Word Frequencies in the Written Texts of Learners of Turkish as a Foreign Language at CEFR Level A

AOBM A Düzeyinde Yabancı Dil Olarak Türkçe Öğrenen Öğrencilerin
Yazılı Metinlerindeki Sözcük Sıklıkları

Derem Çanga*
Middle East Technical University

Abstract

This study aims to list the vocabulary items used by learners of Turkish as a foreign language at CEFR level A and their frequency values, and reveal the factors affecting the differences in the word frequency lists created based on certain demographic criteria. It also aims to compare the lists with the frequency lists of the course book used by the learners and the list of words frequently used by native speakers of Turkish. For these purposes, written texts of 123 students at Ankara University TÖMER were collected and digitized, and using all the words in those texts, a lemma list was created for this learner corpus. Then, separate lists were created to show how certain variables might affect the frequency of words used by learners. The results show that the most significant factor leading to differences in the lists is the level the students were studying at.
Keywords: Turkish as a Foreign Language, corpus linguistics, learner corpus, word frequency list

Öz

Bu çalışma Avrupa Ortak Başvuru Metni (AOBM) A düzeyinde Yabancı Dil Olarak Türkçe öğrenen öğrencilerin yazılı metinlerinde kullandıkları sözcükleri ve sözcük sıklıklarını listeleyerek demografik değişkenlere göre oluşturulan gruplar arasındaki farkların sebeplerini ortaya çıkarmayı amaçlamaktadır. Ayrıca, bu sıklık listelerini öğrencilerin kullandıkları ders kitabının sıklık listesiyle ve anadili Türkçe olan konuşucuların kullandıkları sözcüklerin sıklık listesiyle de karşılaştırmayı hedeflemektedir. Bunun için Ankara Üniversitesi TÖMER'de öğrenim gören 123 öğrencinin yazılı metinleri toplanmış, bilgisayar ortamına aktarılmış ve öğrenci derlemini oluşturan bu metinlerdeki tüm sözcükler ile bir başsözcük listesi oluşturulmuştur. Daha sonra, belli değişkenlerin öğrencilerin sözcük sıklıklarını nasıl etkilediğini incelemek üzere ayrı listeler oluşturulmuştur. Sonuçlar farklı gruplardaki öğrencilerin listeleri arasındaki farklara sebep olan en önemli etmenin öğrencilerin dil düzeyi olduğunu göstermektedir.
Anahtar Kelimeler: Yabancı Dil Olarak Türkçe, derlem dilbilim, öğrenci derlemi, sözcük sıklık listesi

Introduction

Corpus can be defined as the gathering of written texts and spoken language data that can represent a language based on certain criteria (Burkhanov qtd. in Karaoğlu

181), and the characteristics of a corpus can be listed as follows: Electronic/ Computer readability, representativeness, balance and including language used in natural communication contexts (Gries and Berez 380; O'Donnell). As for different types of corpora, they can be classified into different groups such as balanced and non-balanced, synchronic and diachronic, written and spoken, or L1 and learner corpora. This study focuses on one of these types of corpora, namely *learner corpus*, which is the systematic gathering of written or spoken language of second language learners and storing the data electronically (Callies and Paquot 1). These corpora are of crucial importance in second language acquisition and foreign language teaching as they can provide the necessary learner output that can be analyzed by language software tools (Granger 4). Callies and Paquot (1) stated more specifically that learner corpora can be used for a number of purposes such as describing learner language, designing new teaching methods and tools that target learners' needs, creating new dictionaries, and in assessment and evaluation.

The linguistics methodology that is based on gathering natural texts called corpora electronically is called *corpus linguistics* (Granger 1). The number of studies in corpus linguistics in Turkish is low; the number of ones related to a learner corpus in Turkish is even lower. Although there are numerous learner corpora in many languages such as English, German, French, Italian, Spanish, Dutch, Korean, Chinese, Arabic, and the like, there is not one in Turkish. This study aims to contribute to one being constructed.

One important part of corpus studies is word frequency lists. Aksan and Yaldır define *word frequency* as the numerical value that shows how many times a linguistic item is used in a corpus (378). Frequency lists are lists that show how frequently a word is used in a corpus. To be able to make comparisons between the list of one corpus with the list of another, normalized frequency values need to be calculated as the frequency values on their own will not be reliable (Hoffmann et al. qtd in Aksan and Demirhan 89). The reason for this is because the corpora to be compared could be different in size and the frequencies of the words could be different only because of that. Thus, the frequency values have to be normalized; that is the values should be calculated using the formula below to see what they would be if the corpus were to be of one million words.

Frequency of the word in the corpus/the total number of words in the corpus* 1.000.000

The number of word frequency or vocabulary studies in Turkish are low and they mostly focus on Turkish as a native language (Göz; Pilav; Ölker; Gündoğdu). The number of word frequency studies in Turkish as a foreign language is also rather low. Similar to others in the field, Ercan's (7) study shows that among the 118 theses written between the years 1995 and 2013 regarding teaching Turkish to foreigners, 16.94% was on grammar teaching, 13.56% on materials design and/or evaluation, 9.32% on teaching methods and techniques, 7.63% on vocabulary teaching, 6.77% on difficulties faced while teaching and only 4.24% (5 out of 118

theses) on word frequency/preparing dictionaries. In addition, the few studies focusing on the language as a foreign language were either ones evaluating the language in course books (Karadağ and Kurudayıoğlu; Sevinç; Uslu; Akman; Doğan; Karadağ), or oral language performances (Çelebi). One recent study about word frequencies, for example, focused on the word frequencies used in seven course books used to teach Turkish as a foreign language (Şimşek and Gün). Similarly, in a doctoral thesis study, Bulundu aimed to identify the vocabulary items used in the nine course book series used to teach Turkish as a foreign language.

When studies focusing on Turkish as a foreign language are analyzed, it can be seen that only 17 graduate theses were written between the years 2006-2022. As examples for studies on word frequency among these, Göçen's doctoral dissertation on word frequencies of learners of Turkish as a foreign language and Çetin's master's thesis on the factors affecting the word frequencies in texts written by young learners of Turkish as a second language can be given. Another study focusing on the vocabulary used in B1 level course books and 21 students' written essays is Şimşek's article.

In the light of all these, it can be concluded that this study will shed light on a field that has rarely been focused on. The aims of this study are: To determine the vocabulary of learners of Turkish as a foreign language at CEFR level A; to reveal the factors that might influence how frequently learners use these words; to portray the differences in frequency values between the lists of words used by learners and the list of words in the *Hitit Yabancılar İçin Türkçe* course book used by the learners, and the list of the most frequently used words by native speakers of Turkish; to understand the factors leading to the differences in these lists.

Materials and Methods

Data and Participants

Written texts of 123 learners learning Turkish as a foreign language at the CEFR levels A1 and A2 at TÖMER in Ankara, Turkey were collected and analyzed for this study. These learners were students in four sub-levels, which can be seen in Table 1 below.

Table 1. The number of students in each level

CEFR Level	Number of students
A.1.1	45
A.1.2	32
A.2.1	22
A.2.2	24
Total	**123**

The written texts collected from learners were mostly homework assignments, and in two of the levels, there were also texts they wrote in exams held in class. The numbers of each can be seen in Table 2 below.

Table 2. The number of different types of texts written by learners in each level

CEFR Level	Number of homework texts	Number of texts written in exams	Total number of texts
A1.1	93		93
A1.2	62	31	93
A2.1	69		69
A2.2	67	23	90
Total	**291**	**54**	**345**

Participants

All participants filled out a form with certain demographic information. A summary of this information has been given below.

Table 3. Percentage of learners based on gender

Gender	Percentage of learners
Female	47.4%
Male	52.6%

Table 4. Percentages of learners based on age intervals

Age interval	Percentage of learners
13-20	29.3%
21-30	50%
31-40	12.9%
41-50	4.3%
Not indicated	3.5%

Other demographic information collected from the learners includes their nationalities and the native languages they speak. The participants came from 40 different countries; Palestine, Iraq, Russia and Turkmenistan being the ones from which the majority of learners came, 18.1%, 12%, 5.2% and 5.2% respectively. As for the native languages of the learners, 41.5% spoke Arabic, 9.8% Russian, 7.3% English, 4.9% Turkmen, 4.3% Persian, 4.3% French, and the remaining 27.9% 25 other languages.

Methods

In this study, a corpus-driven approach was adopted. This type of approach claims that corpora are the only resource for a researcher to form theses using corpus data (Tognini-Bonelli 84-85).

First of all, the demographic information collected from learners was transferred to an Excel file, during which learners' names were not used and each learner was given a number instead for anonymity. Each of the texts that were collected was also given a separate number such as 015-01, which means it is the first text written by Learner 15. Then, all the hand-written texts were transferred to a computer and were digitized using Notepad++. For character coding UTF-8 was chosen so that Turkish characters would be shown correctly. An example can be seen below.

Figure 1. A sample learner text digitalized using Notepad++
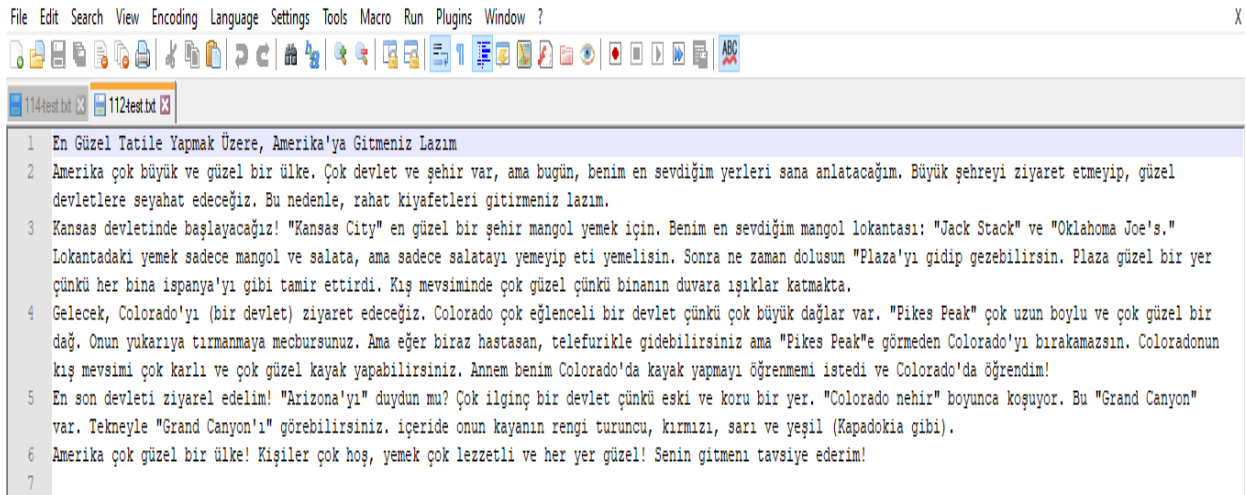
The next step was categorizing each text based on demographic information. To illustrate, a folder was created for all the texts written by women, another one was created for all the texts written by learners between the ages 21-30, and another one for texts written by learners at A.1.1 level. After this, a word list for all the data was created using AntConc. The next step was "a necessary step for computing the lexical density of texts" (Granger and Lefer 27): Lemmatization.

Every linguistic item with a space or a punctuation mark on both sides is called a *token* (Aksan and Yaldır 378). *Type*, on the other hand, is every token that is repeated in a corpus or words that are different than each other and form the corpus (Aksan and Aksan 87). Each type is a different word form; however, some of them are derived from the same root. The words that are the basic forms of those derived ones; that is, words that can be headwords in a dictionary are called lemmas (Aksan and Aksan 87). For this study, a lemma list was created on Excel, which was then loaded to AntConc as the lemma list to be used for the searches. A small part of the list can be seen in Figure 2 below.

Figure 2. A part from the lemma list created for this study

398	basamak	->		basamak	basamaklar	basamaktı	basamakları	basamaklarına										
399	Basil	->		Basil														
400	basit	->		basit	basitti	basittir	basıt	basitçe										
401	basketbol	->		basketbol														
402	basketbolcu	->		basketbolcu	basketbolcuydu	basketbolcuydum												
403	baş	->		baş	başta	baştan	başı	başında	başından	başa	başi	başım	başıma	başına	başinde	başını	başine	başlasında
404	başak	->		başak	başaklarla													
405	başarı	->		başarı	başarıyla	başarılar	başarılığa	başarının	başarısına	başarısıyla	başarıya	başar	başara	başaralar	başarlık			
406	başarılı	->		başarılı	başarili	başarlı	başarıl	başarılılar	başarli	basarlı	başarlıym	başınlı	başırlı					
407	başarısız	->		başarısız	başarsız	başırsız												
408	başar-	->		başar-	başaracağım	başaracağında	başaracaktı	başardı		başariyor	başarıyorlar	başarmak	başarmak					
409	başbakan	->		başbakan														
410	başdede	->		başdede														
411	başka	->		başka	başkas	başk	başkaş	başkalara	başkalard	başkaları		başkalarır	başkalarına	başkalarını		başkalarıy	başkasına	vaşka
412	başkan	->		başkan	başkanı													
413	başkent	->		başkent	başkenti	başkentir	başkentidir	baskent		başkenteş	başkentı	başkentte	başkentten	başkente		başkantıdır		
414	başla-	->		başla-	başlar	başladım	başlarım	başlıyorum	başacağın	başlamış	başliyor	başlıyor	başladığımda	başayıp	başaliyor	başlardı	başladı	başlayaca başlayaca basliyor	başlamışlı başlıyordı
415	başlangıç	->		başlangıç	başlangıcı													
416	başrol	->		başrol	başrolundur													
417	başvur-	->		başvur-	başvuracağim	başvuruldum	başvurulma	başvurmak										
418	başvuru	->		başvuru	başvurusu													
419	bat-	->		bat-	batmak	batımı	batırabilecek											

As a result, the learner corpus that was constructed in this study included 2903 word types and 31897 word tokens in total. The type/token ratio is 0.091. This is the number that is found by dividing the number of types into the number of tokens in a corpus. This value which is represented in percentages is always smaller than 1 as the number of types is always lower than the number of tokens. The bigger the ratio, the higher the number of different words used in a corpus. However, as the size of the corpus gets larger, the ratio will get lower as function words will be repeatedly used and the number of content words will not get that much higher (Aksan and Yaldır 379).

After lemmatization and uploading the lemma list to AntConc, frequency lists were created for each category by uploading only the texts written by a certain group of learners to AntConc. That is, separate frequency lists were created for texts written by learners between the ages 21-30 and those written by the ones between 31-40, for example. Also, separate lists were created for male and female learners' texts, and the texts written by learners at the level A.1.1 and those of the ones at A.1.2, and the like. Then, to be able to make comparisons between these, the normalized frequency value of each word was also calculated. Then, different types of statistical tests were done to see if the differences are statistically significant or not.

Results

The list of the most frequently used fifty words and their equivalent(s) in English in parentheses next to each word together with the frequency and normalized frequency values can be seen in Table 5 below.

Table 5. The 50 words most frequently used in the corpus

	Freq.	Word	Nor. Freq.		Freq.	Word	Nor. Freq.
1	1353	ve (and)	42417.78	26	190	en (most)	5956.67
2	1157	ben (I)	36273.00	27	176	istemek (to want)	5517.76
3	999	çok (very/many)	31319.56	28	172	yaşamak (to live)	5392.36
4	670	bir (one/a/an)	21005.11	29	168	gelmek (to come)	5266.95
5	614	gitmek (to go)	19249.46	30	167	aile (family)	5235.60
6	508	sonra (then/later/after)	15926.26	31	165	ad (name)	5172.90
7	502	o (he/she/it/that)	15738.16	32	163	şehir (city)	5110.20
8	492	var (there is)	15424.65	33	159	biz (we)	4984.79
9	429	yapmak (to do/to make)	13449.54	34	157	şey (thing)	4922.09
10	395	her (every)	12383.61	35	148	insan (human/person)	4639.93
11	330	yemek (food/meal)	10345.80	36	147	iki (two)	4608.58
12	297	ev (house)	9311.22	37	146	yaş (age)	4577.23
13	290	bu (this)	9091.76	38	144	tatil (holiday)	4514.53
14	287	güzel (beautiful/nice)	8997.71	39	143	büyük (big)	4483.18
15	275	gün (day)	8621.50	40	141	iyi (good)	4420.48
16	270	ama (but)	8464.75	41	138	daha (more)	4326.43
17	270	saat (hour)	8464.75	42	133	almak (to take/to get)	4169.67
18	260	için (for)	8151.24	43	128	ora (there)	4012.92
19	251	arkadaş (friend)	7869.08	44	125	deniz (sea)	3918.86
20	245	etmek (to do/to make)	7680.97	45	124	anne (mother)	3887.51

21	237	olmak (to be/to become)	7430.17	46	121	çalışmak (to study/to work)	3793.46
22	231	çünkü (because)	7242.06	47	120	yer (place)	3762.11
23	212	zaman (time)	6646.39	48	109	hafta (week)	3417.25
24	211	yemek (to eat)	6615.04	49	103	kardeş (sibling)	3229.14
25	198	sevmek (to love)	6207.48	50	100	hayat (life)	3135.09

Most of the words in this list and the order in which they appear in this list are similar to the ones in the frequency list of *Hitit Yabancılar İçin Türkçe*, the course book used by the students, and the frequency list of Turkish as a native language. Below in Table 6 is a list of the most frequent fifty words in the learner corpus on the left and the course book on the right.

Table 6. The 50 words most frequently used in the corpus and *Hitit Yabancılar İçin Türkçe* (the course book used in classes)

Learner Corpus								Course book*							
	Freq.	Word	Nor.Freq.		Freq.	Word	Nor.Freq.		Freq.	Word	Nor.Freq.		Freq.	Word	Nor.Freq.
1	1353	ve (and)	42417.78	26	190	en (most)	5956.67	1	826	bir (one/a/an)	24606.04	26	158	iş (work/job)	4706.72
2	1157	ben (I)	36273.00	27	176	istemek (to want)	5517.76	2	811	imek (to be)	24159.19	27	154	almak (to take/to get)	4587.57
3	999	çok (very/many)	31319.56	28	172	yaşamak (to live)	5392.36	3	566	ve (and)	16860.79	28	151	insan (person/human)	4498.20
4	670	bir (one/a/an)	21005.11	29	168	gelmek (to come)	5266.95	4	519	çok (very/many)	15460.69	29	150	gelmek (to come)	4468.41

4140.73	5	614	gitmek (to go)
4021.57	6	508	sonra (then/later/after)
3693.88	7	502	o (he/she/it / that)
3515.15	8	492	var (existent)
3485.36	9	429	yapmak (to do/to make)
3455.57	10	395	her (every)
3366.20	11	330	yemek (food/meal)
3336.41	12	297	ev (house)
3306.62	13	290	bu (this)
3247.04	14	287	güzel (beautiful / nice)
3127.89	15	275	gün (day)

19249.46	30	167	aile (family)
15926.26	31	165	ad (name)
15738.16	32	163	şehir (city)
15424.65	33	159	biz (we)
13449.54	34	157	şey (thing)
12383.61	35	148	insan (person/ human)
10345.80	36	147	iki (two)
9311.22	37	146	yaş (age)
9091.76	38	144	tatil (holiday)
8997.71	39	143	büyük (big)
8621.50	40	141	iyi (good)

5235.60	5	469	ben (I)
5172.90	6	464	bu (this)
5110.20	7	435	mı (question particle)
4984.79	8	384	da (too/also)
4922.09	9	289	için (for)
4639.93	10	282	var (existent)
4608.58	11	275	o (she/ he/it)
4577.23	12	253	yapmak (to do/to make)
4514.53	13	242	olmak (to be/to become)
4483.18	14	239	ne (what)
4420.48	15	215	ama (but)

13971.22	30	139	zaman (time)
13822.28	31	135	değil (not)
12958.38	32	124	en (most)
11439.13	33	118	güzel (beautiful /nice)
8609.13	34	117	çocuk (child)
8400.61	35	116	biz (we)
8192.08	36	113	nere (where)
7536.72	37	112	yemek (food/ meal)
7209.03	38	111	çıkmak (to go out of)
7119.66	39	109	yemek (food/meal)
6404.72	40	105	arkadaş (friend)

3127.89	bura (here)	105	41	6255.77	ev (home)	210	16	4326.43	daha (more)	138	41	8464.75	ama (but)	270	16
3127.89	televizyon (TV)	105	42	6106.82	gitmek (to go)	205	17	4169.67	almak (take/to get)	133	42	8464.75	saat (hour)	270	17
3127.89	yıl (year)	105	43	5898.30	iyi (good)	198	18	4012.92	ora (there)	128	43	8151.24	için (for)	260	18
3098.10	evet (yes)	104	44	5719.56	sen (you)	192	19	3918.86	deniz (sea)	125	44	7869.08	arkadaş (friend)	251	19
3038.52	nasıl (how)	102	45	5421.67	sonra (then/later/after)	182	20	3887.51	anne (mother)	124	45	7680.97	etmek (to do/to make)	245	20
3008.73	çünkü (because)	101	46	5123.77	gün (day)	172	21	3793.46	çalışmak (to study/to work)	121	46	7430.17	olmak (to be/to become)	237	21
3008.73	yer (place)	101	47	5064.20	istemek (to want)	170	22	3762.11	yer (place)	120	47	7242.06	çünkü (because)	231	22
2919.36	tamamlamak (to complete)	98	48	4915.25	daha (more)	165	23	3417.25	hafta (week)	109	48	6646.39	zaman (time)	212	23
2919.36	yok (non-existent)	98	49	4855.67	saat (hour)	163	24	3229.14	kardeş (sibling)	103	49	6615.04	yemek (to eat)	211	24
2800.20	gibi (like)	94	50	4736.51	siz (plural you)	159	25	3135.09	hayat (life)	100	50	6207.48	sevmek (to love)	198	25

* The list was created by adding the frequency values of words in the A1 and A2 level books listed in Göçen's doctoral dissertation (2149-2150).

It can be observed that 31 out of these 50 words are the same in the two lists. However, there are certain differences in the frequencies they were used at. Most of these differences between these two lists stem from the fact that the learner corpus includes texts written as answers to prompts related to basic topics such as family, daily activities, holidays, and the like. This is why the use of the pronouns "I" (nor.freq.: 36273) and "we" (nor.freq.: 4984) is more frequent in this list than in the course book (nor. freq.s: 13971 and 3366 respectively). Also, verbs such as to do, to make, to go, to eat, to want, to come, and to study are all used more frequently in the learner texts to explain what they do in a day, on holidays, and the like. In the course book, on the other hand, there are dialogues; thus, question words such as "ne", "nere", nasıl", the question particle "mı", the pronouns "sen" and "siz" (you), and responses such as "evet" (yes) are used numerous times in addition to verbs such as "tamamla" (to complete) that are used in instructions. None of these appears in the list of the most frequent fifty words used by the learners.

As mentioned before, the most frequently used words by the learners are mostly similar to the ones used by native speakers of Turkish too. Below in Table 7 are the two lists.

Table 7. The 50 words most frequently used in the corpus and TNC (Turkish National Corpus)**

Learner Corpus								TNC**							
	Freq.	Word	Nor.Freq.		Freq.	Word	Nor.Freq.		Freq.	Word	Nor.Freq.		Freq.	Word	Nor.Freq.
1	1353	ve (and)	42417.78	26	190	en (most)	5956.67	1	1402146	bir (a/an)	27494.67	26	182309	yer (place)	3574.90
2	1157	ben (I)	36273.00	27	176	istemek (to want)	5517.76	2	1312151	ol (to be/become)	25729.96	27	179439	bul (to find)	3518.62
3	999	Çok (very/many)	31319.56	28	172	yaşamak (to live)	5392.36	3	1178818	ve (and)	23115.43	28	178382	iç (inside; to drink)	3497.89
4	670	bir (one/a/an)	21005.11	29	168	gelmek (to come)	5266.95	4	989142	bu (this)	19396.08	29	176269	ama (but/yet/still)	3456.46

Ranks 5–15

Rank	Word	Frequency	per million	Word	Frequency	per million
5	da (too/also)	881669	17288.64	gitmek (to go)	614	19249.46
6	et (to do/to make; meat)	511423	10028.49	sonra (then/later/after)	508	15926.26
7	o (she/he/it)	492063	9648.86	o (he/she/it/that)	502	15738.16
8	yap (to do/to make)	358192	7023.78	var (existent)	492	15424.65
9	al (to take/get)	312199	6121.91	yapmak (to do/to make)	429	13449.54
10	de (to say)	321612	6306.49	her (every)	395	12383.61
11	ben (I, me)	316570	6207.62	yemek (food/meal)	330	10345.80
12	gel (to come)	285485	5598.07	ev (house)	297	9311.22
13	için (for)	282208	5533.81	bu (this)	290	9091.76
14	çok (very; much/many/a lot of)	261699	5131.65	güzel (beautiful/nice)	287	8997.71
15	ver (to give)	236298	4633.57	gün (day)	275	8621.50

Ranks 30–40

Rank	Word	Frequency	per million	Word	Frequency	per million
30	zaman (time)	160656	3150.30	aile (family)	167	5235.60
31	her (each/every)	160771	3152.56	ad (name)	165	5172.90
32	şey (thing)	167699	3288.41	şehir (city)	163	5110.20
33	kadar (until; as much as)	154881	3037.06	biz (we)	159	4984.79
34	ara (break/distance to search)	152895	2998.12	şey (thing)	157	4922.09
35	git (to go)	156933	3077.30	insan (person/human)	148	4639.93
36	en (most; width)	148762	2917.07	iki (two)	147	4608.58
37	yıl (year)	154813	3035.73	yaş (age)	146	4577.23
38	değil (be not)	149658	2934.64	tatil (holiday)	144	4514.53
39	çalış (to work)	150234	2945.94	büyük (big)	143	4483.18
40	iste (to want)	147628	2894.84	iyi (good)	141	4420.48

Rank	Word	Frequency	Per million
41	geç (to pass; late)	142923	2802.58
42	bil (to know)	146004	2862.99
43	biz (we/us)	143327	2810.50
44	gün (day)	136776	2682.04
45	insan (human)	140522	2755.49
46	anla (to understand)	134073	2629.04
47	ya (either)	134512	2637.64
48	el (hand)	124409	2439.53
49	ki (who/which-conjunction)	126887	2488.13
50	kal (to stay/remain)	121834	2389.04

Rank	Word	Frequency	Per million
16	ne (what)	243489	4774.57
17	daha (more)	227608	4463.16
18	gibi (like)	231717	4543.74
19	gör (to see)	216069	4236.89
20	kendi (self)	218435	4283.29
21	mi (question particle)	225051	4413.02
22	var (existent; to arrive at)	211317	4143.71
23	çık (to go out of)	198227	3887.03
24	ile (with/together)	202389	3968.64
25	sonra (after/later)	191249	3750.20

Word			
daha (more)	4326.43	8464.75	138
almak (to take/to get)	4169.67	8464.75	133
ora (there)	4012.92	8151.24	128
deniz (sea)	3918.86	7869.08	125
anne (mother)	3887.51	7680.97	124
çalışmak (to study/to work)	3793.46	7430.17	121
yer (place)	3762.11	7242.06	120
hafta (week)	3417.25	6646.39	109
kardeş (sibling)	3229.14	6615.04	103
hayat (life)	3135.09	6207.48	100

Word		
ama (but)	270	16
saat (hour)	270	17
için (for)	260	18
arkadaş (friend)	251	19
etmek (to do/to make)	245	20
olmak (to be/to become)	237	21
çünkü (because)	231	22
zaman (time)	212	23
yemek (to eat)	211	24
sevmek (to love)	198	25

**The list was taken from Aksan et al.

To identify any statistical differences between the two corpora the Mann Whitney U test was used and the results were shown in Table 8 below.

Table 8. The results of the statistical analysis on the learner corpus and TNC

	Count	Mean	Standard Deviation	Minimum	Maximum	Median	U value	Sig.
Group 1	27	7410,86	7383,06	2682,04	27494,67	4143,71	191,000	,003*
Group 2	27	12096,81	10214,19	3762,11	42417,78	8151,24		

*$p<0.05$

As there are 27 words common in the two lists, the comparison was done using the frequency values of these words only. The mean and median of the normalized frequency values in TNC were higher than the ones of the learner corpus. In addition, there is a statistically significant difference between the normalized frequency values belonging to the two corpora ($p>0.05$). This indicates that native speakers of Turkish use these words at significantly higher frequency values than the learners of Turkish as a foreign language at CEFR levels A1 and A2.

As mentioned before, among the 50 most frequently used words by the learners of Turkish and native speakers, 27 are the same. However, how frequently these 27 words were used differed. To illustrate, the conjunctions "ve" (and) and "ama" (but) seem to be used more frequently (nor.freq.: 42417, 8464 respectively) by the A1 level learners of Turkish compared to native speakers (nor.freq.: 23115, 3456 respectively). The reason for this could be that at this level of basic proficiency, the number of conjunctions learners learn and can use is limited and they prefer to use the first ones they learn more frequently than they are used by the native speakers of the language. Similarly, it can be seen in Corefl Learner Corpus, learner corpus of 530,392 words (Lozano, Díaz-Negrillo & Callies, 2020) that L2 learners of English at the CEFR levels A1 and A2 used the word "and" frequently (nor.freq.: 50361 at A1 and 44870 at A2). In the same corpus, it can be seen that the same groups of learners used the word "but" frequently as well (nor.freq.: 7309 at A1 and 9391 at A2). When these results are compared to the results from the BNC (British National Corpus), a corpus of texts in English written by native speakers of the language, it can be observed that the use of the conjunctions "and" and "but" by native speakers of English is less frequent than learners of English at low CEFR levels (nor.freq.: 23471 and 3996 respectively). All these results seem to be similar to the ones obtained from the learner corpus created for this study and TNC.

Another difference in frequency values related to the proficiency level of learners is the use of the pronouns "ben" (I) and "biz" (we). Learners used these more frequently (nor.freq.: 36273, 4984 respectively) than native speakers (nor.freq.: 6207, 2810 respectively) probably because they had limited vocabulary to be able to give information about themselves, and some of the questions they were supposed to answer in the texts they wrote asked them to do so. The normalized

frequency values of these pronouns in BNC (7860 and 3140 respectively) seem to be similar to that of TNC; however, they are much lower in Corefl (4225 at A1 and 5925 at A2, and 1014 at A1 and 984 in A2 respectively) than in the learner corpus in this study. One reason for this is probably the difference between the questions or topics the learners were asked to write about. Whereas they were asked to write about their own lives, and their friends and family in this study, the learners were asked to write about a famous person or a film in Corefl and thus, did not have to use these pronouns when writing about these. Another important reason for this difference is probably because lemmatization was not done in Corefl. Thus, the words "I" and "my" are considered as separate words whereas they are listed under the same lemma "ben" in this study.

In the lists above, it was detected that the learners also used some verbs more frequently than native speakers. For instance, the verbs "git" (to go), "yap" (to do/to make), "iste" (to want), "çalış" (to study/to work) were all used more frequently by the learners (nor.freq.s: 19249, 13449, 5517, 3793 respectively) than native speakers (nor.freq.s: 3077, 7023, 2894, 2945 respectively). Some other verbs such as "ol" (to be/to become), "et" (to do/to make), "bul" (to find), "geç" (to pass) and "bil" (to know), on the other hand, were used more frequently by the native speakers. There might be different reasons for these. The verb "ol" (to be/become) is one that is difficult to master in levels A1 and A2. The frequency value for the word "et", on the other hand, includes the use of the same word in its noun form "et" (meat) as well and this may be one of the reasons why the frequency is higher in TNC. As for the verbs "bul" (to find), "geç" (to pass) and "bil" (to know), these three do not appear in the list of the most frequently used fifty words by the learners at all. Similarly, they do not appear in the list of the most frequently used fifty words in the coursebook either. Thus, it might be concluded that the reason why the learners did not use these verbs as frequently as the native speakers could be because they were not exposed to these in classes as much as the ones listed in Table 4 above.

In addition to the comparisons between the frequency list constructed based on the learner corpus and the coursebook and TNC, the sub-corpora created were also compared. Below in Table 9 are the lists according to the age categories.

Table 9. Frequency lists based on the age groups of learners

	Below 20				21-30				31-40				Above 40	
	Nor. Freq.	Word	Freq.		Nor. Freq.	Word	Freq.		Nor. Freq.	Word	Freq.		Nor. Freq.	Word

1	358	ve	44660.68	604	ben	40651.50	290	ve	51300.19	84	Ve	38461.54
2	322	ben	40169.66	575	ve	38699.69	212	çok	37502.21	76	Çok	34798.53
3	189	O	23577.84	479	çok	32238.52	162	ben	28657.35	70	Ben	32051.28
4	181	çok	22579.84	329	bir	22142.95	114	gitmek	20166.28	48	Bir	21978.02
5	173	bir	21581.84	284	gitmek	19114.28	104	bir	18397.31	38	Gitmek	17399.27
6	162	sonra	20209.58	227	sonra	15277.96	93	var	16451.44	38	Sonra	17399.27
7	161	Gitmek	20084.83	224	var	15076.05	83	o	14682.47	30	Güzel	13736.26
8	125	var	15593.81	217	her	14604.93	72	sonra	12736.60	30	Yapmak	13736.26
9	112	yapmak	13972.06	206	yapmak	13864.58	70	ama	12382.81	29	Saat	13278.39
10	93	saat	11601.80	205	o	13797.28	68	güzel	12029.01	28	Yemek	12820.51
11	86	her	10728.54	150	bu	10095.57	67	her	11852.11	23	Ev	10531.14

12	84	yemek	10479.04	148	yemek	9960.96	62	yapmak	10967.63	20	Gün	9157.51
13	81	ev	10104.79	147	ev	9893.66	55	etmek	9729.35	20	Yaşamak	9157.51
14	72	yaş	8982.04	141	gün	9489.84	54	gün	9552.45	18	Arkadaş	8241.76
15	67	ad	8358.28	128	için	8614.89	53	daha	9375.55	18	etmek	8241.76
16	67	arkadaş	8358.28	126	ama	8480.28	52	yemek	9198.66	18	var	8241.76
17	64	ye	7984.03	120	saat	8076.46	51	olmak	9021.76	17	çünkü	7783.88
18	61	güzel	7609.78	118	arkadaş	7941.85	50	sevmek	8844.86	15	her	6868.13
19	60	bu	7485.03	117	etmek	7874.55	50	çünkü	8844.86	15	tatil	6868.13
20	60	için	7485.03	112	güzel	7538.03	48	arkadaş	8491.07	15	şehir	6868.13
21	52	etmek	6487.03	108	olmak	7268.81	47	zaman	8314.17	14	bu	6410.26
22	52	sevmek	6487.03	106	çünkü	7134.20	44	ev	7783.48	14	gelmek	6410.26

23	51	şey	100	zaman	6730.38	7429.68	istemek	42	14	kalmak (to stay)	6410.26
24	50	ama	94	ye	6326.56	7075.89	için	40	14	mutlu	6410.26
25	50	en	84	insan	5653.52	6722.09	bu	38	14	ye	6410.26
26	50	olmak	84	sevmek	5653.52	6191.40	aile	35	13	ad	5952.38
27	50	zaman	83	en	5586.22	6014.51	en	34	13	en	5952.38
28	48	iki	81	istemek	5451.61	6014.51	gelmek	34	13	olmak	5952.38
29	48	kardeş	79	biz	5317.00	5660.71	saat	32	13	on	5952.38
30	43	aile	79	şey	5317.00	5660.71	şehir	32	12	ama	5494.51
31	43	almak	78	yaşamak	5249.70	5130.02	biz	29	12	biz	5494.51
32	42	gün	77	tatil	5182.39	4953.12	ye	28	11	istemek	5036.63
33	41	yaşamak	75	gelmek	5047.79	4599.33	almak	26	11	okumak (to read)	5036.63

5036.63	Televizyon (TV)	11	4599.33	iki	26	4913.18	iyi	73	5114.77	öğrenci	41	34
5036.63	uzun (long/tall)	11	4422.43	önce	25	4845.87	büyük	72	4865.27	gelmek	39	35
5036.63	yaş	11	4245.53	yer	24	4778.57	şehir	71	4865.27	çünkü	39	36
5036.63	çocuk	11	4245.53	çalışmak	24	4576.66	ad	68	4615.77	büyük	37	37
4578.75	büyük	10	4068.64	bazen (sometimes)	23	4576.66	aile	68	4491.02	anne (mother)	36	38
4578.75	ders	10	4068.64	insan	23	4576.66	ora	68	4491.02	biz	36	39
4578.75	insan	10	4068.64	iyi	23	4240.14	yer	63	4491.02	istemek	36	40
4578.75	için	10	4068.64	ora	23	4105.53	deniz	61	4366.27	deniz	35	41
4578.75	o	10	4068.64	çocuk (child)	23	4038.23	anne	60	4116.77	akşam	33	42
4578.75	sabah (morning)	10	3891.74	yaz (summer)	22	4038.23	iki	60	3992.02	on (ten)	32	43
4578.75	zaman	10	3891.74	yaşamak	22	4038.23	kahvaltı	60	3992.02	tatil	32	44

45	31	dönmek (to return)	3867.27	59	çalışmak	3970.92	21	hava (weather)	3714.84	10	çay (tea)	4578.75
46	31	iyi	3867.27	58	spor (sport)	3903.62	21	lazım (necessary)	3714.84	9	aile	4120.88
47	30	hafta	3742.51	56	almak	3769.01	21	mutlu (happy)	3714.84	9	gezmek (wander)	4120.88
48	30	şehir	3742.51	56	yıl (year)	3769.01	20	hafta	3537.94	9	sevmek	4120.88
49	29	çalışmak	3617.76	54	hafta (week)	3634.41	19	deniz	3361.05	9	çalışmak	4120.88
50	28	içmek (to drink)	3493.01	54	hayat (life)	3634.41	18	akşam	3184.15	8	almak	3663.00

The Kruskal Wallis test was applied to detect if there is a significant difference between the normalized frequency values of the words used by students in different age groups. The results of this analysis can be seen Table 10 below.

Table 10. The results of the statistical analysis done on the frequency lists of words used by learners of different age groups

Age	Count	Mean	Standard Deviation	Minimum	Maximum	Median	Chi-Square	Sig.
Below 20	50	9413,67	8624,07	3493,01	44660,68	6237,52		
21-30	50	9215,24	8319,32	3634,41	40651,50	5653,52	,192	,979
31-40	50	9372,01	8899,10	3184,15	51300,19	6456,75		
Above 41	50	9029,30	7747,27	3663,00	38461,54	6181,32		

The mean of the normalized frequency values of words used by students under the age of 20 is higher than that of the other age groups. The median is the highest in the group of students between the ages 31-40. There is no significant difference between the normalized frequency values of the words used by different age groups ($p>0.05$).

When the lists are analyzed, it can be seen that the first five most frequent words are the same in all age categories except for one. The pronoun "o" (he/she/it/that) was used more frequently by the learners below 20. This can be explained not by the difference in age but the fact that most of these learners were in the A.1.1 level where most of the assignments required them to write about their family members, friends or neighbors and what they do in a day or in their free times, all of which necessitate the use of the pronoun. Another difference that can be explained by looking at the levels rather than ages is the use of "ama" (but) and "çünkü" (because) less frequently by the learners below 20 years of age. The word "ama" has a normalized frequency of 6237 in the texts written by students below 20 years of age compared to a normalized frequency of 8480 in the texts written by learners of ages 21-30 and 12382 in ones written by learners of ages 31-40. As for the word "çünkü", the normalized frequency values for the age groups are 4865, 7134, 8844 and 7783. These conjunctions are used more frequently by students at higher levels of language proficiency, who also happen to be students of greater ages in this study.

In addition to age groups, sub-corpora based on gender were also created. The lists below in Table 11 show the differences between the two genders.

Table 11. Frequency lists based on the gender of learners

	Women							Men						
	Nor. Freq.	Word	Freq.		Nor. Freq.	Word	Freq.		Nor. Freq.	Word	Freq.		Nor. Freq.	Word	Freq.
1	41383.12	ve	675	26	6130.83	aile	100	1	43500.58	ve	678	26	5774.41	şey	90
2	33903.50	ben	553	27	6008.22	en	98	2	38752.73	ben	604	27	5517.77	sevmek	86
3	33290.42	çok	543	28	6008.22	gelmek	98	3	29257.03	çok	456	28	5453.61	ad	85
4	20538.29	bir	335	29	5946.91	yaşamak	97	4	21493.65	bir	335	29	4812.01	yaş	75

5	298	gitmek	18269.88	30	94	şehir	5762.98	5	316	gitmek	20274.61	30	75	yaşamak	4812.01
6	275	var	16859.79	31	93	ye	5701.67	6	306	sonra	19633.00	31	73	almak	4683.69
7	247	o	15143.15	32	90	biz	5517.75	7	255	o	16360.84	32	70	gelmek	4491.21
8	208	yapmak	12752.13	33	86	daha	5272.52	8	221	yapmak	14179.39	33	69	biz	4427.05
9	202	sonra	12384.28	34	86	insan	5272.52	9	217	var	13922.75	34	69	istemek	4427.05
10	199	her	12200.36	35	80	ad	4904.67	10	196	her	12575.39	35	69	iyi	4427.05
11	155	güzel	9502.79	36	79	iki	4843.36	11	181	yemek	11612.99	36	69	tatil	4427.05
12	153	bu	9380.17	37	78	büyük	4782.05	12	153	saat	9816.50	37	69	şehir	4427.05
13	149	yemek	9134.94	38	77	anne	4720.74	13	149	ev	9559.86	38	68	iki	4362.89
14	148	ev	9073.63	39	75	tatil	4598.12	14	144	gün	9239.06	39	68	çalışmak	4362.89
15	134	ama	8215.31	40	72	iyi	4414.20	15	137	bu	8789.94	40	67	aile	4298.73

16	132	için	8092.70	41	71	yaş	4352.89	16	136	ama	8725.78	41	66	yer	4234.57
17	131	arkadaş	8031.39	42	67	şey	4107.66	17	132	güzel	8469.14	42	65	büyük	4170.41
18	131	gün	8031.39	43	66	deniz	4046.35	18	128	için	8212.50	43	63	okumak	4042.09
19	129	etmek	7908.77	44	66	ora	4046.35	19	120	arkadaş	7699.22	44	62	hafta	3977.93
20	127	olmak	7786.16	45	60	almak	3678.50	20	118	ye	7570.90	45	62	insan	3977.93
21	124	çünkü	7602.23	46	56	mutlu	3433.27	21	116	etmek	7442.58	46	62	ora	3977.93
22	117	saat	7173.07	47	54	akşam	3310.65	22	110	olmak	7057.62	47	60	son	3849.61
23	112	sevmek	6866.53	48	54	hayat	3310.65	23	110	zaman	7057.62	48	59	deniz	3785.45
24	107	istemek	6559.99	49	54	yer	3310.65	24	107	çünkü	6865.14	49	59	lazım	3785.45
25	102	zaman	6253.45	50	53	kardeş	3249.34	25	92	en	5902.73	50	52	daha	3336.33

The Mann Whitney U test was used to find if there is a significant difference between the normalized frequency values of the words used by students from each gender. The results are shown in Table 12 below.

Table 12. The results of the statistical analysis done on the frequency values of the words used by female and male learners

	Count	Mean	Standard Deviation	Minimum	Maximum	Median	U value	Sig.
Female	50	9061,37	7997,65	3249,34	41383,12	6192,14	1207,000	,767
Male	50	9276,27	8537,03	3336,33	43500,58	5838,57		

The mean of the normalized frequency values of the words used by male students is higher and the median of the normalized frequency values of the words used by female students is higher. There is no significant difference between the normalized frequency values of the words used by the two groups ($p>0.05$).

When the lists are analyzed descriptively, it can be seen that in these two lists, the first five most frequently used words are the exact same words in the same order. The next five words are also the same words, only in a different order. However, in the rest of the list there are a few significant differences. To illustrate, the word "saat" (hour) is the twelfth most frequently used word (nor.freq.: 9816) in texts written by men whereas it is the twenty-second (nor.freq.: 7173) in texts written by women. Similarly, the word "yaş" (age) is the twenty-ninth (nor.freq.: 4812) most frequently used word in the texts written by men; however, it is the forty-first (nor.freq.: 4352) most frequent word in the texts written by women. Here again, rather than the gender, the level of the students seems to be an important factor. Both these words are required to be used to answer the questions assigned in the A.1.1 level, and among the forty-five A.1.1 students, twenty-seven were men. Another such example is the word "daha" (more) which is listed as the thirty-third (nor.freq.: 5272) most frequent word used by women and the fiftieth (nor.freq.: 3336) by men. The reason for this difference seems to lie in the fact that there were 19 women and 12 men in the A.1.2 level in which one of the prompts asked students to compare two cities, which requires the use of the comparative adverb.

Finally, the lists based on the 4 different levels learners were studying at can be seen below in Table 13.

Table 13. Frequency lists based on the level the learners were studying

	Freq.		A.1.1 Word	Nor. Freq.	Freq.	A.1.2 Word	Nor. Freq.	Freq.		A.2.1 Word	Nor. Freq.	Freq.	A.2.2 Word	Nor. Freq.
1	357		ben	60963.11	369	ve	47325.89	279		Ve	49652.96	492	ve	41961.62

	35650.32	çok	418	38796.94	Ben	218	44632.55	çok	348	36372.95	ve	213	2
	21577.83	bir	253	33279.94	sonra	187	44504.30	ben	347	29883.88	o	175	3
	20042.64	ben	235	31322.30	Bir	176	28857.25	gitmek	225	22711.75	saat	133	4
	16801.71	var	197	28474.82	Çok	160	18981.66	var	148	21004.10	gitmek	123	5
	14925.37	her	175	25271.40	gitmek	142	16801.33	bir	131	20491.80	sonra	120	6
	14925.37	için	175	21712.05	O	122	14364.50	güzel	112	18784.15	bir	110	7
	14754.80	bu	173	18330.66	yapmak	103	14107.99	sonra	110	17418.03	ev	102	8
	13816.63	yapmak	162	15661.15	Her	88	11799.41	ev	92	17076.50	yaş	100	9
	13304.90	yemek	156	15305.21	etmek	86	10901.63	ama	85	15198.09	ad	89	10
	11343.28	en	133	14237.41	gün	80	10388.61	yemek	81	15198.09	yapmak	89	11
	10916.84	çünkü	128	14059.44	arkadaş	79	10260.36	her	80	12978.14	var	76	12

13	73	çok	12465.85	80	o	10260.36	74	olmak	13169.60	125	o	10660.98
14	57	kardeş	9733.61	78	gün	10003.85	71	var	12635.70	124	gitmek	10575.69
15	54	arkadaş	9221.31	75	yapmak	9619.08	69	ev	12279.77	119	güzel	10149.25
16	52	her	8879.78	70	biz	8977.81	57	bu (this)	10144.15	109	olmak	9296.38
17	50	ye	8538.25	70	şehir	8977.81	54	para (money)	9610.25	99	ama	8443.50
18	48	öğrenci (student)	8196.72	68	arkadaş	8721.30	54	zaman	9610.25	95	şey	8102.35
19	47	iki	8025.96	65	deniz	8336.54	53	ama	9432.28	92	gün	7846.48
20	46	kahvaltı (breakfast)	7855.19	57	ora	7310.50	51	demek (to say)	9076.35	91	sonra	7761.19
21	46	yemek	7855.19	56	saat	7182.25	47	istemek	8364.48	89	lazım	7590.62
22	46	çalışmak	7855.19	55	çünkü	7054.00	47	yemek	8364.48	84	etmek	7164.18
23	42	anne	7172.13	53	aile	6797.49	44	almak	7830.57	84	insan	7164.18

24	40	kız (girl)	6830.60	53	ye	6797.49	44	saat	7830.57	84	istemek	7164.18
25	40	on	6830.60	50	tatil	6412.72	44	şey	7830.57	83	sevmek	7078.89
26	40	yaşamak	6830.60	49	anne	6284.47	43	gelmek	7652.61	82	yer	6993.60
27	39	okumak	6659.84	49	daha	6284.47	42	için	7474.64	80	ye	6823.03
28	37	baba	6318.31	49	gelmek	6284.47	41	sevmek	7296.67	79	zaman	6737.74
29	36	öğrenmek (to learn)	6147.54	48	büyük	6156.21	38	vermek (to give)	6762.77	79	şehir	6737.74
30	35	ders	5976.78	47	sevmek	6027.96	38	yardımsever (helpful/friendly)	6762.77	73	tatil	6226.01
31	35	sabah	5976.78	47	zaman	6027.96	37	güzel	6584.80	72	iyi	6140.72
32	33	aile	5635.25	45	yaşamak	5771.45	36	hafta	6406.83	71	yaşamak	6055.44
33	33	ama	5635.25	44	bu	5643.20	34	çünkü	6050.90	69	daha	5884.86
34	32	etmek	5464.48	44	otel (hotel)	5643.20	33	biz	5872.93	68	sağlıklı (healthy)	5799.57

Rank	Word	Freq.	n	Word	Freq.	n	Word	Freq.	n	Word	Freq.	n
35	zaman	5464.48	32	etmek	5514.94	43	son	5694.96	32	ülke (country)	5202.56	61
36	dokuz (nine)	4781.42	28	olmak	5514.94	43	aile	5517.00	31	hayat	5117.27	60
37	uyumak (to sleep)	4781.42	28	kalmak (to stay)	5386.69	42	dönmek	5517.00	31	gelmek	4861.41	57
38	sevmek	4610.66	27	hava	5130.18	40	ye	4983.09	28	büyük	4776.12	56
39	kalkmak (to get up)	4439.89	26	için	4488.91	35	yok (there isn' t)	4449.19	25	önemli (important)	4776.12	56
40	oturmak (to sit)	4439.89	26	mutlu	4488.91	35	iki	4271.22	24	ora	4434.97	52
41	ödev (homework)	4439.89	26	yıl	4488.91	35	insan	4271.22	24	spor	4349.68	51
42	almak	4269.13	25	iki	4360.65	34	iyi	4271.22	24	aile	4264.39	50
43	gün	4269.13	25	insan	4360.65	34	kahvaltı	4271.22	24	arkadaş	4264.39	50
44	uyanmak (to wake)	4269.13	25	yaz	4360.65	34	kız	4271.22	24	içmek	4179.10	49
45	üniversite (university)	4269.13	25	önce	4360.65	34	iş (job/work)	4093.26	23	bayram (holiday)	4008.53	47

46	23	buçuk	3927.60	31	ad	3975.89	21	de	3737.32	46	yeni (new)	3923.24
47	23	hafta	3927.60	31	hafta	3975.89	20	da	3559.35	44	mevsim (season)	3752.67
48	23	Istemek	3927.60	30	okul (school)	3847.63	20	konuşmak (to speak)	3559.35	43	almak	3667.38
49	22	beş (five)	3756.83	30	yüzmek (to swim)	3847.63	20	söylemek	3559.35	42	deniz	3582.09
50	22	içmek	3756.83	30	üniversite	3847.63	20	yıl	3559.35	42	iki	3582.09

To compare the four groups, the Kruskal Wallis test was used. The results are shown in Table 14 below.

Table 14. The results of the statistical analysis done on the frequency lists of the words used by learners studying at different CEFR levels

Level	Count	Mean	Standard Deviation	Minimum	Maximum	Median	Chi-Square	Sig.
A.1.1	50	10430,33	10135,23	3756,83	60963,11	6830,60		
A.1.2	50	10109,02	10104,07	3847,63	47325,89	6348,60	,616	,893
A.2.1	50	11254,67	9953,57	3559,35	49652,96	7741,59		
A.2.2	50	9303,20	7471,71	3582,09	41961,62	7036,25		

The highest mean of normalized frequency values is the mean of the frequency values of the words used by learners in the A.2.1 level. The highest median value belongs to the same group. There is no significant difference between the normalized frequency values of the words used by the four groups of students in different levels (*p*>0.05).

However, when the lists are analyzed, some differences in the order of the words in these lists can be spotted and these can be attributed to two reasons: Some of them seem to be based on the topics the students were asked to write about in each level. For instance, the word "saat" (hour) was used more frequently (nor.freq.: 22711) by the students in level A.1.1 because they were asked to write about what

they do at what time in a day. Similarly, the words "sağlıklı" (healthy), "hayat" (life), "önemli" (important), "bayram" (religious/national holiday) and "mevsim" (season) used by the students in A.2.2 were used this frequently (nor.freq.: 5799, 5117, 4776, 4008, 3752 respectively) because the assignments in this level asked them to write about these specifically. The other reason why there are more differences in the lists is basically the difference between level of proficiency. For example, the conjunctions "çünkü" (because) and "ama" (but) are acquired later and thus are used more frequently by students in the higher levels. Another important difference in the lists is the use of the word "şey" (thing). This word can only be seen among the most frequent fifty words in A.2.1 (nor.freq: 7830) and A.2.2 (nor.freq.: 8102) again because this is acquired later. Actually, it is one of the most frequently used words by native speakers, especially in spoken language (nor.freq.: 3288 in TNC).

Discussion

As confirmed by Şimşek (167), the vocabulary studies in the field of teaching Turkish as a foreign language mostly focus on the vocabulary used in printed materials. Nevertheless, the vocabulary used by the learners of the language should also be considered. For this reason, this study aims at listing the words used by learners with their frequency values and comparing the results with the vocabulary used in the course book the learners used and the TNC. In addition, the reasons for the differences in the frequency lists of learners in different categories were also tried to be accounted for. For these purposes, using the texts written by 123 A1 and A2 level learners of Turkish as a foreign language, a small-scale learner corpus of 2903 word types and 31897 word tokens in total was constructed. Based on this corpus, the list of the most frequently used words by learners was made. This list was then compared to the list of the most frequently used words in the course book the learners used to learn the language. Some of the differences in these two lists were attributed to one being the list of words learners used to answer certain prompts, and one being the list of words used in a coursebook designed to teach a language and thus included dialogues and instructions, which could not be seen in a written learner corpus.

In addition, the word frequency list of the learner corpus was also compared to the most frequently used fifty words by the native speakers of Turkish based on the lists provided in the TNC. The differences between these two lists were mainly attributed to the difference between the proficiency level of the participants; one group being learners at A1 and A2 levels, the other being native speakers of the language. As for the statistical analysis, only between these two lists was there a statistically significant difference that shows that native speakers of Turkish and learners of Turkish as a foreign language at CEFR levels A1 and A2 use these words at significantly different frequencies. Some of the differences in these two lists were also in accordance with the results from an L2 English learner corpus Corefl and BNC.

Also, separate lists based on certain criteria such as age, gender, and proficiency level were created. Statistical analysis revealed that there was no significant difference between the normalized frequency values of the most frequently used 50 words in the lists created for each group of learners: learners at different ages, female vs male learners and learners at different proficiency levels. When these lists were analyzed descriptively, it was concluded that the main factors leading to differences in the words used or the frequency values of those words observed between students in different categories were the proficiency level of the learners and the writing tasks that were assigned to them.

As stated by Karatay (256), "the vocabulary needs of learners are different from those whose native language is Turkish". Thus, studies such as this one will provide researchers, course book writers and teachers with the knowledge of which vocabulary items learners use at which level and how these differ from those used in the course books they use to learn the language, and from those used by native speakers. All these will indicate the needs of the learners; that is, what is lacking and on what more focus should be put in the curriculum. Using this information, curriculum designers could make changes in the curriculum, the syllabus, and the materials to meet those needs so that teachers could do so in the classroom when using those materials, and following those syllabi when teaching the language.

The limitations of this study were that the corpus was constructed using data collected at one single institution from students at the beginner levels only. It is believed that if this study is conducted at more institutions and with more students at higher proficiency levels, it will give a more thorough understanding of the vocabulary used by learners of Turkish as a foreign language, and how similar it is to that of native speakers. The results of such a study could also help to achieve the aims of constructing a Turkish learner corpus that will help learners as a resource during their studies. Utilizing such a corpus, teachers can adopt data-driven learning as a teaching methodology and design classroom activities using concordances that can be achieved through learner corpora and native speaker corpora. Through these, learners will have the opportunity to see sample uses of the target language by learners, and common mistakes made by learners, and to compare these to the language used by native speakers.

Such a learner corpus can also be a resource for a learner dictionary of Turkish, and for grade reader books in Turkish. Using such information as the active vocabulary and sentences used by learners at each different language level from the corpus, these resources could be divided into materials suitable for each CEFR level. To illustrate, a learner dictionary of Turkish at the CEFR level A and another one at the CEFR level B, or a grade reader book in Turkish for CEFR level B and another for CEFR level C could be prepared.

Works cited

Akman, Tuba. *Ortaöğretim Dil Ve Edebiyat Kitaplarında Kelime Sıklığı Çalışması*. 2013. Necmettin Erbakan University, Master's thesis.

Aksan, Yeşim, and Yılmaz Yaldır. "Türkçe Sözvarlığının Nicel Betimlemesi." 24. Ulusal Dilbilim Kurultayı, 17 May 2010, Ankara.

Aksan, Mustafa, and Demirhan, Umut Ufuk. "Bakmak Eylemi ve Söylem İşlevleri: Eşdizimlilik Örüntülerinin Gösterdikleri." *Dil ve Edebiyat Dergisi*, vol. 14, no. 2, July 2017, pp. 85–107. ded.mersindilbilim.info/tr/download/article-file/414576.

Aksan, Yeşim and Mustafa Aksan. "Türkçenin derlem-temelli sıklık sözlüğü: Temel ilkeler ve uygulama." *Uluslararası Sözlük Bilim Sempozyumu Bildiri Kitabı*, 2016.

Aksan, Yesim, et al. *A Frequency Dictionary of Turkish: Core Vocabulary for Learners (Routledge Frequency Dictionaries)*. 1st ed., Routledge, 2016.

The British National Corpus, version 2 (BNC World). Distributed by Oxford University Computing Services on behalf of the BNC Consortium, 2001, http://www.natcorp.ox.ac.uk. Accessed 1 August 2022.

Bulundu, Ahmet Cihan. *Türkçenin Yabancı Dil Olarak Öğretiminde Seviyelere Göre Öncelikli Söz Varlığı – Derlem Temelli Bir Uygulama*. 2022. İnönü University, Master's thesis.

Callies, Marcus, and Magali Paquot. "Learner Corpus Research: An Interdisciplinary Field on the Move." *International Journal of Learner Corpus Research,* vol. 1, no.1, 2015, pp. 1-6.

Çelebi, Hatice. *Extracting and Analyzing Impoliteness in Corpora: A Study Based on The British National Corpus and The Spoken Turkish Corpus*. 2012. Orta Doğu Teknik University, Master's thesis.

Çetin, Sadık Ahmet. *Suriye Uyruklu Ortaokul Öğrencilerinin Türkçe Yazılı Anlatımlarında Kelime Hazinesi*. 2017. Akdeniz University, Master's thesis.

Doğan, Faruk. *Türkçe Dersi Ortaokul Öğrenci Çalışma Kitaplarındaki Etkinliklerin Söz Varlığı Üzerine bir Araştırma.* 2016. Siirt University, PhD dissertation.

Ercan Güven, Aliye Nur. "Yabancılara Türkçe Öğretimi Üzerine Yapılmış Lisansüstü Tezlerin Çeşitli Değişkenler Açısından Analizi." *Türkçenin Eğitimi-Öğretimine Yönelik Çalışmalar*, edited by Mustafa Volkan Coşkun, et al., Pegem, 2015, pp. 2-12.

Göçen, Gökçen. *Yabancılar için Hazırlanan Türkçe Ders Kitaplarındaki Söz Varlığı ile Türkçeyi Yabancı Dil Olarak Öğrenenlerin Yazılı Anlatımlarındaki Söz Varlığı*. 2016. Sakarya University, PhD dissertation.

Göz, İlyas. *Yazılı Türkçenin Kelime Sıklığı Sözlüğü*. Türk Dil Kurumu Yayınları, 2003.

Granger, Sylviane. "A Bird's Eye View of Learner Corpus Research." *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching,* edited by

Sylviane Granger, Joseph Hung and Stephanie Petch-Tyson, Benjamins, 2002, pp. 3-33.

Granger, Sylviane, and Marie-Aude Lefer. "Corpus-based Translation and Interpreting Studies: A forward-looking Review." *Extending the Scope of Corpus-based Translation Studies. Bloomsbury Advances in Translation series,* edited by Sylviane Granger and Marie-Aude Lefer, Bloomsbury, 2022, pp. 13-41.

Gries, Stefan Th. and Berez, Andrea L. "Linguistic Annotation in/for Corpus Linguistics." *Handbook of Linguistic Annotation*, edited by Nancy Ide and James Pustejovsky, Springer, 2017, pp. 379-409.

Gündoğdu, Ayşe Eda. *Türk Çocuk Yazınında Söz Varlığı: Derlem Tabanlı bir Uygulama.* 2012. Mersin University, PhD dissertation.

Karadağ, Özay and Mehmet Kurudayıoğlu. "Türkçe Programına Göre Hazırlanmış İlköğretim Birinci Kademe Türkçe Ders Kitaplarının Kelime Hazinesi." XV. Ulusal Eğitim Bilimleri Kongresi, 13-15 September 2006, Muğla Üniversitesi, Muğla.

Karadağ, Özay. "Aynı Sınıf Düzeyi İçin Hazırlanan Türkçe Ders Kitaplarının Ortak Söz Varlığı Açısından Karşılaştırılması." *Ana Dili Eğitimi Dergisi,* vol. 7, no. 4, 2019, pp. 1130-1140.

Karaoğlu, Serdar. "Türkçe Çevrimiçi Derlemler Üzerine." *KMÜ Sosyal ve Ekonomik Araştırmalar Dergisi,* vol. 16, no. 1, 2014, pp. 181-188.

Karatay, Halit. "Türkçenin Yabancı Dil Olarak Öğretiminde Söz Varlığı ve Geliştirilmesi." *Türkçenin Yabancı Dil Olarak Öğretimi El Kitabı*, edited by Halit Karatay, Pegem Akademi, 2020, pp. 253-270.

Lozano, Cristobal, Ana Díaz-Negrillo and Marcus Callies. "Designing and Compiling a Learner Corpus of Written and Spoken Narratives: COREFL." *What's in a Narrative? Variation in Story-Telling at the Interface between Language and Literacy,* edited by Christiane Bongartz and Jacopo Torregrossa, Peter Lang, 2020, pp. 21-46.

O'Donnell, Matthew Brook. "Principles of corpus construction." University of Liverpool Corpus Linguistics Summer Institute, 2008, Lancaster.

Ölker, Gökhan. *Yazılı Türkçenin Kelime Sıklığı Sözlüğü*. Kömen Yayınları, 2011.

Pilav, Salim. *Üniversite Birinci Sınıf Öğrencilerinin Söz Varlığı Üzerine bir Araştırma*. 2008. Gazi University, PhD dissertation.

Sevinç, Melek. *Anahtar Sözcüklerin Söz Varlığını Geliştirmeye Etkisi*. 2008. Abant İzzet Baysal University, PhD dissertation.

Şimşek, Ramazan. "B1 Düzeyi Ders Materyallerinin Söz Varlığının ve B1 Düzeyi Öğrencilerin Yazılı Söz Varlığının Çeşitli Değişkenler Açısından Derlem Tabanlı İncelenmesi." *RumeliDE Journal of Language and Literature Studies,* vol. 7, 2020, pp. 152-171.

Şimşek, Ramazan, and Mesut Gün. "A Corpus-Based Proposal for the Vocabulary to be Taught at A1 Level in the Teaching of Turkish as a Foreign Language." *Educational Policy Analysis and Strategic Research,* vol. 16, no. 1, 2021, pp. 356-373.

Tognini-Bonelli, Elena. *Corpus Linguistics at Work.* John Benjamins Publishing, 2001.

Uslu, Kemalettin. *MEB 9. Sınıf Türk Edebiyatı Ders Kitabındaki Metinlerin Söz Varlığı Açısından Değerlendirilmesi.* 2010. Marmara University, PhD dissertation.