Proceedings of the Globalex Workshop on Linked Lexicography @LREC2022, pages 75–84
Marseille, 20 June 2022

Time Travel in Turkish: WordNets for Modern Turkish

Ceren Oksal[♡], **Hikmet Nur Oğuz**[♡], **Mert Çatal**[♡], **Nurkay Erbay**[♡], **Aslı Duvarcı**[♡], **Özgecan Yüzer**[♡]
İpek Binnaz Ünsal[♡], **Oğuzhan Kuyrukçu**[♡], **Arife Betül Yenice**[♡], **Aslı Kuzgun**[♡], **Büşra Marşan**[♡]
Ezgi Sanıyar[♡], **Bilge Nas Arıcan**[♡], **Merve Doğan**[♡], **Özge Bakay**[♠], **Olcay Taner Yıldız**[◇]
Starlang Yazılım Danışmanlık[♡], University of Massachussetts Amherst[♠] Ozyegin University[◇]
Istanbul, Turkey
{ceren, hikmet, mert, nurkay, asli, ozgecan, ipek}@starlangyazilim.com
{oguzhan, arife, asli, busra, ezgi, bilge, merve}@starlangyazilim.com, obakay@umass.edu, olcay.yildiz@ozyegin.edu.tr

Abstract

Wordnets have been popular tools for providing and representing semantic and lexical relations of languages. They are useful tools for various purposes in NLP studies. Many researches created WordNets for different languages. For Turkish, there are two WordNets, namely the Turkish WordNet of BalkaNet and KeNet. In this paper, we present new WordNets for Turkish each of which is based on one of the first 9 editions of the Turkish dictionary starting from the 1944 edition. These WordNets are historical in nature and make implications for Modern Turkish. They are developed by extending KeNet, which was created based on the 2005 and 2011 editions of the Turkish dictionary. In this paper, we explain the steps in creating these 9 new WordNets for Turkish, discuss the challenges in the process and report comparative results about the WordNets.

Keywords: WordNet, Turkish, Modern Turkish

1. Introduction

Wordnets are large online lexical databases that are created for various machine related uses. WordNets include the lexical units and the relations that these units have between each other in a relational semantic network. Usually, they are created for general purposes by including as many words as possible, but they can be domain specific as well, such as WordNets specific for tourism, architecture etc. WordNets mostly contain open-class words like nouns, verbs, adjectives and adverbs. There can also be closed-class words such as prepositions, pronouns and conjunction. In WordNets, synsets are created by grouping the word senses with their synonyms. These synsets are representations of unique senses and they enable us to combine the relevant senses.

Linking synsets by making use of nodes provides the relational semantic networks in WordNets. Relations between the nodes in WordNets can be of two kinds; semantic or lexical. This means that WordNets are able to make both semantic and lexical information available. Because of this, WordNets have been a common tool in Natural Language Processing (NLP) studies. These tools can be used for machine translation, word sense disambiguation, information retrieval and sentiment analysis. Wordnets are incredibly useful for these fields since they provide data in an organized way and they are accessible. This also explains their popularity in recent years. The first development of the WordNet which is the Princeton WordNet (PWN) established at Princeton University was in English (Miller, 1995). Over the years, various WordNets have been created for different languages and new and improved versions have been released for existing ones. Moreover, thanks to multilingual WordNets, multiple languages have been linked to each other in multilingual WordNets.

Building a WordNet, or even extending an existing one, is a time-consuming process with multiple steps that requires both human and machine labor. In this paper, we offer a time travel journey on Modern Turkish by presenting a comparative analysis on 9 WordNets on Modern Turkish. For this study, we have taken the first 9 editions of the Turkish dictionary and created the WordNets for these editions (Türk Dil Kurumu Yayınları, 1944; Türk Dil Kurumu Yayınları, 1955; Türk Dil Kurumu Yayınları, 1959; Türk Dil Kurumu Yayınları, 1966; Türk Dil Kurumu Yayınları, 1969; Türk Dil Kurumu Yayınları, 1974; Türk Dil Kurumu Yayınları, 1983; Türk Dil Kurumu Yayınları, 1988; Türk Dil Kurumu Yayınları, 1998). We compared these new WordNets to the comprehensive WordNet KeNet (Bakay et al., 2021; Ehsani et al., 2018), which was created based on the last two editions of the Turkish dictionary. All of these WordNets are online, free and available for 7 different programming languages[1]. The outline of this paper is as follows. In Section 2, we present a literature review on WordNets for various languages, including those on Turkish. We give information about the structure of Turkish in Section 3. In Section 4, we describe and explain the steps that we have taken for the creation of our WordNets of Modern Turkish. In Section 5, we summarize the challenges and interest-

[1] https://github.com/StarlangSoftware/TurkishWordNet
https://github.com/StarlangSoftware/TurkishWordNet-Py
https://github.com/StarlangSoftware/TurkishWordNet-Cy
https://github.com/StarlangSoftware/TurkishWordNet-C#
https://github.com/StarlangSoftware/TurkishWordNet-CPP
https://github.com/StarlangSoftware/TurkishWordNet-Js
https://github.com/StarlangSoftware/TurkishWordNet-Swift

ing cases that we faced during this process. In Section 6, we present the statistical results from the WordNets and finally, in Section 7, we conclude with a discussion on the possible uses of our WordNets.

2. Literature Review

Research on WordNets was pioneered by G.A. Miller when he created the first WordNet, the Princeton WordNet (PWN) on English (Miller, 1995). After this, many other researchers started working on different WordNets for different languages. French WordNet WOLF (Sagot and Fiser, 2008), Arabic WordNet (AWN) (ElKateb et al., 2006), Polish Word-Net (Derwojedowa et al., 2008), Japanese WordNet (Isahara et al., 2008), Finnish WordNet FinnWord-Net (Linden and Carlson, 2010), NorwegianWord-Net (Fjeld and Nygaard, 2009) and Danish WordNet (Pedersen et al., 2009) are a few examples of these works. There are also projects that link WordNets of different languages to create a multilingual WordNet such as EuroWordNet (EWN) (Vossen, 2007), MultiWordNet (Pianta et al., 2002) and BalkaNet (Tufis et al., 2004). MultiWordNet includes Italian, Spanish, Portuguese, Hebrew, Romanian and Latin. BalkaNet consists of Bulgarian, Czech, Greek, Romanian, Serbian and Turkish.

Regarding Turkish WordNets, TR-WordNet of BalkaNet (Bilgin et al., 2004) is the first Turkish WordNet. It includes 14,626 synsets and 19,834 intralingual semantic relations. BalkaNet was constructed by automatically extracting synonyms, antonyms and hyponyms from a core set of lemmas that are common across different languages. The other WordNet for Turkish is KeNet which is the more recent and more comprehensive one (Bakay et al., 2021; Ehsani et al., 2018). Rather than starting from a core set like the BalkaNet, KeNet was created with a bottom-up approach. KeNet was prepared by starting with the whole set of lemmas in the two latest editions of the Turkish Dictionary, the 2005 and 2011 editions (Türk Dil Kurumu Yayınları, 2005; Türk Dil Kurumu Yayınları, 2011). It includes 77,110 synsets and has 107,839 intralingual semantic relationships such as hypernymy, meronymy and antonymy. It is also integrated to the Princeton WordNet through interlingual relationships (Bakay et al., 2019). In our study, we create 9 new WordNets for different historical versions of Modern Turkish by expanding KeNet. This study, to our knowledge, is the first to link WordNets that are based on earlier editions of the Turkish dictionary with KeNet, which is created based on later editions.

3. Turkish

In this chapter, we present a brief overview of Turkish in relation to our current work. Turkish has subject-object-verb (SOV) order and it is an agglutinative language (Göksel and Kerslake, 2005). Morphologically complex words have the "ROOT-SUFFIX1-SUFFIX2-. . ." structure.

Inflectional suffixes in Turkish mark grammatical features. They include those that mark the voice features of verbs such as active, passive, reciprocal and causative. For example, passive voice is formed by attaching the -Il or –(I)n suffixes to verbs. While *açmak* is the active form of the verb meaning "to open", its passive form is *aç*<b>ıl</b>*mak*, where "-mAk" is the infinitive suffix. Causative voice has four different morphemes: -DIr, -(I)t, -(I)r, -Ar. An example is *gül*<b>dür</b>*mek* "to make somebody laugh". Derivational suffixes, on the other hand, can change the meaning as well as the grammatical category of words. For instance, the suffix -CI forms nouns from nouns. An example is *av*-**cı** "hunter" in which *av* means "prey". Another exemplary suffix that changes the category of the word is -sIz, which forms adjectives out of nouns. An example is *anlam*-**sız** "meaningless" in which *anlam* means "meaning".

Spelling rules in Turkish have changed over the years. Some of these changes are a result of the attempts to adapt the phonological structure of borrowed words to that of Turkish. Turkish does not allow consonant clusters at the beginning of words (Göksel and Kerslake, 2005). For example, for borrowed words such as *plan* or *tren*, [i] is inserted in between the first two consonants during articulation, and this inserted vowel was sometimes included in the spelling of these words in Turkish. Another example is circumflex ˆ. It is used in borrowed words from Arabic and Persian with [a] and [u] that occurred after [k] and [g], e.g., *hâl*. It is also used to indicate longer vowels as in *âdet*. However, the use of the circumflex was abandoned from time to time; although it is now used in Turkish orthography, it is not commonly used by Turkish speakers anymore. Turkish Dictionaries reflects these spelling changes.

4. Steps in Creating Turkish Wordnets

There are currently 11 editions of the Turkish dictionary that were written in the Latin alphabet; the 1944, 1955, 1959, 1966, 1969, 1974, 1983, 1988, 1998, 2005 and 2011 editions[2]. All of these dictionaries are prepared by the Turkish Language Association. 1944 edition is composed of around 15.000 entries and this number increases with each subsequent dictionary. In this study, we present 9 new WordNets that we have created for the first 9 editions. None of these editions are available digitally. These new WordNets were created based on KeNet, which was created with the 2005 and 2011 editions. Thus, we provide a complete picture for the comparative analysis on different editions of the Turkish dictionary.

For the annotation of the first dictionary, i.e., the 1944 edition, an Excel sheet with the entries in KeNet was prepared. This excel sheet had 7 columns. The first column was named "R" and this column was used to indicate whether or not an entry in KeNet occurred in the 1944 edition of the dictionary. We wrote "1" for the words that were in the dictionary, "0" for the ones

[2] www.tdk.gov.tr

R	WORD	ID	POS	DEFINITION	SYNSET	EXAMPLE SENTENCE
1	abaküs	TUR10-0670670	NOUN	Basit sayma ve hesap işleri yapmakta kullanılan, her teline onar boncuk geçirilmiş hesap aracı	abaküs sayı boncuğu çörkü	
1	abanmak	TUR10-0000380	VERB	Birine yük olarak onun sırtından geçinmeye bakmak	abanmak	Ekonomik ihtiyaçları için tamamen annesine abandı.
1	abramak	TUR01-0100170	VERB	Yönetmek; idare etmek	abramak	
1	cerrar	TUR07-0300100	ADJECTIVE	Zorla para alan	cerrar	
1	cahilane	TUR07-0300020	ADJECTIVE	Cahilce, cahile yakışır	cahilane	
1	cahilane	TUR10-0130440	ADVERB	Öğrenim görmemiş veya bir konuda bilgisi olmayan kimseye yakışır biçimde	cahilane	
1	firari	TUR01-0701240	NOUN	NO DEFINITION	firari	

Table 1: Example seven entries in the annotation sheet

that were not. Other columns included words, IDs of the senses, definitions, synset members and exemplary sentences. Each letter had its own sheet where entries were alphabetically ordered. The letters were distributed among 13 linguistically-informed annotators. Annotators went through the dictionary and the Excel sheet. Additionally, we created a list of the words that were included in KeNet but not in the dictionary in the same format. This list was created to check the differences between the different editions. These steps were followed for each edition of the dictionary. For later editions, the final version of the Excel sheet, i.e., the one for the previous version of the dictionary, was used. This was because we expected less changes to occur between consequent editions. Table 1 presents an example of seven entries from an Excel sheet for one of the dictionaries.

Checking each word in the dictionaries took the longest time. We also checked whether the definitions and the POS tags in the dictionaries were the same as those in the Excel sheets. If the POS tags were different, we put a new ID for that word. We marked the words that were present in the Excel sheet as well as the dictionary as “1”, and the rest as “0”. If a word in the dictionary was absent in the Excel sheet, we checked the list of words from KeNet that were left out in the earlier edition(s). If the entry occurred in that list, we added it to our original excel sheet; if not, we highlighted it to add later. This was because our priority was to use definitions from KeNet. We did not change the definitons based on those in the dictionaries as it would unnecessarily complicate the process.

KeNet had meaning IDs that started with “TUR10-“. We kept the IDs the same unless there was a new meaning in the dictionary that we added. For these new meanings, an ID that corresponded to the different editions of the dictionaries were created. For example, if a new word was added to the first version, it has the ID starting with “TUR01-01…”. The first ”01” indicates the edition number and the second “01” the first letter of the added word. That is, if a new word starting with the fifth letter “d” was added in the third edition, the ID started with “TUR03-05”.

Next, we took only the entries marked with “1” and sorted them alphabetically. If there were accidental additions of the same rows, they were deleted. Once we had the full list, we created the new version of the WordNet. At this stage, the words with the same IDs are combined. To further check our WordNet, we got a new list with potential mistakes. For example, if two meanings had the same IDs but different definitions or POS tags, we corrected them. Or, if there were words with different IDs but the same definitions, we made sure that they had the same ID. After this, we combined the synsets.

When we completed the first edition, in later ones we made sure to compare and check the new version with the previous ones. In this stage, we compared the meanings and listed the versions that had more than 80% of its words matching with each other. We went through this list to see if there were cases where we could match IDs. This also helped us find cases where, for example, a meaning in the 1944 version was lost in 1955 but was found again in a later edition. There were only a few cases of this type for each dictionary. After this stage, the new version of the WordNet was completed. Only in this new version, we were able to get the statistics of the data such as how many of each POS tags or how many examples of usage there were.

To prepare examples of usage, we made use of the previous versions of the WordNets. We pasted the sentences of previous synsets onto the new words that we added. These sentences were taken from previous versions of WordNets as close in time as possible for historical considerations. However, we still needed to check them for mistakes. We first morphologically analyzed them. Then, we deleted the words that did not appear in the relevant dictionaries. After this step, we had words that were compatible with their dictionaries in terms of spelling rules. If there was a new meaning added for a word, we compared this new meaning with those in the other versions to see if the new meaning was actually distinct from the others. Overall, the number of these kinds of mistakes was around 100, which is very small in comparison to our comprehensive WordNets. Lastly, we morphologically analyzed the words in the definitions to correct any mistakes.

In our process to create new versions of our WordNets, we also matched the meanings of synonymous words in the dictionaries. Moreover, for the 1974 edition, we checked the examples of usage of words with more than one meaning. We did this to make sure that the sentences exemplified the correct meanings. Finally, we had examples of usage for the synsets. However, since these sentences were automatically pasted, as it was stated previously, we had to adjust them for each word in the synset. If there were words that did not appear in the dictionaries, we did not paste those sentences.

5. Challenges and Interesting Cases

During the creation of our WordNets, we encountered many interesting cases and challenges. These include some issues with how the dictionaries were constructed and some cases pertaining to the historical conditions in the time of the editions. We had to overcome these challenges to make sure that our WordNets were consistent but also accurately reflected these dictionaries.

First of all, in all of the dictionaries we made use of the multiple entries of verbs with passive and causative voice such as *yapılmak* “to be made” and *yaptırmak* “to have it made” . Following (Bakay et al., 2021), the passive and causative forms of verbs were excluded from our WordNets.

In some cases, dictionaries had the noun versions of verbs such as *cay-ma* “act of giving up” and *caymak* “to give up”. The definitions of these noun versions were always given as *caymak eylemi* ”act of giving up”. (Böler, 2006) reports that there are multiple entries of this type in the dictionary, but these noun versions have not gained different meanings from their verb meanings. These cases of the Turkish Dictionary have also been noted as problematic by (Uzun, 2003). Thus, we only entered verbs and excluded their noun forms.

Additionally, we did not include parentheses in the definitions. We deleted the phrase inside the parentheses when it conflicted the POS of the entry. In the fourth example in Table 1, the definition lacks the word in parenthesis that was present in the original dictionary. The original entry was “Zorla para alan (kimse)/(Someone) who takes money by force”. However, keeping “kimse” would cause the definition to be that of a noun whereas deleting it makes it the definition of an adjective.

There were also a couple of cases with mistakes regarding the POS of an entry which we corrected in our WordNet. For example, the word *fırlatmak* “to throw” was categorized as noun in the 1944 version but we coded it as verb.

One of the most frequent problems we faced was that dictionaries lacked the entries of some words that were given as synonyms or used in the definitions of other entries. Even though there were many meanings with only one single word explanations, those meanings did not have their corresponding entries. For example, the entry for *ıstırap* “anguish” had the meaning “acıştırmak” in the 1944 dictionary which did not have its own entry in the same dictionary. Same was also true for some synonyms given in the definitions of some words. This meant that we were not able to group such words into our synsets. Moreover, for some words, dictionaries would not define the word itself, but rather only mention the idiom that it is used in as the definition of those words. Such words seemed to not have a meaning on their own, rather they were a part of the phrase. One such entry is given below:

küldür: Paldır küldür deyiminde geçer.
mell: It is used in the phrase “pell-mell”.

With regards to the POS tags of words, there were a lot of differences between especially the older editions and KeNet. POS tags of profession words were one prominent example. While words denoting professions with the derivational “-cI” suffix -today, this suffix derives nouns from nouns- were given as adjectives in the 1944 edition, in later versions and KeNet they were tagged as nouns. For example, *gazeteci* “journalist” and *gemici* “sailor” were categorized as adjectives in 1944 but as nouns in other versions. Another interesting example is that some words were given with two POS tags. In 1983 dictionary, the word *cahilane* “ignorant” is both an adjective and an adverb, which is reflected in the definition as well.

cahilane: Cahilce, cahile yakışır (biçimde)
ignorant: Ignorantly, befittingly of an ignorant.

In the definition above, *biçimde* “befittingly” is given inside a parenthesis because it gives the meaning of the adverb, without it the definition describes an adjective. For such cases, if KeNet included only one version but not the other, we added it with a new ID. For example, KeNet only had the adverb version, so we added the adjective version of *cahilane*. Here, one thing we made sure was that the definition would correspond to the adjective meaning of the word. This meant that we excluded the word in parenthesis above.

There also were discrepancies between the POS tags of synonyms within the dictionaries. For instance, in the 1944 edition the entry for *firari* “escapee” has the synonym *kaçak*, yet the first one is categorized as a noun whereas the latter as an adjective. In such cases, we tagged them with the appropriate POS tag and wrote “NO DEFINITION”.

One interesting thing to note was that the dictionaries were quite influenced by the political tendencies and other sociological factors of their time. For example, in the 1944 edition, a very long and detailed definition of *Güneş - Dil teorisi* “Sun Language Theory” is given. This theory suggests that all languages originated from the so-called proto-Turkish, the first language that humans ever spoke. This can be correlated with the nationalistic ideas that were popular at that time. Other examples include *şeriatçı* “follower of sharia”, *kürt* “kurd”, *şapka* “hat” where their definitions might be reflecting the political discourse of the time. One other intriguing example is the idiom *kızını dövmeyen dizini döver* “spare the rod and spoil the child”. If it is translated literally, this idiom says “someone who does not beat their daughter beats their knees”. However, in older versions, this idiom is given as *evladını dövmeyen dizini döver* which means, translated literally, “someone who does not beat their child beats their knees”. Throughout the years, the idiom seems to have changed and gained a more “sexist” meaning.

Within the definitions, especially in the 1944 edition, there were multiple examples of the relative clauses with the complementizer “ki”, which is borrowed from Persian. However, ”ki” is not a very common way of relativization among Turkish speakers today. Also, looking through the dictionaries, the effects that other languages had on Turkish and the efforts to find Turkish counterparts for foreign words can be seen as well. All the dictionaries that we used were prepared after the Turkish Language Association was established. This association was expected to clear the “yoke of the foreign tongues” (Tachau, 1964). One clear example is words borrowed from Arabic. To introduce the Turkish counterparts of these words, sometimes both the Arabic version and the Turkish version of a word are given. In addition, borrowed words from Arabic that were plural were given with their plural meanings such as *dost-lar* ”budd-ies”, the plural form of *dost*, for *ahibba*. Also, in older versions of the Turkish dictionary, there were a lot of cases of “-î” which is the nisba suffix borrowed from Arabic. This letter was later changed to “-i”.

Regarding spelling, foreign words are spelled in accordance with the phonological structure of Turkish. For example, both “Fransızca” and “Fıransızca” is present in the dictionary where in the latter “ı” is inserted in between two consonants as Turkish does not allow initial consonant clusters. Another ortographic case was that of the suffix *-ile* “with”. In older versions, this suffix did not undergo vowel harmony when attached to stems with the third person possesive suffix as in *araba-s-iyle* ”car-3SG.POSS-with” whereas in other forms it does as in *araba-yla* ”car-with”. Today, both in spelling and articulation the suffix *-ile* always undergoes vowel harmony. Similarly, in older editions, vowels before suffixes that start with the “y” consonant were spelled as close vowels, “ı/i”;, as in *olm***ı***yan,* or *gösterilm***i***yen*. However, today these vowels are not necessarily spelled as close vowels as in *olm***a***yan* or *gösterilm***e***yen*. All these cases in addition to others are presented in the spelling dictionaries of the related years and an overview of those can be found in (Demirtürk, 2019).

6. Results

In this section, we show and explain the various statistical results that we got from these WordNets and their comparison with KeNet. These statistics can show the changes through the editions in different years while highlighting some interesting cases.

6.1. Synsets

First of all, Table 2 shows the total number of synsets for each WordNet. In this and the following tables, we refer to KeNet as the WordNet of 2020 since it was created in this year based on two different editions. It is not surprising that there has been an increase in the number of synsets over the years.

WordNet	# of Synsets
1944	31,762
1955	34,438
1959	35,802
1966	36,353
1969	37,327
1974	42,876
1983	55,161
1988	57,902
1998	67,347
2020	78,311

Table 2: Number of synsets in each WordNet

However, it should be noted that there are differences between the growth rates of two consequent years. The least amount of increase occurred between the 1959 and 1966 WordNets by 1.5%. This is surprising because the two dictionaries that these WordNets are based on have 7 years apart which is not the least number of years between any two consequent WordNets. For example, 1966 and 1969 WordNets are the closest to each other since they have only 3 years apart, but there is a 2,7% increase in the number of synsets, which is still a bit more than the 1959 and 1966 editions. The

WordNet	Literals	Distinct Literals	% Increase
1944	41,855	31,427	
1955	44,813	34,220	%9
1959	46,591	35,670	%4
1966	47,103	36,005	%1
1969	47,439	36,051	%0
1974	54,798	41,610	%15
1983	72,456	51,684	%24
1988	75,786	53,957	%4
1998	87,550	63,053	%17
2020	110,236	82,135	%30

Table 3: Number of literals in each WordNet

Year	Number of Words			
	1	2	3	4
1944	24,466	5,831	814	248
1955	24,502	7,528	1,317	518
1959	25,552	7,816	1,376	577
1966	25,628	8,016	1,401	601
1969	25,729	7,952	1,411	599
1974	28,557	10,272	1,735	658
1983	33,667	14,599	2,218	765
1988	33,643	16,454	2,463	853
1998	37,048	21,647	2,806	939
2020	48,704	28,417	3,556	910

Table 4: Number of words in literals in each WordNet

largest two increases are between the 1969 and 1974, and the 1974 and 1983 editions. The increase for these two comparisons are 14,8% and 28,6%, respectively.

6.2. Literals

Secondly, we have the results from the total number of literals in the WordNets, as given in Table 3. These numbers include all the different definitions that a word has. That is, if a word has 10 meanings, all of them are included in the number of literal. If we divide these numbers by the total numbers of synsets of the related WordNets, we get the average number of literals in a synset. This operation gives similar results for each WordNet, which is somewhere around 1.3 - 1.4. Table 3 also shows the number of distinct literals in each WordNet. Here, the numbers do not include the different definitions a word has; rather, even if a word has 10 meanings, the number of its distinct literals is 1. With respect to the distinct literals, while there is little change between the years of 1944, 1955, 1959, 1966 and 1969, there are larger increases in the following years, 1974, 1998 and 2020.

Table 4 shows the number of literals containing 1, 2, 3 and 4 words. There are literals containing up to 11 words. It is expected that 1-word literals are the most common ones in all WordNets and as the number of words goes up, the number of literals goes down. Literals with 2 and more words are usually idioms. 2-word literals are the second most common ones and they may also contain compound words since they were sometimes written as two separate words or sometimes as one single word. There is very little increase in the number of 1-word literals up until the 1974 WordNet. Between 1969 and 1974, there is an 11% increase and between 1974 and 1983 the increase is 18%. It seems that these more recent WordNets have more increase in the overall results. There is a 30% increase between the 1998 and 2020 WordNets, which could have been higher since there is a large gap in years between the two dictionaries. With respect to 2-word literals, there seems to be a high rise in their number between the first two WordNets. However, in the following ones until 1969, there is not a notable change in numbers. There is even a slight drop in the 1969 WordNet. While in 1969 the number of 2-word literals was 7952, in 1974 WordNet this number increased to 10,272 with a 29% increase. Moreover, there is even a larger change in the following WordNets; an increase of 42% between the 1974 and 1983 WordNets, and a 33% increase from the year of 1998 to 2020. It seems that the number of 3 and 4-word literals grew substantially between the first two WordNets, the 1944 and 1955 ones. Especially the 4-word ones increased almost by 100%. There are also increases larger than 10% between the WordNets of 1955 and 1959, those of 1969 and 1974, and those of 1974 and 1983. As it was stated before, these 3 - 4 or more worded literals are comprised of idioms. This means that especially after the 1944 dictionary idioms were more commonly entered into the dictionaries. However, KeNet has less 4-word literals than the previous WordNet, 1983, which shows that the recent dictionaries may not include longer idioms as much as the 1983 version, but still a close number to the WordNets of the years between 1959 and 1974. However, the total number of the 3- and 4-word literals within each WordNet seem to increase by each WordNet although the ratios between these two groups of literals may vary. Since these literals are usually idiom entries, it shows that by each new dictionary the number of idioms has increased.

6.3. Part of Speech Tags

When it comes to the POS tags, it can be clearly seen from Table 5 that NOUN, VERB and ADJECTIVE tags were the three most common ones in all the WordNets. Within these three categories, there was not much difference between the numbers up until the 1974 WordNet. To exemplify, in the NOUN tags, from the 1969 WordNet to the 1974 one, there has been a 13% increase and from 1974 to 1983 there is an increase of 27%. The VERB and ADJECTIVE tags in these same WordNets have similar percentages of increase; 22% in VERBS and 29% in ADJECTIVES in the former, 11% in VERBS and 35% in ADJECTIVES in the latter. Similar changes occur with the ADVERB tag; it seems to stay close in number until the 1974 when it

Pos	1944	1955	1959	1966	1969	1974	1983	1988	1998	2020
NOUN	17,022	18,224	19,017	19,256	20,013	22,700	28,794	30,110	36,151	43,869
VERB	7,359	7,993	8,157	8,291	8,583	10,469	13,526	14,188	15,947	17,772
ADJECTIVE	5,729	5,800	6,051	6,163	6,123	6,787	9,194	9,696	10,835	12,410
ADVERB	. 978	1,072	1,147	1,165	1,145	1,390	1,864	1,952	2,349	2,549
INTERJECTION	526	1,174	1,244	1,287	1,263	1,322	1,576	1,751	1,848	1,552
CONJUNCTION	74	83	69	70	70	72	81	79	79.	61
PRONOUN.	36	43	46	50	59	62	66	68	77	68
PREPOSITION	. 38	49	71	71	71	74	60	58	61	30

Table 5: Part of Speech Distribution in the Synsets

increases by 23% and in the next WordNet by 32%. The number of INTERJECTION tagged words have different results. While there were 526 words of this tag in the 1944 WordNet, it increased to 1,174 in 1955, which is an increase larger than 100%. While almost all other tags seem to gradually increase over the years, there is a decrease for the INTERJECTION tag in the 2020 KeNet, which is 15% less than the WordNet preceding it. Lastly, there are also some differences in the numbers of CONJUNTION, PREPOSITION and PRONOUN tags, but these are small differences compared to those in the other tags. These three tags also occur the least in all the WordNets. These are expected especially because PREPOSITION and PRONOUN tags are the ones with functional words.

6.4. Examples of usage

Table 6 shows the number of synsets with examples of usage. However, in interpreting this table, it is important to refer back to the steps that we took in the creation of WordNets. In each WordNet, while we put sentences from the respective dictionaries into our WordNets, we also added the examples of usage from 2020 into the relevant synsets. However, we kept only the sentences containing words that appeared in the relevant dictionary to make sure that the WordNets were historically accurate. Thus, these numbers include sentences from the dictionaries and the 2020 WordNet KeNet.

WordNet	# of SynSets
1944	10,505
1955	11,750
1959	11,859
1966	11,958
1969	12,528
1974	14,239
1983	19,095
1988	19,806
1998	21,942
2020	23,626

Table 6: Number of synsets with examples of usage

The number of examples of usage per synset is similar across the WordNets. The ratio in each WordNet ranges from 0.28 to 0.34, with an average of 0.33. So, there was not much of a change in this respect. However, when we compare the number of examples of usage between consequent WordNets, we get different results. Between the years of 1944 and 1955, the number of examples of usage per synset increased by 11%. After this increase, there does not seem to be much of a change in the WordNets of 1959, 1966 and 1969. The increase is 12% between 1969 and 1974, and 17% between 1974 and 1983. A 10% increase is observed between 1988 - 1998 and 1998 - 2020.

6.5. New Synsets

As it was stated previously, while preparing the WordNets we took the 2020 WordNet KeNet as our basis. When we encountered new words in the dictionaries that did not exist in KeNet, we added them with new IDs. The IDs with "TUR10" belonged to the words in KeNet whereas the IDs with "TUR01/02/03/04/05/06/07/08/09" belong to the new words that are added to the WordNets for the years from 1944 to 1998. Table 7 shows the number of words that are added in each WordNet.

Since we started creating the WordNets in chronological order, there are empty slots for each WordNet except the last one. It is also clear for each new WordNet that the number of synsets that were included from the previous WordNets decreases. This is because some of the meanings that were included in earlier editions are not included in later versions as they are not used by Turkish speakers anymore. For example, although with the 1944 dictionary we added 4605 new meanings, only 3658 of them occurred in the 1955 WordNet and in the following WordNets this number kept decreasing. The largest decrease occurred in the 1969 and 1983 WordNets. This may be due to the deletion of old meanings or the adding of more entries from KeNet instead of the old ones. In other words, as we come closer in time to the 2020 version, the differences between the earlier and later editions become less.

In the first four WordNets, namely those for 1944, 1955, 1959 and 1966, the addition of new meanings declined from 4,394 to 186. This is not surprising given the possibility that a meaning in the 1955 dictionary

	1944	1955	1959	1966	1969	1974	1983	1988	1998
TUR01	4,394	3,455	3,347	3,304	2,693	2,483	1,701	1,548	1,392
TUR02	-	1,483	1,416	1,383	1,151	1,068	733	686	623
TUR03	-	-	450	431	348	326	214	200	193
TUR04	-	-	-	186	154	146	114	100	94
TUR05	-	-	-	-	742	652	512	479	442
TUR06	-	-	-	-	-	1,055	676	601	556
TUR07	-	-	-	-	-	-	1,505	1,288	1,192
TUR08	-	-	-	-	-	-	-	567	518
TUR09	-	-	-	-	-	-	-	-	1,214
TUR10	27,363	29,498	30,588	31,048	32,239	37,146	49,706	52,433	61,123

Table 7: Distribution of new and old synsets in each WordNet

may have been already added during the annotation of the 1944 WordNet. However, there is an increase in the number of new synsets between the 1966 and 1969 WordNets. This is interesting given that there are only three years in between these two dictionaries. Moreover, the 1974 and 1983 WordNets also added more new meanings than their previous years. One thing that stayed the same through the WordNets is the increase in the number of new definitions from KeNet. Here, again, the largest increase occurred in the 1974 and 1983 WordNets.

7. Discussion

In this section, we summarize the process that we followed in the creation of our WordNets for Modern Turkish based on dictionaries from different years and discuss the potential uses of those WordNets. Our study overall showed that what has been done with KeNet can be extended to the new WordNets as well. Since our WordNets represent various times in the history of Modern Turkish, they have the potential to exhibit interesting historical facts about it.

In this paper, we presented our WordNets for Modern Turkish that we created with the first 9 editions of the Turkish Dictionary. These new historical WordNets were prepared by making use of the comprehensive Turkish WordNet KeNet (Bakay et al., 2021; Ehsani et al., 2018). Throughout the creation process of our WordNets, we tried to eliminate mistakes as much as possible. Also, we tried to make sure that our WordNets reflected the relevant dictionaries. To do this, we, for example, used the same spelling rules as those used in the dictionaries. We also added examples of usage for some literals from both the 2020 WordNet KeNet and the dictionaries themselves. We also reported statistical results from these WordNets. These statistics showed some changes between the WordNets in terms of the number of synsets, literals, distinct literals, number of words in the literals, POS tags, synsets with examples of usage and the number of new added words to the WordNets. Overall, these comparisons revealed that WordNets that are based on later editions are comprehensive than those that are created with earlier editions, as predicted.

In the process, we also faced some challenges. Some of them were related to the problems in the dictionaries. For example, some words that were used in the definitions or examples were not present in the dictionaries as separate entries, there were more than one POS tag for a single entry or different morphological forms of the same word were included. Other challenges were expected given the changes in the language over time. Those challenges were mostly related to the changes in orthographic rules for Turkish or the policies in using borrowed words or forms in Turkish.

Lastly, these WordNets can be used in future studies. Previously, multiple different studies and projects have been done bu using KeNet. For example, in 2020 a new version of Turkish PropBank, TRopBank, has been created (Kara et al., 2020). A PropBank (Bonial et al., 2014; Kingsbury and Palmer, 2002; Kingsbury and Palmer, 2003; Palmer et al., 2005) brings syntax and semantics together by annotating the argument structures of predicates. With TRopBank this annotation process in which both the arguments and adjuncts of verbs were included was completed on Turkish. This semantic resource can be extended to our new WordNets as well, which could be useful for future works on the historical analyses of Modern Turkish. Another semantic resource that our WordNets could be useful for is FrameNet (Baker et al., 1998; Fillmore and Atkins, 1998; Johnson et al., 2001; Lowe, 1997). This is another tool for coding semantic information of predicates. A FrameNet for Turkish has been done previously by (Marşan et al., 2021). This work can be extended to our WordNets. Moreover, these WordNets may enable us to conduct new studies on the Turkish language that investigate the historical change of the language. Overall, such studies and projects could help to demonstrate different semantic and typological features and interesting historical facts about Modern Turkish.

8. Bibliographical References

Bakay, Ö., Ergelen, Ö., and Yıldız, O. T. (2019). Integrating Turkish WordNet KeNet to Princeton Word-

Net: The case of one-to-many correspondences. In *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 1–5.

Bakay, Ö., Ergelen, Ö., Sarmış, E., Yıldırım, S., Arıcan, B. N., Kocabalcıoğlu, A., Özçelik, M., Sanıyar, E., Kuyrukçu, O., Avar, B., and Yıldız, O. T. (2021). Turkish WordNet KeNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 166–174, University of South Africa (UNISA), January. Global Wordnet Association.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 86–90, Montreal, Quebec, Canada, August. Association for Computational Linguistics.

Bilgin, O., Cetinoglu, O., and Oflazer, K. (2004). Building a wordnet for Turkish. *Romanian Journal of Information Science*, 7:163–172.

Böler, T. (2006). Türkçe sözlük (tdk) ile örnekleriyle Türkçe sözlük'ü (meb) karşılaştırma denemesi. *Sosyal Bilimler Araştırmaları Dergisi*, 1:101–118.

Bonial, C., Bonn, J., Conger, K., Hwang, J. D., and Palmer, M. (2014). PropBank: Semantics of new predicate types. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3013–3019, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Demirtürk, C. (2019). Türkçe yazım kılavuzlarının gelişimi Üzerine bir İnceleme.

Derwojedowa, M., Piasecki, M., Szpakowicz, S., Zawisławska, M., and Broda, B. (2008). Words, Concepts and Relations in the Construction of Polish WordNet. In *Proceedings of GWC 2008*, pages 162–177.

Ehsani, R., Solak, E., and Yildiz, O. (2018). Constructing a WordNet for Turkish Using Manual and Automatic Annotation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(3).

ElKateb, S., Black, W., Rodríguez, H., Alkhalifa, M., Vossen, P., Pease, A., and Fellbaum, C. D. (2006). Building a wordnet for Arabic. In *LREC*.

Fillmore, C. J. and Atkins, B. T. (1998). Framenet and lexicographic reference. In *First International Conference on language resources & evaluation: Granada, Spain, 28-30 May 1998*, pages 417–426. European Language Resources Association.

Fjeld, R. V. and Nygaard, L. (2009). Nornet - a monolingual wordnet of modern Norwegian. In *NODALIDA 2009 workshop: WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies*, pages 13–16.

Göksel, A. and Kerslake, C. (2005). *Turkish: A Comprehensive Grammar*. Routledge, New York, USA.

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., and Kanzaki, K. (2008). Development of the japanese wordnet. 01.

Johnson, C. R., Fillmore, C. J., Wood, B. M., Ruppenhofer, J., Urban, M., Petruck, M. R. L., and Baker, C. F. (2001). The framenet project: tools for lexicon building.

Kara, N., Aslan, D. B., Marşan, B., Bakay, Ö., Ak, K., and Yıldız, O. T. (2020). TRopBank: Turkish PropBank v2.0. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2763–2772, Marseille, France, May. European Language Resources Association.

Kingsbury, P. and Palmer, M. (2002). From treebank to propbank. In *LREC*. European Language Resources Association.

Kingsbury, P. and Palmer, M. (2003). Propbank: The next level of treebank. In *Proceedings of Treebanks and Lexical Theories*, Växjö, Sweden.

Linden, K. and Carlson, L. (2010). Construction of a FinnWordNet. *Nordic Journal of Lexicography*, 17:119 – 140.

Lowe, J. B. (1997). A frame-semantic approach to semantic annotation.

Marşan, B., Kara, N., Özçelik, M., Arıcan, B. N., Cesur, N., Kuzgun, A., Sanıyar, E., Kuyrukçu, O., and Yildiz, O. T. (2021). Building the Turkish FrameNet. In *Proceedings of the 11th Global Wordnet Conference*, pages 118–125, University of South Africa (UNISA), January. Global Wordnet Association.

Miller, G. A. (1995). Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, nov.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. *Comput. Linguist.*, 31(1):71–106, March.

Pedersen, B. S., Nimb, S., Asmussen, J., Sørensen, N. H., Trap-Jensen, L., and Lorentzen, H. (2009). Dannet: the challenge of compiling a wordnet for Danish by reusing a monolingual dictionary. *Language Resources and Evaluation*, 43:269–299.

Pianta, E., Bentivogli, L., and Girardi, C. (2002). Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January.

Sagot, B. and Fiser, D. (2008). Building a free French wordnet from multilingual resources. 05.

Tachau, F. (1964). Language and politics: Turkish language reform. *The Review of Politics*, 26(2):191–204.

Tufis, D., Cristeau, D., and Stamou, S. (2004). Balkanet: Aims, methods, results and perspectives – a general overview. *Romanian Journal of Information Science and Technology Special Issue*, 7:9–43, 01.

Türk Dil Kurumu Yayınları. (1944). *Türkçe Sözlük (1st ed.)*. Türk Dil Kurumu, Ankara, Turkey.

Türk Dil Kurumu Yayınları. (1955). *Türkçe Sözlük (2nd ed.)*. Türk Dil Kurumu, Ankara, Turkey.

Türk Dil Kurumu Yayınları. (1959). *Türkçe Sözlük (3rd ed.)*. Türk Dil Kurumu, Ankara, Turkey.

Türk Dil Kurumu Yayınları. (1966). *Türkçe Sözlük (4th ed.)*. Türk Dil Kurumu, Ankara, Turkey.

Türk Dil Kurumu Yayınları. (1969). *Türkçe Sözlük (5th ed.)*. Türk Dil Kurumu, Ankara, Turkey.

Türk Dil Kurumu Yayınları. (1974). *Türkçe Sözlük (6th ed.)*. Türk Dil Kurumu, Ankara, Turkey.

Türk Dil Kurumu Yayınları. (1983). *Türkçe Sözlük (7th ed.)*. Türk Dil Kurumu, Ankara, Turkey.

Türk Dil Kurumu Yayınları. (1988). *Türkçe Sözlük (8th ed.)*. Türk Dil Kurumu, Ankara, Turkey.

Türk Dil Kurumu Yayınları. (1998). *Türkçe Sözlük (9th ed.)*. Türk Dil Kurumu, Ankara, Turkey.

Türk Dil Kurumu Yayınları. (2005). *Türkçe Sözlük (10th ed.)*. Türk Dil Kurumu, Ankara, Turkey.

Türk Dil Kurumu Yayınları. (2011). *Türkçe Sözlük (11th ed.)*. Türk Dil Kurumu, Ankara, Turkey.

Uzun, N. E. (2003). Modern dilbilim bulguları ışığında türkçe sözlüğe bir bakış. In *Dil ve Edebiyatı Araştırmaları Sempozyumu 2003, Mustafa Canpolat Armağanı*, pages 281–293.

Vossen, P. (2007). EuroWordNet: A multilingual database for information retrieval. In *DELOS workshop on Cross-Language Information Retrieval*.