Journal of Language Education and Research

7(1), 2021

Journal of Language Education and Research

Research Article

A Lexicological Approach to Look-up Frequency of Turkish Sign Language Dictionary Users

Erdem Önal* **Bahtiyar Makaroğlu****

ARTICLE INFO

Received 14.08.2020
Revised form 09.04.2021
Accepted 13.04.2021
Doi:10.31464/jlere.780752

Keywords:
Log files
Corpus
Frequency list
Online dictionaries
Sign Language

ABSTRACT

From a lexicography perspective, the phenomenon of dictionary look-up frequencies can answer many critical questions on the dictionary users' needs and the relationship between corpus data and look-up tendencies. Although numerous studies have been carried out on the theoretical and functional aspects of look-up frequencies in the field of spoken language to date, there are relatively a few attempts to describe the characteristics of look-up behaviors in sign languages. This paper investigates the relationship between the objective frequency list of the Turkish Sign Language Corpus (Makaroğlu, 2021) and The Contemporary Turkish Sign Language Dictionary (Makaroğlu & Dikyuva, 2017) look-up occurrences. Using an approach to compare corpus data and log files, the results demonstrate that there are very limited similarities between the two and that there is no positive relationship for the percentage of the highest ranked items. In addition, the current study concludes that the corpus-based objective frequency list is a comparatively poor predictor of the dictionary look-up frequency and this difference could be direct result of the morphological structure of the words/signs in TİD (rather simultaneous realization) compared to Turkish (rather agglutinative morphology).

Acknowledgments The authors thank the two anonymous reviewers for their comments and suggestions and editors for their guidance in improving this paper.

Statement of Publication Ethics The authors declare that the current research was carried out according to the principles of the Declaration of Helsinki.

Authors' Contribution Rate E.Ö. and B.M. equally contributed to the design of the research. E.Ö. performed the data collection. B.M. developed the theoretical framework and drafted the manuscript. All authors discussed the results and commented on the manuscript.

Conflict of Interest The authors declare that there is no conflict of interest.

* Master Student, ORCID ID: https://orcid.org/0000-0002-5041-6915, Ankara University, Department of Turkish Sign Language and Deaf Studies, erdemonal22@gmail.com

** Res. Assist. Dr., ORCID ID: https://orcid.org/0000-0002-7641-6665, Ankara University, Department of Linguistics, makaroglu@ankara.edu.tr

ISSN: 2149-5602

Introduction

Today, the scope of dramatic changes in the digital media is broadening its limits, and this rapid evolution is beginning to affect lexicography including dictionary users and publishers (Lew & De Schryver, 2014). The fact that dictionaries are increasingly being proposed and accessed in digital format is clear. As part of these advances, conventional printed dictionaries are now being replaced with digital dictionaries that use various innovative tools on digital platforms. Thus, dictionary users can also benefit effectively from the opportunities provided by digital dictionaries (Tan & Woods, 2008). On the other hand, it is well known that although dictionaries have evolved considerably in terms of content and format in recent years and subfields such as pedagogical lexicography have emerged, they have found it difficult to meet higher-level information needs (Frankeberg-Garcia, 2020). Because of this, there is an increasing need for databases, including detailed annotated data, which can provide various characteristics with lexical and grammatical filtering. In addition, this digital transformation ensures that dictionaries are accepted as a widely used tool, unlike the publications accepted as authority (Lew & De Schryver, 2014). So, today's digital dictionaries being conceived as a lexical tool which provides users with information on all lexical aspects of the language is thought to support this view.

Along with studies on spoken languages, sign languages (henceforth SL) with visual-spatial modality also greatly benefit from the rapidly developing technology. Although SL dictionaries are one of the least-resourced language dictionaries (Prinsloo, 2012), certain modality-specific problems in the lexicography of SL have been rapidly solved thanks to the digital transformation of recent years (see: Zwitserlood, 2010; Mckee & Vale, 2017). Thus, multimedia SL dictionaries can serve both L1 and L2 sign language users. As it is well known, SL dictionaries serve two objectives: (i) to describe the language and protect and define the collected language data and (ii) to provide educational language information to individuals who wish to learn a SL (Schermer, 2006). Despite its significance in terms of language policy and education, there is no dictionary in which the basic core vocabulary of many SLs studied by researchers is described (Fenlon, Cormier & Schembri, 2015). Among the most significant factors for this are (i) the lack of a standard and accepted notation systems specifically to SLs, (ii) corpus studies being limited and still at the initial stage, and (iii) difficulties in the data collection and annotation.

Although look-up frequency studies for online dictionaries in spoken languages such as English, German and Danish (see: Bergenholtz & Johnson, 2005; Verlinde & Binon, 2010; Lew & Schryver, 2014; Koplening, Meyer & Müller-Spitzer, 2014; Müller-Spitzer, Wolfer & Koplening, 2015; De Schryver, Wolfer & Lew, 2019;) have made great progress in particular over the last 15 years, as far as is known, apart from the preliminary research on the New Zealand Sign Language (NZSL) Dictionary (Vale, 2015; Vale, 2017) there is no study on online sign language dictionaries carried out in this context in SL literature. Furthermore, there is no comprehensive study in literature on the usability of SL dictionaries (Kristoffersen & Troelsgård, 2012).

Literature review

Web-based online dictionaries not only create a more efficient environment for users, but also provides researchers with the opportunity to perform different types of search on log files where dictionary users' requests are saved. A logfile for an online dictionary is a machine-readable record automatically generated from the user's interaction with the website-based dictionary. It should also be noted that as far as printed dictionaries are concerned, electronic dictionaries are used more frequently and allow quicker access to the required information. As far as internet technology is concerned, the user is not so much the human trying to use the dictionary, but, in other words, the browser client that the human being user applies to communicating with the server hosting the dictionary. By examining the log files, it is possible to obtain important information about dictionary searches of users such as language structures on which they concentrate more, and the frequency of the searches performed (Tarp, 2009). In this way, log files provide complete information on how users access the existing online dictionary, as well as their contribution to the compilation of the dictionary. It also provides researchers with the possibility to record the actual behaviors of users in their natural environment, distinguishing it from other lexical methods (Müller-Spitzer, Wolfer & Koplening, 2015). However, the method is constrained because the scholar – as is the case for all observation methods – does not control the stages of the research.

What information online dictionary users are looking for is mainly of interest for lexicographers whose objective is generally to best meet their users' information needs. Knowing what users of the dictionary are interested in when a new dictionary is compiled is also critical. Following the functional perspective, this study examines the look-up properties and frequencies of Turkish Sign Language Dictionary users using the log files and compares these appearances with the objective lexical frequency list[1] of the Turkish Sign Language corpus (Makaroğlu, 2021). Similar to other studies in literature (see: Koplening, Mayer & Müller-Spitzer, 2014; Müller-Spitzer et al., 2015; De Schryver, Wolfer & Lew, 2019), the main purpose of this research is to describe the relationship between the TİD Corpus frequency and the dictionary look-up and to create a user look-up frequency list for potential new SL dictionaries. In the current study based on the log files on the look-up behaviors in the "written form" in the online The Contemporary Turkish Sign Language Dictionary (Makaroğlu & Dikyuva, 2017), (i) the relationship between the corpus-based word frequency list and the dictionary look-up frequency list, (ii) the effect of dictionary visits on the distribution of look-up patterns, and (iii) the power of the word category within the frequency list are analyzed.

[1] The studies carried out so far on objective word frequency lists of SLs is limited to only a few sign languages: the Swedish Sign Language word frequency list consists of 44,786 tokens collected from 42 signers (Börstell, Hörberg & Östling 2016); the British Sign Language word frequency list consists of 24,823 tokens collected from 249 signers (Fenlon, Schembri, Rentelis, Vinson & Cormier, 2014); the Australian Sign Language word frequency list consists of 55,859 tokens collected from 109 signers (Johnston, 2012); the New Zealand Sign Language word frequency list consists of 100,000 tokens collected from 80 signers (Mckee & Kennedy, 2006) and the American Sign Language word frequency list consists of 4,111 tokens collected from 27 signers (Morford & MacFarlane 2003).

The log file is a text file independent of the current server, that is entered by the web server the moment the user requests a source. In terms of lexicography, it is a type of record that stores the interaction of online dictionary searches of users, that can be read in digital environments and is created automatically. Log files for online dictionaries are machine-readable and automatically records the relationship between the online web-based dictionary and the user. By means of the linguistic information provided by these systematically stored digital data, the habits of online dictionary users are determined, and a guiding database is formed for new dictionaries to be created (Mayr, 2004). Moreover, due to the large data provided, log files provide a wide area of research for online behavioral studies of dictionary users (Nicholas et al., 1999). As mentioned above, in terms of technology, the term "user" is the browser client trying to communicate with the server hosting the dictionary, rather than the person using the dictionary (De Schryver et al., 2019). Therefore, interpreting the requests from each IP address as a single user can result to misleading data, and a general search pattern regardless of the IP address will provide more efficient results (see: Santos & Frankenberg-Garcia, 2007).

A comprehensive review of the relationship between dictionary searches and corpus frequency in literature has been carried out by De Schryver, Joffe, Joffe and Hillewaert (2006), and the findings show that the relationship between the corpus frequency and dictionary search frequency is weak. For this reason, even the frequency lists derived from natural language data are suggested to be a weak indicator in detecting the search tendencies of dictionary users. On the other hand, there are adverse opinions in literature, and in studies examining approximately 30 million dictionary searches in the online Danish dictionary, it is seen that the frequency pattern obtained from the corpus follows dictionary search lists significantly (Trap-Jensen, 2014; Trap-Jensen, Lorentzen & Sørensen, 2014). Therefore, the preliminary data to be obtained in this study are of considerable importance, particularly for subsequent comparative linguistic studies.

Methodology

The Contemporary Turkish Sign Language Dictionary

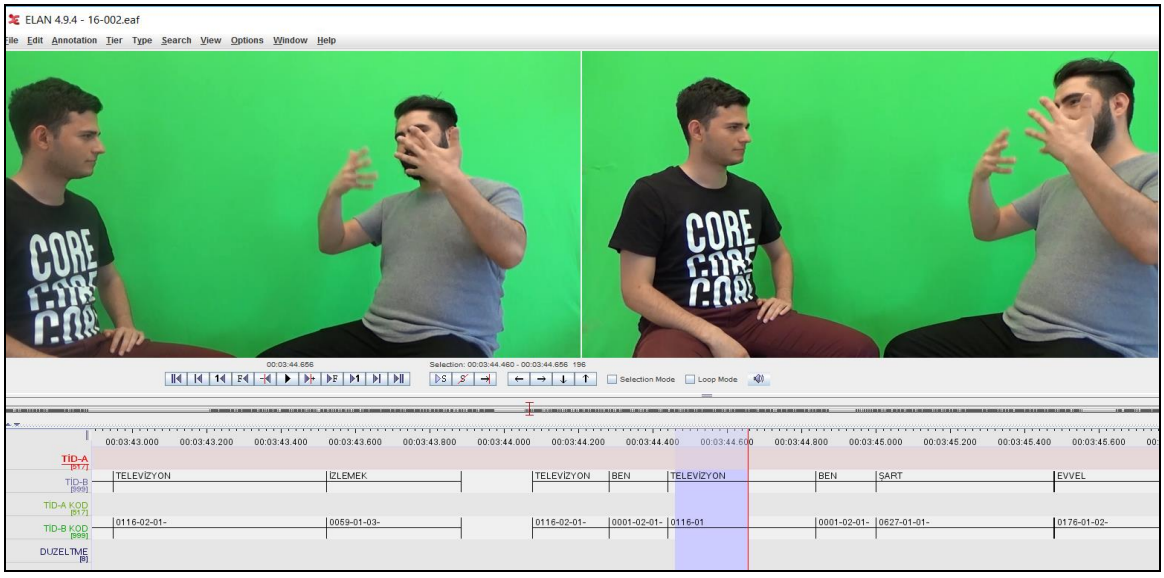
The Contemporary Turkish Sign Language Dictionary (Makaroğlu & Dikyuva, 2017) was developed using the Turkish Sign Language Corpus obtained from 116 native signers of TİD from 26 different provinces. The corpus contains about 104 hours of digital video recording in TİD. Following Schembri, Fenlon, Rentelis, Reynolds & Cormier (2013), to ensure the representation of the region in which deaf participants live, the participants in the corpus study were chosen from those who have been lived or worked in the region/city for 10 years or more. Moreover, individuals who were exposed to Turkish Sign Language at an early age (between 0 and 5 years) were selected in pursuit of the idea that the age of SL exposure in childhood can affect SL proficiency in their future (see: Emmorey, 2002; Cormier, Schembri, Vinson & Orfanidou, 2012). Due to both financial and methodological constraints, less than 30% of the corpus around 800,000 (approximately 240,000 sign tokens) has been annotated up until now (April 2021).

All the content of the dictionary developed in an online video environment is presented in Turkish Sign Language. For this reason, it is designed in both languages,

combining a mother tongue dictionary for the deaf and a translation dictionary for the hearing users. When looking at online SL dictionaries in general, the average number of lemmes is between 2000 and 5000 (Zwitserlood, 2010). In this regard, it is similar to other SL dictionaries in literature in terms of number of lemmes. Because of the number being most of the time more than ten thousand in spoken language dictionaries, it may be said that SL dictionaries are limited in this context. In terms of display type, the video format is seen as the most appropriate format for the nature of sign, in contrast to fixed images in which the movement is represented by arrows or the written form of the phonological realization of the movement.

This corpus-based dictionary consists of 2,000 words with the highest frequency in the Turkish Sign Language Corpus and all the data in the dictionary were annotated by deaf researchers using the ELAN annotation program (Sloetjes & Wittenburg, 2018). The data were first coded as Turkish words, and then in step two of the statistical analysis, they were annotated using the "triple sequential coding" system. As can be seen in figure (1), in this triple coding, (i) at the first step, all words were encoded according to their frequency number in the corpus, (ii) the second, each variable (ID-gloss) was numbered to define the variables of the signs and (iii) at the last step, the meaning of the relevant sign was determined from the context and each meaning was given a number (see: Makaroğlu, 2019). Thus, the relationship between the frequency of ID-gloss and the meaning of each sign acquires a statistically analyzable characteristic.

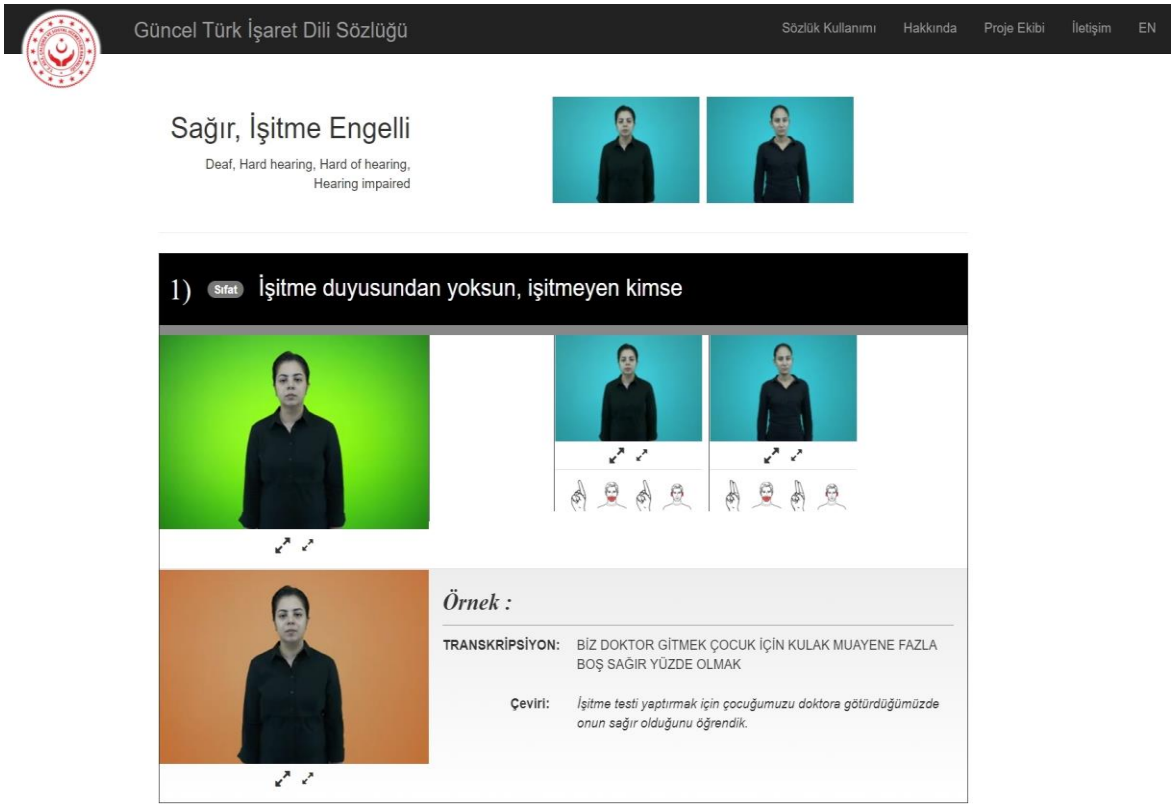
Figure 1. Coding Turkish Sign Language Dictionary data with "Triple Sequential Coding" (Makaroğlu, 2019)



There are three types of videos in the Contemporary Turkish Sign Language Dictionary such as ID-gloss, meanings and examples that total up to 11,428 videos, all videos are prepared using different background colors for each type. In the dictionary where the visual design is prepared entirely on the frequency, the lexical variables (ID-glosses) and meanings on the query screen are listed according to their frequency (Makaroğlu, 2019). Thus, the content and organization of the dictionary were determined

according to the linguistic usage of the native TİD signers living in different regions (see: Figure 2).

Figure 2. The interface of Contemporary Turkish Sign Language Dictionary: DEAF (http://tidsozluk.ailevecalisma.gov.tr)



As it is known, though there are many notation systems for SLs that have visual-spatial modalities suggested in the literature (e.g., The Stokoe Notation, Signwriting and the Hamburg Notation System–HamNoSys), the notation forms used remain in a phonological structure and mostly convey physical forms of the sign. (see for more information, Frishberg, Hoiting & Slobin, 2012). As a result of this restriction, particularly in lexical and grammatical studies, signs in the vocabulary are often shown with the closest semantic word referring to the spoken language in the same geography where sign language is used. As a result, when SL learners want to look at the dictionary for the meaning of a sign that they have encountered, it is important that they are able to complete their request using the phonological parameters (handshape, movement, location, orientation, nonmanual) of the sign (see: Brentari, 1998). In this context, the Contemporary Turkish Sign Language Dictionary includes 3 different search engines for a more effective access for users with different characteristics: (i) in the "word" search engine, the desired words/signs are listed by the nearest equivalent/translation into Turkish or English. (ii) a far as the "sign" search engine is concerned, the user can request the sign using the characteristics of the handshape and the location[2] that are among the

[2] According to the Prosodic Model (Brentari, 1998) handshape and location make up the core structure of words in SLs.

phonological parameters of the sign, (iii) in the "alphabetical" search engine, the Turkish and English equivalents of the word are enumerated in alphabetical order.

Queries made by dictionary users using three different search engines are stored in different categories in log files. Since one of the main purposes of the current study is to describe the relationship between the objective frequency list and the dictionary look-up frequency, the research was limited to only the data of the Turkish "word" search engine.

Data collection and analysis

The Contemporary Turkish Sign Language Dictionary stores all searches done using three different search engines into the log file. The data recorded in the Tab-Separated Value (TSV) format contains information such as the IP address of the requester, the time and date of the visit, and the word searched for. During the data collection, all 36-month searches carried out between May 2017, when the dictionary was published, and June 2020, were transferred to an Excel format and so the database comprised of 1,972,904 words using the word search engine was created. One of the greatest problems in analyzing raw log files is the use of typos and fonts for various languages. Consequently, each inaccurate dictionary search is encoded as a new single word and thus the number of single words increases unexpectedly. To resolve this problem, all the dictionary look-up frequency was scrutinized according to the determined criteria and necessary corrections were made (see: Example 1).

(1)

Raw Search	**Corrected Search**	**Type**
Tasarim	Tasarım	Foreign characters
Temmux	Temmuz	Typing error
Geece	Gece	Extra letter
Herşy	Her ey	Missing letter
İlahiyat5	İlahiyat	Adding an extra number
Kilo,	Kilo	Adding an extra punctuation
Kangrubu	Kan grubu	Incorrect ligatures
Kayın peder	Kayınpeder	Incorrect separate spelling

Results

In lexicographic research, the main aim of the frequency analysis created using a corpus is to identify the candidate lemma to be included in general use dictionaries (De Schryver, 2013; Hanks, 2012). If the candidate lemma in the frequency list can exceed the specified frequency limit, they may be included in the dictionary. Thus, words with more frequent appearances are considered more important than words with less frequent appearances. As seen in Table (1), when looking at the distribution of searches per 1 million visits, while there are 75,579 types in the total corpus, the top 5,000 types cover nearly 90% of the total database. The results provided support previous studies as a small number of types represent the large proportion of the corpus (see: Koplenig et al., 2014; Müller et al., 2015).

Table 1. Distribution of TİD Dictionary Searches per 1 Million Visits

Category (per 1 million visits)	The percentage of the dictionary log files
1	%0,4
2-10	%2,0
11-49	%6,1
50-500	%33,3
501-5000	%47,2
5000+	%10,8
Total	%100

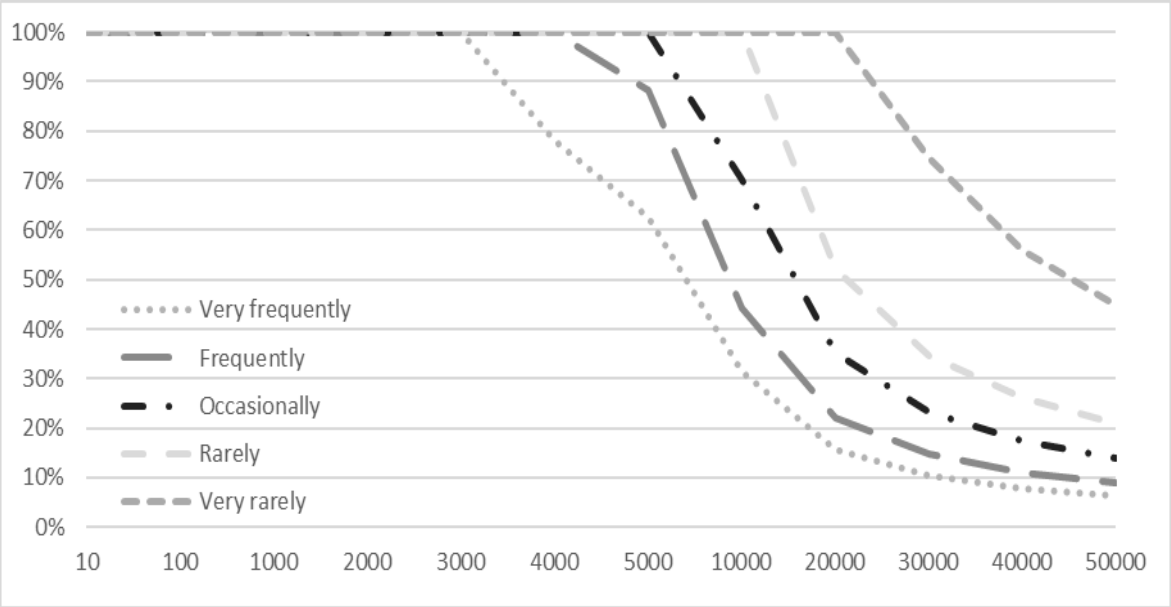
To better interpret the log files and analyze the gradual distribution more clearly, it is often necessary to divide the frequency rates into categories (see: Koplenig et al., 2014). In this sense, lookups which appear at least 50 times out of a million are defined as “very frequently”, those between 25-49 are defined as “frequently,” those between 10-24 are defined as “occasionally,” and those between 5-9 are defined as “rarely,” and those between 2-4 are defined as “very rarely” (see: Table 2). While no specific classification method has been followed, different classifications are observed in the various studies. For example, Müller et al., (2015) classified a view per visit into 1 million, "regularly"; 2 views were classified as "frequently" and more than 11 views as "very frequently". In the current study, classification containing larger numbers was preferred due to the size of the database and the appearance of the vocabulary of two languages with different modalities.

Table 2. The Proportion of Very Rarely, Rarely, Occasionally, Frequently and Very Frequently Visited Entries and The Relationship to the Number of Dictionary Lookup Ranks

Rank	Very Rarely (>2)	Rarely (>5)	Occasionally (>10)	Frequently (>25)	Very Frequently (>50)
10	%100	%100	%100	%100	%100
100	%100	%100	%100	%100	%100
1000	%100	%100	%100	%100	%100
2000	%100	%100	%100	%100	%100
3000	%100	%100	%100	%100	%100
4000	%100	%100	%100	%100	%78,12
5000	%100	%100	%100	%88,16	%62,52
10000	%100	%100	%70,13	%44,08	%31,26
20000	%100	%52,06	%35,07	%22,04	%15,63
30000	%74,44	%34,71	%23,38	%14,69	%10,42
40000	%55,83	%26,03	%17,53	%11,02	%7,82
50000	%44,66	%20,82	%14,03	%8,82	%6,25

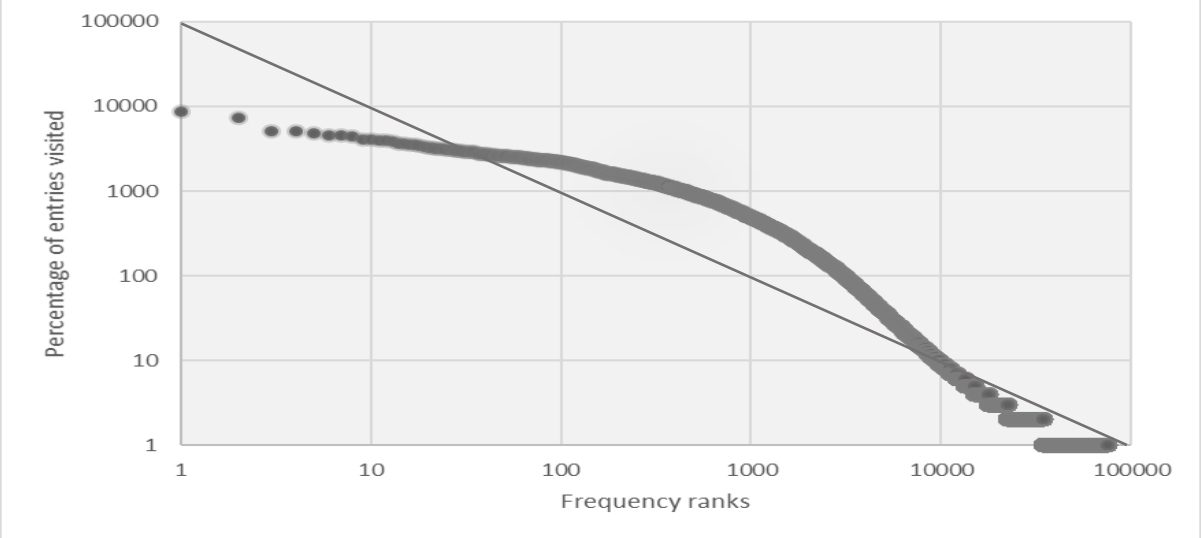
As shown in Table (3), it is seen that the most frequent 2,000 words have more than 100 searches in the corpus, and the decrease across all categories is 10,000 to 20,000. Considering that frequency-based dictionaries are typically based on top ranked 2,000 words, all the lexical units which have the appearance of "very frequently" may be seen in this category. This analysis demonstrated that by digging deeper into the frequency list, apparently more and more less popular items are also being looked up and the percentages of the five classes gradually decline.

Table 3. The Proportion of the Visited Entries in the Online TİD Dictionary



One of the interesting characteristics of human language is the systematic distribution of the frequency of words. This mathematical aspect is commonly described as Zipf's power law (Zipf, 1945/1949). There exists a systematic and reliable frequency relation in the vocabulary of human languages and the distribution of word frequencies for random generated texts is very similar to Zipf's law. This law states that the most frequent word is about twice the frequency of the second order word, and it is about 3 times more than the third most frequent word. Although the distribution of the objective frequency list of the Turkish Sign Language Corpus roughly follows Zipf's law (1st word 8398 tokens, 2nd word 5729 tokens, 3rd word 1494 tokens) (Makaroğlu, 2021), it is noted that when the word frequency of dictionary searches is examined in the log-log scale, as seen in table (4), it does not clearly follow Zipf's law (1st word, "para" (money) 8696 tokens, 2nd word "anne" (mother) 7381 tokens, 3rd word, "en" (most) 5136 tokens).

Table 4. Log-log Scale Distribution of Word Frequency in the Database Consisting of TİD Dictionary Searches



When considering the look-up frequency of the dictionary consisting of 1,972,904 tokens and 75,579 types, it should be noted that the resulting list of frequencies is similar to the objective of the Turkish Sign Language frequency list (Makaroğlu, 2021) in limited aspects. However, there are so many points on which it differs from one another. According to Makaroğlu (2021), the top 100 words/ID-glosses with the highest frequency on the frequency list account for 52.64% of the corpus. However, in this study, the top 100 entries visited in the dictionary cover only 14.84 of the dictionary look-up corpus. Also, while the most frequent words in the objective frequency list are mostly functional words such as pointing having a deictic and discursive function, it is clear that the content words are at the top of the dictionary look-up frequency. Table (5) lists the first 20 words that appear most frequently in dictionary look-ups (Appendix shows the top 300 entries).

Table 5. The Most Frequently Searched 20 Words in the Contemporary Turkish Sign Language Dictionary (approximately 4.5% of the total searches)

Rank	Word (Turkish)	Word (English)	N	Per 1000 (N)	Database (%)	Cum. (%)
1	Para	Money	8696	4.41	0,44	0,44
2	Anne	Mother	7381	3.74	0,37	0,81
3	En	Most	5136	2.60	0,26	1,08
4	Merhaba	Hello	5034	2.55	0,26	1,33
5	Baba	Father	4744	2.40	0,24	1,57
6	Gün	Day	4608	2.34	0,23	1,80
7	Ama	But	4515	2.29	0,23	2,03
8	İstemek	To want	4409	2.23	0,22	2,26
9	Her	Every	4111	2.08	0,21	2,47
10	Var	Have	4021	2.04	0,20	2,67
11	Aile	Family	3930	1.99	0,20	2,87
12	Yapmak	To do	3927	1.99	0,20	3,07
13	Ne	What	3883	1.97	0,20	3,26
14	Olmak	To be	3653	1.85	0,19	3,45
15	Bilmek	To know	3651	1.85	0,19	3,63
16	Almak	To take	3549	1.80	0,18	3,81
17	Okul	School	3500	1.77	0,18	3,99
18	Nasılsın	How are you	3469	1.76	0,18	4,17
19	Çok	Many	3313	1.68	0,17	4,34
20	Çünkü	Because	3275	1.66	0,17	4,50
21	Renk	Color	3205	1.62	0,16	4,66
22	Öğretmen	Teacher	3201	1.62	0,16	4,83
23	Hep	All	3174	1.61	0,16	4,99
24	Günaydın	Good morning	3137	1.59	0,16	5,15
25	Yok	No	3105	1.57	0,16	5,30
26	Ve	And	3080	1.56	0,16	5,46
27	Abla	Sister (older)	3072	1.56	0,16	5,62
28	Etmek	To make	2996	1.52	0,15	5,77
29	Amca	Uncle	2991	1.52	0,15	5,92
30	Son	Last	2980	1.51	0,15	6,07
31	Teyze	Aunt	2938	1.49	0,15	6,22
32	Hiç	None	2890	1.46	0,15	6,36
33	Nasıl	How	2888	1.46	0,15	6,51
34	Ak	White	2873	1.46	0,15	6,66

35	Arkadaş	Friend	2865	1.45	0,15	6,80
36	Ben	I	2818	1.43	0,14	6,94
37	Üniversite	University	2779	1.41	0,14	7,09
38	Gelmek	To come	2775	1.41	0,14	7,23
39	Sevmek	To like	2764	1.40	0,14	7,37
40	Zaman	Time	2764	1.40	0,14	7,51
41	Doğru	Right	2750	1.39	0,14	7,65
42	Başlamak	To start	2673	1.35	0,14	7,78
43	Vermek	To give	2669	1.35	0,14	7,92
44	Anlamak	To understand	2666	1.35	0,14	8,05
45	Kalmak	To stay	2649	1.34	0,13	8,19
46	İyi	Good	2627	1.33	0,13	8,32
47	Sonra	After	2621	1.33	0,13	8,45
48	Ders	Lesson	2615	1.33	0,13	8,58
49	Neden	Why	2590	1.31	0,13	8,72
50	Önemli	Important	2590	1.31	0,13	8,85
51	Kim	Who	2582	1.31	0,13	8,98
52	Pazar	Sunday	2578	1.31	0,13	9,11
53	Boş	Empty	2569	1.30	0,13	9,24
54	Ocak	January	2551	1.29	0,13	9,37
55	İle	With	2544	1.29	0,13	9,50
56	Ay	Month	2505	1.27	0,13	9,62
57	İş	Work	2498	1.27	0,13	9,75
58	Çalışmak	To work	2496	1.27	0,13	9,88
59	Gitmek	To go	2490	1.26	0,13	10,00
60	Üzülmek	To be sad	2489	1.26	0,13	10,13
61	Bulmak	To find	2488	1.26	0,13	10,26
62	Abi	Brother (older)	2486	1.26	0,13	10,38
63	Dayı	Uncle	2436	1.23	0,12	10,51
64	Kadar	To	2435	1.23	0,12	10,63
65	Aramak	To search	2433	1.23	0,12	10,75
66	Evet	Yes	2423	1.23	0,12	10,87
67	Öğrenmek	To learn	2421	1.23	0,12	11,00
68	Çocuk	Child	2408	1.22	0,12	11,12
69	Güzel	Beautiful	2405	1.22	0,12	11,24
70	Sarı	Yellow	2395	1.21	0,12	11,36
71	Konuşmak	To talk	2393	1.21	0,12	11,48
72	Öğrenci	Student	2392	1.21	0,12	11,61
73	Yeni	New	2356	1.19	0,12	11,72
74	Aşk	Love	2355	1.19	0,12	11,84
75	Zor	Difficult	2343	1.19	0,12	11,96
76	İnanmak	To believe	2342	1.19	0,12	12,08
77	Ev	House	2341	1.19	0,12	12,20
78	Şimdi	Now	2332	1.18	0,12	12,32
79	Kolay	Easy	2329	1.18	0,12	12,44
80	Hasta	Ill	2323	1.18	0,12	12,55
81	Eylül	September	2319	1.18	0,12	12,67
82	Geri dönmek	To return	2318	1.17	0,12	12,79
83	Yaş	Age	2307	1.17	0,12	12,91
84	Kardeş	Sibling	2300	1.17	0,12	13,02
85	Akraba	Relative	2298	1.16	0,12	13,14
86	Yarın	Tomorrow	2298	1.16	0,12	13,26

87	Kaç	How much	2286	1.16	0,12	13,37
88	Yıl	Year	2283	1.16	0,12	13,49
89	Aynı	Same	2279	1.16	0,12	13,60
90	Yanlış	Wrong	2264	1.15	0,11	13,72
91	Çıkmak	To get out	2256	1.14	0,11	13,83
92	Görmek	To see	2245	1.14	0,11	13,95
93	Erken	Early	2239	1.13	0,11	14,06
94	Anlatmak	To explain	2225	1.13	0,11	14,17
95	Bu	This	2225	1.13	0,11	14,29
96	Geçmek	To pass	2213	1.12	0,11	14,40
97	Bazen	Sometimes	2212	1.12	0,11	14,51
98	Doktor	Doctor	2210	1.12	0,11	14,62
99	Bugün	Today	2205	1.12	0,11	14,73
100	Mavi	Blue	2204	1.12	0,11	14,84

Examining Table (5) it can be seen that 14 of the top 100 visited are functional words (eg., *but, because, I* etc.). While 22 of the 100 most frequent words in the objective Turkish Sign Language Frequency list are functional words (Makaroğlu, 2021), 24 of the 100 most frequent words in the Turkish Frequency list (Aksan, Aksan, Mersinli & Demirhan, 2016), 56 of the 100 most frequent words in the British National Corpus (Leech, Rayson & Wilson 2001) and 24 of the 100 most frequent words in the Australian Sign Language Frequency list (Johnston, 2012) are classified as functional words. This distribution implies that morphological formation has a direct effect on the distribution of the word categorization (i.e., functional or content) in the frequency lists of spoken and sign languages. It is expected that Turkish, which is one of the most typical examples of agglutinative languages (Göksel & Kerslake 2005), and SLs where morphological categories are coded as simultaneous morphology (Aronoff, Meir & Sandler, 2005), are expected to be distinctly different from English. In other words, because of modality-specific aspects, SLs seem to favor simultaneous internal modification of signs, rather than concatenation of morphemes. The most typical examples of this are classifying structures and verb agreement, which often exhibit simultaneous formation rather than different linguistic units.

As is well known, log-likelihood is widely used in corpus linguistics research to test for significant difference in language databases of two different sizes (Rayson & Garside, 2000). Likewise, in this study, the most common words/ID-glosses in both databases were compared in terms of log-likelihood (LL) ratio, since the corpus consisting of dictionary searches is approximately 19 times larger than the TİD corpus (see: Table 6). Based on the 6.63 LL ratio ($p < 0.01$), there are statistically great significant differences between the 10 highest words found in both corpuses that share the highest ranks. This suggests that the statistical distribution of the signs/words included in the frequency list of native TİD signers are distinctly different from the look-up frequency.

Table 6. Comparison of the 10 Most Frequent Words Sharing the Highest Ranks in the Look-up Frequency List and the Objective Frequency List in Terms of Log-likelihood Ratio

Rank	Word	Dictionary Corpus *(n = 1.972.904)*	TİD Corpus *(n = 103.087)*	
		n	*n*	*LL*
1	Have	4021	1133	471820.59
2	To do	3927	661	193263.42
3	What	3883	939	371882.59
4	To be	3653	1088	512107.81
5	To know	3651	1157	559856.88
6	Always	3174	258	21902.38
7	No	3105	1032	98514.44
8	Never	2890	455	168188.45
9	Friend	2865	328	76856.45
10	I	2818	8398	4690025.00

Comparing the distribution patterns of the frequency lists brings new findings, particularly in terms of comparative studies. Look at Table (7)[3] which shows the first 20 words of the dictionary look-up frequency list and the objective frequency list (Makaroğlu, 2021), it is clear that only five words have taken place in both lists and that these words also have significant differences in terms of occurrence. For instance, while pointing (PT) using as a function word (e.g., pronoun, location, etc.) in the core lexicon of SLs (Brentari & Padden, 2001) is at the top of the objective TİD frequency list, it is not included in the look-up frequency list given.

Table 7. Frequency List of the Top 20 Words in the Two Corpora

Rank	Dictionary Corpus *(n = 1.972.904)*			TİD Corpus *(n = 103.087)*		
	Word	n	%	ID-gloss	n	%
1	Money	8696	0,44	PT:PRONOUN1	8398	8,1%
2	Mother	7381	0,37	PT:PRONOUN2/3	5729	5,6%
3	Most	5136	0,26	PT:LOCATION	1494	1,4%
4	Hello	5034	0,26	GO	1333	1,3%
5	Father	4744	0,24	TO KNOW	1157	1,1%
6	Day	4608	0,23	HAVE	1133	1,1%
7	But	4515	0,23	TO BE	1088	1,1%
8	To want	4409	0,22	NO	1032	1,0%
9	Every	4111	0,21	WHAT	939	0,9%
10	Have	4021	0,20	DEAF1/2	873	0,8%
11	Family	3930	0,20	TO UNDERSTAND	844	0,8%
12	To do	3927	0,20	TO LOOK	801	0,8%
13	What	3883	0,20	GOOD	783	0,8%
14	To be	3653	0,19	TO SAY	759	0,7%
15	To know	3651	0,19	SIGN	705	0,7%
16	To take	3549	0,18	YES1	687	0,7%
17	School	3500	0,18	TO GIVE	685	0,7%
18	How are you	3469	0,18	TO COME	677	0,7%

[3] The areas in grey show the words that are found in both corpus lists

19	Many	3313	0,17	TO DO1/2	661	0,6%
20	Because	3275	0,17	FOR	589	0,6%

When Table (8), which presents the highest 10 verbs in both frequency lists is examined, it is observed that 5 words are included in the two lists. Considering that the Turkish light verb "*yapmak*" (to make) (see: Akşehirli, 2013) which does not have its exact equivalent in TİD, and the lexical sign "FINISH" (see: Karabüklü, 2016), which has the aspect role in TİD are included in the two distinct lists. In fact, it is clearly seen that many of the overlapping verbs that have distinct functional roles in two lexicons. Besides, there are substantial differences in the proportion of the verbs in the frequency lists. For instance, while the word "to start", which is ranked 10th in the dictionary look-up list in terms of verb category, it is classified 42nd on the entire frequency list. As for the verb "FINISH" which is ranked 10th in the objective frequency list in terms of verb category but it is ranked 25th in the full frequency list. This proportional difference shows that contrary to the dictionary look-up list, words in the verb category are listed above in the objective frequency list in TİD.

Table 8. The Frequency of the Top 10 Verbs in Two Corpuses

N	Dictionary Look-up List *(n = 1.972.904)*				Objective Frequency List *(n = 103.087)*			
	Rank	Verb	% Database	% Cum.	Rank	Verb	% Database	% Cum.
1	8	To want	0.22%	0.22%	4	GO	1,29%	1,29%
2	12	To do	0.20%	0.42%	5	KNOW	1,12%	2,41%
3	14	To be	0.19%	0.61%	7	TO BE	1,06%	3,47%
4	15	To know	0.19%	0.80%	11	UNDERSTAND	0,82%	4,29%
5	16	To take	0.18%	0.98%	12	LOOK	0,78%	5,07%
6	28	To make	0.15%	1.13%	14	SAY	0,74%	5,81%
7	38	To come	0.14%	1.27%	17	GIVE	0,66%	6,47%
8	39	To like	0.14%	1.41%	18	COME	0,66%	7,13%
9	42	To start	0.14%	1.55%	19	DO1/2	0,64%	7,77%
10	43	To give	0.14%	1.69%	25	FINISH	0,47%	8,24%

When looking at Table (9), which presents the highest 10 nouns in two frequency lists, is examined, it is observed that only 4 words are included in both lists. Though the overlap of the words in the noun category in both lists is relatively less than the overlap in the verb category, the place of the words in the 10th rank in the noun category in the full frequency list appears to be more similar.

Table 9. The Frequency of the First 10 Words in the Noun Category in Both Corpuses

N	Dictionary Corpus *(n = 1.972.904)*				TİD Corpus *(n = 103.087)*			
	Rank	Noun	% Database	% Cum.	Rank	Noun	% Database	% Cum.
1	1	Money	0,44%	0,22%	10	DEAF1/2	0,85%	0,85%
2	2	Mother	0,37%	0,81%	15	SIGN	0,68%	1,53%
3	5	Father	0,24%	1,05%	21	FATHER	0,54%	2,07%
4	11	Family	0,20%	1,25%	31	MOTHER	0,43%	2,50%
5	17	School	0,18%	1,43%	32	FOOD	0,42%	2,92%
6	21	Colour	0,16%	1,59%	33	PERSON	0,42%	3,34%
7	22	Teacher	0,16%	1,75%	34	SCHOOL	0,41%	3,75%
8	27	Sister	0,16%	1,91%	35	NAME1/2	0,41%	4,16%

9	29	Uncle	0,15%	2,06%	37	CAR	0,40%	4,56%
10	30	Last	0,14%	2,20%	39	MONEY	0,39%	4,95%

From a sociolinguistic perspective, there are expected to be certain differences in cultural norms between the Deaf and the hearing society (see: Lucas, 2001). Because of the modality of SL and the distinctive appearances of Deaf society, these characteristics also have a place in the SL lexicon. In the study by Makaroğlu (2021), it was suggested that these cultural characteristics account for a large proportion in terms of objective frequency in Turkish Sign Language: the sign DEAF 1/2 in the 10th rank represents 0.85 of the corpus with 873 tokens, and the 15th sign SIGN has 705 tokens, as for the sign INTERPRETER in the 99th rank makes up 0.15 of the corpus with 159 types. In this study examining the dictionary look-up behavior, the word "deaf" in the 771st rank was found to make up 0.032 of the corpus with 638 tokens, the word "hearing impaired" in the 852nd rank with 575 tokens which representing 0.29 of the corpus, the word "interpreter" at the 557th rank, with 849 tokens consisting of 0.043 of the corpus and the word "sign" in the 142nd rank made up 0.09 of the corpus with 1853 token appearances. The fact that substantial differences in terms of the ratio of words related to Deaf Culture in both frequency lists suggest that the Contemporary Turkish Sign Language Dictionary is used more extensively by the hearing society.

Discussion

In recent years, extensive research on spoken languages has been carried out on the behaviors and trends of users of the online dictionary (see: Frankenberg-Garcia, 2020; De Schryver et al., 2019; Koplenig et al., 2014). The main objective of these studies is to effectively determine the users' needs for the ideal dictionaries and learn about the user's profiles. For lexicographers, the most important contribution of examining the log files is that they offer the opportunity to test whether users of the dictionaries they create are using it efficiently. Hence, the newly prepared online dictionaries earned an interactive feature that can be shaped according to the tendencies of the users, contrary to a distribution and content envisaged only by the lexicographers. Under conventional methods, the development of dictionaries has been mostly based on the intuitions of lexicographers. While the dictionary user is in passive position, only the dictionary editor/researcher is in active position. Thus, the meaning of influence is from the dictionary editor to the dictionary use. However, the use of modern methods such as log files in the preparation of the online dictionary ensures a bidirectional relation. On the other hand, examining the log system files from another aspect provides scholars with a very limited method, and the researcher has no control over the data collection process. So, it gives limited access on the language profile and demographic characteristics of dictionary users[4]. Consequently, it is important to be aware of the limits when analyzing a log file.

The findings of this study show that when the distribution of dictionary look-up per million visits is taken as a reference, the first 5,000 types cover almost 90% of the entire

[4] Website performance monitoring services such as Google Analytics provide demographic information such as age and geography via IP and Google account.

corpus, even though there are 75,579 types in the total corpus. Secondly, from a statistical point of view, although descriptive statistical methods are useful for presenting the overall distribution, they may have numerous problems, particularly in corpus frequency data. (see: Baayen, 2008: 229). The most prominent of these is that the lexical units that appear in very limited numbers in the frequency lists occupy a very large ratio in terms of proportion of types. Similarly, in this study, 41,235 types were observed only once in the dictionary look-up list. Although these queries cover only 2.09% of the tokens in the dictionary look-up corpus, it accounts for 54.56% of the types. In terms of word distribution, according to Makaroğlu (2021), the top 100 words in the objective frequency list make up 52.64% of the corpus, however, in the current study, the highest 100 words in the online dictionary only constitute 14.84 of the look-up frequency list. This condition is supposed to be that Turkish having a typically agglutinating morphology offers many word formations (e.g., *içinde, içindeki, içini, içinizde, içinize* vb.).

Conclusion

As mentioned above, corpus-based objective word frequency lists provide a very poor estimate for determining the characteristics of dictionary look-up behaviors (see: De Schryver & Joffe, 2004; Verlinde & Binon, 2010). It is also assumed that the reason for this difference may be due to the morphological structure of the sign in Turkish Sign Language (which allows for a more simultaneous appearance) being different from Turkish (which contains a more agglutinating morphology). On the other hand, the results of the study show that online dictionary users do not look-up more frequently for high-frequency words in the objective frequency list based on naturalistic data of the native TİD signers. Although a more detailed investigation is needed, since different groups of dictionary users have different needs, important lexicographic questions cannot be answered with log files alone. However, it would help researchers improve online dictionaries.

References

Aksan, Y., Aksan, M., Mersinli, Ü. & Demirhan, U. U. (2016). *A frequency dictionary of Turkish*. Taylor & Francis. https://doi.org/10.4324/9781315733302

Akşehirli, S. (2013). Türkçede et-katkısız eyleminin sözlüksel işlevleri [Lexical functions of the light verb -et in Turkish]. *Electronic Turkish Studies*, 8(9). https://doi.org/10.7827/TurkishStudies.5476

Baayen. R. H. (2008). Analyzing Linguistic Data. *A Practical Introduction to Statistics Using R*. Cambridge. UK: Cambridge University Press.

Börstell, C., Hörberg, T. & Östling, R. (2016). Distribution and duration of signs and parts of speech in Swedish Sign Language. *Sign Language & Linguistics*, 19(2), 143–196. https://doi.org/10.1075/sll.19.2.01bor

Bergenholtz, H. & Johnson, M. (2005). Log files as a tool for improving internet dictionaries. *Hermes*, 34. 117–141. https://doi.org/10.7146/hjlcb.v18i34.25802

Brentari, D. (1998). *A prosodic model of sign language phonology*. Cambridge. MA: MIT Press.

Brentari, D. & Padden, C. (2001). Native and foreign vocabulary in American Sign Language: a lexicon with multiple origins. In D. Brentari (Ed.), *Foreign vocabulary in sign languages* (pp. 87–120). Mahwah, NJ: Lawrence Erlbaum Associates.

Cormier. K., Schembri, A., Vinson, D. & Orfanidou, E. (2012). First language acquisition differs from second language acquisition in prelingually deaf signers: Evidence from sensitivity to grammatical judgement in British Sign Language. *Cognition*, 124(1). 50–65. https://doi.org/10.1016/j.cognition.2012.04.003

De Schryver, G. M. (2013). Tools to support the design of a macrostructure. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (Eds.), *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography* (pp. 1384–1395). Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110238136.1384

De Schryver, G. M. & Joffe, D. (2004). On how electronic dictionaries are really used. In G. Williams & S. Vessier (Eds.), *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6-10, 2004* (pp. 187–196). Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

De Schryver, G. M., Joffe, D., Joffe, P. & Hillewaert, S. (2006). Do dictionary users really look up frequent words? – On the overestimation of the value of corpus-based lexicography. *Lexikos*, 16. 67–83. https://doi.org/10.5788/16-0-648

De Schryver, G. M., Wolfer. S. & Lew, R. (2019). The relationship between dictionary look-up frequency and corpus frequency revisited: A log-file analysis of a decade of user ınteraction with a Swahili-English Dictionary. *GEMA Online Journal of Language Studies*, 19. 1–27. https://ejournal.ukm.my/gema/article/view/34284/9995

Emmorey, K. D. (2002). *Language cognition and the brain: Insights from sign language research*. Mahwah. NJ: Lawrence Erlbaum Associates.

Fenlon, J., Schembri, A., Rentelis, R., Vinson, D. & Cormier, K. (2014). Using conversational data to determine lexical frequency in British Sign Language: The influence of text type. *Lingua*, 143. 187–202. https://doi.org/10.1016/j.lingua.2014.02.003

Fenlon, J., Cormier, K. & Schembri, A. (2015). Building BSL SignBank: The lemma dilemma revisited. *International Journal of Lexicography*, 28(2), 169–206. https://doi.org/10.1093/ijl/ecv008

Frankenberg-Garcia, A. (2020). Combining user needs, lexicographic data and digital writing environments. *Language Teaching*, 53(1), 29–43. https://doi.org/10.1017/S0261444818000277

Frishberg. N., Hoiting. N. & Slobing. D. (2012). Transcription. In R. Pfau, M. Steinbach, & B. Woll (Eds.). *Sign Language. An international handbook* (pp. 1045–1075). Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110261325.1045

Hanks, P. (2012). Corpus evidence and electronic lexicography. In S. Granger, & M. Paquot (Eds.), *Electronic Lexicography* (pp. 57–82). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199654864.001.0001

Johnston, T. (2012). Lexical frequency in sign languages. *Journal of Deaf Studies and Education* 17, 163–193. https://doi.org/10.1093/deafed/enr036

Karabüklü, S. (2016). *Time and aspect in Turkish Sign Language (TİD): Manual and nonmanual realization of 'FINISH'* [Unpublished master's thesis]. Boğaziçi University, İstanbul.

Koplenig, A., Mayer, P. & Müller-Spitzer, C. (2014). Dictionary users do look up frequent words. A log file analysis. In C. Müller-Spitzer (Ed.). *Using Online dictionaries* (pp. 229–250). Berlin, New York: de Gruyter. https://doi.org/10.1515/9783110341287.229

Kristoffersen, J. H. & Troelsgård, T. (2012). The electronic lexicographical treatment of sign languages: The Danish Sign Language Dictionary. In S. Granger, & M. Paquot (Eds.), *Electronic Lexicography* (1st ed., pp. 293–318). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199654864.001.0001

Leech, G., Rayson, P. & Wilson, A. (2001). *Word frequencies in written and spoken English based on the British national corpus*. London: Longman.

Lew. R. & De Schryver. G. M. (2014). Dictionary users in the digital revolution. International *Journal of Lexicography*. 27(4). 341–359. https://doi.org/10.1093/ijl/ecu011

Lucas, C. (Ed.). (2001). *The sociolinguistics of sign languages*. Cambridge University Press.

Makaroğlu, B. (2021-in press). What the frequency list can teach us about Turkish Sign Language?. *Poznań Studies in Contemporary Linguistics*, 56 (4).

Makaroğlu. B. (2019, February 1). *Methodological issues in the corpus based Turkish Sign Language dictionary*. LSC Corpus Group Meetings. University of Pompeu Fabra, Barcelona, Spain.

Makaroğlu. B. & Dikyuva. H. (2017). *Güncel Türk İşaret Dili Sözlüğü*. Ankara: Aile ve Sosyal Politikalar Bakanlığı. http://tidsozluk.ailevecalisma.gov.tr

Mayr, P. (2004). Website entries from a web log file perspective – a new log file measure. In *Proceedings of the AoIR-ASIST 2004 - Workshop on Web Science Research Methods*.

McKee, R. & Vale, M. (2017). Sign language lexicography. In P. Hanks & G. De Schryver (Eds.), *International Handbook of Modern Lexis and Lexicography* (pp. 1–22). Germany: Springer-Verlag GmbH.

McKee, D. & Kennedy, G. (2006). The distribution of signs in New Zealand Sign Language. *Sign Language Studies*, 6(4). 372–390. http://www.jstor.org/stable/26191007

Morford, J. & MacFarlane, J. (2003). Frequency characteristics of American Sign Language. *Sign Language Studies*, 3(2). 213–225. http://www.jstor.org/stable/26204871

Müller-Spitzer. C., Wolfer, S. & Koplenig, A. (2015). Observing online dictionary users: studies using wiktionary log files. *International Journal of Lexicography*, 28. 1–26. https://doi.org/10.1093/ijl/ecu029

Nicholas, D., Huntington, P., Williams, P., Lievesley, N., Dobrowolski, T. & Withey, R. (1999). Developing and testing methods to determine the use of web sites: case study newspapers. *Aslib Proc*., 51. 144–154. https://doi.org/10.1108/EUM0000000006973

Prinsloo, D. J. (2012). Electronic lexicography for lesser-resourced languages: The South African context. In S. Granger & M. Paquot (Eds.) *Electronic Lexicography* (pp. 119–144). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199654864.001.0001

Rayson, P. & R. Garside. (2000). Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora, Volume 9* (pp. 1–6). Association for Computational Linguistics. https://doi.org/10.3115/1117729.1117730

Santos, D. & Frankenberg-Garcia, A. (2007). The Corpus, its users and their needs: A user-oriented evaluation of COMPARA. *International Journal of Corpus Linguistics*, 12(3). 335–374. https://doi.org/10.1075/ijcl.12.3.03san

Schembri, A., Fenlon, J., Rentelis, R., Reynolds, S. & Cormier, K. (2013). Building the British Sign Language corpus. *Language Documentation & Conservation*, 7. 136–154.

Schermer, G. M. M. (2006). Sign Language: Lexicography. In *Encyclopedia of Language and Linguistics, 2nd Edition*. Amsterdam: Elsevier Ltd.

Sloetjes, H. & Wittenburg, P. (2018). ELAN (version 5.2). Nijmegen: Max Planck Institute for Psycholinguistics. Retrieved from https://tla.mpi.nl/tools/tla-tools/elan/

Tan, K. H. & Woods, P. C. (2008). Media-related or generic-related features in electronic dictionaries: learners' perception and preferences. *GEMA Online® Journal of Language Studies*, 8(2). 1–17.

Tarp, S. (2009). Reflections on lexicographical user research. *Lexikos*, 19. 275–296. https://doi.org/10.5788/19-0-440

Vale, M. (2015). A study of the users of an online sign language dictionary. In I. Kosem, M. Jakubíček, J. Kallas, & S. Krek (Eds.), *Electronic lexicography in the 21st century: linking lexical data in the digital age - Proceedings of the eLex 2015 conference, 11-13 August 2015, Herstmonceux Castle, United Kingdom* (pp. 281–303). Ljubljana / Brighton: Trojina, Institute for Applied Slovene Studies / Lexical Computing Ltd.

Vale, M. (2017). *Folk definitions as a model for sign language dictionary definitions: A userfocused case study of the Online Dictionary of New Zealand Sign Language* [Unpublished doctoral thesis]. Victoria University of Wellington, Wellington.

Verlinde, S. & Binon, J. (2010). Monitoring dictionary use in the electronic age. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV Euralex International Congress* (pp. 1144–1151). Leeuwarden: Fryske Akademy.

Zwitserlood, I. (2010). Sign language lexicography in the early 21st century and a recently published dictionary of Sign Language of the Netherlands. International Journal of Lexicography, 23(4). 443–476. https://doi.org/10.1093/ijl/ecq031

Appendix

Rank	Word (Turkish)	Word (English)	N	Per 1000 (N)	Database (%)	Cum. (%)
1	Para	Money	8696	4.41	0,44	0,44
2	Anne	Mother	7381	3.74	0,37	0,81
3	En	Most	5136	2.60	0,26	1,08
4	Merhaba	Hello	5034	2.55	0,26	1,33
5	Baba	Father	4744	2.40	0,24	1,57
6	Gün	Day	4608	2.34	0,23	1,80
7	Ama	But	4515	2.29	0,23	2,03
8	İstemek	To want	4409	2.23	0,22	2,26
9	Her	Every	4111	2.08	0,21	2,47
10	Var	Have	4021	2.04	0,20	2,67
11	Aile	Family	3930	1.99	0,20	2,87
12	Yapmak	To do	3927	1.99	0,20	3,07
13	Ne	What	3883	1.97	0,20	3,26
14	Olmak	To be	3653	1.85	0,19	3,45
15	Bilmek	To know	3651	1.85	0,19	3,63
16	Almak	To take	3549	1.80	0,18	3,81
17	Okul	School	3500	1.77	0,18	3,99
18	Nasılsın	How are you	3469	1.76	0,18	4,17
19	Çok	Many	3313	1.68	0,17	4,34
20	Çünkü	Because	3275	1.66	0,17	4,50
21	Renk	Color	3205	1.62	0,16	4,66
22	Öğretmen	Teacher	3201	1.62	0,16	4,83
23	Hep	All	3174	1.61	0,16	4,99
24	Günaydın	Good morning	3137	1.59	0,16	5,15
25	Yok	No	3105	1.57	0,16	5,30
26	Ve	And	3080	1.56	0,16	5,46
27	Abla	Sister (older)	3072	1.56	0,16	5,62
28	Etmek	To make	2996	1.52	0,15	5,77
29	Amca	Uncle	2991	1.52	0,15	5,92
30	Son	Last	2980	1.51	0,15	6,07
31	Teyze	Aunt	2938	1.49	0,15	6,22
32	Hiç	None	2890	1.46	0,15	6,36
33	Nasıl	How	2888	1.46	0,15	6,51
34	Ak	White	2873	1.46	0,15	6,66
35	Arkadaş	Friend	2865	1.45	0,15	6,80
36	Ben	I	2818	1.43	0,14	6,94
37	Üniversite	University	2779	1.41	0,14	7,09
38	Gelmek	To come	2775	1.41	0,14	7,23
39	Sevmek	To like	2764	1.40	0,14	7,37
40	Zaman	Time	2764	1.40	0,14	7,51
41	Doğru	Right	2750	1.39	0,14	7,65
42	Başlamak	To start	2673	1.35	0,14	7,78
43	Vermek	To give	2669	1.35	0,14	7,92
44	Anlamak	To understand	2666	1.35	0,14	8,05
45	Kalmak	To stay	2649	1.34	0,13	8,19
46	İyi	Good	2627	1.33	0,13	8,32
47	Sonra	After	2621	1.33	0,13	8,45
48	Ders	Lesson	2615	1.33	0,13	8,58
49	Neden	Why	2590	1.31	0,13	8,72

50	Önemli	Important	2590	1.31	0,13	8,85
51	Kim	Who	2582	1.31	0,13	8,98
52	Pazar	Sunday	2578	1.31	0,13	9,11
53	Boş	Empty	2569	1.30	0,13	9,24
54	Ocak	January	2551	1.29	0,13	9,37
55	İle	With	2544	1.29	0,13	9,50
56	Ay	Month	2505	1.27	0,13	9,62
57	İş	Work	2498	1.27	0,13	9,75
58	Çalışmak	To work	2496	1.27	0,13	9,88
59	Gitmek	To go	2490	1.26	0,13	10,00
60	Üzülmek	To be sad	2489	1.26	0,13	10,13
61	Bulmak	To find	2488	1.26	0,13	10,26
62	Abi	Brother (older)	2486	1.26	0,13	10,38
63	Dayı	Uncle	2436	1.23	0,12	10,51
64	Kadar	To	2435	1.23	0,12	10,63
65	Aramak	To search	2433	1.23	0,12	10,75
66	Evet	Yes	2423	1.23	0,12	10,87
67	Öğrenmek	To learn	2421	1.23	0,12	11,00
68	Çocuk	Child	2408	1.22	0,12	11,12
69	Güzel	Beautiful	2405	1.22	0,12	11,24
70	Sarı	Yellow	2395	1.21	0,12	11,36
71	Konuşmak	To talk	2393	1.21	0,12	11,48
72	Öğrenci	Student	2392	1.21	0,12	11,61
73	Yeni	New	2356	1.19	0,12	11,72
74	Aşk	Love	2355	1.19	0,12	11,84
75	Zor	Difficult	2343	1.19	0,12	11,96
76	İnanmak	To believe	2342	1.19	0,12	12,08
77	Ev	House	2341	1.19	0,12	12,20
78	Şimdi	Now	2332	1.18	0,12	12,32
79	Kolay	Easy	2329	1.18	0,12	12,44
80	Hasta	Ill	2323	1.18	0,12	12,55
81	Eylül	September	2319	1.18	0,12	12,67
82	Geri dönmek	To return	2318	1.17	0,12	12,79
83	Yaş	Age	2307	1.17	0,12	12,91
84	Kardeş	Sibling	2300	1.17	0,12	13,02
85	Akraba	Relative	2298	1.16	0,12	13,14
86	Yarın	Tomorrow	2298	1.16	0,12	13,26
87	Kaç	How much	2286	1.16	0,12	13,37
88	Yıl	Year	2283	1.16	0,12	13,49
89	Aynı	Same	2279	1.16	0,12	13,60
90	Yanlış	Wrong	2264	1.15	0,11	13,72
91	Çıkmak	To get out	2256	1.14	0,11	13,83
92	Görmek	To see	2245	1.14	0,11	13,95
93	Erken	Early	2239	1.13	0,11	14,06
94	Anlatmak	To explain	2225	1.13	0,11	14,17
95	Bu	This	2225	1.13	0,11	14,29
96	Geçmek	To pass	2213	1.12	0,11	14,40
97	Bazen	Sometimes	2212	1.12	0,11	14,51
98	Doktor	Doctor	2210	1.12	0,11	14,62
99	Bugün	Today	2205	1.12	0,11	14,73
100	Mavi	Blue	2204	1.12	0,11	14,84
101	Özlemek	To miss	2184	1.11	0,11	14,96

102	Genç	Young	2183	1.11	0,11	15,07
103	Salı	Tuesday	2174	1.10	0,11	15,18
104	Mart	March	2158	1.09	0,11	15,29
105	Kitap	Book	2147	1.09	0,11	15,39
106	Yaşamak	To live	2147	1.09	0,11	15,50
107	Sınav	Exam	2143	1.09	0,11	15,61
108	Biz	Us	2141	1.09	0,11	15,72
109	İnsan	Human	2116	1.07	0,11	15,83
110	Kar	Snow	2106	1.07	0,11	15,93
111	Banka	Bank	2096	1.06	0,11	16,04
112	Mutlu	Happy	2091	1.06	0,11	16,15
113	Beğenmek	To like	2081	1.05	0,11	16,25
114	Daha	More	2057	1.04	0,10	16,36
115	Yenge	Aunt (Paternal)	2055	1.04	0,10	16,46
116	Pazartesi	Monday	2054	1.04	0,10	16,56
117	Ekim	October	2044	1.04	0,10	16,67
118	Söylemek	To say	2031	1.03	0,10	16,77
119	Bakmak	To look	2018	1.02	0,10	16,87
120	Önce	Before	2003	1.02	0,10	16,98
121	İl	County	1999	1.01	0,10	17,08
122	Yalan	Lie	1997	1.01	0,10	17,18
123	Okumak	To read	1974	1.00	0,10	17,28
124	Kötü	Bad	1967	1.00	0,10	17,38
125	Hala	Still	1961	0.99	0,10	17,48
126	Eski	Old	1946	0.99	0,10	17,58
127	Yemek	Food	1938	0.98	0,10	17,67
128	Geç	Late	1937	0.98	0,10	17,77
129	Araba	Car	1917	0.97	0,10	17,87
130	Öğle	Afternoon	1917	0.97	0,10	17,97
131	Gerçek	Real	1905	0.97	0,10	18,06
132	Ankara	Ankara	1900	0.96	0,10	18,16
133	Farklı	Different	1894	0.96	0,10	18,26
134	İçin	For	1886	0.96	0,10	18,35
135	Ara	Call	1885	0.96	0,10	18,45
136	Benzemek	Look alike	1880	0.95	0,10	18,54
137	Yakın	Close	1878	0.95	0,10	18,64
138	Kullanmak	To use	1864	0.94	0,09	18,73
139	Hatırlamak	To remember	1862	0.94	0,09	18,83
140	Beklemek	To wait	1854	0.94	0,09	18,92
141	İşaret	Sign	1853	0.94	0,09	19,01
142	Sabah	Morning	1852	0.94	0,09	19,11
143	Unutmak	To forget	1850	0.94	0,09	19,20
144	Memleket	Hometown	1847	0.94	0,09	19,29
145	Yardım	Help	1844	0.93	0,09	19,39
146	Mayıs	May	1839	0.93	0,09	19,48
147	Haber	News	1826	0.93	0,09	19,57
148	Kız	Girl	1814	0.92	0,09	19,67
149	Gibi	Like (something)	1800	0.91	0,09	19,76
150	Hayat	Life	1789	0.91	0,09	19,85
151	Başka	Other	1779	0.90	0,09	19,94
152	Ad	Name	1770	0.90	0,09	20,03
153	Tatil	Holiday	1748	0.89	0,09	20,12

154	Yalnız	Alone	1740	0.88	0,09	20,20
155	Gülmek	Laugh	1739	0.88	0,09	20,29
156	Şubat	February	1731	0.88	0,09	20,38
157	Düşünmek	To think	1730	0.88	0,09	20,47
158	Erkek	Boy	1725	0.87	0,09	20,56
159	Televizyon	Television	1725	0.87	0,09	20,64
160	Akşam	Evening	1709	0.87	0,09	20,73
161	Yeşil	Green	1696	0.86	0,09	20,82
162	Dünya	World	1693	0.86	0,09	20,90
163	Kızmak	To get angry	1687	0.86	0,09	20,99
164	Meyve	Fruit	1687	0.86	0,09	21,07
165	Dönmek	Turn	1683	0.85	0,09	21,16
166	Hayal	Imagination	1682	0.85	0,09	21,24
167	Güneş	Sun	1679	0.85	0,09	21,33
168	Kasım	November	1663	0.84	0,08	21,41
169	İstanbul	İstanbul	1660	0.84	0,08	21,50
170	Tamam	Okay	1659	0.84	0,08	21,58
171	Tarih	Date	1654	0.84	0,08	21,66
172	Sağlık	Health	1632	0.83	0,08	21,75
173	Davranış	Behavior	1630	0.83	0,08	21,83
174	Tekrar	Again	1629	0.83	0,08	21,91
175	Dün	Yesterday	1626	0.82	0,08	21,99
176	Özel	Private	1625	0.82	0,08	22,08
177	Uzun	Long	1624	0.82	0,08	22,16
178	Beraber	Together	1621	0.82	0,08	22,24
179	Uzak	Far	1619	0.82	0,08	22,32
180	Açık	Open	1613	0.82	0,08	22,41
181	Hızlı	Fast	1608	0.82	0,08	22,49
182	Hayır	No	1602	0.81	0,08	22,57
183	Kurs	Course	1600	0.81	0,08	22,65
184	Acı	Pain	1597	0.81	0,08	22,73
185	At	Horse	1596	0.81	0,08	22,81
186	Sevgili	Boyfriend/ Girlfriend	1594	0.81	0,08	22,89
187	Bilgi	Information	1593	0.81	0,08	22,97
188	Nisan	April	1591	0.81	0,08	23,05
189	Tanımak	To know	1588	0.80	0,08	23,13
190	Artık	No longer	1586	0.80	0,08	23,21
191	Aç	Hungry	1585	0.80	0,08	23,29
192	Çarşamba	Wednesday	1584	0.80	0,08	23,37
193	Hangi	Which	1570	0.80	0,08	23,45
194	Spor	Sport	1570	0.80	0,08	23,53
195	Elma	Apple	1563	0.79	0,08	23,61
196	Yol	Road	1562	0.79	0,08	23,69
197	Kişi	Person	1556	0.79	0,08	23,77
198	Aralık	December	1554	0.79	0,08	23,85
199	Nerede	Where	1546	0.78	0,08	23,93
200	İzlemek	Watch	1541	0.78	0,08	24,01
201	Dede	Grandad	1540	0.78	0,08	24,08
202	Sınıf	Class	1533	0.78	0,08	24,16
203	İlk	First	1532	0.78	0,08	24,24
204	Hepsi	All of	1529	0.77	0,08	24,32
205	Hafta	Week	1524	0.77	0,08	24,39

206	Müdür	Manager	1518	0.77	0,08	24,47
207	Tam	Complete	1516	0.77	0,08	24,55
208	Hemen	Instantly	1515	0.77	0,08	24,63
209	Ağustos	August	1513	0.77	0,08	24,70
210	Mor	Purple	1512	0.77	0,08	24,78
211	El	Hand	1504	0.76	0,08	24,85
212	Eğitim	Education	1503	0.76	0,08	24,93
213	On	Ten	1498	0.76	0,08	25,01
214	Söz	Promise	1497	0.76	0,08	25,08
215	Siz	You (plural)	1495	0.76	0,08	25,16
216	Mezun	Graduate	1493	0.76	0,08	25,23
217	Bitmek	Finish	1489	0.75	0,08	25,31
218	Temiz	Clean	1482	0.75	0,08	25,38
219	Sen	You	1481	0.75	0,08	25,46
220	Saat	Hour	1472	0.75	0,07	25,53
221	Türkiye	Turkey	1472	0.75	0,07	25,61
222	Belki	Maybe	1471	0.75	0,07	25,68
223	Büyük	Big	1470	0.75	0,07	25,76
224	Gezmek	To travel	1470	0.75	0,07	25,83
225	Bir	One	1467	0.74	0,07	25,91
226	Şehir	City	1467	0.74	0,07	25,98
227	İhtiyaç	Need	1463	0.74	0,07	26,06
228	Suç	Crime	1461	0.74	0,07	26,13
229	Birlikte	Together	1456	0.74	0,07	26,20
230	Heyecan	Excitement	1455	0.74	0,07	26,28
231	Hediye	Gift	1454	0.74	0,07	26,35
232	Ayrılmak	To break up	1450	0.73	0,07	26,42
233	Meslek	Career	1450	0.73	0,07	26,50
234	Kavga	Argument	1448	0.73	0,07	26,57
235	Buluşmak	To meet up	1445	0.73	0,07	26,64
236	Polis	Police	1442	0.73	0,07	26,72
237	Temmuz	July	1439	0.73	0,07	26,79
238	Düşmek	To fall	1438	0.73	0,07	26,86
239	Şey	Thingy	1437	0.73	0,07	26,94
240	Enişte	Aunt’s husband	1435	0.73	0,07	27,01
241	Beyaz	White	1432	0.73	0,07	27,08
242	Ses	Sound	1419	0.72	0,07	27,15
243	Selam	Hi	1413	0.72	0,07	27,22
244	Kaybetmek	To lose	1412	0.72	0,07	27,30
245	Eş	Spouse	1410	0.71	0,07	27,37
246	Küsmek	To be cross	1410	0.71	0,07	27,44
247	Ekmek	Bread	1405	0.71	0,07	27,51
248	Yeter	Enough	1405	0.71	0,07	27,58
249	Ağrı	Pain	1401	0.71	0,07	27,65
250	Sıkılmak	Get bored	1395	0.71	0,07	27,72
251	İzin	Permission	1394	0.71	0,07	27,79
252	Fazla	Too much	1392	0.71	0,07	27,86
253	Vazgeçmek	To give up	1389	0.70	0,07	27,94
254	Alışmak	To get used to	1387	0.70	0,07	28,01
255	Bebek	Baby	1375	0.70	0,07	28,08
256	Haziran	June	1373	0.70	0,07	28,14
257	Benim	Mine	1370	0.69	0,07	28,21

258	Oturmak	To sit	1370	0.69	0,07	28,28
259	Oyun	Game	1369	0.69	0,07	28,35
260	Umut	Hope	1367	0.69	0,07	28,42
261	Girmek	To get in	1365	0.69	0,07	28,49
262	Merak	Curiosity	1365	0.69	0,07	28,56
263	Ağaç	Tree	1363	0.69	0,07	28,63
264	Ne zaman	When	1361	0.69	0,07	28,70
265	Misafir	Guest	1355	0.69	0,07	28,77
266	Ülke	Country	1354	0.69	0,07	28,84
267	Cuma	Friday	1350	0.68	0,07	28,90
268	Ömür	Life span	1348	0.68	0,07	28,97
269	Gündüz	Day	1343	0.68	0,07	29,04
270	Olmaz	Unlikely	1343	0.68	0,07	29,11
271	Bölüm	Section	1341	0.68	0,07	29,18
272	Değil	Not	1340	0.68	0,07	29,24
273	Yaramaz	Naughty	1337	0.68	0,07	29,31
274	Ek	Addition	1336	0.68	0,07	29,38
275	İşaret dili	Sign language	1332	0.68	0,07	29,45
276	Kalp	Heart	1332	0.68	0,07	29,52
277	Şaşırmak	To be surprised	1331	0.67	0,07	29,58
278	Mevsim	Season	1330	0.67	0,07	29,65
279	An	Moment	1329	0.67	0,07	29,72
280	Bekâr	Single	1328	0.67	0,07	29,78
281	Bütün	All of	1321	0.67	0,07	29,85
282	Kazanmak	To win	1316	0.67	0,07	29,92
283	Deniz	Sea	1314	0.67	0,07	29,99
284	Hava	Air	1304	0.66	0,07	30,05
285	Kalem	Pen	1303	0.66	0,07	30,12
286	Kelime	Word	1303	0.66	0,07	30,18
287	Özlem	Longing	1303	0.66	0,07	30,25
288	Bir şey	Something	1296	0.66	0,07	30,32
289	Yasak	Prohibition	1294	0.66	0,07	30,38
290	Damat	Groom	1293	0.66	0,07	30,45
291	Böyle	Like this	1292	0.65	0,07	30,51
292	Et	Meat	1286	0.65	0,07	30,58
293	Rahat	Comfortable	1283	0.65	0,07	30,64
294	Kadın	Woman	1282	0.65	0,06	30,71
295	Korkmak	To be afraid	1282	0.65	0,06	30,77
296	Perşembe	Thursday	1281	0.65	0,06	30,84
297	Kıskanmak	To be Jealous	1280	0.65	0,06	30,90
298	Ölmek	To die	1277	0.65	0,06	30,97
299	Baş	Head	1275	0.65	0,06	31,03
300	Rüya	Dream	1275	0.65	0,06	31,10