Proceedings of the Globalex Workshop on Linked Lexicography @LREC2022, pages 68–74
Marseille, 20 June 2022

Morpholex Turkish: A Morphological Lexicon for Turkish

Bilge Nas Arıcan♡, Aslı Kuzgun♡, Büşra Marşan♡, Deniz Baran Aslan♡, Ezgi Sanıyar♡
Neslihan Cesur♡, Neslihan Kara♡, Oğuzhan Kuyrukçu♡, Merve Özçelik♡
Arife Betül Yenice♡, Merve Doğan♡, Ceren Oksal♡, Gökhan Ercan♠, Olcay Taner Yıldız♢
Starlang Yazılım Danışmanlık♡, Işık University♠, Ozyegin University♢
Istanbul, Turkey
{bilge, asli, busra, deniz, ezgi, neslihanc, neslihank, oguzhan, merve}@starlangyazilim.com
gokhan.ercan@isik.edu.tr, olcay.yildiz@ozyegin.edu.tr

Abstract

MorphoLex is a study in which root, prefix and suffixes of words are analyzed. With MorphoLex, many words can be analyzed according to certain rules and a useful database can be created. Due to the fact that Turkish is an agglutinative language and the richness of its language structure, it offers different analyzes and results from previous studies in MorphoLex. In this study, we revealed the process of creating a database with 48,472 words and the results of the differences in language structure.

Keywords: MorphoLex, Turkish MorphoLex

1. Introduction

Turkish, which has many meaningful words, has a very rich content for Natural Language Processing. With DDI, texts, sounds and data in a language can be analyzed by a computer. For DDI, the structures of words are important as well as their meanings. Morphemes are formed from the meaningful root of a word. The word is divided into its suffixes and descended to the correct root that forms it. In polysemous languages such as Turkish, it is very difficult to find the root of the word. Examples of morpholex studies, which have not been studied much internationally, can be found in English and French. Although the structures of these languages are different from Turkish, the basic work is done in a similar way. After the root of a word is obtained, similar words derived from that word can be determined and even new words can be created.

Since Turkish is an agglutinative language, it always uses suffixes in word processing, unlike the languages studied in MorphoLex before. Words of Turkish origin do not have a prefix, but words of foreign origin can have a prefix. The structure of Turkish has made the analysis part of the MorphoLex study quite different from other languages. For this reason, it is important to understand Turkish structurally in order for the study to be understandable. In this way, the difference in the content of the study will be shown and it will be a pioneer in the studies to be carried out in agglutinative languages such as Turkish.Turkish MorphoLex

This paper is organized as follows: We first give a very brief review of Turkish in Section and discuss the relevant literature on MorphoLexes in Section. We explain how we generated the Turkish MorphoLex. The statistics and experimental results regarding this MorphoLex are given in Section. Lastly, we conclude in Section.

2. Literature Review

Currently, there are two morpholex studies in English and French. These are MorphoLex (Mailhot et al., 2018) and MorphoLex-FR (Mailhot et al., 2020). The English work, Morpholex, has a volume of 68,624 words formed by root words from the English Lexicon Project. It contains six new variants for affixes and three for roots. In the study, it was seen that root density and length, root family size, suffix family size and suffix frequency had a facilitating effect. Suffix length is important and the group in which an affix is included is also important in terms of separating other words. On the other hand, MorphoLex-FR (Mailhot et al., 2020) focused on approximately 70,000 words taken from English. Although the study in English is an important example for Morpholex-FR, the differences between languages also affected the content of the root distinction. In English, two different words can be combined to form a new word, adjectives can be used as verbs in sentences. In French, there are few cases of zero derivation, which relies on derivational processes. To reveal these typological differences, MorpholexFR based on 38,840 words of the French Dictionary Project is presented, using procedures similar to those used in English for segmentation and calculation of morphological variables.

The same inconsistencies were reached in both studies. Although the role of root frequency and the interaction of family size with word frequency are controversial for French, there is extensive evidence for the influence of root frequency on morphological processing in French. Meunier and Segui show that root-sum frequency modulates the effect of whole word frequency on the LD delays of suffixed words (Meunier and Segui, 1999). It is also claimed to modulate the effects of whole word frequency, root frequency and morphological root family size on LD delays, but this effect is only found for

suffixed words (Cole et al., 1989). There are many different methodological studies in Morpholex-FR, what is tried to be shown here is to make reliable comparisons between studies.

Studies have been carried out in the field of vocabulary for many years. It can be said that the studies and methods used in fields such as word recognition form an important basis in terms of linguistics and affect current studies. (Morton, 1969) logogen model is an important example for word recognition. Similar studies on the use of words have also been studied on a smaller scale for Turkish (Cetinkaya et al., 2016). (Bagriacik et al., 2019) and (İbrahim Delice, 2009) also did a Turkish study on affixes and prefixes.

Turkish, which belongs to a different language family, is structurally different from English and French. (Akbaba, 2007) work on verbs is important to see its difference from European languages. Although this difference has limited the similarity between the studies, basically the aim and the result are the same. English and French morpholex studies have been an important source for Turkish MorphoLex. The method applied with these sources has been transferred to Turkish, and a comprehensive morpholex study has been put forward.

3. Turkish MorphoLex

Turkish is an end-to-end language group regarding structure among world languages. It is quite easy to derive new words and terms in additive languages. The most common sentence structure is in the form of subject-object-verb. Transitional sentences are frequently used in daily life. Short narration in Turkish is in the foreground. It is one of the agglutinative languages. In Turkish, all inflectional changes are built on the roots, which remain unchanged. Suffixes follow this structure in specific rules. Derivational changes allow one to make dozens of new words from a single root. There are no prefixes (articles) and no grammatical gender in Turkish grammar. Therefore, there is no change in sentences due to gender differences. When word derivation and conjugation performed with the suffixes, no change occurs on roots. For example, there is a difference between the third-person possessive suffix -(s)I, which is added to nouns to indicate possessiveness, and the compound marker, CM, -(s)I, which is used to form lexicalized noun compounds by specifying their basic semantic and structural differences. (Aslan and Altan, 2006) The richness and diversity of the appendices are remarkable. Regarding the relevance of the elements that make up the sentence, sentences are set up as a natural hierarchy of completed thought, not in the order of developing thoughts.

KeNet (Bakay et al., 2021) is a Turkish Lexicon Project containing 77,330 synsets, 109,049 synset members and 80,956 distinct synset members KeNet has both in-tralingual semantic relations and is linked to PWN through interlingual relations. The fact that KeNet, which was used in the creation of Turkish morpholex, is rich in the number of nouns and verbs, has been a very important resource for the study. Before finding root, the words and their meanings were taken from KeNet.

The words are divided into meaningful units with the data received over KeNet and ordered based on the suffix of the word. According to (Goksel and Kerslake, 2005), almost all suffixes in Turkish have more than one form. The first consonant in some suffixes and the vowels in almost all suffixes depend on the consonant or vowel that precedes them. For example, the suffixes of the words optician and bookstore were considered. The root of the word gözlükçü (optician) is göz (eye), the second word derived from it is gözlük (glasses), and the third word is gözlükçü (optician). A similar derivation applies to the word kitapçı (bookstore). The word kitapçı (bookseller) derives from the word kitap (book).

After all the words were sorted and checked according to their meanings according to their suffixes, a second control stage was carried out. In this second stage, the words were sorted according to their roots, so that the group that a root belongs to and the words derived from this root is seen. In the second control phase, the meaning of the word was a major factor in determining the roots.

In Turkish, when determining the root of a word, taking the smallest semantically meaningful unit of that word as a basis does not produce an accurate result. For example, while the word ab (water) is a meaningful word on its own, it cannot be thought that the root of the word aba (a type of fabric) is ab (water). In Turkish, which is a very rich language, words can have more than one meaning. Therefore, reaching the root of the word by evaluating it semantically has revealed a healthier result.

When examining words in Turkish MorphoLex, it is seen that the ratio of suffixes is much higher than prefixes due to the structure of the language. In languages where prefixes are used frequently, when a prefix at the beginning of a word is considered, the ratio between prefixed and pseudo-prefixed words starting with the same spelling sequence is in favor of prefixed words. (Laudanna et al., 1994) Since Turkish is an agglutinative language, new words are generally not derived with prefixes. These few examples are mostly encountered in reinforced adjectives and examples of foreign origin. For example, the word çare (help) is prefixed and turns into the word biçare (wretched).

Turkish is an agglutinative language. The roots of the words do not change in Turkish, there are stems derived from these roots and construction and inflectional suffixes added to the root stems. Since Turkish is an agglutinative language, it always uses suffixes in word derivation. Originally, there is no prefix in Turkish. But, Turkish has been under the influence of foreign languages throughout its history. Firstly, Arabic and Persian and then French and English. There are also prefixed words among these words. This situa-

Word	Definition	Prefix	Root
anormal (abnormal)	Those who are against the general, customary and rule, abnormal - Those who have lost their minds	a	normal
anormalleşmek (become abnormal)	Become abnormal	a	normal
anormalleştirmek (abnormalize)	Make abnormal	a	normal
anormallik (abnormality)	State of being abnormal	a	normal

Table 1: Derivations of the word "normal" and its "prefixes".

Word	Definition	Prefix	Root
antialerjik (antiallergic)	Characteristics of drugs used in the prevention or treatment of allergies - Non allergic	anti	alerji
antiasit (antacid)	Contains alkali	anti	asit
antibakteriyel (antibacterial)	antibacterial	anti	bakteri

Table 2: Examples of "anti" prefix.

Word	Definition	Prefix	Root	Suffix
apacı (veri hot)	Very hot	ap	acı	
apaçık (obvious)	Very clear, very obvious	ap	Aç	yHk

Table 3: Examples of prefixes in Turkish intensive adjectives.

tion has led to the use of prefixed words in Turkish. Also, in the studies of finding correspondences to foreign words, while transforming the prefixed words into Turkish, compound words were formed. There compound words in Turkish were sometimes perceived as prefixed words.

Table 1 shows the word "normal" and its derivatives, along with their definitions, prefixes and roots. It comes from the French word abnormal. The French word is derived from the French word "normal" with the prefix an+. It is a suitable example of words taken

Word	Definition	Root1	Root2	Suffix
biyoekonomi (bioeconomics)	All economic activities related to research, development, production, trade and consumption of plants, animals and all other living things.	biyo	ekonomi	
biyoelektrik (bioelectricity)	Electricity produced by living things	biyo	elektrik	
biyoelektronik (bioelectronics)	The part of molecular biology that studies the electrostatic forces between the molecules that enter the structure of cells.	biyo	elektron	ik

Table 4: Examples of double-root words.

Word	Definition	Root1	Comb. Letter	Root2
adedimürettep	Fractional number - The number that is agreed upon for singles that make up a whole	adet	i	mürettep
esericedit	Large writing paper used in official correspondence	eser	i	cedit

Table 5: Examples of words of Arabic and Persian origin.

from the languages that Turkish is influenced by. It can take a prefix because it is a word of foreign origin.

Anti is also a prefix used in Turkish with words from other languages. It means "against" in Turkish too. Table 2 contains examples of words with the prefix "anti".

One of the prefix structure used in Turkish is prefixes that are used to derive intensive adjectives. Most of them are formed by ending the first syllable of the word with one of the P, R, M or S consonants. Table 3 shows

Word	Definition	Root1	Comb. Letter	Root2	Suffix
açıkgözlük (astuteness)	Taking advantage by being vigilant, taking advantage of opportunities shrewdly or behavior befitting this situation	aç	yHK	göz	lük
gerilimölçer (tensiometer)	Instrument for measuring stresses related to steam decomposition, surface, etc.	geril	Hm	ölç	Ar

Table 6: Examples of combinative letter.

Word	Definition	Root	Suffix1	Suffix2	Suffix3	Suffix4	Suffix5	Suffix6
akışkanlaştırıcılık	Having the property of making so mething fluid	Ak	Hş	GAn	lAş	DHr	HCH	lHk
ölümsüzleştirilme	to be immortalized	öl	yHm	sHz	lAş	DHr	Hl	mA
şekillendirilebilir	that can be put into a certain format	şekil	lAn	dHr	Hl	yAbil	Hr	

Table 7: Examples of suffixes.

examples of intensive adjectives.

Some foreign-origin words can be considered as double-rooted. As the example shows, ”bio” is not considered as a prefix. Instead, they are considered words consisting of a combination of two roots. In addition, although ”biyo” (bio) is not a root in Turkish, it has been accepted as a root in the study. This is due to the large number of words starting with ”biyo” (bio). Table 4 shows examples of double-rooted words.

This is also seen in words from Arabic and Persian (Table 5). However, there is a difference in these words. These words have combinative letters that combine two roots.

These combinative letters are also found in words of Turkish origin (Table 6). While the roots of words formed by the combination of two words are separated, the suffix of the first root is accepted as a combinative letter. It should be added that the combinative letters in these examples are actually suffixes. Certain roots in words of foreign origin are standard in Turkish. For example, the suffix -loji (logy) is frequently encountered in words of foreign origin. This is also important in terms of distinguishing word origins. Although there was no original logy root in Turkish, -loji (logy) was accepted as a suffix due to the excess of words of foreign origin.

Suffixes are mostly used in Turkish. These suffixes can derive a new noun from the noun, a verb from the noun, a verb from the verb, or a noun from the verb. The number of these suffixes is more than sixty.

The three words with the most suffixes in Turkish MorphoLex are shown in the Table 7. In the work, suffixes are separated according to the specific format shown in the example. For example, the first suffix in the ”ölümsüzleştirilme” example is taken as yHm, not -üm. These rules ensure a certain order between suffixes. This order is very important for the consistency

# of suffixes	# of # of suffixes
6	2
5	28
4	327
3	2,169
2	9,373
1	16,618
0	19,954

Table 8: Number of number of suffixes.

of the study. An annotator can easily understand what the main word is just by looking at the root and suffixes. Also, while deciding on the root, it is very important to check the meaning of the main word. In this way, the same root words with different meanings are easily separated from each other. And the annotater can easily understand what the root is. This significantly increases the accuracy of root words and the prefixes and suffixes they take.

4. Statistics

It is important to give some statistics to reveal the details of the study. For this, we extracted the statistics of different values such as the number of prefixes, the number of suffixes, the number of roots.

As can be seen in Table 8, almost 40% of the words in the study do not have suffixes. However, words without this suffix often form the root of other suffixed words. As was given before, the word “göz” (eye) has no suffixes. However, the root of the word “gözlük” (glasses) is “göz” (eye) and has one suffix (which is -lük), the root of the word “gözlükçü” (optician) is also “göz” (eye) and has two suffixes (which is -lük, -çü).

In the study, there are a total of 458 prefixed words. The most common prefixes, how many words these prefixes are in and examples of these words are shown in Table

Prefix	Number	Example
mü	55	mütedavül, mütehevvür
anti	42	antiserum, antitoksik
gayri	28	gayriresmi, gayrisafi
a	20	anormal, amoralist
na	19	namert, namüsait
bi	19	biseksüel, bizat, bibaht
re	19	reprodüksiyon, rekreasyon
poli	13	polietilen, poligami
oto	13	otomobil, otokontrol

Table 9: Most common prefixes.

# of Roots	# of # Roots
4	1
3	72
2	5,345
1	43,053

Table 10: Number of roots.

Total # of roots	# of Distinct Roots
53,963	19,115

Table 11: Number of total roots and distinct roots.

Root Form	# of Root Form
Baş	296
Et	246
Hane	183
Bil	157
Kara	127
Ol	117
Ot	114
Metre	101
Taş	99
Göz	98

Table 12: Most common root words.

9. Comparing the number of suffixes and prefixes, it can be seen that the number of prefixes is very minimal. Table 10 shows how many roots a word has. The vast majority of them are words with one root. And these one root words are divided into two among themselves. Some get at least one suffix, while others get no suffix. Tables 11 and 12 show the total root numbers, distinct root numbers and the most common root words. There are a total of 53802 roots, of which 19369 are different from each other. The fact that the most common root word is the root of 295 words reveals how rich a language Turkish is and that its meaning should be taken into account when finding a root word.

Tables 13, 14 and 15 show the total number of suffixes, the number of distinct suffixes, the number of most used suffixes and their description. It should be stated again that the number of suffixes used in Turkish is more than sixty.

# of suffixes	# of distinct suffixes
43263	286

Table 13: Number of suffixes.

Suffix	# of suffix
mAk	5,051
lHk	4,847
CH	3,384
lH	3,158
mA	2,266
sHz	2,200
lA	1,944
sH	1,836
lAş	1,535
CA	958
DHr	903
lAn	884
yHm	872
yHk	714
Hl	526
lAr	500
Hn	499
Ht	499
HcH	455
Hş	452

Table 14: Most common suffixes.

5. Conclusion

This study is about MorphoLex, which has not been studied in Turkish before. The study was based on the Turkish Dictionary Project KeNet (Bakay et al.) and the words used in the study were taken from KeNet. The fact that each word has its own meaning in KeNet has been very useful when creating the database. In Turkish, the meaning of the word is also very important when deciding what the root of a word is. Without the meaning of the word, annotater can never be sure of the correctness of a root. Therefore, the meaning of the word should be related to the main word and root and the analysis should be made accordingly.

When the literature is examined, it is seen that both the English MorphoLex and the French MorphoLex were created for basically the same purposes but using different methods. In Turkish MorphoLex, words are obtained from dictionary projects, just like in English and French versions. But unlike the other two studies, all analysis is done manually. Manual annotating has been an appropriate choice for a comprehensive language such as Turkish. The annotater evaluated word meanings with words and analyzed accordingly. In addition to the word meanings, the second annotating was also important in terms of ensuring accuracy. In the first annotating, the annotater started from the end of the prefix-root-suffix sequence and in the second annotat-

Suffix	Description
mAk	Form nouns: Ekmek 'bread', çakmak 'lighter'.
lHk	Nouns from nouns, adjectives or adverbs to indicate: Krallık 'kingship', sağırlık 'deafness'.
CH	A productive suffix: Güreşçi 'wrestler', palavracı 'liar'.
lH	A productive suffix: Atlı 'horseman', hızlı 'rapid'.
mA	Form nouns: Kıyma 'minted meat', inme 'paralysis'. Adjectives: Dökme 'of metal cast'.
sHz	Productive suffix added to nouns to form adjectives: Parasız 'peniless'. Nouns and pronouns to form adverbs denoting the non-involment in an event of whatever is: Arabasız 'without the car'.
lA	Attaches to nouns to designate a place associated with the concept in the root: Yayla 'plateau', tuzla 'salt mine'.
sH	Expresses approximation to particular quality. Added only to nouns to form adjectives: Kadınsı 'feminine'.
lAş	Added to adjectives of quality to form intransitive verbs that indicate the process of attaining that particular quality: Güzelleş- 'become beautiful'.
CA	A productive suffix which creates adjectives from nouns: Çocukça 'childish'. From the pluralized form of a round numeral: Binlerce 'thousands of'. / Creates nouns, adjectives or adverbs denoting a language from nouns of nationality: Japonca 'in Japanese'.
DHr	Indicates intensive or repetitive action: Araştır- 'investigate'.
lAn	Passive/reflexive, added to adjectives: Avlan- 'hunt'.
yHm	Forms nouns from underived verb roots: Bölüm 'department'.
yHk	Forms nouns: Konuk 'guest', kayık 'boat'.
Hl	Forms nouns: Okul 'school', kural 'rule'.
lAr	The plural suffix. Çocuklar 'children', kediler 'cats'.
Hn	Forms nouns: Basın 'press', yayın 'publication'.
Ht	Forms nouns: Geçit 'crossing', umut 'hope'.
HcH	A person practising a certain profession or having a certain occupation: Koruyucu 'guardian'. A tool, machine or substance performing a particular function: Yazıcı 'printer'.
Hş	Form nouns: Direniş 'resistance', giriş 'entrance'.

Table 15: Description of most common suffixes.

ing, the annotater followed the opposite path. The double control system has increased the accuracy of roots and prefix-suffixes.

At the end of the study, a database consisting of 48,472 roots emerged. It is seen that a very small part of these 48,472 roots have prefixes, most of them have suffixes and most of them are root only. As the statistics show, the fact that Turkish is a language rich in suffix has been one of the reasons that made the analysis work difficult. The study shows a result both showing that Turkish has a different structure when considering English and French studies, and the values that emerge when creating a database based on this different structure. Turkish, which is an agglutinative language, has quite a lot of suffixes compared to other languages. Statistics show the differences between languages and the effect of the differences on the prefix-root-suffix.

We believe that this database contains most of the Turkish roots and has been properly analysed. In this way, we think that automatic analysis can be done with MorphoLex and this database will be useful in modern technology.

6. Bibliographical References

Akbaba, D. E. (2007). Compound verbs with a noun-verb in structure in Turkish. *Journal of Language Studies*, 12.

Aslan, E. and Altan, A. (2006). The role of -(s)i in Turkish indefinite nominal compounds. *Journal of Language*, pages 57–75, 03.

Bagriacik, M., Goksel, A., and Ralli, A. (2019). Two Turkish suffixes in pharasiot. pages 116–147, 04.

Bakay, O., Ergelen, O., Sarmis, E., Yildirim, S., Kobalcioglu, A., Arican, B. N., Ozcelik, M., Saniyar, E., Kuyrukcu, O., Avar, B., and Yildiz, O. T. (2021). Turkish wordnet kenet. In *Proceedings of GWC 2021*, 01.

Cetinkaya, G., Ulper, H., and Bayat, N. (2016). Analysing errors reference to use of connectives. *Journal of Theoretical Educational Sciense*, pages 198–213, 04.

Cole, P., Beauvillain, C., and Segui, J. (1989). On the representation and processing of prefixed and suffixed derived words: A differential frequency effect. *Journal of Memory and Language*, pages 1–13, 01.

Goksel, A. and Kerslake, C. (2005). *Turkish: A Comprehensive Grammar*. Routledge, 1st edition.

Laudanna, A., Burani, C., and Cermele, A. (1994). Prefixes as processing units. *Language and Cognitive Processes*, pages 295–316, 01.

Mailhot, H., Sanchez-Gutiérrez, C. H., Deacon, S. H., Wilson, M. A., and Macoir, J. (2018). Morpholex: A derivational morphological database for 70,000 English words. 08.

Mailhot, H., Sanchez-Gutiérrez, C. H., Deacon, S. H., Wilson, M. A., and Macoir, J. (2020). Morpholex-fr: A derivational morphological database for 38,840 French words. 06.

Meunier, F. and Segui, J. (1999). Frequency effects in auditory word recognition: The case of suffixed words. *Journal of Memory and Language*, 01.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, page 165, 03.

İbrahim Delice, H. (2009). Prefixes in Turkish and structure which have prefixes. *International Periodical For the Languages, Literature and History of Turkish or Turkic*, 01.