, given a word parsed as *Stem* + *a* + *b* + *c* + *d*, the pairs that are derived are (*Stem*, (*a*, *b*, *c*)) and (*Stem*, (*b*, *c*, *d*)). Similarly with suffix collocate pairs, the stems and the trigrams do not have to be adjacent, they just have to cooccur in a single word. An imposed restriction is that among all the trigrams that were derived, only 10 of the ones identified in Durrant (2013) are considered

here. The reason for this restriction is that including all of the possible trigrams would have resulted in an inordinate amount of processing time. This procedure derived 3421 stem-trigram pairs. Then risk ratio was calculated for each stem-trigram pair. The values presented here are calculated with the stem as the conditioning variable, but another set of risk ratios with the trigram as the conditioning variable was also calculated.

The first step in this part of the analysis is to observe if there is a significant portion of stem-trigram pairs with risk ratio above 1. The distribution of risk ratios of stem-trigram pairs is depicted in Figure 8. As can be seen, 2476 (49%) stem-trigram pairs have risk ratios above 1. This suggests that formulaicity may hold between stems and suffixes as well. Similar to suffix collocate pairs, the majority of stem-trigram pairs have risk ratios below 1. These may represent more novel inflection on verb stems. Not only that, there are fewer cases of asymmetric association between stems and trigrams compared to suffix collocate pairs. Expressed in another way, risk ratio values in both contingency table orientations tend to be close for stem-trigram pairs. In fact, the two sets of risk ratio values are highly correlated (Pearson's $r = 0.97$).

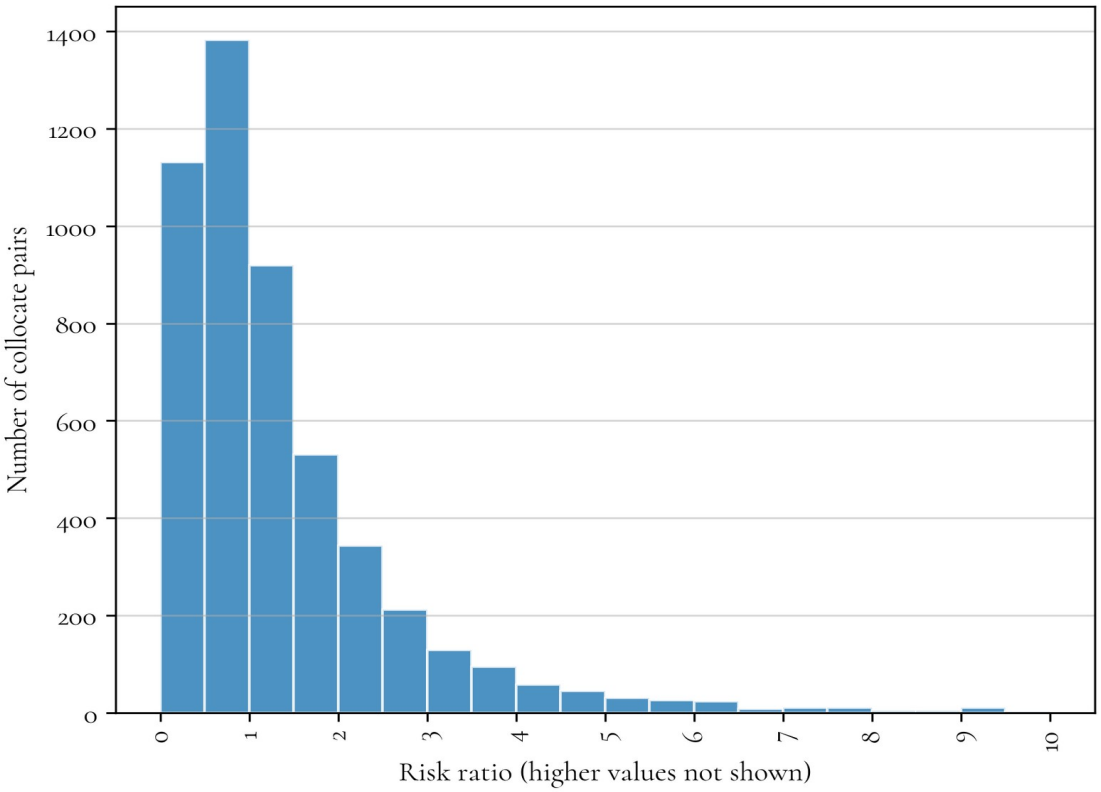


Figure 8: Distribution of risk ratio of stem-trigram pairs

To probe further the patterns of association of stem-trigram pairs, one approach is to examine the number of stems that the trigrams are associated with. Expressed in another way, the data of interest is for a given trigram, how many stem-trigram pairs are there such that the pair contains that trigram and that pair has a risk ratio above 1. As can be seen in Table 20, each suffix trigram is associated with some verb stems but also not associated with a significant portion of verb stems. This may indicate that suffix sequences are not neutral with regard to the lexical items that they attach to. In addition to association or non-association, there may be negative association or repulsion as well between stems and trigrams. Log risk ratios were calculated to accentuate the difference between risk ratio values. The ranges of log risk ratio of stem-trigram pairs are displayed in Table 20 as well. As

can be seen, low risk ratio values of stem-trigram pairs range into low negative numbers, which suggests the existence of disassociation. In addition to that, there is variation among these trigrams; some trigrams have stronger disassociation with stems than other trigrams do.

Table 20: Stem-wise frequencies of trigrams

Trigram	Risk ratio of stem-trigram pairs		Log risk ratio range
	Up to 1	Above 1	
('PastPart→Noun', 'P3pl', 'Acc')	100 (29%)	250 (71%)	-2.46 - 5.99
('PastPart→Noun', 'P3sg', 'Dat')	137 (35%)	256 (65%)	-2.56 - 4.77
('Inf2→Noun', 'P3sg', 'Dat')	411 (59%)	282 (41%)	-4.74 - 4.23
('Neg', 'PastPart→Noun', 'P3sg')	247 (52%)	225 (48%)	-3.78 - 4.11
('PastPart→Noun', 'P3sg', 'Acc')	349 (51%)	341 (49%)	-5.1 - 3.82
('Pass→Verb', 'PastPart→Noun', 'P3sg')	281 (52%)	261 (48%)	-4.34 - 3.79
('FutPart→Noun', 'P3sg', 'Acc')	252 (47%)	284 (53%)	-4.05 - 3.7
('Inf2→Noun', 'P3sg', 'Acc')	347 (51%)	333 (49%)	-4.69 - 2.96
('Pass→Verb', 'Inf2→Noun', 'P3sg')	402 (62%)	244 (38%)	-6.97 - 2.64

Only nine of the trimorphemic bundles identified in Durrant (2013) are shown here because the trigram ('Able→Verb', 'Neg', 'Aor') was not found in the dataset, likely because the morphological parser uses the morpheme 'Unable' to cover both 'Able→Verb' and 'Neg'.

Given these patterns of association, grammatical reasons can be conjectured for them via manual inspection. First, among the trigrams containing a subordinate phrase marker (i.e.: 'FutPart', 'Inf2', 'PastPart'), trigrams containing the subjunctive subordinate ('Inf2') have lower association rates than trigrams containing the other subordinate phrase markers. This could be due to a preference for factuality in embedded clauses in language use. Second, among the trigrams containing subordinate phrase markers, the trigrams that also contain the accusative ('Acc') have higher association rates than trigrams that also contain other case markers (e.g., 'Dat', 'Abl'). This could be due to the fact that the

accusative requires the subordinate phrases to be a complement of transitive verbs, of which there are many in Turkish. In contrast, for the dative and the ablative to be used on a subordinate phrase, the phrase must be the complement of verbs that require such cases. Often this class of verbs is idiosyncratic in the sense that an ablative or dative verb in Turkish may not be an ablative or dative verb in other languages. Third, there is particularly low association with trigrams containing the passive suffix, particularly ('Pass→Verb', 'Inf2→Noun', 'P3pl'). This could be due to the fact that the passive suffix can only be used on transitive verbs, limiting the number of verb stems such trigrams can be associated with. The passive suffix's relatively restricted usage then constrains the magnitude of formulaicity of suffix trigrams containing it. In conclusion, the number of verb stems that suffix trigrams are associated with may be a product of the distributional properties of each member suffix in the trigram.

To conclude this section, the preceding analysis suggests that there is association between suffix trigrams and verb stems such that suffix trigrams associate with certain verb stems but not others. This, in turn, may suggest that formulaicity holds between suffixes and stems, and not just among suffixes. This may also suggest that grammatical forms are not neutral with respect to the lexical items they attach to. The data and analysis here may support the notion that there is not a definite separation between grammar and lexicon (Stefanowitsch & Gries 2003). The relationship between suffixes and lexical items shown here may pose a challenge for certain psycholinguistic models of morphological processing discussed in Section 5.2.

Chapter 5. Discussion

This discussion section consists of the following in order: 1) a recap of the results, 2) discussion of the theoretical implications of the results, and 3) discussion of a possible application of the results.

5.1. Recap

The following is a recap of the main results of this study based on the data consisting of risk ratio values on collocate pairs of suffixes from verbs in the Turkish National Corpus. First, the corpus data suggest that formulaicity among affixes does exist and it is a major phenomenon in Turkish. This is implied by the non-trivial proportion of collocate pairs being associated as measured by risk ratio. This conclusion was further supported by additional analyses based on the frequency of usage of affix formulas, distinction between registers, and distinction between adjacent and nonadjacent suffixes. In addition to association, the data also exhibit repulsion between suffixes and asymmetric association. Second, the affix formulaicity observed in the data is a gradient rather than a discrete phenomenon. This gradience was found both in terms of the magnitudes of formulaicity being continuous and that affix formulas are gradient structures. Finally, by measuring association between stems and suffix trigrams, formulaicity was also observed to occur between stems and affixes, and not just among suffixes. Additionally, it appeared that suffix

sequences are not neutral with regard to the lexical items they attach to. Generally, the results here are consistent with the findings of Durrant (2013). Moving on, these aspects of the results can serve as the basis of comparison of models of psycholinguistic morphological processing discussed in the following section.

5.2. Implications

The existence of affix formulaicity as shown by this study shows that there may be holistic processing of morpheme sequences, and not just sequences of words. This means that there are sequences of morphemes that are processed as single wholes, despite their analyzability. This also implies the possibility that the lexicon contains preassembled affix sequences. This phenomenon may lend support to certain psycholinguistic models of morphological comprehension over others.

The existence of affix formulaicity poses a challenge for models that have only analytic morphological processing (e.g., Taft & Forster 1975, Pinker 1997). That certain affix sequences recur as if they are single units poses a challenge for such models in which polymorphemic strings have to be fully decomposed in processing. The data in this study also pose a challenge for full-listing models in which any multimorphemic word is holistically processed (e.g. Manelis & Tharp 1977). This is because not all sequences in the data are formulaic. Although affix formulas may be consistent with holistic processing, sequences that were unlikely to be formulaic still need to resort to some analytic processing.

Instead, the findings here occupy a middle ground between the two extremes and are more consistent with models that can accommodate both analytic and holistic processing of multimorphemic words. This is because the data contain both formulaic and non-formulaic affix sequences. However, there are multiple models that can accommodate both types of processing. Some of the models to be considered are the following:

1) Hybrid models (e.g., Marslen-Wilson et al. 1994, Caramazza et al. 1988),

2) Interactive activation models (Taft 1994)

3) Distributed connectionist models (Seidenberg & McClelland 1989)

First, in hybrid models, there are two routes of processing: holistic processing and decomposition. A word may be processed through either of these routes. Particular models differ on what categories of words go through which processing (e.g., inflectional vs. derivational morphemes). Hybrid models also assume that morphemes are discrete entities. Second, in interactive activation models, there are nodes corresponding to various units (e.g. syllables, morphemes, words) that are hierarchically connected and are activated in processing. The connected nodes map low-level units such as graphemes onto meanings or concepts. In this model representations of words have to be stipulated rather than discovered via learning. Third, distributed connectionist models consist of a network of weighted connections between neuron-like processing units. The network learns to map from one domain to another (e.g. sound to meaning) and knowledge is stored as weights on the connections. Representations are patterns of activation in the network that are discovered in learning rather than stipulated. These models do not assume morphemes as discrete entities. Instead, morphemes are the result of regularities that the network picks up in the mappings. Figure 9 provides graphical representations of these models.

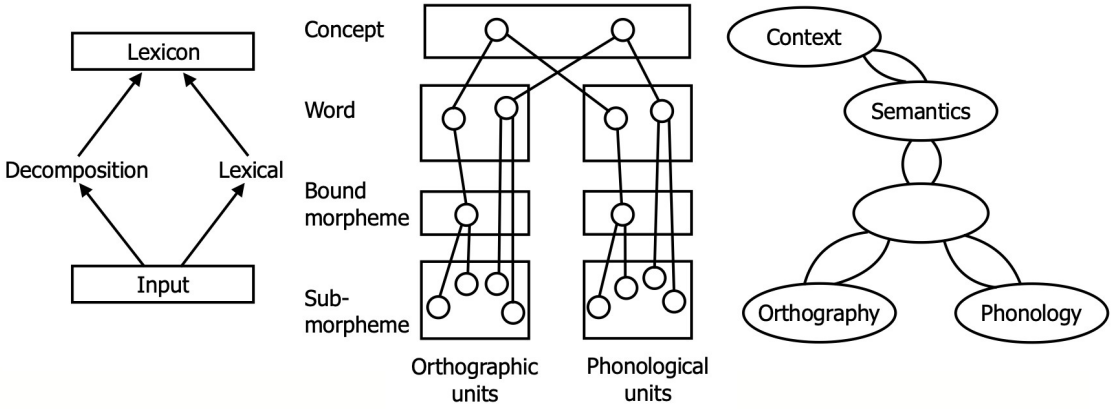


Figure 9: Models of morphological processing (based on Seidenberg & Gonnerman 2000)

From left: 1) Hybrid models, 2) Interactive activation models, 3) Distributed connectionist models.

Next, the following discussion considers which of these models is most likely supported by this study's findings. The basis of this discussion are the three main aspects of the results stated in Section 5.1. The first aspect is the existence of affix formulaicity and what affix sequences were formulaic. This aspect of the results appears to pose a problem for hybrid models. Hybrid models involve a stipulation of what classes of sequences are holistically processed or decomposed. However, some affix sequences were found to be formulaic and some to be non-formulaic in this study. This lack of clear distinction between affix sequences is hard to be accommodated by the dichotomy of hybrid models. Similarly, this aspect of the results appears to pose a problem for the interactive activation model as well. The interaction activation model has to stipulate levels of representation. To accommodate the data, the model needs to somehow stipulate representations for multimorphemic sequences, some of which are intermediate between atomic morphemes and full words. Given the variety of affix formulas, this much stipulation would be difficult. Instead, the findings here do appear to be more consistent with distributed connectionist models. This is because distributed connectionist models do not have to stipulate distinctions between strings or levels of representation. Instead, the feature of feedback into memory can be construed as the pathway by which affix sequences become formulaic. Affix sequences become formulaic through recurrent usage; recurrent sequences are fed back into

the network, and the connection weights therein are adjusted to codify the unitary status of those sequences. Thus the flexibility of distributed connectionist networks means they can accommodate the varied formulaicity data better than the other models.

The second aspect to consider these models by is the gradience of affix formulaicity. This aspect of the results appears to pose a problem for hybrid models. The duality of hybrid models by which strings are either decomposed or holistically processed is challenged by the range of degrees of formulaicity. In contrast, interactive activation models can accommodate gradience in the magnitude of formulaicity. This type of gradience can be encoded as the weights on the connections between nodes in these models. However, structural gradience is a challenge for interactive activation models. This is because the stipulation on levels of representation in these models involves having discrete morphological structures. Instead, gradience of affix formulaicity appears to be more consistent with distributed connectionist network models. This kind of models predicts that association between items would be gradient. Gradience of affix formulaicity reflects how strong is the set of nodes representing one morpheme is connected to the set of nodes representing the other suffix. That strength of connection is codified as the weights on connections, which come in gradient values. Furthermore, the weight on connections involved depends on learning from data, thus the degrees of affix formulaicity need not be stipulated.

The third aspect to consider these models by is the existence of formulas consisting of stems and affixes. This aspect of the results appears to pose a problem for hybrid models. In hybrid models, a string either goes through decomposition or holistic processing. However, the data suggest that stem-affix formulas may not necessarily encompass whole words. Thus, a word can be partially decompositional. This would be difficult to be accommodated by the dichotomy of hybrid models. Similarly, this aspect of the results appears to pose a problem for interactive activation models as well. To accommodate this aspect of the results, these models would need to stipulate nodes corresponding to stem-affix formulas. Given that affix-stem formulas can be intermediate between stems and

morphologically complete words, this would be difficult to accommodate in the model with its segregation between words and bound morphemes. Instead, stem-affix formulas appear to be more consistent with distributed connectionist models. Distributed connectionist models do not stipulate a difference between stems and suffixes. This lack of distinction allows the network to reinforce connections between nodes representing stems and nodes representing affixes. Additionally, the network seems to be flexible enough to codify the unitary status of stem-affix formulas that are less than whole words.

Thus, various aspects of the results on affix formulaicity here pose problems for hybrid models and interactive activation models. In contrast, the aspects of the results highlighted appear to be more readily accommodated by distributed connectionist models. Regardless of which of these three models is truly supported by this study's results, it is apparent that there is a need for both analytic and holistic processing in morphology.

5.3. Application

In addition to theoretical implications as discussed in the previous section, the findings of this study may apply to other areas of language research as well. To exemplify this, an area is chosen that may concern the various aspects of this study's findings, namely the existence of affix formulaicity, gradience in affix formulaicity and types of processing in psycholinguistic processing models. One area suitable for this discussion is aphasia research.

In research on word-level formulaic sequences in aphasics, the general finding is that some formulaic sequences can remain in aphasic language (Wray 2002). This lies in contrast to a general deterioration of the ability to construct novel utterances. A possible interpretation of the resilience of formulaic sequences is that they are stored like single words, thus producing them involves a simpler retrieval of a single lexical item rather than lengthier online construction. A topic of aphasia research by which there is a potential interaction between affix formulaicity and aphasia is agrammatism, a symptom of which is

the omission of inflectional morphemes (Grodzinsky 1984). This study found that there are sequences of affixes that may be formulaic. A possible interpretation of this is that those morpheme sequences are stored in the lexicon, just like words. In parallel to the findings with word-level formulas, a possible prediction is that aphasics may well be able to produce affix formulas while still struggling with the rest of the morphology. Thus formulaic status of certain affix sequences may mean they are as easy to access as single words.

In addition to predicting the resilience of affix formulas in agrammatism, some predictions are also possible about the patterns within that resilience. The first pattern concerns stem-affix formulas. Given that there may also be stem-affix formulas, it is uncertain if resilience in aphasia can be equally expected for both affix-only formulas and stem-affix formulas. A possible reference point for this question is the finding that when it comes to word-level formulas, aphasics struggle more with semi-fixed formulas than with fixed formulas (Wray 2002). Analogically, affix-only formulas are semi-fixed formulas because the elements around them still have to be specified. Thus, it is predicted that aphasics would be better at producing stem-affix formulas and would struggle more with affix-only formulas. However, given that it has been found that affix formulaicity is a gradient phenomenon, it can be expected that the difference between affix-only formulas and stem-affix formulas may not be categorical.

The second pattern concerns the findings from previous studies that found that derivational morphology is affected less than inflectional morphology in aphasia (Badecker & Caramazza 1989, Niemi et al. 1994). It is possible that the previously found inflection vs. derivation distinction is epiphenomenal. Rather, what is truly at work is the formulaic status of polymorphemic sequences, which is related to unitary status in memory. Because derivational morphology is more particular than inflection, it is expected that a derived word is more likely to be formulaic than an inflected word. This resonates elsewhere by that some hybrid models of morphological processing stipulate that inflected words go through decomposition and derived words are holistically processed (Marslen-Wilson et al. 1994). This distinction based on formulaicity rather than affix class should be more explanatory

because the patterns in agrammatism in previous studies (e.g., Badecker & Caramazza 1989, Niemi et al. 1994) are not as categorical as would be expected based on the derivation vs. inflection distinction alone.

The patterns predicted above may have a methodological implication for aphasia research as well. If only decomposition is assumed to be taking place in morphological processing, the correct production of affix sequences may only indicate some normalcy in morphological functioning. If both decomposition and holistic processing are assumed to be taking place in morphological processing, the correct production of affix sequences may be due to some normalcy in morphological functioning, or it may be due to the holistic processing of such affix sequences. In the latter case, the affix sequence may just be accessed like single words. Given this, the resilience of affix formulas then gives a semblance of normal morphological functioning. This could lead to an underestimation of the severity of the aphasia. Further complicating matters, because formulaic status of affix sequences may be due to language experience, as described by distributed connectionist models (Seidenberg & McClelland 1989), what affix sequences are formulaic may differ by language user. Thus, in assessing the severity of aphasia, what counts as a unit in a user's lexicon may need to be relativized.

References

Agresti, Alan. 2019. An Introduction to Categorical Data Analysis. Third edition. Wiley Series in Probability and Statistics. Hoboken, NJ: John Wiley & Sons.

Akın, Ahmet, and Mehmet Akın. 2018. Zemberek-NLP. https://github.com/ahmetaa/zemberek-nlp.

Aksan, Yeşim, Mustafa Aksan, Ahmet Koltuksuz, Taner Sezer, Ümit Mersinli, Umut Ufuk, Hakan Yılmazer, et al. 2012. "Construction of the Turkish National Corpus (TNC)." Proceedings of the 12th International Conference on Language Resources and Evaluation, 3223–27.

Aksan, Yeşim, Mustafa Aksan, Ümit Mersinli, and Umut Ufuk Demirhan. 2017. A frequency dictionary of Turkish: core vocabulary for learners. Routledge frequency dictionaries. London ; New York: Routledge, Taylor & Francis Group.

Alegre, Maria, and Peter Gordon. 1999. "Frequency Effects and the Representational Status of Regular Inflections." Journal of Memory and Language 40 (1): 41–61. https://doi.org/10.1006/jmla.1998.2607.

Badecker, William, and Alfonso Caramazza. 1989. "A Lexical Distinction between Inflection and Derivation." Linguistic Inquiry 20 (1): 108–16.

Becker, Joseph D. 1975. "The Phrasal Lexicon." In Proceedings of the 1975 Workshop on Theoretical Issues in Natural Language Processing - TINLAP '75, 60. Cambridge, Massachusetts: Association for Computational Linguistics. https://doi.org/10.3115/980190.980212.

Beckner, Clay, Richard Blythe, Joan Bybee, Morten H. Christiansen, William Croft, Nick C. Ellis, John Holland, Jinyun Ke, Diane Larsen-Freeman, and Tom Schoenemann. 2009. "Language Is a Complex Adaptive System: Position Paper." Language Learning 59 (December): 1–26. https://doi.org/10.1111/j.1467-9922.2009.00533.x.

Benton, A. L., and R. J. Joynt. 1960. "Early Descriptions of Aphasia." Archives of Neurology 3 (2): 205–22. https://doi.org/10.1001/archneur.1960.00450020085012.

Bertram, R. 2000. "Affixal Homonymy Triggers Full-Form Storage, Even with Inflected Words, Even in a Morphologically Rich Language." Cognition 74 (2): B13–25. https://doi.org/10.1016/S0010-0277(99)00068-2.

Bertram, Raymond, Robert Schreuder, and R. Harald Baayen. 2000. "The Balance of Storage and Computation in Morphological Processing: The Role of Word Formation Type, Affixal Homonymy, and Productivity." Journal of Experimental Psychology: Learning, Memory, and Cognition 26 (2): 489–511. https://doi.org/10.1037/0278-7393.26.2.489.

Biber, Douglas. 2009. "A Corpus-Driven Approach to Formulaic Language in English: Multi-Word Patterns in Speech and Writing." International Journal of Corpus Linguistics 14 (3): 275–311. https://doi.org/10.1075/ijcl.14.3.08bib.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. Corpus Linguistics: Investigating Language Structure and Use. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511804489.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad, and Edward Finegan. 1999. Longman Grammar of Spoken and Written English. Harlow: Longman.

Biskup, Danuta. 1992. "L1 Influence on Learners' Renderings of English Collocations: A Polish/German Empirical Study." In Vocabulary and Applied Linguistics, edited by Pierre J. L. Arnaud and Henri Béjoint, 85–93. London: Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-12396-4_8.

Booij, G. E. 2010. Construction Morphology. Oxford Linguistics. Oxford ; New York: Oxford University Press.

Bouamor, Dhouha, Nasredine Semmar, and Pierre Zweigenbaum. 2012. "Identifying Bilingual Multi-Word Expressions for Statistical Machine Translation." In LREC, 674–679.

Brooke, Julian, Jan Šnajder, and Timothy Baldwin. 2017. "Unsupervised Acquisition of Comprehensive Multiword Lexicons Using Competition in an n -Gram Lattice." Transactions of the Association for Computational Linguistics 5 (December): 455–70. https://doi.org/10.1162/tacl_a_00073.

Bybee, Joan. 2003. "Mechanisms of Change in Grammaticization: The Role of Frequency." In The Handbook of Historical Linguistics, edited by Brian D. Joseph and Richard D. Janda, 602–23. Oxford, UK: Blackwell Publishing Ltd. https://doi.org/10.1002/9780470756393.ch19.

Bybee, Joan L., and Rena Torres Cacoullos. 2009. "The Role of Prefabs in Grammaticization: How the Particular and the General Interact in Language Change." In Typological Studies in Language, edited by Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali, and Kathleen Wheatley, 82:187–218. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/tsl.82.09the.

Caballero, Gabriela, and Alice C. Harris. 2012. “A Working Typology of Multiple Exponence.” In Current Issues in Linguistic Theory, edited by Ferenc Kiefer, Mária Ladányi, and Péter Siptár, 322:163–88. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/cilt.322.08cab.

Caramazza, Alfonso, Alessandro Laudanna, and Cristina Romani. 1988. “Lexical Access and Inflectional Morphology.” Cognition 28 (3): 297–332. https://doi.org/10.1016/0010-0277(88)90017-0.

Chomsky, Noam. 1965. Aspects of the Theory of Syntax. 3. paperback print. Massachusetts Institute of Technology (Cambridge, Mass.). Research Laboratory of Electronics. Special Technical Report 11. Cambridge, Mass: M. I. T. Press.

Church, Kenneth, and Patrick Hanks. 1990. “Word Association Norms, Mutual Information, and Lexicography.” Journal of Computational Linguistics 16 (1): 22–29.

Clear, Jeremy. 1993. “From Firth Principles: Computational Tools for the Study of Collocation.” In Text and Technology: In Honor of John SInclair, edited by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 271–92.

Cohen, Jacob. 2013. Statistical Power Analysis for the Behavioral Sciences. Hoboken: Taylor and Francis. http://www.123library.org/book_details/?id=107447.

Comrie, Bernard, ed. 1990. The World’s Major Languages. 1st ed. New York: Oxford Univ. Press.

Conklin, Kathy, and Norbert Schmitt. 2008. “Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers?” Applied Linguistics 29 (1): 72–89. https://doi.org/10.1093/applin/amm022.

———. 2012. “The Processing of Formulaic Language.” Annual Review of Applied Linguistics 32 (March): 45–61. https://doi.org/10.1017/S0267190512000074.

Critchley, Macdonald. 1970. Aphasiology and Other Aspects of Language. London: Arnold.

Crystal, David. 2011. Dictionary of Linguistics and Phonetics. New York, NY: John Wiley & Sons.

Dice, Lee R. 1945. “Measures of the Amount of Ecologic Association Between Species.” Ecology 26 (3): 297–302. https://doi.org/10.2307/1932409.

Durrant, Philip. 2013. “Formulaicity in an Agglutinating Language: The Case of Turkish.” Corpus Linguistics and Linguistic Theory 9 (1): 1–38. https://doi.org/10.1515/cllt-2013-0009.

Ellis, Nick C. 1996. “Sequencing in SLA: Phonological Memory, Chunking, and Points of Order.” Studies in Second Language Acquisition 18 (1): 91–126. https://doi.org/10.1017/S0272263100014698.

Erman, Britt. 2009. “Formulaic Language from a Learner Perspective: What the Learner Needs to Know.” In Typological Studies in Language, edited by Roberta Corrigan, Edith A. Moravcsik, Hamid Ouali, and Kathleen Wheatley, 83:323. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/tsl.83.05erm.

Evert, Stefan. 2005. “The Statistics of Word Cooccurrences : Word Pairs and Collocations.” Universität Stuttgart. https://doi.org/10.18419/opus-2556.

Finlayson, Mark Alan, and Nidhi Kulkarni. 2011. “Detecting Multi-Word Expressions Improves Word Sense Disambiguation.” In Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World, 20–24. MWE ’11. Stroudsburg, PA, USA: Association for Computational Linguistics. http://dl.acm.org/citation.cfm?id=2021121.2021128.

Frauenfelder, Uli H., and Robert Schreuder. 1992. “Constraining Psycholinguistic Models of Morphological Processing and Representation: The Role of Productivity.” In Yearbook of Morphology 1991, edited by Geert Booij and Jaap van Marle, 165–83. Dordrecht: Springer Netherlands. https://doi.org/10.1007/978-94-011-2516-1_10.

Gablasova, Dana, Vaclav Brezina, and Tony McEnery. 2017. “Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence: Collocations in Corpus-Based Language Learning Research.” Language Learning 67 (S1): 155–79. https://doi.org/10.1111/lang.12225.

Gardner, H. 1985. “Loss of Language.” In Language: Introductory Readings, edited by V.P. Clark, P.A Escholz, and A.F. Rosa, 4th ed. New York: St Martin’s Press.

Göksel, Aslı, and Celia Kerslake. 2005. Turkish a Comprehensive Grammar. London; New York: Routledge. http://www.myilibrary.com?id=34877.

Gonnerman, Laura M., Mark S. Seidenberg, and Elaine S. Andersen. 2007. “Graded Semantic and Phonological Similarity Effects in Priming: Evidence for a Distributed Connectionist Approach to Morphology.” Journal of Experimental Psychology: General 136 (2): 323–45. https://doi.org/10.1037/0096-3445.136.2.323.

Grodzinsky, Yosef. 1984. “The Syntactic Characterization of Agrammatism.” Cognition 16 (2): 99–120. https://doi.org/10.1016/0010-0277(84)90001-5.

Hay, J, and R Baayen. 2005. “Shifting Paradigms: Gradient Structure in Morphology.” Trends in Cognitive Sciences 9 (7): 342–48. https://doi.org/10.1016/j.tics.2005.04.002.

Hay, Jennifer. 2001. “Lexical Frequency in Morphology: Is Everything Relative?” Linguistics 39 (6). https://doi.org/10.1515/ling.2001.041.

Heffernan, Kevin, and Yo Sato. 2017. “Relative Frequency and the Holistic Processing of Morphology: Evidence from a Corpus of Vernacular Japanese.” Asia-Pacific Language Variation 3 (1): 67–94. https://doi.org/10.1075/aplv.3.1.04hef.

Jaworski, Adam. 1990. “The Acquisition and Perception of Formulaic Language and Foreign Language Teaching.” Multilingua - Journal of Cross-Cultural and Interlanguage Communication 9 (4): 397–412. https://doi.org/10.1515/mult.1990.9.4.397.

Jiang, Nan, and Tatiana M. Nekrasova. 2007. “The Processing of Formulaic Sequences by Second Language Speakers.” The Modern Language Journal 91 (3): 433–45. https://doi.org/10.1111/j.1540-4781.2007.00589.x.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. “Statistical Phrase-Based Translation.” In Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL ’03, 1:48–54. Edmonton, Canada: Association for Computational Linguistics. https://doi.org/10.3115/1073445.1073462.

Kornfilt, Jaklin. 1990. “Turkish and the Turkic Languages.” In The World’s Major Languages, edited by Bernard Comrie, 619–44. Oxford: Oxford University Press.

———. 2003. Turkish. Descriptive Grammars. London: Routledge.

———. 2006. “Agreement: The (Unique and Local) Syntactic and Morphological Licenser of Subject Case.” In Linguistik Aktuell/Linguistics Today, edited by João Costa and Maria Cristina Figueiredo Silva, 86:141–71. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/la.86.08kor.

Krenn, Brigitte, and Stefan Evert. 2001. “Can We Do Better than Frequency? A Case Study on Extracting PP-Verb Collocations.” In Proceedings of the ACL Workshop on Collocations, 39–46.

Kuiper, Koenraad. 1996. Smooth Talkers: The Linguistic Performance of Auctioneers and Sportscasters. Mahwah, N.J: L. Erlbaum Associates.

———. 2004. “Formulaic Performance in Conventionalised Varieties of Speech.” In Language Learning & Language Teaching, edited by Norbert Schmitt, 9:37–54. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/lllt.9.04kui.

Lambert, Patrik, and Rafael Banchs. 2005. "Data Inferred Multi-Word Expressions for Statistical Machine Translation." In Proceedings of Machine Translation Summit X, 396–403.

Lehtonen, Minna, Toni Cunillera, Antoni Rodríguez-Fornells, Annika Hultén, Jyrki Tuomainen, and Matti Laine. 2007. "Recognition of Morphologically Complex Words in Finnish: Evidence from Event-Related Potentials." Brain Research 1148 (May): 123–37. https://doi.org/10.1016/j.brainres.2007.02.026.

Libben, Maya R., and Debra A. Titone. 2008. "The Multidetermined Nature of Idiom Processing." Memory & Cognition 36 (6): 1103–21. https://doi.org/10.3758/MC.36.6.1103.

Manelis, Leon, and David A. Tharp. 1977. "The Processing of Affixed Words." Memory & Cognition 5 (6): 690–95. https://doi.org/10.3758/BF03197417.

Manning, Christopher D., and Hinrich Schütze. 1999. Foundations of Statistical Natural Language Processing. Cambridge, Mass: MIT Press.

Marslen-Wilson, William, Lorraine K. Tyler, Rachelle Waksler, and Lianne Older. 1994. "Morphology and Meaning in the English Mental Lexicon." Psychological Review 101 (1): 3–33. https://doi.org/10.1037/0033-295X.101.1.3.

Millar, N. 2011. "The Processing of Malformed Formulaic Language." Applied Linguistics 32 (2): 129–48. https://doi.org/10.1093/applin/amq035.

Nelson, Katherine. 1973. "Structure and Strategy in Learning to Talk." Monographs of the Society for Research in Child Development 38 (1/2): 1. https://doi.org/10.2307/1165788.

New, Boris, Marc Brysbaert, Juan Segui, Ludovic Ferrand, and Kathleen Rastle. 2004. "The Processing of Singular and Plural Nouns in French and English." Journal of Memory and Language 51 (4): 568–85. https://doi.org/10.1016/j.jml.2004.06.010.

Newman, David, Nagendra Koilada, Jey Han Lau, and Timothy Baldwin. 2012. "Bayesian Text Segmentation for Index Term Identification and Keyphrase Extraction." In Proceedings of COLING 2012, 2077–2092. Mumbai, India: The COLING 2012 Organizing Committee.

Niemi, Jussi, Matti Laine, and Juhani Tuominen. 1994. "Cognitive Morphology in Finnish: Foundations of a New Model." Language and Cognitive Processes 9 (3): 423–46. https://doi.org/10.1080/01690969408402126.

Pawley, Andrew, and Frances Syder. 1983. "Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency." In Language and Communication, edited by Jack Richards and R.W. Schmidt, 191–226. New York: Longman.

Perkins, Mick. 1999. "Productivity and Formulaicity in Language Development." In Issues in Normal & Disordered Child Language: From Phonology to Narrative., edited by M. Garman, C. Letts, C. Schelletter, and S. Edwards, 51–67. Special Issue of The New Bulmershe Papers. Reading: University of Reading.

Peters, Ann M. 1977. "Language Learning Strategies: Does the Whole Equal the Sum of the Parts?" Language 53 (3): 560. https://doi.org/10.2307/413177.

———. 1983. The Units of Language Acquisition. Cambridge Monographs and Texts in Applied Psycholinguistics. Cambridge [Cambridgeshire] ; New York: Cambridge University Press.

Pinker, Steven. 1997. "Words and Rules in the Human Brain." Nature 387 (6633): 547–48. https://doi.org/10.1038/42347.

Pinker, Steven, and Michael Ullman. 2002. "Combination and Structure, Not Gradedness, Is the Issue." Trends in Cognitive Sciences 6 (11): 472–74. https://doi.org/10.1016/S1364-6613(02)02013-2.

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. "Multiword Expressions: A Pain in the Neck for NLP." In Computational Linguistics and Intelligent Text Processing, edited by Alexander Gelbukh, 2276:1–15. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-45715-1_1.

Schmidt, Carsten Oliver, and Thomas Kohlmann. 2008. "When to Use the Odds Ratio or the Relative Risk?" International Journal of Public Health 53 (3): 165–67. https://doi.org/10.1007/s00038-008-7068-3.

Schmitt, Norbert, Sarah Grandage, and Svenja Adolphs. 2004. "Are Corpus-Derived Recurrent Clusters Psycholinguistically Valid?" In Language Learning & Language Teaching, edited by Norbert Schmitt, 9:127–51. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/lllt.9.08sch.

Schneider, Nathan, Emily Danchik, Chris Dyer, and Noah A. Smith. 2014. "Discriminative Lexical Semantic Segmentation with Gaps: Running the MWE Gamut." Transactions of the Association for Computational Linguistics 2 (December): 193–206. https://doi.org/10.1162/tacl_a_00176.

Seidenberg, Mark S., and Laura M. Gonnerman. 2000. “Explaining Derivational Morphology as the Convergence of Codes.” Trends in Cognitive Sciences 4 (9): 353–61. https://doi.org/10.1016/S1364-6613(00)01515-1.

Seidenberg, Mark S., and James L. McClelland. 1989. “A Distributed, Developmental Model of Word Recognition and Naming.” Psychological Review 96 (4): 523–68. https://doi.org/10.1037/0033-295X.96.4.523.

Semenza, Carlo, Claudio Luzzatti, and Simona Carabelli. 1997. “Morphological Representation of Compound Nouns: A Study on Italian Aphasic Patients.” Journal of Neurolinguistics 10 (1): 33–43. https://doi.org/10.1016/S0911-6044(96)00019-X.

Sereno, Joan A., and Allard Jongman. 1997. “Processing of English Inflectional Morphology.” Memory & Cognition 25 (4): 425–37. https://doi.org/10.3758/BF03201119.

Simmons, Joseph P., Leif D. Nelson, and Uri Simonsohn. 2011. “False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant.” Psychological Science 22 (11): 1359–66. https://doi.org/10.1177/0956797611417632.

Sinclair, John. 1995. Corpus, Concordance, Collocation. 3. impr. Describing English Language. Oxford: Oxford Univ. Press.

Siyanova-Chanturia, Anna, Kathy Conklin, and Walter J. B. van Heuven. 2011. “Seeing a Phrase ‘Time and Again’ Matters: The Role of Phrasal Frequency in the Processing of Multiword Sequences.” Journal of Experimental Psychology: Learning, Memory, and Cognition 37 (3): 776–84. https://doi.org/10.1037/a0022531.

Smadja, Frank, Vasileios Hatzivassiloglou, and Kathleen R McKeown. 1996. “Translating Collocations for Bilingual Lexicons: A Statistical Approach.” Computational Linguistics 22 (1): 38.

Soveri, Anna, Minna Lehtonen, and Matti J. Laine. 2007. “Word Frequency and Morphological Processing in Finnish Revisited.” The Mental Lexicon 2 (3): 359–85. https://doi.org/10.1075/ml.2.3.04sov.

Stefan Th., Gries. 2017. “Corpus Approaches.” In Cambridge Handbook of Cognitive Linguistics, edited by Barbara Dancygier, 590–606. Cambridge: Cambridge University Press.

Stefanowitsch, Anatol, and Stefan Th. Gries. 2003. “Collostructions: Investigating the Interaction of Words and Constructions.” International Journal of Corpus Linguistics 8 (2): 209–43. https://doi.org/10.1075/ijcl.8.2.03ste.

Stevick, Earl W. 1989. Success with Foreign Languages: Seven Who Achieved It and What Worked for Them. Prentice Hall International Language Teaching Methodology Series. New York: Prentice Hall.

Taft, Marcus. 1994. "Interactive-Activation as a Framework for Understanding Morphological Processing." Language and Cognitive Processes 9 (3): 271–94. https://doi.org/10.1080/01690969408402120.

Taft, Marcus, and Kenneth I. Forster. 1975. "Lexical Storage and Retrieval of Prefixed Words." Journal of Verbal Learning and Verbal Behavior 14 (6): 638–47. https://doi.org/10.1016/S0022-5371(75)80051-X.

Thanopoulos, Aristomenis, Nikos Fakotakis, and George Kokkinakis. 2002. "Comparative Evaluation of Collocation Extraction Metrics," 6.

Tremblay, Antoine, Bruce Derwing, Gary Libben, and Chris Westbury. 2011. "Processing Advantages of Lexical Bundles: Evidence From Self-Paced Reading and Sentence Recall Tasks: Lexical Bundle Processing." Language Learning 61 (2): 569–613. https://doi.org/10.1111/j.1467-9922.2010.00622.x.

Ullman, Michael T. 2001. "A Neurocognitive Perspective on Language: The Declarative/Procedural Model." Nature Reviews Neuroscience 2 (10): 717–26. https://doi.org/10.1038/35094573.

Underhill, Robert. 1976. Turkish Grammar. Cambridge, Mass: MIT Press.

Underwood, Geoffrey, Norbert Schmitt, and Adam Galpin. 2004. "The Eyes Have It: An Eye-Movement Study into the Processing of Formulaic Sequences." In Language Learning & Language Teaching, edited by Norbert Schmitt, 9:153–72. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/lllt.9.09und.

Van Lancker, Diana. 1988. "Nonpropositional Speech: Neurolinguistic Studies." In Progress in the Psychology of Language, edited by Andrew Ellis, 49–118. Hillsdale, NJ: Lawrence Erlbaum Associates.

Van Lancker, Diana, and Gerald J. Canter. 1981. "Idiomatic versus Literal Interpretations of Ditropically Ambiguous Sentences." Journal of Speech, Language, and Hearing Research 24 (1): 64–69. https://doi.org/10.1044/jshr.2401.64.

Van Lancker, Diana, Gerald J. Canter, and Dale Terbeek. 1981. "Disambiguation of Ditropic Sentences: Acoustic and Phonetic Cues." Journal of Speech, Language, and Hearing Research 24 (3): 330–35. https://doi.org/10.1044/jshr.2403.330.

Van Lancker, Diana Roupas, and Daniel Kempler. 1987. "Comprehension of Familiar Phrases by Left- but Not by Right-Hemisphere Damaged Patients." Brain and Language 32 (2): 265–77. https://doi.org/10.1016/0093-934X(87)90128-3.

Vechtomova, Olga. 2005. "The Role of Multi-Word Units in Interactive Information Retrieval." In Advances in Information Retrieval, edited by David E. Losada and Juan M. Fernández-Luna, 3408:403–20. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-31865-1_29.

Wray, Alison. 2002. Formulaic Language and the Lexicon. Cambridge: Cambridge University Press.

———. 2012. "What Do We (Think We) Know About Formulaic Language? An Evaluation of the Current State of Play." Annual Review of Applied Linguistics 32 (March): 231–54. https://doi.org/10.1017/S026719051200013X.

Wray, Alison, and Tess Fitzpatrick. 2008. "Why Can't You Just Leave It Alone? Deviations from Memorized Language as a Gauge of Nativelike Competence." In Phraseology in Foreign Language Learning and Teaching, edited by Fanny Meunier and Sylviane Granger, 123–47. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/z.138.11wra.

Wurm, Lee H. 1997. "Auditory Processing of Prefixed English Words Is Both Continuous and Decompositional." Journal of Memory and Language 37 (3): 438–61. https://doi.org/10.1006/jmla.1997.2524.

Appendix A: Verb stems used for queries on the Turkish National Corpus (TNC)

Frequency rank	Verb stem	Translation	Frequency rank	Verb stem	Translation
1	ol	be, become	41	tut	hold
2	et	do, make	42	at	throw
3	yap	do, make	43	sür	drive, last
4	al	take, get	44	bırak	leave
5	de	say	45	oku	read
6	gel	come	46	sev	love, like
7	ver	give	47	duy	hear
8	gör	see	48	tanı	recognize, be acquainted
9	çık	go out (of)	49	bekle	wait (for)
10	bul	find	50	uygula	apply
11	git	go	51	değiş	change
12	çalış	work	52	koy	put, place
13	iste	want	53	kaç	escape
14	geç	pass	54	dön	turn
15	bil	know	55	art	increase, remain
16	anla	understand	56	belir	appear
17	kal	stay, remain	57	öl	die
18	söyle	say, tell	58	kazan	win
19	bak	look at	59	ayır	separate
20	ye	eat	60	otur	sit
21	başla	begin, start	61	taşı	carry
22	i	be (defective)	62	say	count
23	yaşa	live	63	bit	finish, end
24	gir	enter	64	açıkla	explain, disclose
25	kullan	use	65	karşıla	meet
26	gerek	be necessary	66	kat	add
27	düşün	think	67	kes	cut
28	getir	bring	68	gül	laugh
29	aç	open	69	gerçekleş	materialize, happen
30	yaz	write	70	ulaş	reach
31	göster	show	71	işle	commit, work
32	konuş	talk, speak	72	yak	ignite, light
33	oluş	come into being	73	yürü	walk
34	çek	pull	74	kalk	stand up
35	geliş	develop	75	öğren	learn
36	düş	fall	76	düzenle	arrange, organize
37	dur	stop	77	bin	ride, get on
38	sor	ask	78	yarat	create
39	sağla	provide	79	izle	follow, watch
40	kur	set up	80	oyna	play

Frequency rank	Verb stem	Translation	Frequency rank	Verb stem	Translation
81	sık	squeeze	126	kaybet	lose
82	koru	protect	127	boz	damage
83	hazırla	prepare	128	koş	run
84	belirle	determine	129	tak	attach, wear
85	topla	gather, collect	130	çöz	solve, untie
86	seç	select, choose	131	kon	put, land
87	doğ	be born	132	çevir	rotate, turn, translate
88	inan	believe	133	kap	grab, seize, catch
89	değerlen	gain value	134	kır	break
90	tart	weigh	135	çal	steal, play an instrument
91	daya	lean	136	yakala	catch
92	kapa	close, shut	137	yaklaş	approach
93	dol	get full, fill	138	ağla	cry
94	benze	resemble	139	öğret	teach
95	bas	step on, print, press	140	giy	put on, wear
96	büyü	grow	141	tanımla	define
97	unut	forget	142	yansı	echo
98	in	descend, get off	143	güven	trust
99	gönder	send	144	uza	get longer
100	etkile	affect, influence	145	sar	wrap, roll
101	incele	analyze	146	birleş	unite
102	san	suppose, imagine	147	uğra	stop by
103	sat	sell	148	aş	cross over
104	kaldır	lift, raise	149	ekle	add
105	dinle	listen	150	içer	include, contain
106	sun	present, submit	151	tamamla	complete
107	yat	lay down	152	sok	insert
108	kar	mix, blend	153	yay	spread
109	ele	sieve	154	harca	spend
110	götür	take	155	evlen	marry
111	yüksel	rise	156	dök	pour
112	hatırla	remember	157	çağır	call, invite
113	hisset	feel, perceive	158	yönel	head towards
114	dönüş	turn into	159	gez	stroll, go around
115	öde	pay	160	kurtul	be rescued
116	yet	suffice	161	dene	try, test
117	üret	produce	162	ilerle	go forward
118	bağla	tie	163	ilgilen	show interest, care
119	vur	hit, strike	164	destekle	support
120	yetiş	catch	165	çiz	draw
121	azal	lessen, decrease	166	davran	treat
122	yerleş	settle	167	dolaş	stroll, tangle
123	uy	fit	168	gözle	watch, observe
124	kork	fear	169	şaş	amaze
125	savun	defend	170	paylaş	share

Frequency rank	Verb stem	Translation	Frequency rank	Verb stem	Translation
171	uyan	wake up	216	yayınla	publish, broadcast
172	uç	fly	217	algıla	perceive
173	kurtar	save, rescue	218	sil	wipe up
174	yararlan	benefit from	219	alış	get used
175	aktar	transfer	220	düzel	become
176	vurgula	emphasize	221	atla	jump, skip
177	kop	break off	222	kavuş	rejoin
178	bağır	shout	223	patla	burst, explode
179	çarp	bump, crash	224	er	attain, reach
180	kay	slide	225	sus	remain silent
181	ak	flow	226	yükle	load
182	oy	carve	227	beğen	like, admire
183	vazgeç	give up	228	yen	beat
184	ata	appoint	229	başvur	apply
185	bahset	mention	230	dokun	touch
186	planla	plan	231	öngör	foresee
187	zorla	force	232	gülümse	smile
188	engelle	obstruct	233	benimse	adopt
189	uyu	sleep	234	yıka	wash
190	uğraş	struggle	235	buyur	command
191	besle	feed	236	örgütle	organize
192	değ	touch	237	kaybol	disappear
193	başar	succeed in	238	dağıt	distribute
194	sapta	determine	239	genişle	widen
195	öner	propose	240	it	push
196	kaynaklan	originate	241	yayımla	publish
197	yık	demolish	242	öp	kiss
198	rastla	run into	243	yorumla	interpret
199	önle	prevent	244	salla	wave, shake
200	kaydet	record	245	suçla	accuse
201	bula	cover with	246	uyar	warn
202	uzan	lie down	247	eğ	lean, bend
203	sakla	hide, conceal	248	sırala	line up
204	yolla	send	249	es	blow
205	seyret	watch	250	sergile	exhibit
206	böl	divide	251	kuru	dry, get dry
207	üz	upset, sadden	252	parçala	smash up, tear
208	dik	plant, erect	253	sığ	fit into
209	ek	plant, spread	254	ısın	warm, heat up
210	yit	disappear, vanish	255	anımsa	remember
211	hesapla	calculate, estimate	256	kapsa	contain
212	uzaklaş	leave	257	eleştir	criticize
213	yönet	manage, rule, govern	258	don	freeze
214	sözleş	agree	259	as	hang, suspend
215	yağ	rain, snow	260	seslen	call out

Frequency rank	Verb stem	Translation	Frequency rank	Verb stem	Translation
261	kavra	comprehend	306	titre	shiver, tremble
262	güçlen	get strong	307	gizle	hide, conceal
263	döv	beat	308	piş	be cooked
264	yönlen	direct towards	309	sal	release, set free
265	çök	collapse	310	tekrarla	repeat
266	imzala	sign	311	gözük	appear
267	reddet	refuse	312	onayla	approve
268	ört	cover, hide	313	yenile	renew
269	ölç	measure	314	yanıtla	answer, reply
270	üstlen	undertake	315	kanıtla	prove
271	yarış	race, compete	316	özen	imitate
272	kutla	congratulate, celebrate	317	keşfet	discover
273	dinlen	rest, relax	318	zannet	suppose, guess
274	çoğal	increase	319	yakış	suit
275	boya	paint, dye	320	hızlan	speed up
276	göç	migrate	321	sön	die down
277	dağıl	scatter	322	canlan	light up, refresh
278	amaçla	aim, intend	323	tüket	consume
279	boşal	become empty	324	pazarla	market
280	gecik	be late	325	çak	nail
281	bat	sink	326	sürükle	drag
282	fırla	pop out	327	ilet	convey, deliver
283	sınırla	limit	328	utan	be ashamed
284	yargıla	judge	329	bulun	keep
285	temizle	clean	330	yapış	stick, cling
286	yoğunlaş	become dense, thicken	331	süz	filter, strain
287	sap	turn to	332	özetle	summarize
288	çözümle	analyze	333	akla	clear
289	yor	tire	334	rahatla	feel better
290	adlan	name	335	denetle	check
291	ez	crush, mash	336	işit	hear
292	iyileş	get better	337	özle	miss
293	hedefle	aim	338	göm	bury
294	sorgula	interrogate	339	um	hope, expect
295	derle	compile	340	soy	peel, undress
296	eri	melt, dissolve	341	yarala	injure
297	yapılan	structure	342	barın	take shelter
298	sula	water	343	sonuçlan	result
299	birik	accumulate	344	sakın	avoid
300	doy	be satisfied	345	bütünleş	become
301	kayna	boil	346	saldır	attack
302	yasakla	forbid, ban	347	hoşlan	like
303	boşa	divorce	348	yaygınlaş	become
304	boğ	choke, suffocate	349	diren	resist
305	kapla	cover	350	kolaylaş	get easy

Frequency rank	Verb stem	Translation	Frequency rank	Verb stem	Translation
351	yanıl	be mistaken	396	zayıfla	lose weight
352	sez	sense, perceive	397	ser	spread
353	em	suck, absorb	398	kıyasla	compare
354	dola	wind, roll	399	yut	swallow
355	eriş	reach, attain	400	ertele	postpone
356	ayarla	adjust	401	okşa	caress
357	tasarla	plan	402	yar	split realize
358	örnekle	exemplify	403	sin	pervade
359	aydınlat	enlighten	404	tara	comb
360	arın	become clean	405	sök	pull out
361	sars	shake	406	kov	drive away, fire
362	toparla	gather	407	aldat	cheat
363	eğlen	have fun	408	cevapla	answer
364	özelleş	specialize	409	kuşat	surround guard
365	aydınlan	become clear	410	nitele	modify
366	kok	smell	411	devret	transfer wave
367	ger	tighten, irritate	412	sav	get rid of
368	danış	consult	413	bayıl	faint
369	katlan	tolerate	414	kucakla	hug
370	kızar	turn red	415	programla	program
371	gözlemle	observe	416	varsay	assume
372	yığ	pile up	417	ör	knit
373	koşulla	condition	418	kalkış	try
374	daral	become	419	sanayileş	become
375	gerile	regress	420	tıka	plug, block
376	öv	praise	421	bağışla	forgive
377	küçül	shrink	422	kabullen	accept
378	uzlaş	compromise	423	becer	accomplish
379	süsle	decorate	424	sıçra	jump
380	şekillen	form	425	farklılaş	differentiate
381	tüken	be exhausted	426	eğit	educate, train harvest
382	varol	exist	427	bürü	cover up
383	sız	leak	428	hallet	handle, solve punctuate
384	kana	bleed	429	sıyır	strip off
385	tırman	climb	430	faydalan	benefit arrange
386	eyle	act, do	431	çırp	scramble, flap
387	elle	touch	432	önemse	care about
388	kısıtla	restrict, limit	433	dışla	exclude light up
389	doğrula	verify	434	kabar	swell
390	bulaş	smudge, smear, infect	435	borçlan	get into debt
391	soğu	cool	436	yaşlan	grow old together
392	küreselleş	globalize	437	nitelen	be modified
393	tutukla	arrest	438	ayaklan	rebel
394	parla	shine, flare	439	üşü	feel cold
395	çabala	make great effort	440	farket	notice

Frequency rank	Verb stem	Translation	Frequency rank	Verb stem	Translation
441	çürü	rot, decay	486	terket	abandon
442	şiş	swell, inflate	487	kirlen	get dirty
443	sınıflan	classify	488	abar	exaggerate
444	yırt	tear	489	selamla	greet
445	yinele	repeat	490	kararla	be
446	affet	forgive	491	türe	derive
447	gözet	consider	492	kıvır	curl
448	yücel	become great	493	bük	bend
449	dalgalan	fluctuate	494	işaretle	mark
450	bilgilen	be informed	495	heyecanlan	get
451	zenginleş	become wealthy	496	ürk	wince, flinch
452	sinirlen	get angry	497	düşle	imagine
453	kıs	reduce, lessen	498	fısılda	whisper
454	del	drill	499	havalan	be ventilated
455	görevlen	assign	500	belgele	document
456	kastet	mean, imply	501	yakınlaş	become nearer
457	serp	sprinkle	502	kısal	shrink
458	odaklan	focus	503	yabancılaş	become alienated
459	irdele	examine	504	kına	condemn
460	olgunlaş	mature	505	yuvarla	roll
461	emret	order	506	avla	hunt
462	sınıfla	classify	507	çevrele	surround
463	zorlaş	become difficult	508	sınırlan	limit
464	biç	cut into shape	509	kavur	roast
465	noktala	end, punctuate	510	pekiş	reinforce
466	oyla	vote	511	kovala	chase
467	alkışla	applaud	512	yum	shut eyes or mouth
468	yalvar	beg	513	kilitle	lock
469	indirge	degrade	514	bağdaş	be compatible
470	kirala	rent	515	parala	tear into
471	çiğne	chew	516	sark	hang down
472	yavaşla	slow down	517	aşağıla	insult
473	doğrul	straighten out	518	kokla	smell, sniff
474	yanaş	draw near	519	çatla	crack
475	haberleş	communicate	520	özdeşleş	identify with
476	yumuşa	soften	521	terle	sweat
477	depola	store	522	gereksin	need
478	donat	equip	523	inle	moan
479	yasla	lean	524	derinleş	deepen
480	küçümse	look	525	arala	space
481	kıskan	envy	526	gevşe	loosen
482	yokla	inspect	527	kus	vomit
483	haykır	shout	528	oyala	delay
484	kıpırda	move	529	aksa	hamper
485	bık	get bored by	530	döşe	furnish

Frequency rank	Verb stem	Translation	Frequency rank	Verb stem	Translation
531	üre	produce	576	eşle	pair, match
532	aşıla	vaccinate	577	kaşı	scratch
533	betimle	describe	578	ödüllen	reward
534	tap	worship	579	savrul	be thrown away
535	hafifle	lighten	580	susa	get thirsty
536	ısır	bite	581	arzula	wish, desire
537	sürt	rub	582	ürper	shiver
538	hastalan	get ill	583	yadırga	find odd
539	yeğle	prefer	584	öfkelen	rage
540	acık	get hungry	585	üfle	blow
541	sapla	stick	586	konakla	stay
542	resimle	illustrate	587	belirginleş	become apparent
543	modernleş	become	588	fışkır	gush out
544	küs	sulk, be cross	589	dengele	balance
545	ağırla	host	590	mırılda	murmur
546	simgele	symbolize	591	fethet	conquer
547	zehirle	poison	592	esirge	protect
548	öt	coo	593	biçimlen	form
549	zedele	injure	594	yatış	calm down
550	kirlet	dirty, pollute	595	katlet	massacre
551	sömür	exploit	596	ilişkilen	relate
552	naklet	transport	597	umursa	care
553	fotoğrafla	photo	598	devral	take over
554	sarar	turn yellow	599	alıkoy	withhold
555	çıldır	go mad	600	yalanla	deny
556	onar	repair, restore	601	ispatla	prove
557	ateşle	set fire	602	kurgula	build
558	soyutla	abstract	603	hapset	imprison
559	kışkırt	provoke	604	duyumsa	feel
560	ilaçla	apply medicine	605	yozlaş	degenerate
561	yadsı	deny, reject	606	irkil	be startled
562	bunal	be distressed	607	aşın	erode
563	tık	cram	608	yıpran	wear out
564	ezberle	memorize	609	kımılda	move
565	iliş	catch on	610	sik	make love
566	sızla	ache, hurt	611	biçimle	format
567	yala	lick, brush	612	coş	overflow
568	ayıkla	clean, pick	613	tıkla	click, knock
569	bombala	bomb	614	bilinçlen	become conscious
570	uzmanlaş	specialize	615	güçleş	become difficult
571	zıpla	jump, bounce	616	saçmala	talk nonsense
572	uğurla	bid farewell	617	ısmarla	order, request
573	ağırlaş	slow down	618	kesinleş	become definite
574	demokratikleş	democratize	619	karala	cross out, scratch
575	hükmet	rule	620	katla	fold over, double

Frequency rank	Verb stem	Translation	Frequency rank	Verb stem	Translation
621	yaşar	become wet	666	netleş	become apparent
622	haşla	boil	667	yardımla	help
623	diril	revive	668	çınla	ring
624	yeşer	leaf out, become green	669	sarmala	wrap
625	güzelleş	become beautiful	670	sabret	be patient
626	kıvran	agonize	671	körükle	stir up
627	kodla	encode	672	özgürleş	liberate
628	sahiplen	embrace	673	kurcala	tamper
629	gölgele	overshadow	674	temellen	gain ground, establish
630	görüntüle	film	675	tüt	smoke
631	azarla	reprehend	676	çeşitle	vary
632	yağla	oil, grease	677	kuvvetlen	gain strength
633	yüzleş	face	678	paketle	pack
634	solu	breathe	679	devşir	gather
635	sakinleş	calm	680	silk	shake out
636	delir	go mad	681	yudumla	sip
637	esinle	inspire	682	duraksa	hesitate
638	genelle	generalize	683	sarfet	spend
639	süpür	sweep	684	öyküle	narrate
640	duygulan	be affected	685	küfret	swear
641	eksil	lessen	686	somutlaş	embody
642	doğra	chop	687	kurumsallaş	institutionalize
643	yankılan	echo	688	sıç	excrete, mess up
644	büyüle	charm	689	yıprat	wear out
645	durakla	pause	690	vedalaş	say
646	kentleş	urbanize	691	tazele	freshen
647	hareketlen	move	692	alçal	go down
648	isimlen	name	693	eklemle	joint, join
649	beze	adorn	694	şükret	thank god, praise
650	bıçakla	stab	695	endişelen	worry
651	incit	hurt	696	kuşkulan	suspect
652	püskür	spew	697	dona	decorate
653	otla	graze	698	salgıla	excrete
654	zikret	cite, mention	699	yapılaş	structure
655	sertleş	get harder	700	buharlaş	evaporate
656	sırıt	grin	701	usan	get bored
657	tükür	spit	702	hıçkır	hiccup, sob
658	anlamlan	make meaningful	703	arzet	present
659	tep	kick	704	demle	brew
660	büz	pucker	705	yeral	appear in, take part
661	atfet	attribute	706	şüphelen	suspect
662	gözetle	watch	707	lafla	chat
663	boşver	ignore	708	sonlan	be over, end
664	kemir	gnaw	709	esne	yawn, stretch
665	sıva	plaster	710	çağla	splash

Frequency rank	Verb stem	Translation	Frequency rank	Verb stem	Translation
711	sek	hop			
712	buruş	wrinkle			
713	eşitle	equalize			
714	meşrulaş	legitimize			
715	düğümle	knot, tie			
716	ağar	dawn, whiten			
717	ışı	shine, radiate			
718	aldan	be deceived			
719	tuzla	salt			
720	imren	envy			
721	mahvet	destroy			
722	kamaş	be dazzled			
723	resmet	portray			
724	kak	poke			
725	tetikle	trigger			
726	belle	memorize			
727	güncelle	update			
728	hırpala	treat roughly			
729	öksür	cough			
730	yama	patch			
731	çalkala	shake, rinse			
732	başkaldır	rebel			