Selcen Erten-Johansson, Valtteri Skantsi, Sampo Pyysalo & Veronika Laippala

Linguistic variation beyond the Indo-European web: Analyzing Turkish web registers in TurCORE

2024

https://doi.org/10.1075/rs.24002.ert

Final draft

Erten-Johansson, Selcen, et al. Linguistic Variation beyond the Indo-European Web: Analyzing Turkish Web Registers in TurCORE. Register Studies, vol. 6, no. 1, pp. 60–90. https://doi.org/10.1075/rs.24002.ert

Linguistic variation beyond the Indo-European web: Analyzing Turkish web registers in TurCORE

Selcen Erten-Johansson, Valtteri Skantsi, Sampo Pyysalo, Veronika Laippala

Abstract
A register, defined as a text variety with specific situational characteristics and a communicative purpose (Biber & Conrad 2019), is also recognized as a cultural construct (Biber & Egbert 2023). Registers merit thorough investigation due to their pivotal role in reflecting linguistic and cultural landscapes. However, existing studies predominantly focus on Indo-European languages. This study investigates Turkish web registers through the introduction of the Turkish Corpus of Online Registers (TurCORE). Comprising 2,780 web texts, TurCORE was manually annotated using a register taxonomy targeting the entire unrestricted web and identifying 24 web register categories. By employing Text Dispersion Keyword Analysis (Egbert & Biber 2019), the research examines the register characteristics with a specific focus on news reports, interactive discussions, and recipes, drawing comparisons with their English equivalents. Results reveal parallels between Turkish and English news reports while Turkish interactive discussions and recipes exhibit distinctive language- and culture specific features.

Keywords: Web registers; Turkish; manual register annotation; Text Dispersion Keyword Analysis; linguistic analysis of web registers

1. Introduction

Register, indicating whether a text is a news report, interactive discussion, or recipe, significantly influences linguistic variation, and shapes our interpretation of the text (Biber 2012). While traditionally associated with situational and linguistic characteristics, registers are alternatively viewed as cultural constructs, given that language and culture are structured based on cultural frameworks (Biber & Egbert 2023).

The web provides access to a wide range of registers, both similar to those in the printed world — such as song lyrics — and those specific to the web — such as different blogs and discussion fora. While the set of all the web registers has been an open question and corpora featuring them have been scarce, the introduction of the Corpus of Online Registers of English (CORE) (Egbert, Biber, & Davies 2015) marks a milestone, representing the entire searchable web and enabling inspection of linguistic characteristics of the full range of web registers in English. Studies have examined lexico-grammatical features strongly associated with register variation (e.g., Egbert et al. 2015; Biber & Egbert 2016, 2018). The CORE register taxonomy developed to cover the entire searchable web has also been adapted to other languages, including Swedish and French with SweCORE and FreCORE (Repo, Skantsi, Rönnqvist, Hellström, Oinonen, Salmela, Biber, Egbert, Pyysalo, & Laippala 2021), and Finnish with FinCORE (Laippala, Kyllönen, Egbert, Biber, & Pyysalo 2019; Skantsi & Laippala 2023). However, these studies presenting these corpora have focused on natural language processing (NLP) and register identification, leaving the linguistic variation across registers, languages, and cultures unexplored.

Understanding of web registers in languages like Turkish is notably lacking. Turkish, as a non-Indo-European language with a complex morphology and a different cultural context, can be

challenging to investigate. Its agglutinative nature results in lengthy words that can convey meanings that would otherwise be expressed in entire sentences in languages with less morphological complexity (Biber 1995; Lewis 2000). Current knowledge on Turkish web registers is inadequate, providing limited insight into their linguistic characteristics. While pioneering research on web registers has primarily concentrated on Indo-European languages, particularly English, expanding the linguistic analysis to languages with more morphological complexity and cultural distinctiveness will enrich our comprehension of web registers.

In this paper, following the principles set for CORE, we developed TurCORE: the Turkish Corpus of Online Registers. Our primary aims are (1) to explore the full range of registers found on the Turkish web, and (2) to conduct a detailed analysis of the linguistic characteristics of news reports, interactive discussion, and recipe registers. Our focus is primarily drawn to these registers due to their notable prevalence in TurCORE as well as their comparability with their English counterparts in CORE. We can compare the analyses of TurCORE with those of CORE, benefitting from the extensive research on English. However, since FinCORE has not been studied linguistically, the comparison between TurCORE and FinCORE is limited to the register distribution within their respective corpora.

We examine the linguistic characteristics of the registers with Text Dispersion Keyword Analysis (Egbert & Biber 2019). Keyword analysis aims at identifying the prominent words in a corpus reflecting its topic and style. Typically, keyness is measured by comparing the words in a target corpus with a reference corpus (Scott 1997; Baker 2004). However, this approach overlooks the distribution of words across texts and risks producing keywords prominent only in a fraction of the target corpus texts (Egbert & Biber 2019). Hence, we opt for text dispersion keyness, which is based on the dispersion of the words across texts. Our analysis categorizes keywords based on semantic and grammatical criteria. Given that most grammatical features are distributed in very different ways across registers (Staples, Egbert, Biber, & Conrad 2015), our grammatical analysis enables us to delve into the functional aspects beyond lexical information.

This article commences with a review of the existing literature on web registers, followed by the methodology section. The results are presented in two sections: Section 4 provides an overview of each register within TurCORE, while Section 5 focuses on three specific registers, offering linguistic analyses that reveal cultural aspects related to Turkish. Finally, the conclusion summarizes the overall contribution of the article.

2. Previous Work

2.1. Registers

Various terms have been employed to describe different situational, linguistic, and cultural variations, including ‘styles’, ‘genres’, and ‘registers’. We adopt the term ‘register’, as a register approach allows us to characterize text samples from diverse linguistic contexts (Biber & Conrad 2019), delve into linguistic features of the text samples (Egbert et al. 2015), and treat registers as cultural categories (Biber & Egbert 2023).

Registers are based on their situational characteristics, but they can be analyzed through linguistic features that are functionally tied to the situation (Biber 1988). This distinction is further discussed in Biber (1995), where it is framed as a distinction between the linguistic features versus the situational profiles of a text. Another perspective to this has been presented in Sharoff (2021), highlighting the distinction between text-internal and text-external aspects within registers. The

text-internal perspective focuses on the linguistic features in a text, whereas the text-external perspective considers the function a text serves in communication. Recent studies have primarily focused on online registers in restricted settings. For example, Berber-Sardinha (2018) employed multidimensional analysis to investigate blogs, webpages, Facebook, Twitter, and e-mails, revealing significant differences among the web registers examined. Research on register variation has extended to platforms like Reddit (Liimatta 2019, 2022).

Furthermore, research has examined the distinct linguistic features of specific online registers. For instance, Li, Li, Song, Li, Huang, and Ye (2021) studied inquiries on an academic Question & Answer platform, showing linguistic variations across disciplines. Candarli (2022) analyzed online academic discussions, revealing differences based on language backgrounds, forum post categories, and academic performance levels. Alazzawie (2022) investigated informal language patterns in WhatsApp text messages. These studies sample texts by selecting registers for analysis and then locating texts that fit the categories, thus not encompassing the full range of registers found on the web.

The advent of CORE by Egbert et al. (2015) has facilitated a comprehensive, unrestricted understanding of linguistic variation across the full range of online registers. Biber and Egbert (2018) analyzed the lexico-grammatical features of the CORE registers. Nonetheless, a gap exists in the understanding of web register variation in less-studied languages. Similarly, the emergence of cultural aspects associated with registers remains largely unknown. As exceptions, Olfert (2023) investigated language external factors that predominantly affect German heritage-language retention across intimate and formal registers, finding that factors that affect the retention vary depending on different registers.

2.2. Turkish

Turkish, as an agglutinative language, features an extensive range of suffixes, facilitating the formation of lengthy words through the expansion of stems. Biber (1995) notes that a single Turkish word often corresponds to many words in English, primarily due to suffixation, where new words are generated by adding suffixes to the right of a root. Many suffixes can be appended to a single root, with derivational suffixes typically preceding inflectional ones, which convey grammatical information such as case, person, and tense (Göksel & Kerslake 2005).

The example below shows how agglutination operates in Turkish allowing single-word sentences corresponding to many words in English. (All abbreviations used in the grammatical annotations are provided in the appendix.)

(1) Çalış-tır-ıl-ma-malı-y-mış.
work-caus-pass-neg-nec-cop-inf
“They say that s/he ought not to be made to work.”
(Lewis 1967; gloss from Comrie 1997: 14)

Previous research on Turkish registers has primarily focused on specific registers, including legislative language, recipes, and news reports. For example, Özyıldırım (2011) examined legislative language and compared it to various registers, revealing its predominantly nominal and expository characteristics. She found that legislative language presents straightforward and concise information, making it the least narrative among the registers studied. Koçak (2013) investigated Turkish recipes, comparing cookery books from different years to discern potential linguistic shifts

over time, revealing distinctive linguistic and discoursal features characterizing Turkish recipes. In a recent study, Kaya and Yağlı (2023) analyzed innocence claims made by a football club, employing critical discourse analysis to examine the registers formal speech and news reports. Their findings revealed that news reports either recontextualize or disregard the claims in alignment with ideological stances.

Previous studies have compared registers in Turkish with their English equivalents, revealing culture-specific characteristics. For example, Akbas (2014) and Can and Cangır (2019) examined academic language in Turkish and English dissertations. The analysis uncovered significant differences in the use of self-mentions. Turkish dissertations tended to downplay authorial identity to maintain objectivity, whereas English dissertations exhibited increased authorial involvement through greater use of self-mentions. Additionally, Can and Hatipoğlu (2023) investigated the conceptualization of congratulatory happy events in Turkish and English newspapers from a lexical perspective, revealing both similarities and differences.

3. Methodology

3.1. Data and Annotation Process

The documents in TurCORE are retrieved from the CommonCrawl datasets from the years 2019-2021 (https://commoncrawl.org). The CommonCrawl corpus includes petabytes of massively multilingual raw web page data, collected by systematic crawlers and updated regularly since 2008. Typically used in NLP for the development of large language models, the crawls target snapshots of the entire web without focusing on a particular language or domain. The data follows the WET format, where the documents have dedicated lines for metadata including the document URLs and the language, followed by the actual document content (the text) (Common Crawl - Blog - Navigating the WARC file format).

To ensure that the documents in TurCORE do not retain any structure from the original CommonCrawl datasets, we extracted the TurCORE documents from the 2019-2021 datasets through multiple rounds of random sampling. First, within each dataset, we focused on the documents identified as Turkish and randomly selected a subset using the random module in Python (https://docs.python.org/3/library/random.html). Then, we combined the datasets into one, and randomly chose a smaller sample that formed the basis of the datasets to be annotated.

CommonCrawl data are typically very noisy due to the automatic processes involved in crawling and compilation. To obtain the cleanest data possible, we retrieved the documents in HTML format from the available URLs. Then, we performed boilerplate removal using Trafilatura (Barbaresi 2021) and ran deduplication using Onion (Pomikalek 2011). Finally, the annotation process commenced on the refined texts.

Annotation was conducted on the annotation tool Prodigy (https://prodi.gy). Initially, the TurCORE dataset consisted of 3,767 unique web texts, totalling 1,512,183 words. However, to ensure the dataset quality for linguistic analysis, approximately 27 % of these web texts were excluded because they did not include entire coherent texts but only short list items and incoherent text. This left us with 2,780 texts, comprising 1,181,572 words.

The process of annotating the data was conducted by a supervisor and a trained annotator with a background in corpus linguistics and Turkish. We followed the taxonomy set for CORE (Egbert et al. 2015) and FinCORE (Laippala et al. 2019; Skantsi & Laippala 2023), which were developed in a hierarchical and data-driven manner covering full range of registers and linguistic

variation online. However, instead of the full taxonomy used in the English CORE, we used a simplified taxonomy based on CORE and FinCORE, which excludes registers that have been found infrequent and vaguely defined in previous studies (Repo et al. 2021; Skantsi & Laippala 2023).

Our annotation process involved a decision tree, where the annotator first assessed the mode of the text being spoken or written, determined whether the text is interactive or non-interactive, and identified the overall register of the text. Finally, the most accurate sub-register was chosen. In the cases where the text seemed to fit more than one register category, the annotator chose two or more registers, resulting in 'hybrid' registers. Instances of uncertainty were resolved through collaborative discussions with the annotators who have experience in annotating other languages.

The taxonomy adapted from English CORE and FinCORE to Turkish in this study recognizes 24 registers falling under 9 main register categories. Of the main registers, the Interactive Discussion, the Machine-translation, and the Lyrical do not have sub-registers.

Table 1 displays the registers of TurCORE. The main registers are shown in bold.

Table 1. Main and sub-registers of TurCORE.

Informational Description	**Opinion**	**Narrative**	**Informational Persuasion**	**How-to or Instructions**	**Spoken**	**Machine-translated**	**Interactive Discussion**	**Lyrical**
-Encyclopaedia article -Research article -Description of a thing/person -FAQ -Legal terms and conditions -Other	-Review -Opinion blog -Religious blog/sermon -Advice -Other	-News report -Sports report -Narrative blog	-Description with intent to sell -Editorial -Other	-Recipe -Other	-Interview -Other	(no sub-registers)	(no sub-registers)	(no sub-registers)

Note. Texts not featuring specific characteristics of a sub-register are annotated as “other”.

3.2. Keyword Analysis

The concept of keyness has evolved into a method, termed as keyword analysis, which has been a crucial component of quantitative text analysis to examine the characteristics of texts (Scott 1997). Typically, keyness is measured by identifying and ranking keywords within a target corpus compared to a reference corpus, relying on basic statistics and frequency calculations (e.g., Scott & Tribble 2006). However, this approach often overlooks individual variations within texts.

Egbert and Biber (2019) proposed determining keyness through dispersion, considering the number of documents in which a word appears. Text Dispersion Keyword Analysis (TDK) focuses on the analysis of texts rather than the entire corpus, disregarding word frequency. It is said to surpass traditional frequency-based methods in effectiveness (Gries 2022; Zhang, Jo, & Jhang 2022), and suits well the analysis of large corpora containing many texts (Egbert & Biber 2019).

Like all keyness methods, Text Dispersion Keyword compares the target corpus with a reference corpus. In our study, we created the reference corpus by incorporating all the corpus texts but excluding the texts of the target register. We examined the top 100 keywords with the highest rankings from the register category, aligning with Egbert and Biber (2019).

After identifying the keywords for each register, we organized them into semantic and grammatical groupings. Grammatical groupings were established based on the observation that certain linguistic features tend to co-occur in texts due to their interconnected functions (Biber & Egbert 2018). In terms of semantic categorization, we allocated each of the identified hundred words into relevant semantic categories to facilitate our understanding of their discourse functions. Following Biber and Egbert (2018), semantic categories were made based on the authors' judgments. To address the possibility of multiple senses of a word, we manually examined concordance lines. We found that the keywords in news reports and recipes were context-specific, resulting in clear semantic categories. However, the keywords in interactive discussion exhibited multiple senses, as the semantic categories of this register are more diverse compared to those of news reports and recipes. In such cases, we reviewed all concordance lines for the top 100 keywords in interactive discussions and categorized the words based on the most common sense.

4. Findings: Registers of TurCORE

In this section, we demonstrate all registers of TurCORE with brief descriptions to provide an overview of the registers.

Table 2 shows the main registers and the hybrids with their frequencies and the average word length of the texts (mean scores). The fact that Turkish is an agglutinative language where one word corresponds to several words in English affects the quantitative results, and this may be reflected in the mean scores.

Table 2. Frequencies and mean scores (average word length of texts) in TurCORE

Register	Number of texts	Percentage in TurCORE	Mean
Informational Persuasion	778	27.99 %	317
Narrative	663	23.85 %	249
Informational Description	481	17.30 %	409
Machine-translated	329	11.83 %	480
Opinion	215	7.73 %	502
How-to/Instructions	62	2.23 %	203
Interactive Discussion	50	1.79 %	1,035
Spoken	32	1.15 %	450
Lyrical	16	0.57 %	217
Hybrids	154	5.5. %	1,431
Total	2,780	100	425

As depicted in Table 2, TurCORE exhibits Informational Persuasion as the most frequent register, followed by Narrative and Informational Description. Spoken and Lyrical seem to have the smallest distribution. Hybrids, covering 5.5 % of the texts in TurCORE are primarily two-way, which means that they are the combinations of two registers. The most frequent combinations were found to be as 'Informational Description + Informational Persuasion' and 'Opinion + Informational Persuasion'.

In Figure 1, we compare the register distributions of TurCORE, FinCORE, and the English CORE, ordering registers from the most to the least frequent in Turkish. We approach these differences cautiously, considering that variations in distributions may stem from factors like corpus compilation methods or register annotation criteria.

Nevertheless, certain disparities in register distributions prompts speculation. For instance, Informational Persuasion constitutes the largest proportion of TurCORE, accounting for 28% of the corpus, while in FinCORE, it covers only 11%. In broad terms, approximately every 1 out of 4 texts observed on Turkish-speaking websites aims to persuade readers to purchase a product or service, whereas on Finnish-speaking websites, this ratio decreases to approximately 1 out of 9 texts. This difference could indicate that Finnish culture tends to prioritize directness, as noted by Tanova and Nadiri (2010). Potential customers in Finland are more likely to base their purchase decisions on explicit and straightforward information. In contrast, Turkish culture appears to place more emphasis on persuasion to influence customer choices. This can be because communication in Turkish culture often involves indirect messages and relies on emotional appeals (Tanova & Nadiri 2010). While these observations could apply to TurCORE and FinCORE, it is difficult to draw inferences regarding the minimal occurrence of Informational Persuasion within the CORE. This register ranks as the third least common with 1.6 % of the entire CORE. Roughly 1 in 60 texts attempts to persuade readers to purchase a product or service, which may not be attributed solely to a particular inclination within American culture.

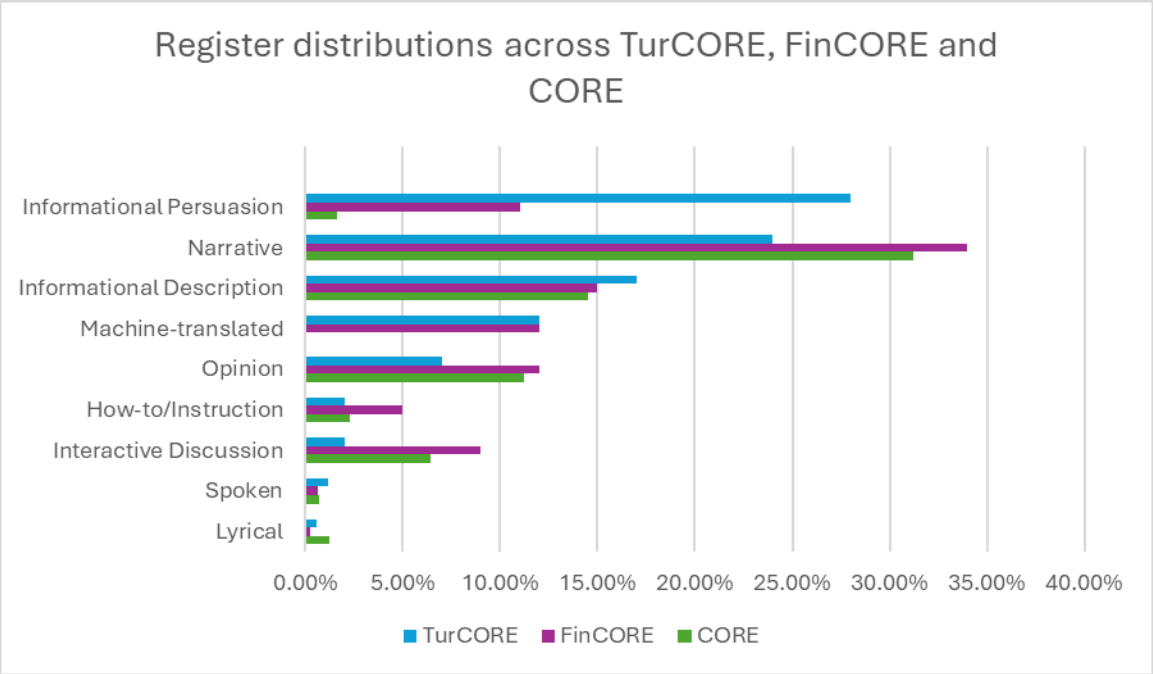

Figure 1. Distributional differences of registers across TurCORE, FinCORE, and CORE

Below, we present the main registers and their sub-registers, excluding Machine-translated due to its technical nature rather than linguistic, and omitting Spoken and Lyrical as they are the least frequent registers in TurCORE. Additionally, we illustrate the top 10 Turkish keywords of each register to highlight the primary content of each textual category with their English translations.

4.1. Informational Persuasion (IP)

Informational Persuasion comprises factual texts designed to persuade or market while simultaneously providing information to the readers. It is the most frequent register in TurCORE.

The most frequent sub-register identified within Informational Persuasion is 'Description with intent to sell', not only within the main category itself, but also across all registers. Texts falling under this category primarily show the contents of professional services, as can be seen from the keywords *our firm*, *quality*, *service of* and *shipment* in Table 3.

Table 3. Frequencies of Informational Persuasion registers and their top 10 keywords

Sub-register	Number of texts	Percentage in IP	Text Dispersion Keywords (Top 10)
Description with intent to sell	645	82.90%	*firmamız*, *kaliteli*, *sizlere*, *profesyonel*, *vermekteyiz*, *hizmeti*, *alabilirsiniz*, *nakliyat*, *fiyatları*, *escort*. *our firm*, *quality*, *to you all*, *professional*, *we provide*, *service of*, *you can buy*, *shipment*, *prices of*, *escort*.
Editorial	93	11.95%	*iktidar*, *ama*, *meselesi*, *bile*, *üstelik*, *karşı*, *yok*, *devlet*, *siyasi*, *ne*.

			rulership, *but*, *matter of*, *even*, *what's more*, *against*, *non-existent*, *state*, *political*, *what*.
Other IP	40	5.14%	*gerçekleştirilecek*, *buluşacak*, *düzenlenecek*, *programı*, *sanat*, *gerçekleştirecek*, *yürüten*, *sergisi*, *sanatın*, *tıklayınız*. *will be fulfilled*, *will meet*, *will be held*, *program*, *art*, *will fulfil*, *that who conducts*, *exhibition of*, *of art*, *please click*.

A significant portion of texts categorized under 'Editorial' in TurCORE were identified as newspaper columns. These columns are designed to persuade readers through opinion content, particularly focusing on political information and marked by stance markers such as *even*, *against*, and *non-existent.*

The remaining texts, classified under the 'Other IP' category, as they do not fall into any of the specific sub-registers, predominantly comprise descriptive texts outlining upcoming events. These can be seen from the keywords of future tense featuring *will be fulfilled*, *will meet*, and *will be held*.

4.2. Narrative (NA)

The primary function of Narrative is to report on past events, typically aimed at an audience with no prior special knowledge of the topic (Skantsi & Laippala 2023). In TurCORE, Narrative, comprising 663 texts, is the second most frequent register.

Table 4. Frequencies of Narrative registers and their top 10 keywords

Sub-register	Number of texts	Percentage in NA	Text Dispersion Keywords (Top 10)
New reports	556	83.86%	*dedi*, *başkanı*, *konuştu*, *söyledi*, *başkan*, *etti*, *kullandı*, *belediye*, *ifadelerini*, *edildi*. *s/he said*, *chairman of*, *s/he spoke*, *s/he said*, *chairman*, *s/he did*, *s/he used*, *municipality*, *expressions*, *was done*.
Narrative blog	52	7.84%	*ben*, *sanki*, *dedim*, *ettim*, *gittik*, *başladım*, *oturdu*, *çıktım*, *beni*, *o*. *I*, *as if*, *I said*, *I did*, *we went*, *I started*, *s/he sat*, *I went out/up*, *me*, *he/she/it*.
Sports reports	30	4.52%	*maçında*, *maçta*, *mağlup*, *deplasmanda*, *haftasında*, *sahasında*, *ikinci*, *maçtan*, *golle*, *berabere*. *in the game of*, *in the game*, *defeated*, *away*, *in the week of*, *in the field of*, *the second*, *from the game*, *with goal*, *all square*.

'News reports' exhibit characteristics of recent events reported by journalists (Biber & Egbert 2018), as seen from the keywords of *s/he said*, *s/he spoke*, *s/he did*, and *s/he used* in Table 4. This register is further analyzed with TDK in Section 5.

'Narrative blogs', one of the registers specific to the web, are personal logs, typically authored by amateur authors who have directly experienced the events (Biber & Egbert 2018). As seen in Table 4, Narrative blogs display a considerable number of personal elements related to the author and the narrative function, the majority of which is expressed with the first person singular and in the simple past tense.

'Sports reports' provide an account of events related to sports, as reflected in phrases such as *in the field of*, adjectives like *defeated* and temporal expressions such as *in the week of*. A deeper analysis of these keywords within the concordance lines reveals that they predominantly appear in football-related contexts, showing football as the most popular sport in Türkiye.

4.3. Informational Description (IN)

The purpose of Informational Description is to describe, inform, or explain something in detail where the authors are not usually indicated. The output can vary greatly from carefully written texts to unedited ones (Skantsi & Laippala 2023). In TurCORE, Informational Description consists of a total of 481 texts.

'Description of a thing or person' in TurCORE is primarily related to illnesses, as seen from the keywords *infections*, *symptoms of*, and *in the patients* in Table 5. The keyword *there is/are* (*vardır* in Turkish) further shows a notable feature. The *-DIr* suffix, commonly used in Turkish grammar to assert certainty, underscores the accuracy and validity of statements (Göksel & Kerslake 2005), particularly in factual descriptions. *Vardır* emerged as one of the prominent keywords in the register description of objects or individuals in TurCORE.

'Legal terms/conditions' concern any topic related to legality. In TurCORE, most of these texts pertains to delivery of purchases and the return policy of the products, as seen from the keywords *cargo* and *return*. This register also displays formal words specific to legal documents such as *hereby*.

The majority of 'Encyclopaedia articles' in TurCORE display biographical descriptions as seen from the keyword *was born* (*doğdu* and *doğmuştur* as two different keywords in Turkish). The years such as *1972*, *1989* and *2004* also display time-related numerals used in 'Encyclopaedia articles'.

Table 5. Frequencies of Informational Description registers and their top 10 keywords

Sub-register	Number of texts	Percentage in IN	Text Dispersion Keywords (Top 10)
Description of a thing or person	124	25.77%	*tedavisi*, *enfeksiyonlar*, *enfeksiyonun*, *belirtileri*, *fizik*, *rol*, *hastalarda*, *2005*, *hastalık*, *vardır*. *treatment of*, *infections*, *of the infection*, *symptoms of*, *physics*, *role*, *in the patients*, *2005*, *disease*, *there is/are*.
Legal terms	105	21.82%	*iade*, *sayılı*, *yasal*, *uyarınca*, *sözleşmesi*, *işbu*, *belirtilen*, *bilgilerin*, *kişisel*, *kargo*.

			return, *numbered*, *legal*, *in accordance with*, *contract of*, *hereby*, *stated*, *of information*, *personal*, *cargo*.
Encyclopaedia article	18	3.74%	*doğdu*, *1972*, *2004*, *1988*, *silmeden*, *evrenselliğe*, *1989*, *doğmuştur*, *adlı*, *rock*. *was born*, *1972*, *2004*, *1988*, *wraparound*, *to universality*, *1989*, *was born*, *with the name of*, *rock*.
FAQ	6	1.24%	*hazırlıyoruz*, *inceleyebilir*, *başvurarak*, *araması*, *bulunabilir*, *yönlendirme*, *çatlatma*, *pod*, *yurdumuz*, *sgkya*. *we are preparing*, *can examine*, *by consulting*, *searching of*, *can be found*, *guidance*, *fracturing*, *pod*, *our homeland*, *to 'sgk'* (social security institution).
Research article	4	0.83%	*sendromu*, *etkilerini*, *araştırmak*, *frekans*, *frekanslı*, *polariteli*, *coronavirüsler*, *ailesidir*, *polarite*, *algınlığından*. *syndrome of*, *its effects*, *to search*, *frequency*, *with frequency*, *with polarity*, *coronaviruses*, *is the family of*, *polarity*, *from the delusion of*.
Other IN	224	46.56%	*kıyasla*, *verilere*, *oldu*, *artış*, *artarken*, *yahoo*, *paratic*, *kemiklerin*, *danıştay*, *tracking*. *by comparison*, *to the data*, *happened*, *increase*, *while increasing*, *yahoo*, *paratic*, *of the bones*, *state council*, *tracking*.

In TurCORE, there are only a few texts categorized as 'FAQs' and 'research articles'. The keywords associated with them are displayed in Table 5.

The sub-register 'Other IN' emerges as the most frequent within the main category of Informational Description. This sub-register accounts for nearly half of all Informational Description texts. The keywords associated with 'Other IN' do not indicate any shared characteristics.

4.4. Opinion (OP)

Opinion expresses subjective viewpoints based on personal opinion of an author or a group of authors (Biber & Egbert 2018). In TurCORE, a total of 215 texts are of opinionated nature which are expressed in five sub-registers shown in Table 6.

Table 6. Frequencies of Opinion registers and their top 10 keywords

Sub-register	Number of texts	Percentage in OP	Text Dispersion Keywords (Top 10)
Review	66	30.69%	*izlenim*, *açıkçası*, *filmde*, *belgeselin*, *kitapta*, *a0*, *android*, *filmleri*, *modelin*, *filmi.* *impression*, *frankly*, *in the film*, *of the documentary*, *in the book*, *a0*, *android*, *films of*, *of the model*, *film.*

Opinion blog	58	26.97%	*mi*, *şey*, *ne*, *hiç*, *ama*, *kim*, *o*, *bunu*, *değil*, *belki.* *mi* (a question particle), *thing*, *what*, *never/any*, *but*, *who*, *he/she/it*, *this*, *not*, *maybe.*
Advice	46	21.38%	*burcu*, *kendinizi*, *burç*, *sizi*, *yaşamınızda*, *duygusal*, *nasıl*, *yorumu*, *dikkat*, *insanlar.* *zodiac of*, *yourselves*, *zodiac*, *you*, *in your life*, *emotional*, *how*, *interpretation of*, *carefulness*, *people.*
Religious blog/sermon	29	13.48%	*allah*, *ey*, *peygamber*, *suresi*, *allahın*, *ayet*, *bakara*, *namaz*, *onu*, *muhammed.* *Allah*, *ey* (an interjection used in poetic contexts), *prophet*, *surah*, *of Allah*, *verse*, *Baqarah*, *prayers*, *him*, *Muhammad.*
Other OP	16	7.44%	*rüyada*, *yorumlanır*, *rüyayı*, *tabir*, *sahibinin*, *rüyasında*, *kimseleri*, *işarettir*, *rüya*, *işaret.* *in dream*, *it is interpreted*, *dream*, *interpretation*, *of owner*, *in dream of*, *folks*, *is a sign*, *dream*, *sign.*

'Review' is an evaluation of a product, or service written with no intent to sell. Within TurCORE, it emerges as the most frequent sub-register of Opinion with typical examples including the evaluations of films, documentaries, and books expressed in the keywords *impression*, *film*, *in the film*, *in the book*, and *of the documentary.*

'Opinion blog' aims at sharing personal viewpoints publicly without seeking to persuade the reader and it includes expressions of evaluation and stance such as *never*/*any*, *but*, and *maybe.* Furthermore, it can have a wide variety of topics.

'Advice' offers recommendation that leads to actions with the aim of solving a particular problem (Biber & Egbert 2018). 'Advice' texts in TurCORE offer guidance to the reader, as seen from the second person pronouns *yourselves* and *you*. The keyword *how* also displays guidance which is expected to lead to actions.

'Religious blog/sermon' consists of any denominational religious text excluding the ones describing a religion (Biber & Egbert 2018). The predominant religion in Türkiye is Islam, and it is reflected in the 'Religious blog/sermon' texts of TurCORE, as seen from the keywords *Allah*, *Baqarah*, and *Muhammad*.

'Other OP' encompasses texts that do not align with the aforementioned sub-registers. In cases where labels were ambiguous between 'Other OP' and 'Editorial', distinctions were made based on the level of argumentation. Typically, 'Other OP' features less substantial reasoning than 'Editorial', such as interpretation of dreams as seen from the keywords *in dream of*, *is a sign*, and *interpretations*.

4.5. How-to/Instructions (HI)

How-to/Instructions gives step-by-step instructions to accomplish a task.

Table 7. Frequencies of How-to/Instructions registers and their top 10 keywords

Sub-register	Number of texts	Percentage in HI	Text Dispersion Keywords (Top 10)
Recipe	40	64.51%	*tarifi*, *afiyet*, *kaşığı*, *malzemeler*, *bardağı*, *tuz*, *yapılışı*, *yoğurt*, *karıştırın*, *suyunu*. *recipe of*, *appetite*, *spoon of*, *ingredients*, *glass of*, *salt*, *making of*, *yoghurt*, *stir*, *water of*.
Other HI	22	35.48%	*tıklayarak*, *tamamlanmasını*, *başlat*, *disk*, *işlemini*, *işaretli*, *bilgisayarınızı*, *komutunu*, *basınız*, *indirdiğiniz*. *by clicking*, *completion of*, *start*, *disc*, *processing*, *marked*, *your computer*, *command of*, *press*, *that you downloaded*.

Recipe, another register further analyzed in Section 5, presents a list of ingredients and a set of instructions for preparing a food product (Asheghi, Sharoff, & Markert 2016; Biber & Egbert 2018). The keywords in Table 7 such as *glass of* as a measurement unit and *yoghurt* as a commonly used ingredient in the Turkish cuisine showcase culture-specific features of this register in TurCORE.

'Other HI' shares similarities with 'Recipe' in providing detailed, step-by-step instructions, although the instructions pertain to how to perform a task. 'Other HI' differs from 'Advice' in that 'Advice' is subjective while 'Other HI' is objective (Biber & Egbert 2018). In TurCORE, examples of 'Other HI' predominantly include processes related to computer usage, as seen from the keywords *by clicking*, *disc*, and *that you downloaded.*

4.6. Interactive Discussion

Interactive Discussion, another register specifically focused in Section 5, is a forum where people have a conversation about a certain topic. This register has an interactive nature where the individual initiating the discussion actively participates after initializing it (Skantsi & Laippala 2023). Texts are not categorized as Interactive Discussion if they primarily comprise reader comments following an article or blog post, which are already visible within the text (Biber & Egbert 2018). To emphasize the distinctive interaction and discussion characteristics inherent in this register, our analysis exclusively considered content written by multiple participants engaging in interactive discussions.

Table 8. Frequency of interactive discussion register and its top 10 keywords

Register	Number of texts	Text Dispersion Keywords (Top 10)
Interactive Discussion	50	*burda*, *bi*, *cok*, *mi*, *dogru*, *suan*, *bende*, *bey*, *bilmem*, *simdi.* *here*, *a/one*, *very*, *mi*, *correct*, *right now*, *I have*, *mister*, *I don't know*, *now*.

5. Findings: Lexical and Grammatical Analyses

In this section, we present the detailed analysis of the three selected registers. For each register, we begin by analyzing the semantic groupings of the keywords, followed by their grammatical groupings, thus offering a comprehensive perspective to the linguistic characteristics reflected by the keywords.

At the end of each sub-section of the register, we present an excerpt from TurCORE exemplifying the characteristics discussed. These excerpts are from the initial sections of the original texts. The English translations were conducted by paying particular attention to preserving the style and tone of the original Turkish texts.

5.1. News Reports

News reports are one of the most primary journalistic materials in media which report on past events. We identified the semantic categories for Turkish news reports in Table 9.

Table 9. Semantic categories of Turkish news reports with examples

Semantic category	**Sample keywords**
administration	*başkan*, *belediye*, *heyet*, *içişleri* *chairman*, *municipality*, *board*, *internal affairs*
communication	*ifade*, *açıklama, demek*, *konuşmak*, *belirtmek* *expression*, *explanation*, *to say*, *to speak*, *to indicate*
action	*gerçekleştirmek*, *düzenlemek*, *yapmak*, *başlamak* *to fulfil*, *to organize*, *to make*, *to start*
emergency	*koronavirüs*, *itfaiye*, *salgın*, *yaralı* *coronavirus*, *fire department*, *epidemic*, *injured*
legality	*soruşturma*, *polis*, *gözaltı*, *jandarma* *investigation*, *police*, *custody*, *gendarme*
time	*devam*, *saat*, *ardından* *continuation*, *time*, *afterward*
journalism	*muhabir*, *basın*, *haber* *reporter*, *press*, *news*
names and abbreviations	*Chp*, *Recep*, *Mehmet*, *AA* *Chp* (Republican party), *Recep*, *Mehmet*, *AA* (A. Agency)
numbers	*bin*, *yüzde*, *milyon*, *19* *thousand*, *percent*, *million*, *19*
other	*şunlar*, *ilişkin*, *tarım*, *inşallah* *those*, *related*, *agriculture*, *God willing*

Our comparative observations reveal similarities between the semantic patterns found in Turkish and English news reports. Categories such as ‘administration’, ‘communication’, ‘journalism’, ‘names and abbreviations’, and ‘numbers’ in TurCORE appear to correspond closely to the categories of “government”, “reporting”, “news features”, “people’s names and titles”, and “figures/details” in the English CORE (Biber & Egbert 2018: 85), respectively.

Grammatical categories are provided in Table 10.

Table 10. Prevalent grammatical categories observed for Turkish news reports with examples

Grammatical categories	Keyword in English	Keyword in Turkish	Grammatical annotation
passive voice	*it was informed* *which was organized* *which is not used*	*bildir-il-di* *düzenle-n-en* *kullan-ıl-ma-yan*	inform+pass+pf.3S organize+pass+part use+pass+neg+part
past tense	*said* *spoke* *indicated*	*de-di* *konuş-tu* *belirt-ti*	say+pf.3S speak+pf.3S indicate+pf.3S
third-person singular	*s/he noted* *s/he participated* *s/he explained*	*kaydet-ti* *katıl-dı* *açıkla-dı*	note+pf.3S participate+pf.3S explain+pf.3S
relative clause	*that underline-d/-s* *that mentione-d/-s* *that told/tells*	*vurgula-yan* *değin-en* *söyle-yen*	underline+part mention+part tell+part

Among the top 100 keywords of news reports, 33 were identified as verbs. Many of these verb forms consistently appear in the simple past tense followed by the third person singular, mirroring a common pattern in English news reports (Biber & Egbert 2018). In Turkish, due to its agglutinative nature where suffixes for voice, tense, and person are appended together, the prevalent pattern of past tense followed by third person singular, occasionally in the passive voice, is particularly noticeable. This grammatical pattern in news reports aligns with the characteristics of the register, which involves recounting past events.

Relative clauses are attributive constructions that modify noun phrases. In Turkish, relative clauses are marked with suffixes and they correspond to the relative pronouns *who*, *which*, *that*, *whom*, *whose*, *where*, etc. in English (Göksel & Kerslake 2005). In Turkish news reports, the majority of non-finite verb forms were found to be formed with *–(y)An* suffix, as seen in Table 10. The use of relative clauses might be a stylistic preference to avoid a monotonous succession of finite clauses and is otherwise expressed in the simple past tense + third person singular pattern in news reports.

Text Sample 1 highlights prevalent lexical and grammatical features of news reports. Lexical features such as communication and journalism words and proper nouns are shown in bold. Grammatical features such as simple past tense and third person singular are underlined.

Text Sample 1a. News report retrieved from www.bursadabugun.com/haber/dsovirusun-kokenini-arastiracak-heyetin-cin-e-girisine-izin-verilmedi-1370574.html

DSÖ, **Covid-19**'un kökenini araştıracak bilim heyetinin **Çin**'e girişine izin verilmediğini **duyur**du**. Dünya Sağlık Örgütü (DSÖ)** Genel Direktörü Dr. **Tedros Adhanom Ghebreyesus**, yeni tip corona virüsün (**Covid-19**) kökenini araştıracak uluslararası uzmanlardan oluşan bilim heyetinin **Çin**'e girişine izin verilmemesinden "büyük hayal kırıklığına" uğradığını **bildir**di. **Ghebreyesus**, **DSÖ**'nün **İsviçre**'nin **Cenevre** kentindeki merkezinde, video konferans yöntemiyle 2021'in ilk **basın** toplantısını **düzenle**di. **Covid-19**'un kökeni araştırmak için bu hafta **Çin**'e gitmesi beklenen uluslararası bilim insanlarının

durumuna ilişkin **konuş**an **Ghebreyesus**, "Bugün, Çinli yetkililerin **Çin'**e gidecek heyet için gerekli izinleri henüz tamamlamadığını **öğren**dik. Ekibin iki üyesinin (**Çin**'e) yolculuklarına çoktan başlamış olması ve diğerlerinin de son dakikada seyahat edememesi nedeniyle bu **haber** beni büyük hayal kırıklığına uğrattı" **de**di. Üst düzey Çinli yetkililerle temas halinde olduğunu **aktar**an **Ghebreyesus**, "virüsün kökenini araştırma misyonu"nun **DSÖ** ve uluslararası heyet için "öncelik" olduğunu Çinli yetkililere bir kez daha açıkça **ilet**tiğini **belirt**ti.

Text Sample 1b. Translation of the news report

The **WHO announce**d that the scientific delegation to investigate the origin of **Covid-19** was allowed to enter **China**. **World Health Organization (WHO)** Director-General Dr. **Tedros Adhanom Ghebreyesus indicate**d that he was "very disappointed" that the scientific delegation of international experts investigating the origin of the novel coronavirus (**Covid-19**) was not allowed to enter **China**. **Ghebreyesus held** the first **press** conference of 2021 via videoconference at the **WHO**'s headquarters in **Geneva**, **Switzerland**. **Speak**ing about the situation of international scientists who are expected to go to **China** this week to investigate the origin of **Covid-19**, **Ghebreyesus said**, "Today, we **learn**ed that the Chinese authorities have not yet completed the necessary permits for the delegation to go to **China**. I am very disappointed by this **news** as two members of the team had already started their journey (to **China**) and the others were unable to travel at the last minute," he said. **Convey**ing that he was in contact with senior Chinese officials, **Ghebreyesus state**d that he, once again, clearly **deliver**ed to them the "mission to investigate the origin of the virus" was a "priority" for the **WHO** and the international delegation.

The regularity in the structure of the finite verbs, characterized by the usage of simple past tense and third person singular form in Turkish news reports, aligns well with the reporting of events. Additionally, the abundance of communication and journalism words, along with proper nouns, resonates with the findings of Biber and Egbert (2018) regarding the English CORE. When semantic and grammatical analyses are considered together, it becomes evident that news reports in Turkish and English share similar linguistic characteristics.

5.2. Interactive Discussion

Interactive discussions involve communication conducted on the Internet, constituting a form of register which is specific to the web. The interactive aspect signifies that participants actively contribute, respond, and collaborate within a digital environment such as a forum or social media platform. For Turkish interactive discussions, we identified the categories in Table 11.

Table 11. Semantic categories of Turkish interactive discussions with examples

Semantic category	**Sample keywords**
place	*burda*, *orda*, *yukarda* *here*, *there*, *above*
time	*şimdi*, *şuan*, *ozaman*, *önce* *now*, *right now*, *then*, *before*
addressing	*bey*, *hanım*, *kardeşim*, *arkadaşlar* *mister*, *mistress*, *my bro*, *guys*

salutation	*merhabalar*, *selam*, *saygılarımla* *hellos*, *hi*, *sincerely*
discourse organizer	*bak*, *neyse*, *inşallah* *look*, *anyways*, *God willing*
opinion	*katılıyorum*, *bence*, *sanıyor* *I agree*, *in my opinion*, *s/he thinks*
evaluation	*doğru*, *güzeldi*, *aynı*, *baya* *correct*, *it was nice*, *same*, *quite/pretty*
other	*bir*, *çok*, *diğer*, *şey* *a/one*, *very*, *other*, *stuff*

Our Turkish-English comparison reveals differences in the semantic categories. In English, these include "forums", "pronouns", "contractions", "abbreviations", "stance", and "interaction" (Biber & Egbert 2018: 188), which differ from those in Turkish, particularly, in the notable prevalence of forum-specific words in English. Similarly, 'abbreviations' and 'stance' words are not as abundant in Turkish. The keywords categorised as 'opinion' and 'evaluation' in TurCORE seem to be similar to the keywords categorized as 'stance' in CORE. We preferred not to merge these two categories as each has their own keywords. In CORE, however, 'stance' also includes certain adverbs which were not identified in Turkish interactive discussions.

Keywords for terms of address are not present in the interactive discussions of CORE. They not only exist in TurCORE but also displays hints of culture. In Turkish culture, when the person who is being talked to is a stranger or hierarchically above the speaker, they are addressed as *mister* or *mistress*. These terms of address using the formal *mister* and *mistress* in Turkish interactive discussions show that spoken language is mimicked, sometimes even by violating the rules of writing, but this spoken language does not have to be informal. The participants of the interactive discussion might still want to keep the formal tone. This formal tone is also visible in the salutation word *sincerely*.

Table 12 displays salient grammatical categories discovered in the top 100 keywords of Turkish interactive discussions.

Table 12. Prevalent grammatical categories observed for Turkish interactive discussions with examples

Grammatical category	**Keyword in Turkish**	**Keyword in English**
contractions	*bi* (*bir*) *oluyo* (*oluyor*) *baya* (*bayağı*)	*a/an*, *one* *happening* *quite*, *pretty*
spelling variants	*şuan* (*şu an*) *ney* (*ne*) *ozaman* (*o zaman*)	*right now* *what* *then*
mental verbs	*bilmiyorum* *anlamadım* *sanıyor*	*I don't know* *I didn't understand* *s/he supposes*
first-person singular pronoun	*benim* *beni* *bana*	*my* *me* *to me*
questions	*kaç*	*how many*

	var mı *değil mi*	*is/are there?* *isn't it?*

The heavy use of contractions (e.g., *baya* instead of *bayağı* "*quite*/*pretty*") and spelling variants (e.g., *ney* instead of *ne* "*what*") seem to be the outcomes of imitating colloquial language.

An analysis of the first 100 keywords further reveals a consistent presence of mental process verbs of *know*, *think*, and *understand* in Turkish interactive discussions, mirroring their occurrence in English. However, in TurCORE, these mental verbs manifest in various verb forms and tenses, often accompanied by the first-person singular, as exemplified in the keywords of *bilmiyorum* (*I don't know*) and *anlamadım* (*I did not understand*). The findings align with previous research conducted by Erten (2019), highlighting a connection between the use of mental verbs and the first-person singular.

Turkish allows mental and state verbs to be used in the imperfective aspect with the suffix *— (I)yor* in an ongoing viewpoint. The use of the imperfective aspect is very common in Turkish compared to languages like English, which does not canonically allow an ongoing temporal viewpoint with verbs expressing perception, cognition, or state. This distinction is evident in interactive discussions *Bilmiyorum* (lit. *I am not knowing*) and *sanıyor* (lit. *s/he is supposing*) are some examples that can be seen in Table 12.

Different forms of questions also seem to be used relatively often in Interactive Discussions. Question words such as *ne* (*what*), question particles such as *mi* and *mı*, which are untranslatable into English in one word, and the tag questions such as *değil mi* (*isn't it*') show the interactive nature of this register. The particle *mI* forms yes/no questions, and is also used in tag questions when the speaker seeks corroboration of a statement that s/he believes to be true (Göksel & Kerslake 2005).

Figure 2 extracted from a forum, displays almost no use of punctuation and capital letters. It also exhibits spelling variants such as *buda* instead of *bu da* (*and this*), and *içinde* instead of *için de* (*also for*), displaying that interactive discussion texts reflect colloquial language.

Figure 2a. Interactive discussion sample extracted from www.chip.com.tr/forum/gtx860-ekran-kartimdan-performans-alamiyorum_t299184.html

CHIP Online > Forum > Test and purchasing > Laptop and Notebook

I can't get performance from my TX860 graphics card

ataberkasdaqs1	31-01-2015, 00:58 | #1
New member Number of thanks: 0 2 messages Sign-up Date: Jan 2015	Hi friends I have monster abra v.1.1 there are those of you who know around 2600 TL it has been almost 7-8 months since I bought it at first it was playing cs go without any problems but right now although I formatted it just yesterday I am experiencing a sudden loss of fps in a game like cs go at the lowest settings and it is a very frustrating situation I bought a fan for under it there is nothing that has changed my computer's features: GTX860 2GB Graphics card 8 Gb ram and i7 the compuder that I was very satisfied with at the beginning is now disappointing me what should I do friends I am waiting for your help. Noteit is not the situation only for cs go it is a problem with almost all games)

Figure 2b. Translation of the interactive discussion sample.

The presence of contractions, spelling variants, and mental verbs predominantly used in first-person singular are in line with the findings of the English interactive discussions presented in CORE (Biber & Egbert 2018). Nonetheless, through a combined examination of semantics and grammar, it becomes apparent that Turkish interactive discussions tend to highlight features specific to Turkish culture (e.g., addressing terms) and language (e.g., ongoing viewpoint in the usage of mental verbs).

5.3. Recipes

Recipes are a well-defined procedural register with specific writing objectives. The typical structure of recipes follows a pattern, including elements such as the title, indications of use, ingredients, preparation and dosage instructions, application guidelines with sometimes assurance of efficacy at the end (Taavitsainen 2001). We observed the semantic categories in Turkish recipes in Table 13.

Table 13. Semantic categories of Turkish recipes with examples

Semantic category	**Sample keywords**
ingredients	*kıyma*, *yoğurt*, *salça*, *nane* *minced meat*, *yoghurt*, *salça* (a sauce), *mint*
pre-preparation	*yağlanmış*, *ısıtılmış*, *doğranmış* *which has been greased*, *which has been warmed*, *which has been chopped*
action	*eklemek*, *dökmek*, *serpmek*, *karıştırmak* *to add*, *to pour*, *to sprinkle*, *to mix*
measurement unit	*su bardağı*, *çay bardağı*, *tatlı kaşığı* *water glass*, *tea glass*, *dessert spoon*
utensil	*tencere*, *tava*, *tepsi*, *kap* *pot*, *pan*, *tray*, *pot*
kitchen appliance	*ocak*, *fırın*, *buzdolabı*

	stove, *oven*, *fridge*
evaluation	*iyice*, *güzelce*, *nefis*, *lezzet* *well*, *well*, *delicious*, *taste*
other	*tarif*, *kıvam*, *yemeklik* *recipe*, *consistency*, *cooking*

Our comparative analysis highlights similarities in recipe-related vocabulary between English and Turkish, with English recipe words grouped under "actions", "ingredients", "tools", "attributes and evaluations", and "time/quantity" (Biber & Egbert 2018: 158). The bulk of Turkish recipe keywords align with their English equivalents, particularly in ingredient words. However, certain ingredients common in Turkish recipes, such as *salça* (a special sauce) and *yoghurt*, are not typically encountered in English recipes, underscoring their Turkish culture specificity. Another distinction lies in the occurrence of action words. As noted in Biber and Egbert (2018), English recipes exhibit a wide range of action word vocabulary, which is not the case in Turkish recipes. This divergence in semantic categories may be attributed to linguistic differences between English and Turkish. While English encompasses a diverse array of action verbs, Turkish seems to rely on a more limited set of verbs expressed through various grammatical structures.

Manual analysis showed that some keywords in Turkish recipes are complemented with other keywords. In the English CORE, the measurement unit for a spoon is typically represented as a tablespoon or teaspoon (Biber & Egbert 2018). Although they also exist within TurCORE, Turkish recipes further demonstrate the unique measurement unit *dessert spoon*. What further distinguishes Turkish recipes is the prevalent use of *water glass* and *tea glass* as measurement units which are different than the word *cup*, commonly employed for similar purposes in American English recipes. This distinct feature in Turkish recipes is closely intertwined with Turkish culture, especially in the case of the *tea glass* unit. Türkiye's tea-centred culture boasts a rich lexicon of words and phrases associated with tea. The ubiquity of tea glasses in the culture underpins their prominent usage in Turkish recipes.

Table 14 displays the prevalent grammatical categories of the Turkish recipe keywords.

Table 14. Prevalent grammatical categories observed for Turkish recipes with examples

Grammatical category	**Keyword in English**	**Keyword in Turkish**	**Grammatical annotation**
case-marked noun phrase	*the ingredient* *to the pot* *in the owen*	*malzeme-yi* *tencere-ye* *fırın-da*	ingredient+acc pot+dat owen+loc
noun compound	*olive oil* *fridge* *black pepper*	*zeytinyağı* *buzdolabı* *karabiber*	olive+oil+nc ice+cupboard+nc black+pepper
action verb as optative	*stir!* *boil!* *let's rest it!*	*karıştır-ın* *haşla-yın* *dinlendir-elim*	stir+opt+2P boil+opt+2P rest+opt+1P
adjectival	*which has been greased* *which has been warmed up* *which has been chopped*	*yağla-n-mış* *ısıt-ıl-mış* *doğra-n-mış*	grease+pass+part warm up+pass+part chop+pass+part

In Turkish, case markers are affixed to the root of the word, resulting in one-word noun phrases. Within the context of recipes, these noun phrases signify distinctive Turkish linguistic features. A prominent presence of case-marked noun phrases was observed in the accusative, dative, and locative cases, indicating instructions regarding the placement or positioning of objects. For instance, the noun *ingredient* (*malzemeyi*) in the accusative case signifies an object that may be placed *to the pot* (*tencereye*) in the dative case and baked *in the owen* (*fırında*) in the locative case. The presence of these cases in Turkish recipes serves to reinforce the instructional nature inherent in recipes.

Noun compounds are word-like units which are made up of two nouns (*–(s)I* compound) or an adjective and a noun (bare compounds) (Göksel & Kerslake 2005). Turkish recipes display both types of noun compounds. Being an adjective + noun compound, *karabiber* (*black pepper*) is an example of bare compounds. Being made up of two nouns, *zeytinyağı* (*olive oil*) and *buzdolabı* (*fridge*; lit. *ice cupboard*) are the examples of *–(s)I* compound.

Finite verb forms in recipes predominantly target the second person plural, emphasizing reader engagement and action on their part. Furthermore, these verb forms are saliently found in optative forms, the majority of which are expressed as imperatives (e.g., *mix!*, *boil!*). This abundant use of the second person plural in imperative forms in Turkish recipes aligns with the observed patterns in recipe keywords in English. Nevertheless, in TurCORE, two optative predicates were detected in the first-person plural, one of which is exemplified in Table 14 with the keyword *dinlendirelim 'let's rest it'*. The use of first-person plural optatives in Turkish recipes show another specific characteristic of Turkish recipe writing. Considering that Turkish culture can often be regarded as collectivist (Ayçiçegi-Dinn & Caldwell-Harris 2011; Arpaci & Baloğlu 2016), the presence of the first-person plural optatives in recipes appears to resonate with this collectivist aspect of Turkish culture.

Three verb forms were identified as attributive constructions within the top 100 keywords: *yağlanmış* (*which has been greased*), *ısıtılmış* (*which has been warmed up*), and *doğranmış* (*which has been chopped*). These verbs take the form of participles, signifying the use of the passive voice. This is interesting, as recipes guide the reader through the cooking or baking process step-by-step by instructing the reader to perform actions in the present tense (Biber & Egbert 2018). However, in instances like *yağlanmış*, *ısıtılmış*, and *doğranmış* in Turkish recipes, it appears that some actions are implied to have been completed prior to the reader's engagement in Turkish recipes.

Text Sample 2 shows the features of Turkish recipes. The first-person plural optatives are shown in bold while action verbs are underlined.

Text Sample 2a. Recipe retrieved from www.yemektarifhane.com/firin-posetinde tavuk-2/

Tavuk bagetleri yıkayıp derin bir kabın içine **alalım**.
Üzerine doğradığımız biberleri ve 5-6 parçaya böldüğümüz patatesleri **dökelim**.
Sonra tereyağını, sıvı yağı, tuzu, kekiği, karabiberi, toz kırmızı biberi ve dövülmüş sarımsağı ekleyip güzelce **harmanlayalım**.
Fırın poşetinin içini **unlayalım**.
Harmanlanan malzemeyi fırın poşetinin içine doldurup ağzını sıkı bir şekilde **kapatalım**.
Bu şekilde buzdolabında en az 1 saat **dinlendirelim**.

Text Sample 2b. Translation of the recipe

Washing the chicken drumsticks, **let's put** them in a deep bowl.
Let's pour the peppers we chopped and the potatoes we cut into 5-6 pieces on the top.
Then adding the butter, oil, salt, oregano, black pepper, powdered paprika and crushed garlic, **let's blend** them well.
Let's flour the inside of the baking bag.
Filling the blended material into the oven bag, **let's close** it tightly.
In this way, **let's rest** it in the refrigerator for at least 1 hour.

When evaluated together, the general semantic and grammatical characteristics reflected by keywords found in Turkish and English recipes demonstrate a degree of similarity. Nevertheless, a closer analysis reveals that the distinct traits of Turkish recipes, compared to their English equivalents, highlight culturally embedded features, which are also apparent in language-specific aspects unique to Turkish.

6. Conclusion

In this article, we followed the principles established for the English CORE (Egbert et al. 2015) in compiling and annotating our corpus TurCORE, thereby creating the first register-annotated corpus for Turkish web language use. By identifying web registers on the Turkish-speaking web and examining their distribution within TurCORE, we conducted an in-depth exploration of the linguistic characteristics exhibited by specific web registers. Our linguistic analyses were based on Biber and Egbert's (2018) studies of lexico-grammatical features in English web registers, and we uncovered cultural aspects as a result of our linguistic analyses, in line with Biber and Egbert (2023).

Our analysis identified 24 web registers categorized into the main registers of Informational Persuasion, Narrative, Informational Description, Machine-translated, Opinion, How-to/Instructions, Interactive Discussions, Spoken, and Lyrical. Despite notable distributional differences between TurCORE and the English CORE, the register categories were similar, with TurCORE featuring a simplified categorization due to the annotation process. Although FinCORE (Laippala et al. 2019; Skantsi & Laippala 2023) is the first web register corpus for a non-Indo-European language following CORE principles, our comparison of TurCORE with FinCORE was limited to distributional differences, because no linguistic analyses of FinCORE exist. Thus, our study contributes to the understanding of linguistic and cultural aspects of web registers in a non-Indo-European language.

We utilized Text Dispersion Keyword Analysis (Egbert & Biber 2019) to examine specific registers of news reports, interactive discussions, and recipes. The study revealed diverse Turkish keywords, some exhibiting lexical and grammatical similarities to their English counterparts. However, Turkish interactive discussions and recipes also displayed certain unique attributes, notably the use of culturally rooted words and Turkish-specific grammatical structures. These findings highlight the multifaceted nature of online registers and provide insights into how linguistic features are shaped by specific communicative objectives.

The findings also enhance our understanding of how linguistic features of registers reflect cultural interpretations. Biber and Egbert (2023) recently incorporated culture into register definitions, and studies like Olfert's (2023) have explored heritage language from a register variation perspective. Our research demonstrates that linguistic analyses of registers can provide cultural insights, emphasizing the intertwined nature of registers and culture. While more culture-related investigations are needed in register studies, our work contributes to this field by elucidating the connection between registers and culture.

We hope that the findings of this study will serve as a basis for future research exploring comprehensive linguistic characteristics and cultural dimensions of web-based communication in lesser-studied non-Indo-European languages. Additionally, we expect our research to stimulate comparative studies across different languages.

Funding

The study was funded by the Eino Jutikkala Fund of the Finnish Academy of Science and Letters, and the Research Council of Finland.

Acknowledgements

We would like to thank our anonymous reviewers for their careful reading of the manuscript and their invaluable comments.

References

Akbas, E. (2014). Are they discussing in the same way? Interactional metadiscourse in Turkish writers' texts. In A. Łyda & K. Warchał (Eds.), *Occupying niches: Interculturality, cross-culturality and aculturality in academic research* (pp. 119-133). Springer International Publishing. https://doi.org/10.1007/978-3-319-02526-1_8

Alazzawie, A. (2022). The linguistic and situational features of WhatsApp messages among high school and university Canadian students. *SAGE Open*, *12*(1). https://doi.org/10.1177/21582440221082124

Arpaci, I., & Baloğlu, M. (2016). The impact of cultural collectivism on knowledge sharing among information technology majoring undergraduates. *Computers in Human Behaviour*, *56*, 65-71. https://doi.org/10.1016/j.chb.2015.11.031

Asheghi, N., Sharoff S., & Markert, K. (2016). Crowdsourcing for web genre annotation. *Language Resources and Evaluation*, *50*(3), 603-641. https://doi.org/10.1007/s10579-015-9331-6

Ayçiçegi-Dinn, A., & Caldwell-Harris, C. (2011). Individualism–collectivism among Americans, Turks and Turkish immigrants to the U.S. *International Journal of Intercultural Relations*, *35*, 9-16. https://doi.org/10.1016/j.ijintrel.2010.11.006

Baker, P. (2004). Querying keywords: Questions in difference, frequency, and sense in keyword analysis. *Journal of English Linguistics*, *32*(4), 346-359. https://doi.org/10.1177/0075424204269894

Barbaresi, A. (2021). Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In *Proceedings* of *the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*: *System Demonstrations* (pp. 122-131). https://doi.org/10.18653/v1/2021.acl-demo.15

Berber-Sardinha, T. (2018). Dimensions of variation across Internet registers. *International Journal of Corpus Linguistics*, *23*(2), 125-157. https://doi.org/10.1075/ijcl.15026.ber

Biber, D. (1988). *Variation across speech and writing.* Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic perspective.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519871

Biber, D. (2012). Register as a predictor of linguistic variation. *Corpus Linguistics and Linguistic Theory*, *8*(1), 9-37. https://doi.org/10.1515/cllt-2012-0002

Biber, D., & Egbert, J. (2016). Register variation on the searchable web: A multi-dimensional analysis. *Journal of English Linguistics*, *44*(2), 95-137. https://doi.org/10.1177/0075424216628955

Biber, D., & Egbert, J. (2018). *Register variation online*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316388228

Biber, D., & Conrad, S. (2019). *Register, genre, and style* (2nd ed). Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108686136

Biber, D., & Egbert, J. (2023). What is a register? Accounting for linguistic and situational variation within – and outside of – textual varieties. *Register Studies*, *5*(1), 1-22. https://doi.org/10.1075/rs.00004.bib

Comrie, B. (1997). Turkic languages and linguistic typology. *Turkic Languages*, *1*, 14-24.

Can, T., & Cangir, H. (2019). A corpus-assisted comparative analysis of self-mention markers in doctoral dissertations of literary studies written in Turkey and the UK. *Journal of English for Academic Purposes*, *42,* 1-14. https://doi.org/10.1016/j.jeap.2019.100796

Can, H., & Hatipoğlu, Ç. (2023). Cultural conceptualization of congratulatory happy events in British English and Turkish: A cross-cultural perspective. *Journal of Cognition and Culture*, *23*(3), 289-309. https://doi.org/10.1163/15685373-12340164

Candarli, D. (2022). Linguistic characteristics of online academic forum posts across subregisters, L1 backgrounds, and grades. *Lingua*, *267*, 103190. https://doi.org/10.1016/j.lingua.2021.103190

Egbert, J., Biber, D., & Davies, M. (2015). Developing a bottom-up, user-based method of web register classification. *Journal of the Association for Information Science and Technology*, *66*(9), 1817-1831. https://doi.org/10.1002/asi.23308

Egbert, J., & Biber, D. (2019). Incorporating text dispersion into keyword analyses. *Corpora*, *14*(1), 77-104. https://doi.org/10.3366/cor.2019.0162

Erten, S. (2019). *Corpus profiles of Turkish mental verbs with reference to Pattern Grammar and Corpus-Assisted Discourse Studies* (Master thesis). Retrieved from https://tez.yok.gov.tr/UlusalTezMerkezi/tezSorguSonucYeni.jsp

Göksel, A., & Kerslake C. (2005). *Turkish: A comprehensive grammar.* London, New York: Routledge.

Gries, S. (2021). A new approach to (key) keyword analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics*, *9*(2), 1-33. https://doi.org/10.32714/RICL.09.02.02

Kaya, E. K., & Yağlı, E. (2023). Recontextualization of the arguments of 'innocence' by a football club on Turkish newsprint media. *Text & Talk*. https://doi.org/10.1515/text-2022-0048

Koçak, A. (2013). *A comparative register analysis of the language of cooking used in Turkish recipes* (Master thesis). Retrieved from https://acikbilim.yok.gov.tr/handle/20.500.12812/465043

Laippala, V., Kyllönen, R., Egbert, J., Biber, D., & Pyysalo, S. (2019). Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics* (pp. 292-297). https://aclanthology.org/W19-6130

Lewis, G. (2000). *Turkish grammar.* Oxford: Oxford University Press.

Li, L., Li, A., Song, X., Li, X., Huang, K., & Ye, E. M. (2023). Characterizing response quantity on academic social Q&A sites: A multidiscipline comparison of linguistic characteristics of questions. *Library Hi Tech*, *41*(3), 921–938. https://doi.org/10.1108/LHT-05-2021-0161

Liimatta, A. (2019). Exploring register variation on Reddit. A multi-dimensional study of language use on social media website. *Register Studies*, *1*(2), 269-295. https://doi.org/10.1075/rs.18005.lii

Liimatta, A. (2022). Do registers have different functions for text length? A case study of Reddit. *Register Studies*, *4*(2), 263-287. https://doi.org/10.1075/rs.22007.lii

Olfert, H. (2023). The concept of register in heritage language retention. *Register Studies*, *5*(1), 52-81. https://doi.org/10.1075/rs.20017.olf

Özyıldırım, I. (2011). A comparative register perspective on Turkish legislative language. In T. Salmi-Tolonen, I. Tukiainen, R. Foley (Eds.), *Law and language in partnership and conflict*. (pp. 79-94). Turku, Finland: Lapland Law Review.

Pomikalek, J. (2011). *Removing boilerplate and duplicate content from web corpora* (Doctoral dissertation), Masaryk University, Faculty of Informatics, Czech Republic.

Repo, L., Skantsi, V., Rönnqvist, S., Hellström, S., Oinonen, M., Salmela, A., Biber, D., Egbert, J., Pyysalo, S., & Laippala, V. (2021). Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, 183–191. https://doi.org/10.18653/v1/2021.eacl-srw.24

Scott, M. (1997). PC analysis of key words – and key words. *System*, *25*(2), 233-245. https://doi.org/10.1016/S0346-251X(97)00011-0

Scott, M., & Tribble, C. (2006). *Textual patterns: Keywords and corpus analysis in language education.* Amsterdam: John Benjamins. https://doi.org/10.1075/scl.22

Sharoff, S. (2021). Genre annotation for the Web: Text-external and text-internal perspectives. *Register Studies*, *3*(1), 1-32. https://doi.org/10.1075/rs.19015.sha

Skantsi, V., & Laippala, V. (2023). Analyzing the unrestricted Web: The Finnish corpus of online registers. *Nordic Journal of Linguistics*, *1*(1), 1-31. https://doi.org/10.1017/S0332586523000021

Staples, S., Egbert, J., Biber, D., & Conrad, S. (2015). Register variation: A corpus approach. In D. Tannen, H. E. Hamilton & D. Schiffrin (Eds.), *The handbook of discourse analysis* (2nd ed) (pp. 505-525). Wiley Blackwell. https://doi.org/10.1002/9781118584194.ch24

Taavitsainen, I. (2001). Middle English recipes: Genre characteristics, text type features and underlying traditions of writing. *Journal of Historical Pragmatics*, *2*(1), 85–113. https://doi.org/10.1075/jhp.2.1.05taa

Tanova, C., & Nadiri, H. (2010). The role of cultural context in direct communication. *Baltic Journal of Management*, *5*(2), 185-196. http://dx.doi.org/10.1108/17465261011045115

Zhang, G., Jo, C., & Jhang, S. (2022). Keyword analysis of maritime legal text: Text-dispersion approach. *Corpus Linguistics Research*, *7*(2), 21-41.

Appendix: Abbreviations in Grammatical Annotations

acc	accusative case	neg	negative
caus	causative	opt	optative
cop	copula	pass	passive
dat	dative case	part	participle
inf	infinitival	pf	perfective
loc	locative case	1P	first person plural
nc	noun compound	2P	second person plural
nec	necessity	3S	third person singular

Address for correspondence
Selcen Erten-Johansson
University of Turku
Arcanuminkuja 1
20014 Turku
Finland
seerte@utu.fi
https://orcid.org/0000-0003-3625-4777

Co-author information
Valtteri Skantsi
University of Oulu
Pentti Kaiteran katu 1
90570 Oulu
Finland
valtteri.skantsi@oulu.fi
https://orchid.org/0000-0002-1230-9983

Sampo Pyysalo
University of Turku
Vesilinnantie 3
20014 Turku

Finland
sampo.pyysalo@utu.fi
https://orcid.org/0000-0002-6279-5000

Veronika Laippala
University of Turku
Arcanuminkuja 1
20014 Turku
Finland
mavela@utu.fi
https://orcid.org/0000-0002-7635-429X