Selcen Erten-Johansson and Veronika Laippala

Utilizing Text Dispersion Keyness on Turkish web registers: The case of Informational Description and Opinion

Abstract: Registers, such as news reports, frequently asked questions (FAQs), and opinion blogs, serve as indicators of linguistic variation that reflect the linguistic characteristics of digital environments. Understanding registers occurring on the web, known as web registers, is crucial in today's digital age. However, despite the abundance of linguistic data available on the internet, a notable gap in knowledge regarding the origins and the linguistic analyses of web registers remains. This especially applies to less-explored non-Indo-European languages. This study focuses on the Turkish web, drawing data from the Turkish Corpus of Online Registers (TurCORE) (Erten-Johansson et al. 2024). The linguistic characteristics of the Informational Description register comprising 481 texts and the Opinion register comprising 215 texts were examined using Text Dispersion Keyword Analysis (Egbert and Biber 2019). The findings highlight disparities not only in discoursal but also in linguistic features between the two registers. Moreover, notable variations in linguistic characteristics were found within each register, notwithstanding their shared objective or subjective discourse. By shedding light on the similarities of and differences between the characteristics of these registers, particularly within the context of a less studied non-Indo-European language, this research contributes to a more comprehensive understanding of language use in the digital environment.

Keywords: web registers, linguistic variation, Turkish Corpus of Online Registers (TurCORE), register analysis, Informational Description, opinion, Text Dispersion Keyword Analysis

Selcen Erten-Johansson, University of Turku, Finland, e-mail: selcen.s.erten@utu.fi
Veronika Laippala, University of Turku, Finland, e-mail: veronika.laippala@utu.fi

https://doi.org/10.1515/9783111434018-002

1 Introduction

The internet is a primary source for seeking and sharing information, with language playing a crucial role in expressing subjective opinions alongside objective descriptions. Understanding linguistic variations, particularly in web contexts, involves understanding how registers – for example, whether a text is informational or opinionated – affect language use and the interpretation of the text. However, analyzing web data poses challenges owing to the absence of essential metadata pertaining to, for example, text purpose or the factual versus opinionated nature of the content (Biber, Egbert, and Davies 2015). This makes the identification of registers on the web and the analysis of their linguistic characteristics difficult.

Prior studies on web registers have mostly focused on text collections with only few different registers, limiting our understanding of linguistic attributes across different registers. The creation of the Corpus of Online Registers of English (CORE) (Egbert, Biber, and Davies 2015) and similar corpora such as FinCORE for Finnish (Laippala et al. 2019; Skantsi and Laippala 2023), SweCORE for Swedish and FreCORE for French (Repo et al. 2021) have expanded our register knowledge, particularly in terms of natural language processing (NLP) and register identification. Despite this progress, CORE remains the only corpus among these corpora that has undergone extensive linguistic analysis, highlighting a research gap for languages like Turkish – a language characterized by its agglutinative nature and rich morphology. To address this gap, the Turkish Corpus of Online Registers (TurCORE) (Erten-Johansson et al. 2024) was developed following the principles established for CORE. In their study, Erten-Johansson et al. (2024) identified all the existing registers on the Turkish web and provided brief descriptions for each. In particular, they analyzed news reports, interactive discussions and recipes from linguistic and cultural standpoints. Since there is limited knowledge about the other registers in TurCORE, this study focuses on examining the linguistic characteristics of the Informational Description and Opinion registers, with a particular focus on descriptive and legal texts versus reviews and opinion blogs. We apply Text Dispersion Keyword Analysis proposed by Egbert and Biber (2019) to investigate how the Informational Description and Opinion registers are presented in Turkish.

Our research questions are as follows:

1. How do Informational Description versus Opinion serve the main registers' communicative purpose when analyzed using Text Dispersion Keyness?
2. What semantic and grammatical characteristics are prevalent in the sub-registers 'description of a thing/person' and 'legal terms and conditions' versus 'review' and 'opinion blog'?

The article proceeds with a review of the current literature on registers in Section 2 followed by the methodology in Section 3. The findings are detailed in Section 4.1 and 4.2. Finally, Section 5 summarizes the article's main points.

2 Previous work

Linguistic research has increasingly emphasized the importance of language use context, particularly when investigating variations in text with different communicative purposes (e.g., Biber 1986; Biber and Conrad 2001). Various terms including *style*, *genre,* and *register* have been used to denote these linguistic variations (Biber and Conrad 2009). As noted by Seoane and Biber (2021), although style and genre traditionally found their roots in literary studies, their scope has expanded to encompass non-literary variations. Aligning with the framework developed over the last three decades (e.g., Biber 1995; Biber et al. 1999; Biber and Conrad 2009), we define registers as a language variety associated with a particular situation of use, including communicative purposes. They are characterized by their typical grammatical features that reflect their linguistic attributes. These features inherently serve a functional purpose when considered from a register perspective, suggesting that linguistic features tend to recur in a register as they are well-suited for the intended purpose and situational context of the register (Biber and Conrad 2009).

Research has typically used pre-defined registers and limited collection of texts, mainly in Indo-European languages to study the linguistic characteristics of registers. For instance, Karapetjana and Lokastova (2015) analyzed the linguistic features of maritime e-mails written by chief engineers, revealing written and spoken registers that created a hybrid form of language use. More recently, Li et al. (2021) explored an academic Question & Answer platform, identifying that certain linguistic characteristics such as second-person pronouns in questions have a positive effect on response quantity, whereas linguistic characteristics such as first-person pronouns have the opposite effect.

Multi-dimensional Analysis has frequently been employed in register studies. For instance, Berber-Sardinha (2018) examined register variation across samples extracted from blogs, webpages, social media platforms and e-mails. Berber-Sardinha (2022) also employed a multi-dimensional perspective to investigate register variation on social media platforms such as Facebook, X (formerly Twitter), Instagram, Reddit, Telegram, and YouTube. Both studies revealed distinct linguistic characteristics among registers. In a similar vein, Liimatta (2019, 2022) extensively explored register variation through a multi-dimensional perspective on the social

media platform Reddit. The former study analyzed the *online subjective production, informational style,* and *instructional focus* dimensions, revealing patterns of register variation within Reddit. The latter study delved into the impact of registers on comment lengths, considering that the same text length may have different functions in different registers.

These studies have contributed to the field of online register studies, particularly in the context of English. However, it is important to note that they have primarily focused on pre-defined registers, failing to encompass the entire spectrum of registers that can be found online. The introduction of CORE (Egbert, Biber, and Davies 2015) marked a notable advancement in this regard, as it addressed the need for the comprehensive coverage of registers and linguistic variations on the internet without relying on pre-established categories. Furthermore, Biber and Egbert (2018) conducted extensive analyses using CORE data and examined a wide range of lexico-grammatical features.

Although CORE (Egbert, Biber, and Davies 2015) represents a major advancement in exploring register variation, and Biber and Egbert's (2018) work has contributed to the linguistic analyses of registers, the pivotal roles played by non-Indo-European languages, which have received less attention in register studies, is important. Turkish, an agglutinative language, has a wide range of suffixes with distinctive features. Most multi-syllabic words in Turkish are complex, resulting in lengthy words that could be translated to entire sentences in English (Lewis 2001). The primary mechanism for word formation in Turkish is suffixation, where a new word is formed by attaching a suffix to the end of a root. A considerable number of suffixes can be affixed to a root. Derivational suffixes creating new words typically precede inflectional suffixes, which provide grammatical information such as case, person, and tense (Göksel and Kerslake 2005: 43).

Previous studies on register variation in Turkish have primarily focused on specific registers. For instance, Özyıldırım (2011) investigated legislative language and compared it with various registers, concluding that it is the least narrative in comparison with research articles, men's/women's magazines, newspaper feature articles and television commercials. Koçak (2013) examined the lexico-grammatical and discoursal features of Turkish cooking recipes from two cookery books published in 1974 and 2011 to investigate whether any linguistic and discoursal differences existed between them. The findings revealed that although the recipes in the two books show similar discoursal characteristics, such as explicit reference discourse, they differ in terms of their linguistic features, including the use of the second-person pronoun. News articles and editorials have also been studied. Çarkoğlu, Baruh and Yıldırım (2014) explored news articles and editorial columns collected from various newspapers to investigate press-party parallelism in the 2011 national elections and found out linguistic variations across the newspapers. Using a genre-

based approach, Aksan and Aksan (2015) examined the differences between informative and imaginative texts with Turkish multi-word units. They identified distinct lexical patterns and linguistic features prevalent in fiction and non-fiction texts. For instance, person references were frequently found in fictional imaginative texts, whereas they appeared much less often in non-fiction informative ones.

Similar to many studies conducted for English, these investigations have focused on pre-defined registers and restricted collection of texts without covering linguistic variations in their entirety. This underscores the need for developing a corpus that encompasses the linguistic variations present in Turkish web content and analyzing their linguistic characteristics. The creation of the Turkish Corpus of Online Registers (TurCORE) and in-depth analyses of various registers within it such as news reports, interactive discussions, and recipe texts (Erten-Johansson et al. 2024) address this need, thereby highlighting the necessity to explore the Informational Description and Opinion registers within the Turkish web.

3 Methodology

3.1 TurCORE

The compilation, cleaning, and annotation processes of TurCORE are presented in Erten-Johansson et al. (2024). TurCORE comprises texts randomly sampled from the CommonCrawl dataset that contains web documents. The corpus contains 2,780 unique web texts, comprising 1,026,253 tokens. The average token length of the texts varies, from a minimum of 203 to a maximum of 1,431 (Erten-Johansson et al. 2024). As noted by Erten-Johansson et al. (2024), the cleaning process involved fetching the documents in HTML format from the URLs, followed by boilerplate removal with Trafilatura (Barbaresi 2021), and deduplication using Onion (Pomikalek 2011). Next, manual annotation was performed on Prodigy (https://prodi.gy, last accessed 14 February 2025). The process involved collaboration between a supervisor and a trained annotator with a corpus linguistics background.

In TurCORE, a taxonomy based on that of CORE (Egbert, Biber, and Davies 2015) and FinCORE (Laippala et al. 2019; Skantsi and Laippala 2023) was utilized to cover the full range of online registers. However, this taxonomy was simplified, by excluding registers that were found to be low-frequent and vaguely defined in previous studies (Biber, Egbert, and Davies 2015; Skantsi and Laippala 2023). The adapted taxonomy was developed in a hierarchical manner. This enabled the identification of the basic situational characteristics for each web document by classifying them into main register categories, which then led to the development of specific *sub-reg-*

isters. The taxonomy used for the Turkish data recognizes 9 main registers along with a total number of 24 sub-registers (Erten-Johansson et al. 2024). The main registers are Informational Persuasion, Narrative, Informational Description, Machine-translated, Opinion, How-to/Instruction, Interactive Discussion, Spoken, and Lyrical. The documents assigned to more than one category were annotated as hybrids.

This study centers on the examination of the main registers Informational Description and Opinion, along with their sub-registers. The Informational Description register delivers factual details, whereas the Opinion register provides subjective and interpretive content. Understanding their linguistic expressions is crucial for distinguishing between them on the internet.

The Informational Description register comprises 481 texts, totaling 196,624 tokens, whereas the Opinion register comprises 215 texts, with 107,828 tokens in total. Table 1 illustrates the total number of texts, the token counts, and the distributions of each sub-register within its corresponding main register. Texts that do not exhibit the specific characteristics of a sub-register are labelled as 'other.' Percentage was calculated in terms of the number of texts.

Table 1: Registers and sub-registers with counts of texts and tokens.

Sub-register of Informational Description (IN)	**No. of texts**	**No. of tokens**	**Percentage (%) in IN**
Description of a thing/person	124	46,273	25.77
Legal terms and conditions	105	52,059	21.82
Encyclopedia article	18	7,979	3.74
FAQs	6	2,447	1.24
Research article	4	1,693	0.83
Other	224	86,173	46.56
Total	**481**	**196,624**	**99.96**
Sub-register of Opinion (OP)	**No. of texts**	**No. of tokens**	**Percentage (%) in OP**
Review	66	20,933	30.69
Opinion blog	58	32,351	26.97
Advice	46	17,122	21.39
Religious blog/sermon	29	26,695	13.48
Other	16	10,727	7.44
Total	**215**	**107,828**	**99.97**

The 'other' sub-register within Informational Description constitutes about half of the Informational Description register. Although this finding is intriguing, it falls

outside the scope of this study and would necessitate separate research. Nevertheless, we can state that a variety of informational and descriptive texts such as descriptive reports, course materials, test papers, and meeting minutes are classified under this sub-register.

3.2 Text dispersion keyword analysis

Scott (1997: 236) defines keywords as words that occur with notable frequency in a target corpus compared to a reference corpus. The notion of keyness employs a comparison of a target corpus and a reference corpus to assess the "aboutness" of a text or corpus (Baker 2004: 347). Traditionally, keyness has been determined using log-likelihood statistics (Scott and Tribble 2006), which approached the concept of keyness through frequency. This calculation is called the standard frequency keyword analysis. Although the analysis aims to identify statistically significant words within a collection of texts (Scott 1997; Scott and Tribble 2006), it often overlooks how these words are spread across different texts. Standard frequency keyword analysis operates under the assumption of corpus homogeneity, where words are evenly distributed across the corpus. However, corpus frequency keywords are often abundant in a corpus but not widely dispersed across its texts (Egbert and Biber 2019), marking them inadequate in representing the domain of the corpus in question.

Content-distinctiveness and content-generalizability serve as criteria for assessing the effectiveness of keyword analysis. Content-distinctiveness refers to the strength of the relationship between a keyword and the discourse domain of the target corpus, whereas content-generalizability pertains to the degree to which a keyword represents the discourse across the entire target corpus (Egbert and Biber 2019: 78–79). Content-distinctive keywords should better typify the target discourse domain relative to other domains, whereas content-generalizable keywords should be representative of the entire target corpus. In pursuit of greater content-distinctiveness and content-generalizability, as "keywords should be used by many different writers/speakers.", Egbert and Biber (2019) introduced Text Dispersion Keyword (TDK) Analysis. TDK uses text rather than the corpus as the unit of observation. The frequency of word repetition is not crucial, as words that are repeated often hold significance within a specific text but not necessarily across the entire corpus (Egbert and Biber 2019: 83). In TDK analysis, word frequency is disregarded; and instead, keyword lists are generated based on word dispersion across texts. It has demonstrated its suitability for register studies with large corpora and surpasses traditional frequency-based methods in terms of effectiveness (Gries 2021).

To create the reference corpus, we incorporated all texts from various registers, excluding those from the target register. For instance, when conducting TDK analysis on legal texts, our target corpus comprised all legal texts, and our reference corpus comprised all texts of TurCORE except legal texts. Following the creation of the reference and target corpora, we utilized Python scripts to extract the keywords associated with each sub-register. After identifying the keywords, we categorized those from the sub-registers with highest number of texts. They are 'description of a thing/person' and 'legal terms and conditions' under the main register of Informational Description and 'review' and 'opinion blog' under the main register of Opinion. The keywords were grouped into semantic and grammatical categories based on their semantic and functional similarities. Grammatical categories were established based on the observation that certain linguistic features tend to co-occur in texts owing to their interconnected functions (Biber and Egbert 2018: 46). Semantic categorization involved allocating each identified word into relevant semantic categories. Consistent with Egbert and Biber's (2019) methodology, we examined the top 100 keywords for each register.

4 Findings

We define the sub-registers of Informational Description in Section 4.1. and those of Opinion in Section 4.2. using examples from their top keywords. The top 20 keywords provide sufficient information to illustrate the aboutness of each register. The top 20 Turkish keywords for each register, along with their English translations, are provided in Appendices A and B.

Section 4.1.1. offers a detailed examination of 'description of a thing/person' and Section 4.1.2. delves into 'legal terms and conditions.' Similarly, Section 4.2.1 provides detailed analyses of 'review' and Section 4.2.2. focuses on 'opinion blog.' The analyses entail grouping the top 100 text dispersion keywords of each register into semantic and grammatical categories. Appendix C shows all abbreviations used in grammatical annotations.

4.1 Informational Description

The primary aim of the Informational Description register is to describe or explain information. Authors are typically not specified in the texts of this register. The output can vary depending on the text – from carefully written texts like 'legal terms' to unedited ones such as 'description of a thing' (Skantsi and Laippala 2023:

14). A notable characteristics of Informational Description texts is their factual or factually expressed essence.

Below, we cover all sub-registers of Informational Description. As is the case with all keywords, a single Turkish word can often correspond to multiple words in English (Biber 1995). For this reason, when translated into English, the original single-word keywords might be expressed as several words.

The 'description of a thing or person' sub-register involves the depiction of a thing or a person. In TurCORE, this register was observed to predominantly describe a thing rather than a person, with a focus on medical topics indicated by keywords such as *tedavisi* 'treatment of', *enfeksiyonlar* 'infections', *belirtileri* 'symptoms of', *hastalarda* 'in the patients', *hastalık* 'disease', *bağırsak* 'intestine' and *ilaçlar* 'medications.' One of the most frequent keywords *vardır* 'there is/are' highlights a distinctive characteristic. The suffix *-DIr* is a generalizing modality marker in Turkish grammar (Göksel and Kerslake 2005: 80), commonly used to emphasize a generalization or statement of principle in the content being described.

The 'legal terms and conditions' sub-register concerns any topic related to legality where the author is not mentioned (Skantsi and Laippala 2023). One of the notable features of this register, which makes the text easily identifiable, is its official context-specific and formal words (Özyıldırım 2011), such as *işbu* 'hereby', *kanunu* 'act of', *maddesinde* 'in the article of' and *yetkili* 'authorized' that we identified within the top 20 keywords. Specific to online data, privacy policies and cookie descriptions are often present in this register (Biber and Egbert 2018), expressed via the keywords *bilgilerin* 'of data', *kişisel* 'personal' and *korunması* 'protection of'. In the Turkish data, a notable group of words also pertain to the delivery of purchases and the return policy of products, as the keywords *iade* 'return', *kargo* 'cargo' and *kargoya* 'to cargo' exemplify.

A prime example of 'encyclopedia articles' found on the internet is Wikipedia, whose format is the same across languages, making the articles easily identifiable (Biber and Egbert 2018). A considerable number of encyclopedia articles in Turkish feature biographical descriptions, evident from keywords such as *doğdu* 'was born' and *doğmuştur* 'was born'. In addition, temporal markers such as years *1972, 1989* and *2004*, phrases like *yılında* 'in the year of' and geographic location names such as *Ankara* and *Berlin* reflect time- and place-related words typically used in biographical descriptions.

TurCORE has a scarcity of texts categorized as 'FAQs' and 'research articles', necessitating a cautious approach to their interpretation. Nevertheless, FAQs typically comprise a list of commonly asked questions regarding a particular subject, which is typically accompanied by answers, structured in a question-and-answer format (Asheghi, Sharoff, and Markert 2016; Biber and Egbert 2018). FAQs predominantly address products or services offered on a website, with responses usually

provided by company personnel (Skantsi and Laippala 2023), as observed via the top 20 keywords *hazırlıyoruz* 'we are preparing' that suggests product readiness and *sitelerimizi* 'our sites' that indicates website affiliation.

'Research articles' are a form of academic writing detailing a research study, including the motivation for the study, the methodologies employed, and the research findings. They are typically written by an individual or a group of authors affiliated with an academic institution, and are intended for an audience of specialists (Biber and Egbert 2018). On the Turkish web, although infrequent, research articles seem to mainly pertain to medical research, evident from keywords such as *sendromu* 'syndrome of', *coronavirüsler* 'coronaviruses', *algınlığından* 'from the delusion of' and *akciğerlere* 'to the lungs.' Technical terminology such as *nanopartiküller* 'nanoparticles', *iğnesiz* 'mutic' and *nebulizatör* 'nebulizer' was also observed.

4.1.1 Description of a thing or person

In this semantic grouping and the subsequent ones for other registers, variations in keywords that do not affect the meaning or the sentence structure, have been disregarded, and only the nominative-cased noun is used for semantic categorization. For instance, keywords like *tedavisi* 'treatment of' and *tedavisinde* 'in the treatment of' are simplified to *tedavi* 'treatment' and displayed as one keyword instead of three.

Based on the top 100 text dispersion keywords, description of a thing/person texts are categorized into semantic groups of medicine and physiology, description, higher education, time, and other, as illustrated in Table 2.

In texts concerning the description of a thing/person, the presence of keywords related to description and time is expected owing to the communicative purpose of such texts. In Turkish texts, keywords were observed within both the description and time-related semantic categories. However, the inclusion of keywords categorized under the groups of medicine and physiology and higher education seems to reflect the annotation process, which distinguishes description of a thing/person from encyclopedia articles (Erten-Johansson et al. 2024). Although the current register predominantly features description of things rather than persons, these descriptions often pertain to diseases and treatments, as well as to universities and faculties.

Most keywords associated with the description of a thing/person texts in Turkish are nouns, followed by adjectives, verbs, numerals, and adverbs. However, from a more Turkish-specific perspective, they fall under the main categories of copular marker formed with the suffix *-DIr* and aorist with *-(A/I)r*, and adjectives, as seen in Table 3.

Modality is concerned with whether a situation is presented as a directly known fact or in some other way. The modality marking system in Turkish enables differentiation between statements that reflect the speaker's direct experience, knowledge, or observation, and those that make assertions of more general, theoretical ideas, or convey assumptions or hypotheses. Among the various modalized expressions, the markers indicating generalization, general rule, or statement of principle are the aorist forms *-(A/I)r* in verbal sentences and the generalizing modality marker *-DIr* in nominal sentences (Göksel and Kerslake 2005: 294–295). In the description of a thing/person sub-register, both forms were observed within the top 100 keywords. Examples such as *vardır* 'there is/are', *nedir* 'what is' and *denir* 'is called' in Table 3 illustrate this, aligning with the descriptive nature of this register.

Table 2: Semantic categories of description of a thing/person with examples.

medicine and physiology	description	higher education	time	other
treatment	there is/are	faculty	2005	announcement
infection	has gotten	university	2007	south
symptom	personifies	undergraduate	2009	film
patient	is directing	class	1997	
disease	what is	title	1991	
series	is seen	graduate	1973	
medications	can be seen	edu	when it is	
history	that is seen	sciences	generally	
apo	indicates	career		
psychotherapy	is available	program		
addiction	they pass	literature		
cell	is called	school		
pain	was born			
receptor	must be done			
histamin	widespread			
auditory	married			
chronic	apparent			
genetic	general			
doctor	expression			
immunity	character			
therapy	family			
hereditary	civil servant			
contagious	child			
disorder	in women			
incidence	role			
diagnosis	in Anatolia			
examination				

medicine and physiology	description	higher education	time	other
physique				
intestine				
tooth				
bone				
kidney				
body				

In Turkish, adjectives have the flexibility to function as nouns and adverbs. When employed as a noun, their identification is straightforward owing to the case markers attached to the noun. However, pinpointing their function as adverbs may not be as straightforward. Nonetheless, by checking the concordance lines, we observed a considerable number of adjectives that can also function as adverbs in description of a thing or person texts, some of which are illustrated in Table 3. Göksel and Kerslake (2005) classify adjectives into two main groups based on whether they contain a productive derivational suffix. Although both types were present within the top 100 keywords of this sub-register, we included those with a derivational suffix in Table 3 to demonstrate the influence of grammar on adjective formation. For instance, the suffix *-GIn* forms *yaygın* 'widespread' from the verb *yay-* 'spread', and *belirgin* 'apparent' from the verb *belir-* 'appear'. Similarly, the suffix *-sAL* forms *kalıtsal* 'hereditary' from the noun *kalıt* 'gene'. Although the function of these keywords may not immediately be apparent without examining concordance lines, the inclusion of adjectives that can also serve as nouns and adverbs in this register enriches descriptions of things or persons by offering detailed characteristics.

Table 3: Grammatical categories of description of a thing/person with samples.

Grammatical category	Keyword in English	Keyword in Turkish	Grammatical annotation
copular marker -DIr and aorist -(A/I)r	there is/are what is is called	var-dır ne-dir de-n-ir	existent + -DIr what+ -DIr say +pass+ -(A/I)r
adjective	widespread apparent hereditary	yay-gın belir-gin kalıt-sal	spread+ -GIn appear+ -GIn gene+ -sAl

4.1.2 Legal terms and conditions

The top 100 keywords of Turkish legal texts were semantically classified into the categories of legality, coordination/relation, online shopping, data protection, number and other, as Table 4 illustrates.

Legal texts are recognizable by their extensive use of formal language, comprising technical legal terms that typically require specialized knowledge for comprehension (Özyıldırım 2011: 85). The keywords grouped under the semantic categories of legality and coordination/relation support the presence of this formal and official terminology. However, it is noteworthy that certain keywords related to the internet, such as data protection and online shopping, which may not require specialized expertise, are also prominent in these texts. Upon manually examining concordance lines, we observed that certain keywords categorized as number, such as *otuz* 'thirty', are commonly used in online shopping texts, referring to the thirty-day period within which the customer can return a product. The presence of legal documents on the internet, intended not only for specialists but also for potential consumers or website users, appears to challenge the conventional perception of legislative language.

Table 4: Semantic categories of legal terms and conditions with examples.

legality	coordination/ relation	online shopping	data protection	number	other
legitimate	thereunder	return	personal	numbered	responsible
accordment	that is designated	cargo	data	no	written
hereby	within	order	protection	6698	registered
act	anticipated	delivery	privacy	third	cancellation
article	pursuant to	address	processing	thirty	responsibility
court	determinated	mail	reserved		cannot be used
commitment	on the basis of	firm	unpermitted		that you are
abolitionary	that might arise	customer	mischief		
legislation	that originates	product	copyright		
declaration	provided that	membership			
obligation	in case of	site			
outher	inflicting	www			
execution	expounded	com			
feasance	regarding				
compatibility	appurtenant				
prerogative	concerned				
parties	with intendments				
demand	or				
right					
contract					
authorized					

We observed patterns of relative clauses, passive voice structures, and plural nouns in legal texts, as seen in Table 5.

Table 5: Grammatical categories of legal terms and conditions with samples.

Grammatical category	Keyword in English	Keyword in Turkish	Grammatical annotation
relative clause	that might arise	doğ-abil-ecek	arise+psb+part
relative clause	that originates	kaynaklan-an	originate+part
passive voice	that is designated	belirt-il-en	designate+pass+part
passive voice	on the basis of	dayan-ıl-arak	base+pass+cv
passive voice	cannot be used	kullan-ıl-a-maz	use+pass+psb+neg.aor
plural noun	obligations	yükümlülük-ler	obligation+pl
plural noun	parties	taraf-lar	party+pl
plural noun	data	veri-ler	datum+pl

The relative clause structure in Turkish is a complex adjectival construction where a modifying clause precedes the head (Kornfilt 1997). The most common type of relative clause is marked by the suffix *-(y)An, -DIk,* or *(y)-AcAk*, corresponding to various relative pronouns including who, which, that, whom, whose and where in English (Göksel and Kerslake 2005). In legal documents, we identified these relative clause structures, and found the *-(y)An* suffix as the most common. The prevalence of various relative clauses in legal documents indicates the aim for precision and unambiguousness within the legal context. The frequent use of relative clauses in legal texts underscores their role in conveying messages that are clearly defined and aimed to be correctly understood in legal discourse.

Another commonly observed structure among the top 100 keywords in legal documents is the passive voice. Some instances of passive voice were embedded within structures featuring non-finite verbs forms such as *belirtilen* 'that is designated' and *dayanılarak* 'on the basis of', whereas others occur finite verb forms such as *kullanılamaz* 'cannot be used', as seen in Table 5. The frequent presence of passive voice in various verb forms seems to indicate an intent to maintain a formal tone. Passive voice serves to centers the attention on actions and consequences without attributing to a specific person or entity.

The plural form in Turkish is created by adding the suffix *-lAr* to the noun. In legal documents, we observed the frequent use of plural nouns such as *yükümlülükler* 'obligations', *taraflar* 'parties' and *veriler* 'data', indicating that legal terms apply to multiple entities, individuals, and situations. This reflects the generaliza-

ble nature of legal discourse, where rules are designed to be broadly applicable. The frequent use of plural-marked nouns for generalizations is both anticipated and intriguing, given the objectives of precision and unambiguousness inherent in legal documents.

4.2 Opinion

The Opinion register expresses subjective viewpoints based on the personal opinions of an individual author or a group of authors (Biber and Egbert 2018). In some cases, the author is mentioned by a pseudonym or by their actual name (Skantsi and Laippala 2023). Below, we explore each sub-register of Opinion with their top 20 keywords.

'Review' entails the evaluations of a product or service written by an individual on a personal, institutional, or commercial website. Although the author may claim to have expertise regarding the product or service under review, they often may have only used the product or service (Biber and Egbert 2018). Reviews hold considerable influence on consumers' purchasing decisions online (Wang et al. 2023). This might explain why the review register emerged as the most prevalent sub-register of Opinion on the Turkish web. Turkish reviews were found to commonly evaluate various forms of media such as films, documentaries, books, and games, with specific mentions such as *Spinoza* and *Minecraft*.

'Opinion blog', a sub-register specific to the internet, serves as a platform for individuals to publicly share personal viewpoints, often involving evaluations and stances. They are commonly written by non-professional authors. According to Biber and Egbert (2018: 107), opinion blogs are one of the least well-defined registers owing to their diverse nature, encompassing a broad spectrum of texts that may not clearly exhibit opinionated elements. Although this register typically includes politics-related topics and may even be written by political figures (Skantsi and Laippala 2023), politics did not emerge as a prominent theme on the Turkish web. This could be attributed to the fact that during the annotation process, texts related to politics were often categorized under Informational Persuasion, indicating an intent to persuade rather than simply express opinion (Erten-Johansson et al. 2024).

'Advice' involves offering recommendations based on personal opinion with the aim of prompting action to solve a particular problem (Biber and Egbert 2018). Often, the authors remain anonymous but claim expertise regarding the problem and its solution. The focus lies on the thoughts and emotions of the reader (Skantsi and Laippala 2023), who could be anyone seeking guidance on addressing a particular problem. In Turkish advice texts, the guidance offered to readers is reflected

through keywords related to second-person markers, such as *kendinizi* 'yourselves', *size* 'to you', *yaşamınızda* 'in your life', *unutmayın* 'don't you forget' and *olabilirsiniz* 'you might be.' In addition, keywords such as *nasıl* 'how' and *dikkat* 'caution' indicate the provision of guidance intended to lead to actions in this register.

'Religious blog/sermon' comprises texts of denominational religious nature, excluding those merely describing a religion (Biber and Egbert 2018). Typically authored by individuals, these texts are often hosted on institutional websites, and some of the website visitors are regular followers. Despite being based on beliefs and opinions, the discourse within religious blog/sermon often adopts an informational description framework, which complicates its communicative purposes (Biber and Egbert 2018). Furthermore, the complexity of this register arises from the occasional inclusion of narrative elements such as stories. Nevertheless, a consistent feature of this register is the utilization of context-specific religious terminology, as observed in Turkish with keywords such as *Allah, Muhammad, peygamber* 'prophet', *namaz* 'prayers', *Hz* 'His holiness' and *ahirette* 'in the afterlife'.

4.2.1 Review

The top 100 keywords found in Turkish review texts were semantically categorized into groups of evaluation, (background) description, material, feature, name and other, as Table 6 illustrates.

A large number of the keywords identified in reviews pertain to the material under review, such as a film, game, or book, providing detailed descriptions and mentioning associated features. In line with this observation, specific names such as the title of a website informing users about new technology *Teknolojioku* or the seventh season of a series *s7* were noted. These items are assessed using various evaluation-related keywords, such as *açıkçası* 'frankly', *düşünülmüş* 'thought-out' and *yargılıydım* 'I was prejudiced.' It is notable that many of the keywords of this register demonstrate close semantic connections with each other, enhancing and complementing their meanings.

In Turkish reviews, we did not observe frequently repeating grammatical patterns, aligning with Biber and Egbert's (2018: 123) findings for English. However, the categories in Table 7 display certain Turkish-specific grammatical features in the context of reviews.

Table 6: Semantic categories of reviews with examples.

evaluation	(background) description	material	feature	name	other
impression	it addresses to	film	layers	Spinoza	honor
frankly	it points at	documentary	model	Minecraft	I will not mention
black	it comes about	book	differences	Sigma	questioning
rationalist	when I saw	game	features	Nurdan	in seasons
shattered	it leaves	cover	vibe	Teknolojioku	living creatures
as much as	in the town	skirt	android	Franz	thought
thought-out	of the town	image	pixel	auto	audience
can be provided	it catches	life	dialogue	news	2010
what we understand	that I used	series	weight	a0	punch
I was biased	from the building	religion	version	s7	
stance	that I know		visual	v1	
which does not resemble	that s/he takes on				
unuseful	identity				
driven	to the front				
fictional	role				
views	I had used				
publicity					
masterpiece					
prominent					
my interest					

Table 7: Grammatical categories of reviews with samples.

Grammatical category	Keyword in English	Keyword in Turkish	Grammatical annotation
negative adjectivals	unuseful irrelevant which does not resemble	kullanış-sız alaka-sız benze-me-yen	use+ *-sIz* relevance+ *-sIz* resemble+neg+ *-(y)An*
adjectives with perfect participle	shattered thought-out	parçala-n-mış düşün-ül-müş	shatter+pass+part think+pass+part
accusative case marked noun	the film the book the documentary	film-i kitab-ı belgesel-i	film+acc book+acc documentary+acc

Adjectives or words functioning as adjectives (adjectivals) within the top 100 keywords were not notably prevalent in Turkish reviews. However, most of the adjectives found were constructed either with the suffix *-sIz* or with the negative marker

-mA. The suffix *-sIz* expresses absence or lack, and is translatable as 'less', 'without' and 'lacking' (Johanson 2021: 489). Illustrated in Table 7, adjectives such as ***kullanışsız*** 'useless' and *alakasız* 'irrelevant' are examples of the pattern where the suffix *-sIz* is added to the nouns ***kullanış*** 'usage', and *alaka* 'relevance.' The keyword ***benzemeyen*** 'which does not resemble' functions as an adjectival, formed with negation marker *-mA* followed by the relative clause marker *-(y)An*. It is noteworthy that despite the limited number of adjectives, the subjectivity in reviews tends to be negative. However, we are cautious not to draw generalizing conclusions from individual words.

Another group of adjectives was found to be formed with the perfect participle. In Turkish, the suffix *-mIş* serves functions, including inference and perception (Johanson 2021: 653). The keywords ***parçalanmış*** 'shattered' and ***düşünülmüş*** 'thought-out' exemplify the role of this suffix as an indicator of perfect participle. The inference and perception functions of the suffix are in line with the nature of reviews, which involve evaluations of a product or service based on the reviewer's experiences.

In Turkish reviews, we observed a greater prevalence of nouns than of adjectives. These nouns varied in their grammatical cases, with one group showing a consistent pattern: the accusative case. The accusative case is used to mark the definite object of a verb, indicating an object specified by its identity as a name or title, or one that has been previously mentioned, such as the use of the definite article in English (Lewis 2001: 35). In reviews, both functions of the accusative case on the object are evident, especially when reviewers assess products or services, as exemplified by the keywords ***filmi*** 'the film', ***kitabı*** 'the book' and ***belgeseli*** 'the documentary'.

4.2.2 Opinion blog

The top 100 keywords extracted from Turkish opinion blog texts were categorized into semantic groups of stance, opinion and evaluation, identity, topics and concepts, action and activity, and other. Table 8 displays these categories.

We observed that the keywords related to opinion blog displayed a more diverse distribution than the keywords from other registers. Accordingly, we created the semantic groups by combining related categories, such as opinion and evaluation, or topics and concepts. Words categorized under stance were found to express negativity, certainty and uncertainty, and probability and improbability, all indicating the markings of stance. In addition, we identified a large number of words belonging to the category other in this register. The lack of clear-cut semantic categories and the presence of numerous keywords in the other category in Turkish

opinion blogs reinforces the observation by Biber and Egbert (2018) regarding the diverse nature of opinion blogs, which cover a wide range of textual content.

Table 8: Semantic categories of opinion blogs with examples.

stance	opinion and evaluation	identity	topics and concepts	action and activity	other
never	I think	who	thing, stuff	to start	mi (a question particle)
but	you are right	s/he, it	pain	to hold	what
not	by thinking	I	life	to do, make	this
maybe	in my opinion	ourselves	happiness	to say	mu (a question particle)
must be	there is/are	individual	years	dance	that it is
in fact	when you look at	my	society	to study, work	how
if only	I say	others	wisdom	to write	because
no, nothing	that I know	man	decisions	writing	homeland
already, anyway	regularly	person	words	to show	then
should/must be	let's not be bothered			sharing	I want
so that				to give	to be
if it happened				to cover	ya (a clitic)
while				to lose	mı (a question particle)
a little					there
like that					that
accordingly					full of
işte (a discourse marker)					team
even					lessons
of course					modern-day
cannot be					the yes sayers
in vain					
a bit					
again					

We observed a relatively frequent pattern of pronouns in opinion blogs, some of which are shown in Table 9. These include *ben* 'I', *benim* 'my' and *kendimize* 'to ourselves'. Most of the pronouns are in first-person singular form, with some also appearing in first-person plural form. This aligns with the subjective nature of opinion blogs, which reflect the personal viewpoints of the bloggers. In Turkish, subject pronouns can be omitted, as the verbs are marked with person suffixes. Despite this, the subject pronoun *ben* 'I' was identified as one of the most frequent keywords within the top 100. This could be attributable to blogger's intentions to

emphasize their personal opinion or to introduce a new topic of discussion in the opening sentence of a paragraph. (Göksel and Kerslake 2005: 241–242).

Modalized utterances are of various kinds, and encompass different functions such as assumptions or hypotheses, possibility or necessity statements, or expressions of desire or willingness for an event or state to occur (Göksel and Kerslake 2005: 294–295). For instance, as demonstrated in Table 9, the keyword *başlarız* 'we would start' indicates an assumption or hypothesis, *olmalı* 'it should/must be' signifies necessity, and *takılmayalım* 'let's not be bothered' exemplifies willingness. The diverse array of modals found in opinion blog appears to reflect the subjective nature of this register, enabling the author to convey their viewpoints by employing a variety of personal perspectives.

Table 9: Grammatical categories of opinion blogs with samples.

Grammatical category	Keyword in English	Keyword in Turkish	Grammatical annotation
pronouns	I I have, my to ourselves	ben ben-im kendimiz-e	I I+gen ourselves+dat
modality	we would start let's not be bothered it should/must be	başla-r-ız takıl-ma-yalım ol-malı	start+aor+1P be bothered+neg+opt+1P be+nec3S
imperfective aspect	I think I want I am saying	düşün-üyor-um ist-iyor-um d-iyor-um	think+impf+1S want+impf+1S say+impf+1S

Considering that aspect expresses a viewpoint from which a situation is presented, the imperfective aspect typically refers to actions or events that are ongoing, habitual, or continuous without any endpoint. In Turkish opinion blogs, we identified several verbs expressed in imperfective aspect. For instance, *düşünüyorum* 'I think', *istiyorum* 'I want' and *diyorum* 'I am saying' are keywords expressed in the imperfective aspect, which is consistent with the content of opinion blogs. This usage of imperfective aspect appears to give a sense of continuous engagement with the topic under discussion. Further, it contributes to a more conversational tone, possibly enhancing the blog's relatability to the reader.

5 Conclusion

In this article, we examined the Informational Description and Opinion registers of TurCORE created by Erten-Johansson et al. (2024). The analysis revealed that Informational Description serves as a register for objectively presenting information, whereas Opinion represents a domain where individuals express subjective viewpoints, in line with Biber and Egbert (2018) and Skantsi and Laippala (2023).

Within Informational Description in Turkish, we found that the 'description of a thing/person' and 'legal terms and conditions' sub-registers exhibit similarities in their descriptive content but differ in their linguistic characteristics. Descriptions typically rely on nouns and adjectives supported by generalizing modality markers. In contrast, texts of 'legal terms and conditions' are distinguishable not only by their vocabulary but also by their frequent use of specific grammatical structures such as the relative clause and passive voice. These linguistic features serve to enhance precision and formality in legal discourse, characteristics not commonly found in descriptions of things or persons. Moreover, legal texts can encompass online-specific contexts, a feature not always observed in descriptions.

Our findings reveal that the sub-registers within Opinion, such as 'review' and 'opinion blog', share similarities owing to their subjective content. However, notable differences exist between the two. Reviews predominantly center around products or services and are authored by individuals who used the products or service. This sub-register is characterized by an abundance of evaluation words reflecting subjective assessment. In contrast, opinion blogs cover a wide range of topics, making them less well-defined. They predominantly express personal viewpoints with emphatic stances, adopting a conversational tone through the frequent use of modality and imperfective aspect markers.

Informational Description and Opinion represent distinct registers, each characterized by unique linguistic features. In today's digital age, where information is sought and shared primarily online, distinguishing between informative content and opinion-based material to enhance media literacy is essential. We anticipate that the work by Erten-Johansson et al. (2024) along with the research presented in this study, will contribute to future studies on the linguistic characteristics of web-based communication in less studied non-Indo-European languages.

6 Funding

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

7 Acknowledgements

We thank our anonymous reviewers for their time in reading the manuscript and providing valuable feedback.

References

Aksan, Mustafa & Yeşim Aksan. 2015. Multi-word expressions in genre specification. *Dil ve Edebiyat Dergisi* 12 (1), 1–42.

Asheghi, Nouhsin Rezapour, Serge Sharoff & Katja Markert. 2016. Crowdsourcing for web genre annotation. *Language Resources and Evaluation* 50 (3), 603–641.

Baker, Paul. 2004. Querying keywords: Questions in difference, frequency, and sense in keyword analysis. *Journal of English Linguistics* 32 (4), 346–359.

Barbaresi, Adrien. 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. 122–131.

Berber-Sardinha, Tony. 2018. Dimensions of variation across Internet registers. *International Journal of Corpus Linguistics* 23 (2), 125–157.

Berber Sardinha, Tony. 2022. A text typology of social media. *Register Studies* 4 (2), 138–170.

Biber, Douglas. 1986. Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language* 62 (2), 384–414.

Biber, Douglas. 1995. *Dimensions of register variation: A cross-linguistic perspective.* Cambridge: Cambridge University Press.

Biber, Douglas & Susan Conrad. 2001. *Variation in English: Multi-Dimensional studies*. London: Routledge.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, Douglas & Jesse Egbert. 2018. *Register variation online*. Cambridge: Cambridge University Press.

Biber, Douglas, Jesse Egbert & Mark Davies. 2015. Exploring the composition of the searchable web: A corpus-based taxonomy of web registers. *Corpora* 10 (1), 11–45.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad & Edward Finegan. 1999. *The Longman grammar of spoken and written English.* London: Longman.

Çarkoğlu, Ali, Lemi Baruh & Kerem Yıldırım. 2014. Press-party parallelism and polarization of news media during an election campaign: The case of the 2011 Turkish elections. *The International Journal of Press/Politics* 19 (3), 295–317.

Egbert, Jesse & Douglas Biber. 2019. Incorporating text dispersion into keyword analyses. *Corpora* 14 (1), 77–104.

Egbert, Jesse, Douglas Biber & Mark Davies. 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.

Erten-Johansson, Selcen, Valtteri Skantsi, Sampo Pyysalo & Veronika Laippala (2024). Linguistic variation beyond the Indo-European web: Analyzing Turkish web registers in TurCORE. *Register Studies* 6 (1), 60–90.

Göksel, Aslı & Celia Kerslake. 2005. *Turkish: A comprehensive grammar.* London & New York: Routledge.

Gries, Stefan. 2021. A new approach to (key) keyword analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics* 9 (2), 1–33.

Johanson, Lars. 2021. *Turkic*. Cambridge: Cambridge University Press.

Karapetjana, Indra & Jelena Lokastova. 2015. Register of electronic communication at sea. *Baltic Journal of English Language, Literature and Culture* 5, 52–61.

Koçak, Aslıhan. 2013. *A comparative register analysis of the language of cooking used in Turkish recipes* MA thesis. Ankara: Hacettepe University.

Kornfilt, Jaklin. 1997. *Turkish*. London & New York: Routledge.

Laippala, Veronika, Roosa Kyllönen, Jesse Egbert, Douglas Biber & Sampo Pyysalo. 2019. Toward multilingual identification of online registers. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, 292–297. Turku, Finland: Linköping University Electronic Press.

Lewis, Geoffrey. 2001. *Turkish grammar*. Oxford: Oxford University Press

Li, Lei, Anruze Li, Xue Song, Xinran Li, Kun Huang & Edwin M. Ye. 2021. Characterizing response quantity on academic social Q&A sites: A multidiscipline comparison of linguistic characteristics of questions. *Library Hi Tech* 41 (3), 921–938.

Liimatta, Aatu. 2019. Exploring register variation on Reddit. A multi-dimensional study of language use on social media website. *Register Studies* 1 (2), 269–295.

Liimatta, Aatu. 2022. Do registers have different functions for text length? A case study of Reddit. *Register Studies* 4 (2), 263–287.

Özyıldırım, Işıl. 2011. A comparative register perspective on Turkish legislative language. In Tarja Salmi-Tolonen, Iris Tukiainen & Richard Foley (eds.), *Law and language in partnership and conflict*. Turku: Oikeustieteiden tiedekunta Lapin yliopisto.

Pomikalek, Jan. 2011. *Removing boilerplate and duplicate content from web corpora*. Dissertation. Brno: Masaryk University.

Repo, Liina, Valtteri Skantsi, Samuel Rönnqvist, Saara Hellström, Mika Oinonen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo & Veronika Laippala. 2021. Beyond the English web: Zero-shot cross-lingual and lightweight monolingual classification of registers. In *Proceedings of the 16th Conference of European Chapter of the Association for Computational Linguistics: Student Research Workshop*, Kiev, 21–23 April.

Scott, Mike. 1997. PC analysis of key words – and key words. *System* 25 (2), 233–245.

Scott, Mike & Christopher Tribble. 2006. *Textual patterns: Keywords and corpus analysis in language education.* Amsterdam: John Benjamins.

Seoane, Elena & Douglas Biber. 2021. Corpus-based approaches to register variation. In Douglas Biber & Elena Seoane (eds.), *Corpus-based approaches to register variation*, 1–18. John Benjamins Publishing.

Skantsi, Valtteri & Veronika Laippala. 2023. Analyzing the unrestricted Web: The Finnish corpus of online registers. *Nordic Journal of Linguistics* 1 (1), 1–31.

Wang, Yiru, Xun Xu, Christina A. Kuchmaner & Ran Xu. 2023. But it was supposed to be healthy! How expected and actual nutritional value affect the content and linguistic characteristics of online reviews for food products. *Journal of Consumer Psychology* 33 (4), 743–761.

Appendices

Appendix A: Top 20 keywords of the sub-registers of Informational Description

Sub-register	Text Dispersion Keywords (Top 20)
Description of a thing/person	*tedavisi* 'treatment of' *enfeksiyonlar* 'infections' *enfeksiyonun* 'of the infection' *belirtileri* 'symptoms of' *fizik* 'physique' *rol* 'role' *hastalarda* 'in the patients' *2005* *hastalık* 'disease' *vardır* 'there is/are' *fakültesi* 'faculty of' *yaygın* 'widespread' *bağırsak* 'intestine' *ifadeyle* 'with expression' *dizisinde* 'in the series of' *ilaçlar* 'medications' *öykü* 'history' *kadınlarda* 'in women' *2007* *almıştır* 'has gotten'
Legal terms and conditions	*iade* 'return' *sayılı* 'numbered' *yasal* 'legitimate' *uyarınca* 'thereunder' *sözleşmesi* 'contract of' *işbu* 'hereby' *belirtilen* 'designated' *bilgilerin* 'of data' *kişisel* 'personal' *kargo* 'cargo' *kanunu* 'act of' *no* 'number' *6698* *korunması* 'protection of' *maddesinde* 'in the article of' *kanun* 'act' *verilerin* 'data of'

Sub-register	Text Dispersion Keywords (Top 20)
Legal terms and conditions (continued)	*kargoya* 'to cargo' *maddesi* 'article of' *yetkili* 'authorized'
Encyclopedia article	*doğdu* 'was born' *1972* *2004* *1988* *silmeden* 'wraparound' *evrenselliğe* 'to universality' *1989* *doğmuştur* 'was born *adlı* 'with the name of' *rock* 'rock' *yılında* 'in the year of' *almıştır* 'has gotten' *düze* 'dose' *filmleriyle* 'with the films of' *Ankara* *Mayıs* 'May' *2003* *Berlin* *kimdir* 'who is' *çini* 'tile'
FAQs	*hazırlıyoruz* 'we are preparing' *inceleyebilir* 's/he can examine' *başvurarak* 'by consulting' *araması* 'searching of' *bulunabilir* 'can be found' *yönlendirme* 'guidance' *çatlatma* 'fracturing' *pod* 'pod' *yurdumuz* 'our homeland' *sgkya* 'to sgk (social security institution)' *alkantra* 'alcantara' *coil* 'coil' *kargolanır* 'is shipped' *likitleri* 'liquids of' *iğnesinden* 'from the needle of' *klomen* 'klomen' *dansite* 'density' *lamine* 'laminated' *süngeri* 'the sponge' *sitelerimizi* 'our sites'

Sub-register	Text Dispersion Keywords (Top 20)
Research article	*sendromu* 'syndrome of' *etkilerini* 'the effects of' *araştırmak* 'to search' *frekans* 'frequency' *frekanslı* 'with frequency' *polariteli* 'with polarity' *coronavirüsler* 'coronaviruses' *ailesidir* 'is the family of' *polarite* 'polarity' *algınlığından* 'from the delusion of' *mers* 'mers' *virüsleridir* 'are the virus of' *zarflı* 'enveloped' *rmit* 'rmit' *akciğerlere* 'to the lungs' *nanopartiküller* 'nanoparticles' *iğnesiz* 'mutic' *nebulizatör* 'nebulizer' *kimyaya* 'to chemistry' *yeo* 'yeo'

Appendix B: Top 20 keywords associated with the sub-registers of Opinion

Sub-register	Text Dispersion Keywords (Top 20)
Review	*izlenim* 'impression' *açıkçası* 'frankly' *filmde* 'in the film' *belgeselin* 'of the documentary' *kitapta* 'in the book' *a0* *android* 'android' *filmleri* 'films of' *modelin* 'of the model' *filmi* 'the film' *kitabı* 'the book' *izleyiciye* 'to the audience' *oyunla* 'with game' *modeli* 'model of' *Spinozanın* 'of Spinoza' *izleyiciyi* 'the audience'

Sub-register	Text Dispersion Keywords (Top 20)
Review (continued)	*siyahi* 'black' *hocalarım* 'my gowns men' *Minecraft* *şerefi* 'honor of'
Opinion blog	*mi* (a question particle) *şey* 'thing, stuff' *ne* 'what' *hiç* 'never' *ama* 'but' *kim* 'who' *o* 'he/she/it/that' *bunu* 'this' *değil* 'not' *belki* 'maybe' *insanın* 'of the person' *düşünüyorum* 'I think' *mu* (a question particle) *yıllar* 'years' *olduğunu* 'that it is' *olmalı* 'should/must be' *diye* 'so that' *olsa* 'if it happened' oysa 'but/while' *biraz* 'a bit'
Advice	*burcu* 'zodiac of' *kendinizi* 'yourselves' *burç* 'zodiac' *size* 'to you' *yaşamınızda* 'in your life' *duygusal* 'emotional' *nasıl* 'how' *yorumu* 'interpretation of' *dikkat* 'caution, attention' *insanlar* 'people' *olacaktır* 'will happen' *olabilir* 'might happen' *unutmayın* 'don't you forget' *gerekiyor* 'it is necessary' *olabilirsiniz* 'you might be' *olun* 'be' (in imperative form) *günlük* 'daily' *aşk* 'love' *yaparken* 'while you do' *sevdiğiniz* 'that you love'

Sub-register	Text Dispersion Keywords (Top 20)
Religious blog/sermon	*Allah* *ey* (an interjection used in poetic contexts) *peygamber* 'prophet' *suresi* 'surah of' *Allahın* 'of Allah' *ayet* 'verse' *Bakara* 'Baqarah' *namaz* 'prayers' *onu* 'him/her' *onun* 'his/her' *Muhammed* 'Muhammad' *ona* 'to him/her' *insanın* 'of the person *Allaha* 'to Allah' *Hz* 'His holiness' *iman* 'faith' *Allahı* 'Allah' *ahirette* 'in the afterlife' *inkâr* 'denial' *dua* 'prayer'

Appendix C: Abbreviations in grammatical annotations

acc	accusative case	opt	optative
aor	aorist	part	participle
cv	converb	pass	passive
dat	dative case	pl	plural
gen	genitive case	psb	possibility
impf	imperfective	1P	first person plural
nec	necessity	1S	first person singular
neg	negative	3S	third person singular