Sezer, T., & Karadağ, Ö. (2025). A Turkish word frequency tool: LexiTR frequency. *Ana Dili Eğitimi Dergisi*, *13*(2), 266-276.

Ana Dili Eğitimi Dergisi
Journal of Mother Tongue Education
www.anadiliegitimi.com

Geliş/Received: 09.02.2025 Kabul/Accepted: 17.03.2025
Araştırma Makalesi / Research Paper

A Turkish Word Frequency Tool: LexiTR Frequency[*]

Taner SEZER[**]
Özay KARADAĞ[***]

Abstract
Word frequency is a fundamental concept in linguistics, computational linguistics, natural language processing (NLP) and language education. Word frequency plays a critical role in understanding the characteristics and usage patterns of a word. This study introduces the "Turkish Word Frequency Tool" (TWFT), developed as part of the LexiTR Project, along with its features. TWFT is based on a balanced corpus consisting of over 193 million words from four distinct text types: academic, social media, fictional, and informative texts. TWFT serves a scalable online platform that provides researchers with the ability to examine word usage trends across different text types. It enables comprehensive analyses through real-time querying, graphical data representation, and both raw and normalized frequency values. Additionally, it provides API support, presenting word frequency information in a structured format. By filling a significant gap in the existing literature, TWFT aims to establish a consistent, transparent, and comprehensive foundation for linguistic research and natural language processing applications.
Keywords: Frequency, lexicon, tokenization, TS Tokenizer, LexiTR

Türkçe Sözcük Sıklığı Aracı: LexiTR Sıklık Aracı

Öz
Sözcük sıklığı, dilbilim, bilişimsel dilbilim, doğal dil işleme (NLP) ve dil eğitimi alanlarında temel bir kavramdır. Sözcük sıklığı bir sözcüğün özelliklerini ve kullanım eğilimlerini anlamada kritik bir rol oynamaktadır. Bu çalışmada, LexiTR Projesi kapsamında geliştirilen "Türkçe Sözcük Sıklığı Aracı (TSSA)" ve özellikleri tanıtılmaktadır. TSSA, akademik, sosyal medya, kurgusal ve bilgilendirici metinler olmak üzere dört farklı türden oluşan 193 milyondan fazla sözcük içeren dengeli bir derleme dayanmaktadır. TSSA, araştırmacılara farklı metin türleri arasında sözcük kullanım eğilimlerini inceleme olanağı sunan, gerçek zamanlı sorgulama, grafiksel veri gösterimi, ham ve normalize edilmiş sıklık değerleri ile kapsamlı analiz imkânı sağlayan ölçeklenebilir bir çevrimiçi platformdur. Ayrıca, sağladığı API desteği ile sözcüğe ilişkin sıklık bilgilerini yapılandırılmış bir formatta sunmaktadır. Mevcut literatürdeki önemli bir boşluğu dolduran TSSA dilbilim araştırmaları ile doğal dil işleme uygulamaları için tutarlı, şeffaf ve kapsamlı bir temel oluşturmayı hedeflenmektedir.
Anahtar Kelimeler: Sıklık, sözcük listesi, birimlendirme, TS Tokenizer, LexiTR

[*] TS Tokenizer was developed by Taner Sezer as part of his PhD research at Hacettepe University under the supervision of Prof. Dr. Özay Karadağ. https://lexitr.tscorpus.com

[**] Öğr. Gör., Mersin Üniversitesi, TTO Ofisi, Mersin, tanersezerr@gmail.com, ORCID: orcid.org/0000-0002-7328-7650

[***] Prof. Dr., Hacettepe Üniversitesi, Eğitim Fakültesi, Türkçe ve Sosyal Bilimler Eğitimi Bölümü, Ankara, ozaykaradag@hacettepe.edu.tr, ORCID: orcid.org/ 0000-0003-4596-1203

Introduction

A Missing Tool for Turkish

This study introduces the Turkish Word Frequency Tool (TWFT), a set of lexical tools for Turkish based on a 193-million-word corpus across four genres, developed as a part of the LexiTR project. The tool presents an online query interface with options and presents both raw and relative values over four genres with graphics.

Frequency is a fundamental concept in linguistics, computational linguistics, natural language processing (NLP), and education. The word frequency studies have implications on language teaching, learning and linguistic research. Understanding the most frequent words enables educators to develop effective learning resources, shape curriculum design, create language proficiency tests and conduct research on language acquisition processes. Sinclair (1991) highlights the importance of word frequency saying "anyone studying a text is likely to need to know how often each different word form occurs in it".

Linguists consider frequency as a multi-dimensional phenomena. The "raw frequency" simply counts the number of instances of a linguistic form that occur in a given body of text*(s*) or corpus (Leech, 2011). It is determined by counting the occurrences of a word within a given text (Popescu et al., 2009). "Relative" or "normalized frequency" counts occurrences of a linguistic form relative to some reference value such as per/million words (Schützler, 2023).

Word frequency analysis has been a subject of interest for decades for researchers seeking to understand the distribution of words in languages, identify patterns and following trends in society or evolving culture — whether contemporary or in a specific period of history.

Thorndike's "Word Book of 10,000 words", published in 1921, is generally referred to as a pioneering study about word frequency (Soliman & Familiar, 2024). This book provides a list of words along with their frequencies, aiming to help teachers select vocabulary for instruction. The very first samples of frequency lists, like Thorndike's, were, of course, printed in hard copy. In 1967, Kucera and Francis marked a milestone with their pioneering study, titled "Computational analysis of present-day American English", which is also called the foundation of modern corpus linguistics (Xu, 2022).

Frequency Studies in Turkish

The literature reports two major studies for Turkish frequency, first by İlyas Göz (2003), covering over 975 thousand words and second by Gökhan Ölker (2011), covering over 929 thousand words (Baş, 2011; Çal, 2015; Çınar and İnce, 2015), focusing on general language in use. For specific domains in language, such as text books (Arslan ve Bay, 2023), books of a specific author (Evler and Aksoy, 2024), a periodical (Gürler and Yıldız, 2024), and students at a specific level (Karadağ, 2005; Kurudayıoğlu, 2005), a great number of studies have been conducted. However, the literature lacks a unified, large-scale corpus that integrates these datasets for real-time analysis and interactive querying. While two notable efforts, the *Turkish Electronic Living Lexicon* (TELL) (Inkelas et al., 2000) and *LexiTürk* (Başaran, 2022), claim to offer online and interactive resources, neither of these tools was accessible at the time of this research.

A Turkish Word Frequency Tool

The Internet has revolutionized the way we interact with data. Data shifted from static, printed materials to dynamic, real-time accessible sources. Printed data, a standard medium for reference and research once, is now outdated due to its limitations in accessibility, searchability and adaptability. For researchers seeking linguistic insights, educators designing curricula or even developers building applications, the ability to query, filter, and visualize data in real time has become an essential requirement. This shift is the main motivation behind TWFT.

Another motive is about the size and coverage of the data presented in the existing studies. Most studies report data sizes that barely approach or slightly exceed one million words, restricting the scope of frequency analyses—especially for a morphologically rich language like Turkish. Turkish is an agglutinative language which allows for the creation of a vast number of inflected and derived word forms. A morphological parser developed by Hankamer highlights this productivity and for a single verb the parser generated over 1.8 million possible valid forms when instructed to generate all possible forms (Hankamer, 1989). Given this complexity, a word frequency tool that only allows queries for the base form might not be sufficient for Turkish, as it would fail to capture the full range of information encoded in different word forms, potentially limiting its usefulness for linguistic research, language learning, and other applications.

This vast morphological productivity is also reflected in existing Turkish NLP resources. For instance, the word list within the Turkish NLP tool Zemberek (Akın and Akın, 2007) includes over 1.2 million unique forms, while the TS Corpus Word List (Sezer, 2021) contains more than 3.2 million. However, beyond sheer numbers, the true significance of data size lies in the diversity of word forms themselves. Each form encodes specific linguistic information and interacts dynamically within different syntactic and semantic contexts. This underlines the need for a word frequency tool capable of capturing not only base forms but also their inflected and derived counterparts, ensuring a more comprehensive and nuanced understanding of Turkish word usage.

And the last motive is using the same data (*corpus*) for multiple lexical tools. The LexiTR project is designed to provide a set of linguistic tools for Turkish that relies on large-scale and balanced corpus. Instead of creating isolated datasets for each tool, LexiTR stands on the same data and integrates the tools, allowing for more efficient data management and cross-referencing. This is expected to supply consistency among tools and useability across different applications.

LexiTR Corpus

The LexiTR Corpus[*] is structured to provide a balanced and representative dataset for Turkish language research, by presenting a selection of texts from four distinct genres: (i) Academic, (ii) Social Media, (iii) Fictional, and (iv) Informative. Each genre has been chosen to capture different linguistic registers and usage patterns, ensuring a broad and reliable lexical database.

LexiTR offers a relatively big corpus, over 193 million words. Table 1 presents the distribution of data by genres and word counts.

Table 1.
Distribution of Genre in LexiTR Corpus and their Representation Rates

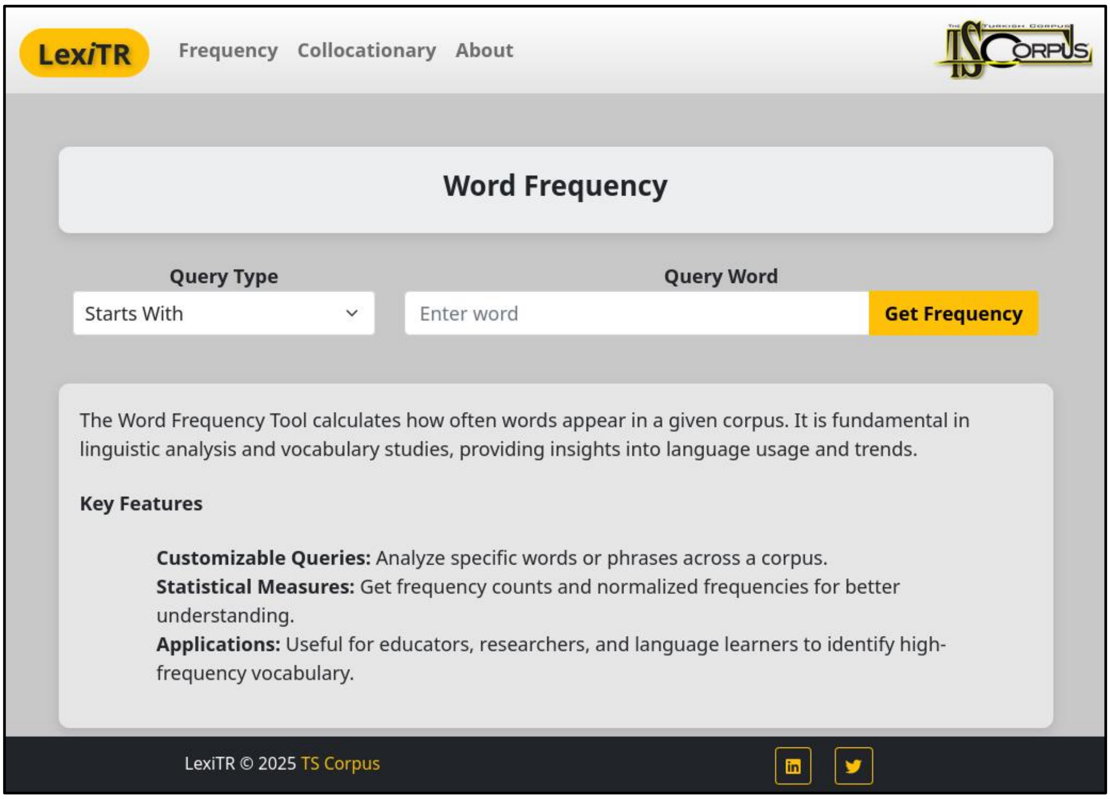
	Percentage in Total	Word Count
Academic	7.94%	15.365.389
Social Media	19.87%	38.461.199
Fictional	25.78%	49.908.578
Informative	46.41%	89.835.094
Total	100%	193.570.260

Academic texts provide insights of formal and technical language. Social media captures informal, evolving, and spoken-like discourse. Fictional texts include creative, literary, and conversational elements such as novels and stories. Informative sources reflect neutral, general-purpose, and widely used language such as news, columns and Wikipedia. This diversity ensures a comprehensive representation of modern Turkish usage across multiple contexts.

Frequency Tool User Interface

The frequency tool offers a user-friendly web interface, enabling researchers, educators, and language learners to explore word frequency data in both raw and normalized formats. Users can perform two types of queries. “Starts With” query type performs a search that returns all the words starting with the given character string. For example, if the query word is “kapı”, kapı, kapılar, kapısı, etc. are fetched. “As Is” query returns only the result that matches exactly with the given query word. Figure 1 presents the query interface.

[*] Creation of the LexiTR corpus will be presented in detail in another paper.



Word Frequency

Query Type: Starts With

Query Word: Enter word

Get Frequency

The Word Frequency Tool calculates how often words appear in a given corpus. It is fundamental in linguistic analysis and vocabulary studies, providing insights into language usage and trends.

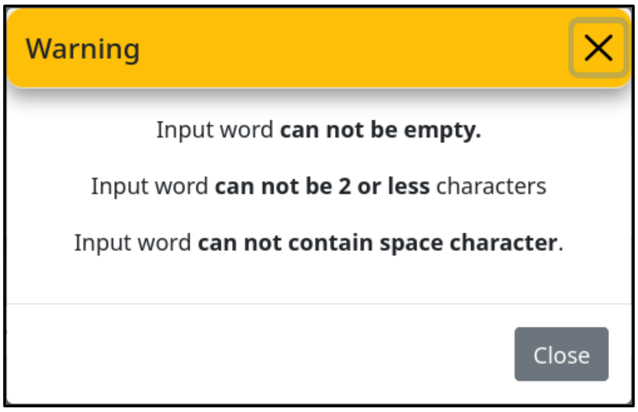
Key Features

Customizable Queries: Analyze specific words or phrases across a corpus.
Statistical Measures: Get frequency counts and normalized frequencies for better understanding.
Applications: Useful for educators, researchers, and language learners to identify high-frequency vocabulary.

LexiTR © 2025 TS Corpus

Figure 1. Word Frequency Tool Query Interface

Users type a query word and click the 'Get Frequency' button to search. A control function checks for a given string. The function controls if the string is given, if the given string is composed of two or more characters and if the string has white-space character. If one of these three conditions is not met, a warning, given in Figure 2, shows up at the screen.



Input word **can not be empty.**

Input word **can not be 2 or less** characters

Input word **can not contain space character**.

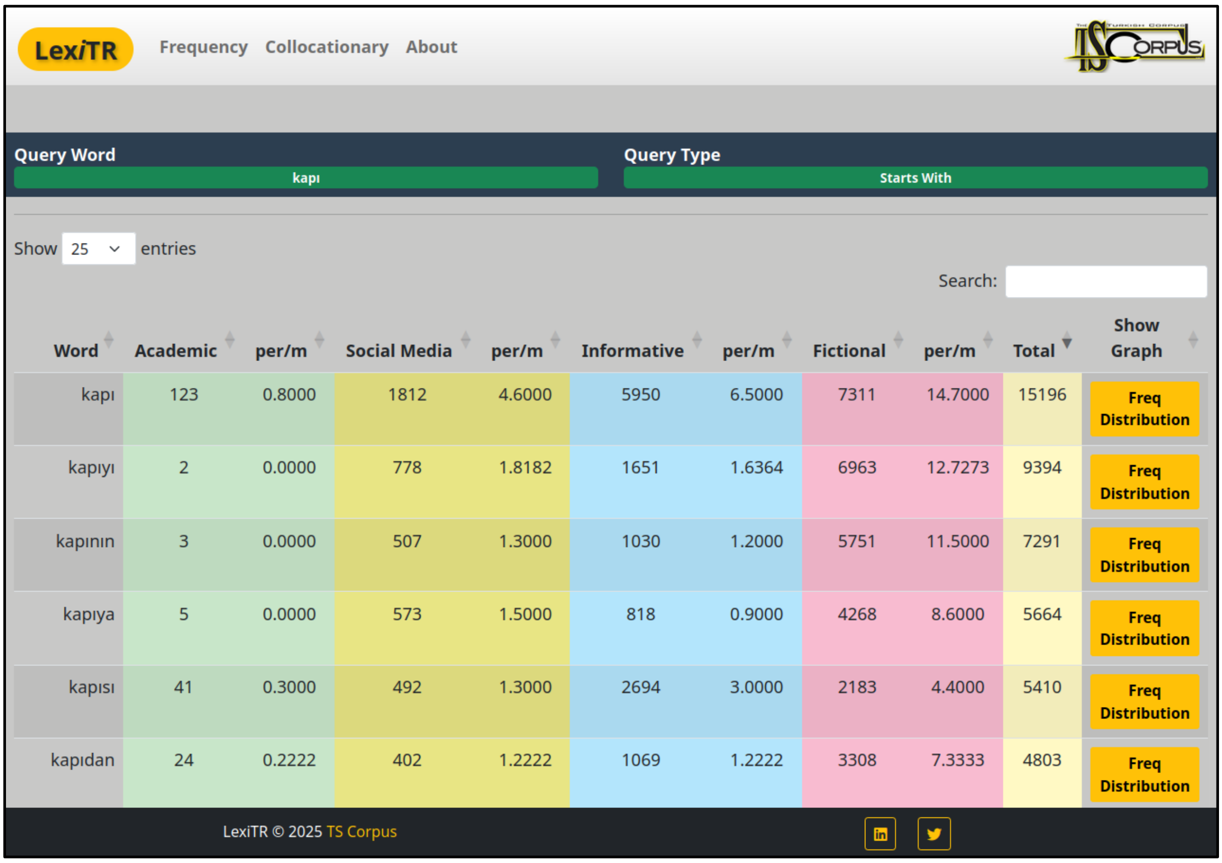
Figure 2. Warning Screen

At the top of the result screen, “the query word” and “the query type” are presented. The top bar is followed by the result table. Result table has a built-in search and sort mechanism. Sorting can be triggered by a single click on the column header. Search function runs as user types in.

Results are presented within an 11-column-interactive table. The first column presents the matching word and is followed by genre distributions. For each genre both the “raw frequency” and “relative frequency” values are presented. The relative value is calculated over per/million words. If a word has less than 5 occurrences in a genre, per/million value is not calculated, therefore it is shown as 0.

On the top left corner a dropdown menu acts as a selector to set the number of results that will be presented in the table. The default value is set to 25 results.

Last two columns are, respectively, the total matches and the button that triggers the "Frequency Distribution Graph". A sample results screen for the word "kapı" is presented in Figure 3.



Query Word: kapı

Query Type: Starts With

Show 25 entries

Search:

Word	Academic	per/m	Social Media	per/m	Informative	per/m	Fictional	per/m	Total	Show Graph
kapı	123	0.8000	1812	4.6000	5950	6.5000	7311	14.7000	15196	Freq Distribution
kapıyı	2	0.0000	778	1.8182	1651	1.6364	6963	12.7273	9394	Freq Distribution
kapının	3	0.0000	507	1.3000	1030	1.2000	5751	11.5000	7291	Freq Distribution
kapıya	5	0.0000	573	1.5000	818	0.9000	4268	8.6000	5664	Freq Distribution
kapısı	41	0.3000	492	1.3000	2694	3.0000	2183	4.4000	5410	Freq Distribution
kapıdan	24	0.2222	402	1.2222	1069	1.2222	3308	7.3333	4803	Freq Distribution

LexiTR © 2025 TS Corpus

Figure 3. Results Screen

The forms presented in the table like "kapıyı", "kapının", "kapıya", "kapısı" are also featuring the frequency of morphological suffixes for the target word. This feature presents interesting results. For instance, for the word "masa", the most frequent form is not the base form but "masaya".

- For each genre a unique color is used for clear distinction. The same color code is also used in distribution graphics.
- At the end of the table the result counter and the pagination buttons are located. Result counter presents the total number of rows (matching results) and it is updated when a search is performed via the search field.
- A polar area graph is generated for each result upon request. Figure 4 presents a sample graph for the word "masaya".

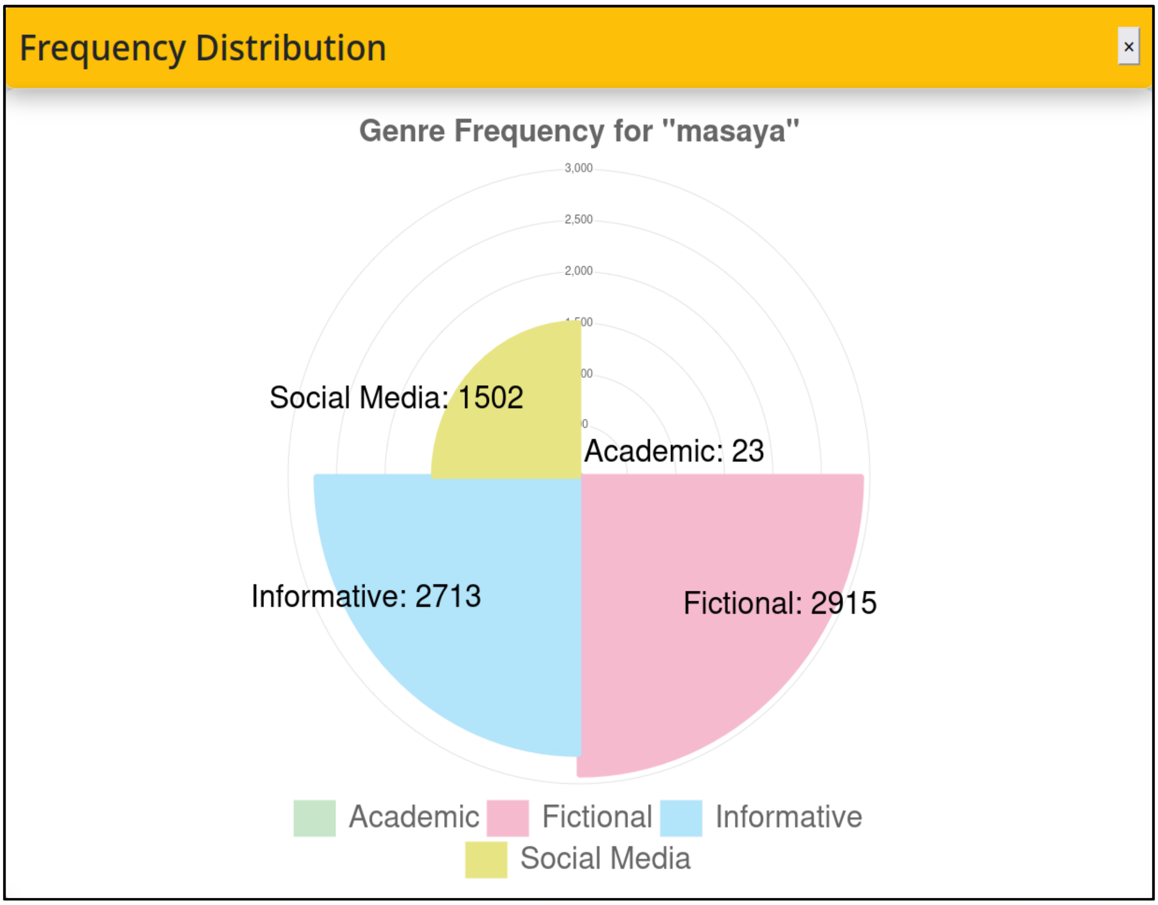


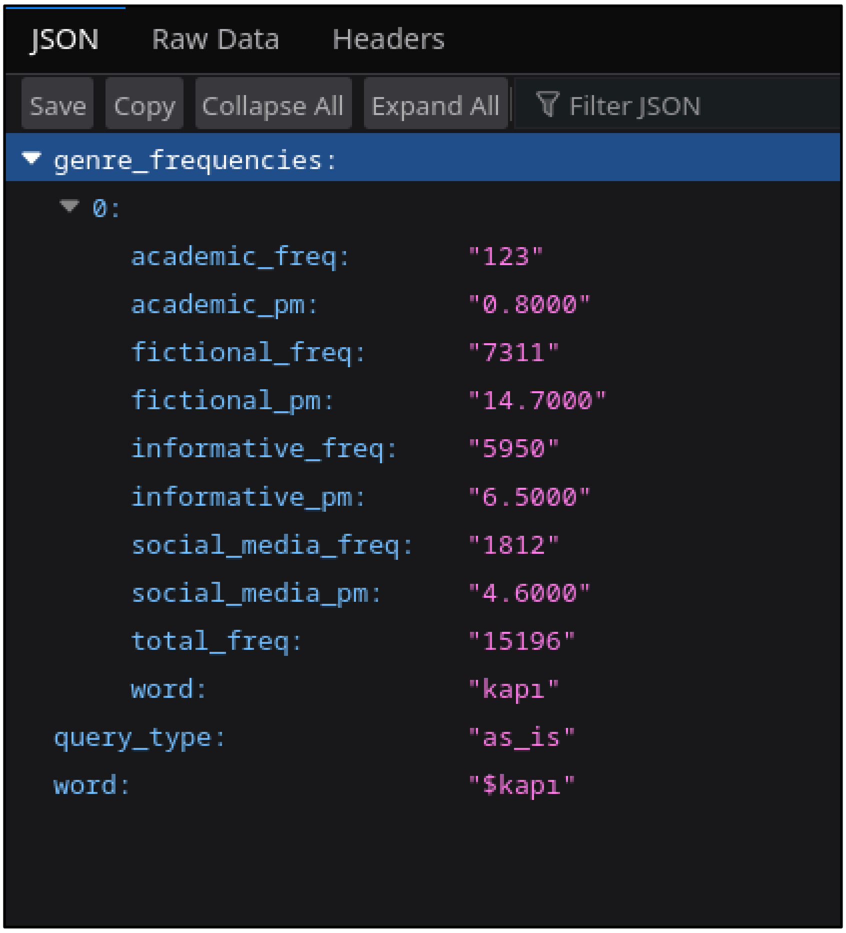
Figure 4. Polar Area Graph For “masaya”

The API Layer

As mentioned above, nowadays dynamic and real-time results are the main expectation for many researchers. Therefore, we have built an API (*Application Programming Interface*) layer that responds in JSON to the word frequency tool. An API is an interface for programming languages that responds in a structured format, which then could be implemented into other pipelines following the intentions of the programmers. The requests could be sent over a browser or over a programming language.

The API layer* gets two parameters, (i) the query word and (ii) query type then responds to the request in JSON format. The API presents data in a structured way, making it easier for users to analyze and visualize the results. Furthermore, the API layer can be used to integrate the word frequency tool with other applications or platforms, which can enhance its functionality and reach.

* https://lexitr.tscorpus.com/freq/frequency?word=$kapı&query_type=as_is



```
▼ genre_frequencies:
  ▼ 0:
      academic_freq:      "123"
      academic_pm:        "0.8000"
      fictional_freq:     "7311"
      fictional_pm:       "14.7000"
      informative_freq:   "5950"
      informative_pm:     "6.5000"
      social_media_freq:  "1812"
      social_media_pm:    "4.6000"
      total_freq:         "15196"
      word:               "kapı"
  query_type:             "as_is"
  word:                   "$kapı"
```

Figure 5. API Response For the Word “kapı”

Figure 5 samples the response for the word “kapı” from the API layer. The response covers the “query word”, “query type”, “the total frequency” as well as both “raw” and “relative” frequency values for each genre.

For future developments, the API layer is planned to have various other information such as if the query word is in the top 1k, 2k or 3k of the most frequent words in Turkish.

Methodology

The methodology of this study is rooted in corpus linguistics and computational linguistics, combining quantitative and qualitative approaches to analyze Turkish word frequency.

TWFT, developed as part of the LexiTR project, relies on a structured and balanced corpus, allowing for dynamic linguistic analysis. The LexiTR Corpus, a collection of over 193 million words, is covering four genres: academic, social media, fictional, and informative. These genres were selected to provide a representative sample of contemporary Turkish language use across different registers and discourse types.

Corpus Design

The design of the LexiTR corpus is dominated by three major factors. The first factor is to provide a solid background for running language over different genres, second is to ensure that the corpus is sufficiently large to allow for meaningful linguistic analysis and the third factor is the availability of textual sources with genre-specific metadata.

Genres are selected mainly by following COCA (The Corpus of Contemporary American English). COCA covers 5 genres; (i) spoken, (ii) fiction, (iii) magazine, (iv) newspaper, (v) academic. However, obtaining textual sources for Turkish is significantly more challenging than it is for English. For instance, Davies (2009) reported that COCA’s spoken genre included 6 million words from CNN transcripts, where a total of over 170 million words from the period 2000–2008 exists. However, we had to exclude the spoken genre due to limited data availability.

We merged magazine and newspaper sources into a single genre as "informative". The main source for this part of the corpus was carried from TimeLine Corpus released under the TS Corpus project. We also used data from "Turkish Dataset for Identification of Author Gender" (Tüfekçi, 2020) and finally we used the Turkish Wikipedia dump of May 2024 for this genre. Data for the fictional genre was extracted from various ebooks using 'pdftotext' v22.12. Data for academic genre collected from "İletisim Abstracts" and "Educational Sciences Abstracts" from TS Corpus project (Sezer & Sezer, 2013) and "Turkish Labeled Text Corpus" by Özturk (Özturk et al., 2014). We also wrote a script to harvest abstracts of publicly available master and PhD thesis from Thesis Center by the Turkish Council of Higher Education. The total size of the academic genre reaches 15.365.389 words. Even though this genre presents the smallest data set in the whole corpus, the importance of its existence is highlighted as it ensures formal and technical language usage in the corpus (Biber, 1995; McEnery and Hardie, 2012). And finally for social media we used "TweetS Corpus" (Sezer, 2016), Covid 19 Tweets (Sezer, 2017, Törenli & Kıyan, 2023). These two dataset provided over 20 million words (12,564,769 from TweetS and 7,561,844 million words from Covid19 Tweets corpus).

Data Processing

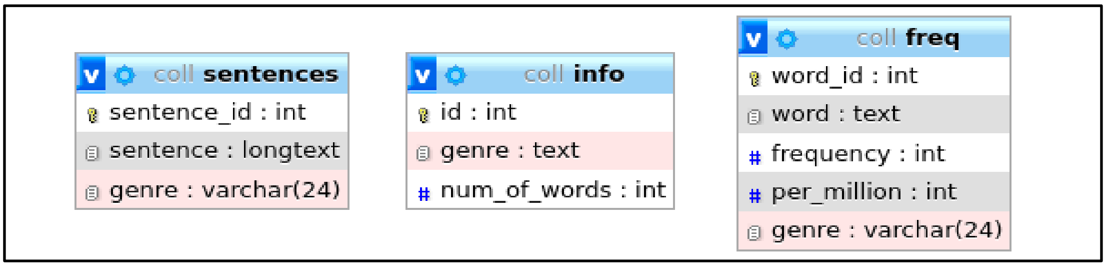
The first step of the data processing was tokenization. Tokenization is splitting text into usable smaller chunks in coherence with expected outputs. Tokenization is called the initial step for text processing (Webster and Kit, 1992; Rychlý and Špalek, 2022). The accuracy of the tokenization directly affects the accuracy of the following steps (Rust et al., 2020).

For tokenization we used TS Tokenizer. TS Tokenizer is a hybrid tokenizer designed to tokenize Turkish texts. It is available for use with MIT Licence and served as a Python package*.

The tokenizer supports different output formats declared by parameters. When no parameter is set, the default option "tokenize" is used and each token is presented in a new line. Another featured parameter is "lines", which is used to keep the structure of the input. Each line in the input file is kept together and tokenization is applied within the same structure.

For each genre, we created a single text file compiled from relevant sources. Then by using a Python script we processed the data and wrote it into a relational database. We preferred using MariaDB as it is an open-source relational database manager.

The database is composed of three tables, "sentences", "info" and "freq". The script first asks the user to specify the genre. The algorithm of the script has three functions. First the script counts the total number of lines in the given text and writes this information to the "info" table. Then the script inserts each line to the "sentences" table with a unique id. And finally, after each new line is added to the database, the script calculates the word frequencies and writes it to the "freq" table.



coll sentences
sentence_id : int
sentence : longtext
genre : varchar(24)

coll info
id : int
genre : text
num_of_words : int

coll freq
word_id : int
word : text
frequency : int
per_million : int
genre : varchar(24)

Figure 6. LexiTR database structure

After the first script completes the task then a second script is used to calculate per/million value for each word and for each genre. This algorithm ensures ease of use and accuracy while allowing LexiTR to scale with new data.

[*] https://pypi.org/project/ts-tokenizer/

Discussions and Results

TWFT serves as a novel and unique resource in corpus-based research for Turkish, addressing multiple gaps in the field.

Unlike previous works that relied on small datasets or printed frequency lists, this tool is hosted on an interactive, scalable online platform, allowing users to query words in real time and explore frequency data across four distinct genres: academic, social media, fiction, and informative texts. With a balanced corpus of over 193 million words, researchers can analyze word frequency variations across genres, enabling deeper linguistic insights. For instance, informal expressions may dominate social media, while academic texts exhibit higher frequencies of domain-specific terms.

One critical challenge in word frequency tools is tokenization inconsistency, which leads to varying results. A study that compared two online and one offline tool for Turkish word frequency calculations and found discrepancies due to differing tokenization methods (Pilten-Ufuk, 2021). TWFTl fixed this issue by implementing a standardized approach using the TS Tokenizer, ensuring a consistent baseline for all data.

Another notable aspect of the tool is its scalable database architecture and data processing scripts, which allows for future expansion in data size, number of genres, and additional linguistic features. This scalable approach facilitates the development of other advanced linguistic tools.

The tool provides both raw and normalized (per-million) frequency values, enabling cross-genre comparisons with statistical accuracy. Besides, its API interface, which provides researchers with direct access to the data, removes limitations set by a graphical interface.

Research and Publication Ethics

All rules specified under the Higher Education Institutions Scientific Research and Publication Ethics Directive have been adhered to in this study. None of the actions listed under the second section of the directive, titled Violations of Scientific Research and Publication Ethics, have been committed.

Author Contribution Rate

The authors’ contributions to the study are equal.

Conflict of Interest

The research does not involve any conflict of interest

References

Akın, A. A. ve Akın, M. D. (2007). Zemberek, an open source NLP framework for Turkic languages. *Structure*, *10*(2007), 1-5.

Arslan, K. ve Bay, Y. (2023). İlkokul Türkçe ders kitaplarının söz varlığı bakımından incelenmesi. *Turkish Journal of Primary Education*, *8*(1), 14-27.

Baş, B. (2011). Söz varlığı ile ilgili çalışmalarda kullanılacak ölçütler. *Türklük Bilimi Araştırmaları*, (29), 27-61.

Başaran, B. (2022). Measuring word frequency in language teaching textbooks using LexiTürk. *International Online Journal of Education and Teaching (IOJET)*, *9*(1), 571-583.

Çal, A. (2015). Türkiye’de farklı dönemlere ait kelime sıklığı çalışmaları üzerine bir değerlendirme. *Turkish Studies: International Periodical for the Languages, Literature and History of Turkish or Turkic*, *10*(8), 715-730.

Çınar, İ. ve İnce, B. (2015). Türkçe ve Türk kültürü ders kitaplarındaki söz varlığına derlem temelli bir bakış. *International Journal of Languages' Education and Teaching*, *3*(1), 198-209.

Davies, M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, *14*(2), 159-190.

Douglas, B. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Evler, D. ve Aksoy, E. (2024). Şermin Yaşar'ın çocuklara yönelik eserlerinde söz varlığı. *SEBED*, *2*(1), 1-15.

Göz, İ. (2003). *Yazılı Türkçenin kelime sıklığı sözlüğü*. Ankara: Türk Dil Kurumu Yayınları.

Gürler, H. ve Yıldız, M. (2024). Doğan Kardeş Dergisinin söz varlığı üzerine bir araştırma. *Milli Eğitim Dergisi*, *53*(242), 969-996.

Hankamer, J. (1989). Morphological parsing and the lexicon. In W. Marslen-Wilson (Ed.), *Lexical representation and process* (pp. 392-408). United States: MIT Press.

Inkelas, S., Küntay, A., Orgun, O. ve Sprouse, R. (2000). Turkish electronic living lexicon (TELL). *Turkic Languages*, *4,* 253-275.

Karadağ, Ö. (2005). *İlköğretim I. kademe öğrencilerinin kelime hazinesi üzerine bir araştırma* (Unpublished doctoral dissertation). Gazi University, Institute of Educational Sciences, Ankara.

Kurudayıoğlu, M. (2005). *İlköğretim II. kademe öğrencilerinin kelime hazinesi üzerine bir araştırma* (Unpublished doctoral dissertation). Gazi University, Institute of Educational Sciences, Ankara.

Leech, G. N. (2011). Frequency, corpora and language learning. In *A taste for corpora: In honour of Sylviane Granger* (pp. 7-32). Netherlands: John Benjamins Publishing Company.

McEnery, T. ve Andrew H. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

Ölker, G. (2011). *Yazılı Türkçenin kelime sıklığı sözlüğü (1945-1950 arası)* (Unpublished doctoral dissertation). Selçuk University, Institute of Social Sciences, Konya.

Popescu, I. I., Mačutek, J. ve Altmann, G. (2009). *Aspects of word frequencies*. Lüdenscheid: RAM-Verlag.

Pilten-Ufuk, Ş. (2021). Derlem dilbilim ve edebiyat çalışmalarının kesişim noktası: Derlem biçem bilimi. Ö. Solak ve S. Doykun (Ed.), *Disiplinlerarası edebiyat çalışmaları* içinde (ss. 145-171). İstabul: Paradigma Akademi Yayın.

Rust, P., Pfeiffer, J., Vulić, I., Ruder, S. ve Gurevych, I. (2020). How good is your tokenizer? On the monolingual performance of multilingual language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing* (pp. 3118-3135). Association for Computational Linguistics, Bangkok.

Rychlý, P. ve Spalek, S. (2022, December). Utok: The fast rule-based tokenizer. In Proceedings of *recent advances in Slavonic natural language processing.* (pp. 149-154). South Moravia: Tribun EU.

Schützler, O. (2023). Frequencies in corpus linguistics: Issues of scaling and visualisation. In *Data visualization in corpus linguistics: Critical reflections and future directions*. Helsinki: Varieng.

Sezer, T., Sezer, B. ve Üniversitesi, M. (2013, May). TS corpus: Herkes için Türkçe derlem. In *Proceedings of the 27th national linguistics conference* (pp. 217-225).

Sezer, T. (2016). Tweets corpus: Building a corpus by social media. *Journal of National Education and Social Sciences*, *210,* 621-633.

Sezer, T. (2017). TS corpus project: An online Turkish dictionary and TS DIY corpus. *European Journal of Language and Literature Studies*, *3*(3), 18-24.

Sezer, T. (2021). *TS Corpus word list* (Version 001) [Data set]. TS Corpus. Erişim adresi: https://doi.org/10.57672/B6M8-8333

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Soliman, R. ve Familiar, L. (2024). Creating a CEFR Arabic vocabulary profile: A frequency-based multi-dialectal approach. *Critical Multilingualism Studies*, *11*(1), 266-286.

Törenli, N. ve Kıyan, Z. (2023). The importance of sustainable communication in the covid-19 period: The case of Turkey. In *SDG18 Communication for All, Volume 2: Regional perspectives and special cases* (pp. 225-246). Springer International.

Tüfekçi, P. (2020). *Turkish dataset for identification of author gender* [Data set]. Mendele Data. https://doi.org/10.17632/8f93rjhgjk.1

Webster, J. J. ve Kit, C. (1992). Tokenization as the initial phase in NLP. In *Proceedings of COLING 1992, Volume 4: The 14th International Conference on Computational Linguistics* (pp. 1106-1110).

Xu, J. (2022). A historical overview of using corpora in English language teaching. In *The Routledge handbook of corpora and English language teaching and learning* (pp. 11-25). England: Routledge.