Universal Journal of Educational Research 6(6): 1199-1206, 2018 http://www.hrpub.org
DOI: 10.13189/ujer.2018.060610

A Corpus Investigation on the Journal of Social Sciences of the Turkic World

İsa Yılmaz

Faculty of Education, Recep Tayyip Erdoğan University, Rize, Turkey

Abstract In recent years, a rapid development in computer technologies has been witnessed and feasibility of data access has been increased. In today's world, restoring documents, or data in general, and transferring them to interested parties are ordinary tasks. The amount of restored documents has also increased expeditiously and this development has required new technologies to emerge for building knowledge from large data sets. Basic applications of text mining include gathering and processing text to extract information that embodies raw data. Thus, basic text mining applications can help researchers to reach valuable knowledge from a mass of documents. This study investigated academic articles published in *bilig* (*Journal of Social Sciences of the Turkic World)* between 1996 and 2017 to find the frequencies of words and letters used in academic Turkish. Basic text mining of 4850817 words in 19437 pages from 81 *bilig* issues was completed using a natural language processing library, *Zemberek* and a programming language, R.

Keywords Journal of Social Sciences of the Turkic World, Text Mining, Document Processing, Turkish Word Frequency, Academic Turkish

1. Introduction

In parallel with technological developments, it can be said that there is a rapid change and progress in different areas. It stands as a major problem to compile, classify, interpret, and access, as required, the information and documents shared on the Internet in accordance with individual or corporate needs. As a result of all these developments, the number and amount of the documents stored in electronic media such as articles, reports, theses, annuals, e-books, e-mails, etc. increases day by day. It is not difficult any longer to inquire, find and transfer the data in an electronic environment making it now easier to save the collected data and perform transactions on the data stacks thanks to the platforms such as Internet [11]. This has brought with the necessity of analysing and processing large textual data with statistical methods. Today, data mining and text mining have gained importance due to the reasons mentioned.

Advances in the field of statistics and artificial intelligence constitute the foundations of data mining. At the same time, data mining is affected by developments in various disciplines and technologies, for example machine learning, which is a sequel of artificial intelligence developments [3]. In Longman Dictionary [14], data mining is defined as the process of using a computer in order for unnoticed details which cannot be seen easily and to examine a large amount of information. Oxford Dictionary [17] defines the term as the application about examination of current large databases to generate new information. Departing from such definitions, the three primary elements in this process were listed by Akküçük [2] as large amounts of data, potentially useful information, and mathematical and statistical techniques.

There may be structured or unstructured documents in databases. Structured documents include information such as title, author, publication date, category, etc. which allows access to information on the contents. However, documents composed of abstract and information flow only are not structured [8]. With the changing and developing technology, much more data has lately begun to be produced than before in most organizations [18]. This, in turn, has raised the issue of making accurate predictions for the future by eliciting meaningful relationships, structures and trends from existing data sets [3, 23].

Data mining applies to fields which relate to more than one sector [22]. Especially in the field of education, it can be used for categorizing students by performance and increasing their levels of adjustment, success and satisfaction [7].

Text mining is also called text data mining or document mining. Text mining works in reference to accessible and useable data. It is comprised of four steps as removing the data to be used; pre-treating the data, selection of the words to represents the text, and forming vector [5, 13]. It allows for not only to analyse large collections of unstructured

documents in order to obtain quality information about text mining [26], but also to identify whether a scientific work is a piece of pilferage or to identify the author of an anonymous text [7]. Developments in computer technologies increase the importance of text mining. Text mining, which is associated with several disciplines such as linguistics, statistics, and machine learning, can be used in many different areas [24]. In addition, text mining is becoming more and more important because it is not easy to analyse large amounts of unstructured data manually and to access necessary information [15].

Although human beings have the ability to understand unstructured data with some of their abilities, they lack computers' ability to process large volumes or high-speed texts [9]. Topics such as automatic summarization of texts examined, extraction of terms or concepts related to text, and clustering of texts by their similarities [21] can be done in a shorter time with developing computer technologies.

Turkish language requires more different text processing techniques compared to English and other languages due to its morphological and phonological characteristics. As a part of preparation, upper-case letters are changed into lower-case letters and punctuation marks are omitted from the text to be processed. Moreover, other preliminary works are undertaken such as throwing unnecessary words and creating keyword lists [12], except that all the words are converted to lower case and punctuation marks are thrown. Below is an illustration of the steps in exploration/extraction of information from databases.

The aim of this study is to find out the frequency of the Turkish letters and words in the papers published in the academic database called *bilig* (*Journal of Social Sciences of the Turkic World*) between 1996 and 2017 and to create the list of academic words in social sciences. For significance of this study, it is reasonable to refer to Pilavcılar's [19] suggestion that manual classification will be replaced by automatic classification in the near future and that text mining and text classification will be useful in management of large data. In addition; Al, Soydal and Yalçın [4] pointed out that more detailed work can be done through *bilig*. For this purpose, answers were sought for the sub-problems mentioned below:

1. What are the most and least frequently used Turkish letters in the bilig corpus?
2. What are the most frequently used Turkish words in the bilig corpus?
3. What are the common academic terms in the bilig corpus?

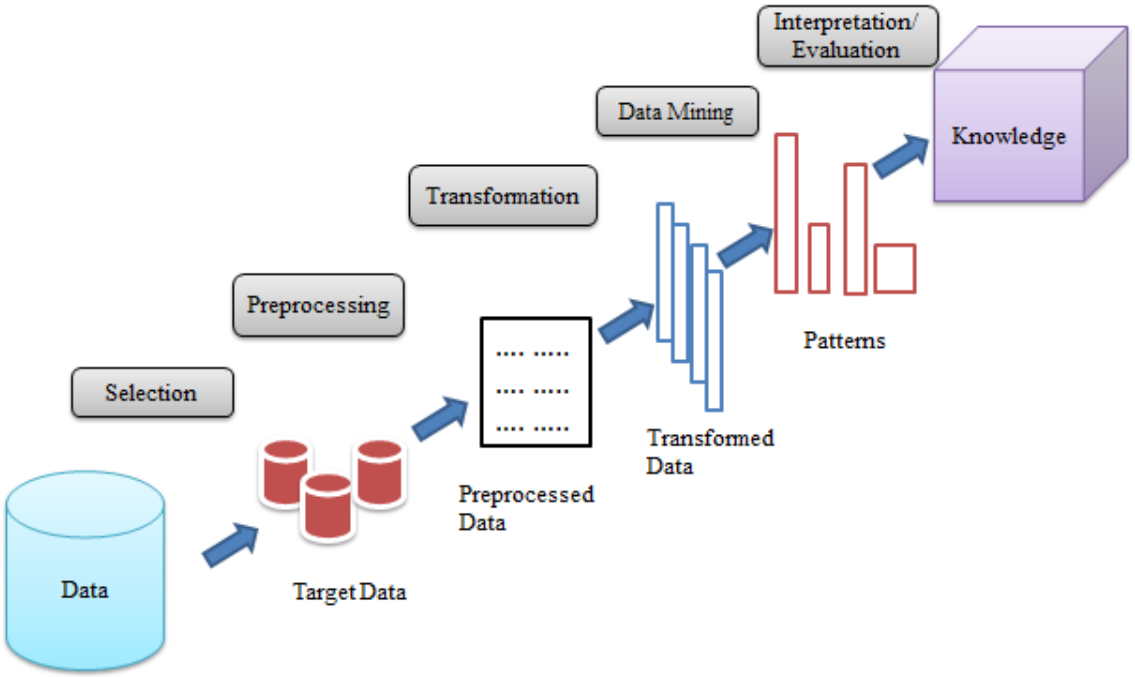

Figure 1. Steps of KDD (knowledge discovery in databases) [10]

2. Method

The aim of this study is to determine some numerical information from the vocabulary contained in papers published in *bilig* from the first till the last issue (81). For this purpose, the study was conducted to find out the frequency of Turkish words in article bodies and to create the list of academic words used in the context of social sciences texts in that review. In this process, a total number of 19,437 pages of document published in the review were scanned using the *pdftools* package [16] *pdf_text* function, which runs on the programming language R [20]. Scanned texts were merged into a single master file. Then, it was separated into words with the help of spaces and punctuation marks, and frequencies related to Turkish words were created. Before calculating word frequencies, *tolower* function provided in the *base* package was used to change all upper-case letters into lower-case letters.

Since words derived from the same root with derivational suffixes or inflectional suffixes increase vector size during analysis, the master file was then processed with Zemberek [1], which is a Natural Language Processing (NLP) library. In using the *Zemberek,* two main amendments were applied on the master file before finding root words frequencies[1]; firstly, non-Turkish sections (e.g. Russian, English) were identified and omitted; then, words separated by hyphen in line breaks were bound automatically. Next, edges of sentences were identified (tokenization) and structural analysis process (morphology) was performed to elicit the roots.

There are limitations that readers must consider before reading the results. Despite being not substantial, these limitations can affect accuracy of the results negatively. Some of the published articles were printed in two columns on the same page, making it difficult to bind the words separated by hyphen. Also it must be remembered that structural analysis function of *Zemberek* can still be improved.

3. Findings

This section is dedicated to results obtained from *bilig* corpus. First, information will be given on the structure of the surveyed corpus followed by the most common word roots and proper nouns found in the same corpus. Table 1 displays the overall structure of *bilig*, and the number of sentences and words used in articles published in all 81 issues from 1996, when the journal started its life, up to year 2017. A total of 4,850,817 Turkish words were found after deducing texts in English and Russian and abstracts.

Table 1. Overall information about the bilig corpus

Corpus	f
No of sentences	310,650
No of words, figures and punctuation marks	6,295,672
No of words without punctuation marks	5,142,818
No of words (without figures, address, etc.)	4,850,817

After cleaning, the corpus was comprised of *31,410,231* letters all in Turkish texts and then surveyed, attaching the highest frequency to letter *a*. It was followed by letters *e, i, n, r, l, and k*. The lowest frequency was found with letter *j*. The other least used letters were seen to be *p* and *f*. The graphic that shows the usage frequency of letters is given in Figure 2.

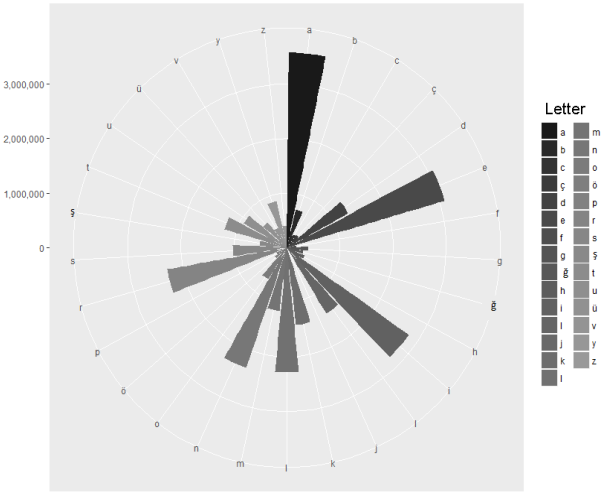

Figure 2. Frequency of letters in the bilig corpus

Table 2 reports the place names in the *bilig* corpus, *Türkiye* is the most often cited place name in *bilig*. *Türkiye* is followed by the city names *İstanbul* and *Ankara*. The third and fourth places are occupied by *Rusya* and *Anadolu,* respectively. They are followed by *Azerbaycan, Asya,* and *Kazakistan*. Apart from these; it is seen that the term *Sovyetler* along with the corresponding abbreviation *SSCB* and *Moskova* has a high frequency across the corpus. It also seems worth to note that the abbreviation *ABD* (the USA) and the word *Amerika* have almost the same frequency. Also, the most often cited Turkish cities were found to be *Konya, Bursa and İzmir* in papers in *bilig*.

1The amendments were not regarded necessary to be used for determining frequency of words in using the program R.

Table 2. Place names in the bilig corpus

Continent, country, city, region	f	Continent, country, city, region	f
Türkiye	10321	Suriye	766
İstanbul	7419	Kafkasya	725
Ankara	6152	Bulgaristan	682
Rusya	3916	Almanya	673
Anadolu	3336	Fransa	658
Azerbaycan	3313	Konya	642
Asya	2580	Hindistan	635
Kazakistan	2464	Özbekistan	632
İran	1985	Kırım	619
Avrupa	1746	Moskova	559
Türkistan	1730	SSCB	536
Kıbrıs	1319	Bursa	506
Altay	1245	İngiltere	494
Sovyetler	1155	Bakü	467
Çin	1113	Kosova	466
Batı	1006	Taşkent	424
ABD	920	Almatı	419
Irak	899	İzmir	408
Kırgızistan	892	Makedonya	407
Amerika	829	Paris	406
Türkmenistan	774	Bosna	405

The list of the most frequently used person' names in *bilig* are reported in Table 3 and it starts with the name *Ali*. It is followed by other names Ahmet and Ahmed, which refer to one single name except for the strong or soft consonant ending in accordance with Turkish or Arabic phonetics, respectively. The same applies to the pair of *Mehmet* and *Mehmed*. Another high-frequency person' name is seen to be *Ahmet Yesevi,* the name of the Great Turkish Sufi. The other prominent person' names in *bilig* include *Mustafa Kemal Atatürk*, the founder of the Republic of Turkey; *Abay Kunanbayev*, Kazakh poet; *Köroğlu*, Turkish minstrel; and *Nasreddin Hoca,* who is an immortal character due to his jokes as well as his identity as a religious scholar.

Table 4 displays the most prevalent nationality names appearing in *bilig*. In this group; while *Türk* and *Osmanlı* were used for 32,576 times (54%), the frequency of the remaining words was calculated as 27,871. Following the first two entries, the most frequently used nationality names are *Rus, Kazak, Türkmen, Sovyet, Arap, Kırgız, Uygur, Rum, Selçuklu, Ermeni, Bulgar, Moğol, Özbek, alman, İngiliz, Fransız, Yunan, Kıpçak, Kafkas, and Arnavut.*

Table 3. Personal names in the bilig corpus

Person's Name	f	Person's Name	f
Ali	2788	Süleyman	814
Ahmet	2655	İsmail	759
Ahmed	1631	Hasan	703
Mehmet	1578	Yusuf	690
Oğuz	1514	Abay	643
Mustafa	1420	Mahmud	512
Mehmed	1245	Abdullah	489
Timur	1139	Köroğlu	486
Yesevi	1125	Nasreddin	486
Muhammed	1022	Stalin	469
İbrahim	881	Mahmut	439
Atatürk	864	Cengiz	419
Hüseyin	820		

Table 4. Nationality names in the bilig corpus

Nationality	f	Nationality	f
Türk (Turkish)	25554	Ermeni (Armenian)	1036
Osmanlı (Ottoman)	7022	Bulgar (Bulgarian)	1019
Rus (Russian)	4765	Moğol (Mongolian)	856
Kazak (Kazakh)	4420	Özbek (Uzbek)	725
Türkmen (Turkmen)	2242	Alman (German)	693
Sovyet (Soviet)	1948	İngiliz (British/English)	604
Arap (Arab)	1803	Fransız (French)	585
Kırgız (Kirgiz)	1686	Yunan (Greek)	489
Uygur (Uyghur)	1359	Kıpçak (Qipchaq)	468
Rum (Greek)	1169	Kafkas (Caucasian)	423
Selçuklu (Seljukian)	1162	Arnavut (Albanian)	419

Table 5 shows language names referred in the articles in bilig, which indicates the order of Türkçe on top of the list. It is followed by *Farsça, Arapça, İngilizce*, and *Rusça*. In summary, the data above could help find out the issues or elements with priority or less important in the corpus surveyed.

Table 5. Language names in the bilig corpus

Language	f
Türkçe (Turkish)	9942
Farsça (Persian)	1333
Arapça (Arabic)	779
İngilizce (English)	763
Rusça (Russian)	537

Two specialists in the area prepared two different lists of academic words by eliciting from *bilig* corpus. The disputes in the lists were resolved through discussion and some amendments were made. Then, the list was finalized by the researcher and this list is depicted in Figure 3. As a

result, *bilig* offered the following items as the most frequently used academic terms:

dönem (period), *ilişki* (relationship), *oluşmak* (to be composed of), *bilgi* (information), *şiir* (poetry), *araştırmak* (to research), *ifade* (expression), *toplum* (society), *kültür* (culture), *edebiyat* (literature), *kaynak* (resource), *bölge* (region), *etki* (effect), *değer* (value), *eğitim* (education), *şair* (poet), *baz* (basis), *sosyal* (social), *sağlamak* (to ensure), *değerlenmek* (to improve), *uygulamak* (to apply), *incelemek* (to examine), *sanat* (arts), *tür* (type/genre), *gelenek* (tradition), *yön* (direction), *sistem* (system), *kavram* (concept), *metin* (text), *politika* (policy), *merkez* (centre), *bilim* (science), *yayınlamak* (to broadcast), *belirlemek* (to determine), *ekonomik* (economic), *sınır* (border), *kurum* (institution), *kısım* (part), *unsur* (factor), *yönetim* (administration), *düzey* (level), *bağım* (dependence), *yaratmak* (to create), *uygun* (suitable), *görüş* (view), *kültürel* (cultural), *faaliyet* (activity), *yayın* (publication), *oran* (ratio), *makale* (paper), *düşünce* (thought), *yaşam* (life), *vergi* (tax), *nitelik* (quality), *tespit (*finding), *çerçeve* (framework), *anlayış* (understanding), *siyasi* (political).

Besides, some collocations of the most frequent words are given as they appear in different texts. Bearing in mind that *bilig* is a prominent academic journal specific to social sciences could be referred in studies related to Academic Turkish, example uses of the words are given below from different contexts.

Klasik dönem İran şairleri, Dede Korkut destanlarının anlatıldığı dönem, IV. dönem milletvekili, III. Selim Dönemi, Mustafa Kemal dönemi, soğuk savaş dönemi, geçiş döneminde beklenmeyen pürüzler, ilk dönem şiirinin temel özelliği.

Kazaklarda akrabalar arası ilişkiler, dil ilişkileriyle uğraşmanın güçlüğü, tabiat ve tasvirle ilişkisi, Sovyet-Alman ilişkileri, öğreticilerin talebeyle samimi bir ilişki kurması, ticari ilişkiler kurmak.

Eser 52 varaktan oluşmakta, yeryüzü şekilleri olarak dağlardan ve ovalardan oluşmakta, bu kelime sapog kökünden oluşmakta, Rumeli'ye gitmiş Yörüklerden oluşmakta, buna karşılık halk dini Şamanizm ekseninde oluşmakta, daha oluşmakta olan Hakas edebiyatında, Kazakça sözcüklerden oluşmakta.

Bu eserdeki bilgiler geliştirilerek daha sonraki yıllarda birçok kere bastırılmıştır, giriş bölümünde eserin teknik bilgileri, geleneği hakkında da bilgi verilmiş, marifetin hakikat bilgisi demek olduğu, bir bilimsel bilgi.

Makedon lirik şiiri, seyahatlerini anlatan öyküsel şiirleri, şiirsel ifadenin ustaca kullanımı, bu dönem şiir ve düzyazısındaki değişim, 1935-1960 şiirin, şiir sanatı, izleksel ve imgesel bakımdan, divan şiiri, şiir dili ve konusu, farklı şiir formları.

Altay halklarının folklorunu araştırmak gerekmekte, meseleyi araştırmak için alandaki uzmanlar, bunların içeriğinden bahseden eserleri araştırmak, yazarın estetik dünyasını derinden araştırmak, Türk gençlerinin evlilik tercihlerini araştırmak, çözüm yollarını araştırmak, Türkçe kullanımına ilişkin öğretmen düşüncelerini araştırmak.

Çelişkili ifadelerle dolu bir rapor, on binlerle ifade edilebilecek, bu uzun ifade tercih edilecekse, bir tek kelimeyle ifade ettiği, farklı bir durumu ifade eden, birçok üniversitede tarih bölümlerinde ders olarak okutulmakta, ifadesiyle, Türkçe kitaplardan yararlandıklarını ifade etmekte.

Kendi toplumunu çok iyi gözlemleyen bir sosyal bilimci, sanayi devrimiyle birlikte ortaya çıkan sanayi toplumu, içinde yaşadığı toplumun yapısı, göçebe toplumlarda çok sık karşılaşılan, eski Türk toplumlarında, toplumumuzu oluşturan bireyler.

Kaybolmaya yüz tutmuş kültür kaynaklarımız, bu farklı yöreleri birbirlerine yaklaştıran kültür elçileri, maddi kültür unsurları, ortak kültür mirası, bozkır kültürü, at kültürü, kültürel değerlerimiz, okuyucunun kültür düzeyi, dünya yazı kültürünün en zengin hat mirasına sahip olan ülkemiz, batı kültürü.

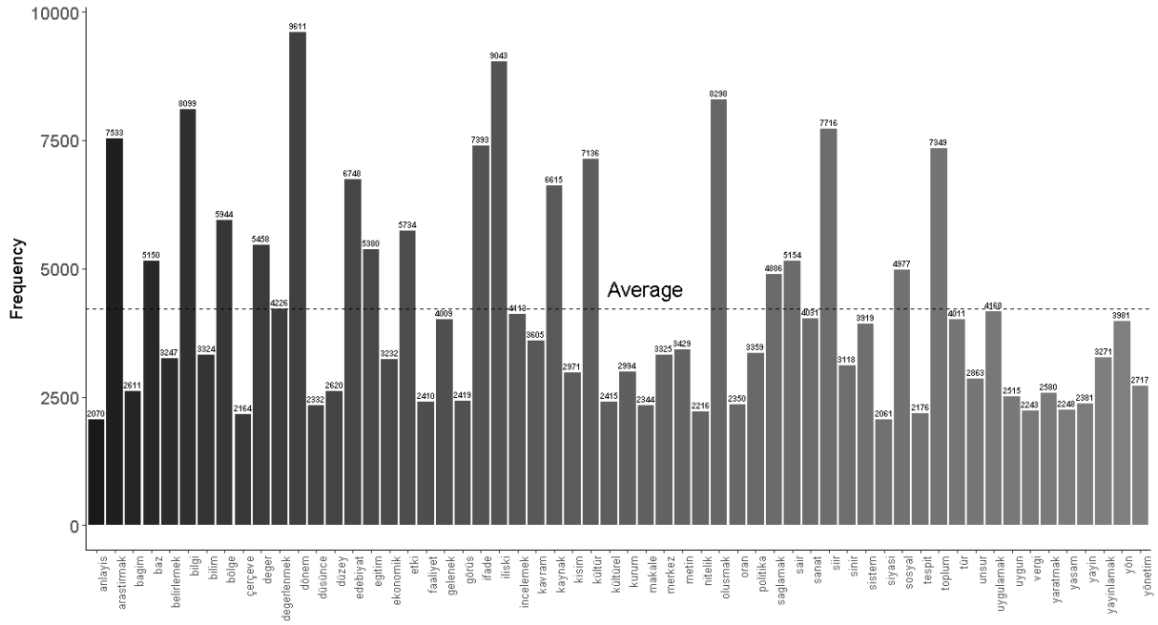

Figure 3. Academic words with the highest frequency in the bilig corpus

4. Conclusions

In this study, the corpus of *bilig*, which covers the period from 1996 to 2017 was analysed. In this framework, a corpus containing 310,650 sentences and 4,850,817 words was analysed. In this scope, the corpus was evaluated from various aspects and interpreted accordingly. Here as stated by Büyük [6], software is available to generate quite accurate and consistent results in text analysis in languages such as English; however, it cannot be applied properly to Turkish language and there are not enough studies in this area.

In this research, the most recurrent letters and words in the articles published in *bilig* were determined. As a result, the letter *a* was found to have the highest frequency followed by *e, i, n, r, l,* and *k*. On the other hand, the least recurrent letters were found to be *p*, *f,* and *j*. In a similar study conducted by Çelikyay [7] on a Turkish text consisting of *118,553* letters, the letter *a* was also found as the most frequently used one followed by *a, e, i, l, n*, and *r*. Again, the letter with the lowest frequency rate was found to be *j*. It is noteworthy that the same results were attained in a corpus which is about as 265 times large as the samples in Çelikyay's [7] survey.

It was demonstrated that the words *Türkiye* and *Türkçe* are among the highest-frequency words in relation to the root *Türk*. In fact, our findings seem to be supported by Bayer [5], in which an application was developed to survey the meaning map of Turkish words and lexical meanings of words could be guessed at accuracy level of 86% thanks to word collocations. In the present study, it was found out that other words associated with frequently used words also have a high frequency.

Also, the most repeated words in this corpus were found as follows: görmek, almak, yapmak, yer, vermek, gibi, o, dil, gelmek, kendi, ara, tarih, yıl, daha, konu, çok, çalışmak, ad, ise, yazmak, orta, bulunmak, şekil, söz, karşı, önem, sonra, demek, taraf, eser, zaman, çıkmak, her, kullanmak, baş, en, üzeri, el, dönem, ilgi, devlet, halk, geçmek, büyük, insan, and göstermek. These most repeated words were identified after removing the stop-words. For example the conjunction *ve* was found to be the most frequently used stop-word. It is followed by *olmak* as an auxiliary stop-verb and other stop-words *bir, bu, etmek, iç, da, de,* and *ile*.

As for the place names, the highest frequency was reported for the country name *Türkiye*. It was followed by *İstanbul, Ankara, Rusya, Anadolu, Azerbaycan, Asya, Kazakistan, İran,* and *Avrupa*.

Another list, person' names, was found to include Ali, Ahmet, Ahmed, Mehmet, Oğuz, Mustafa, Mehmed, Timur, Yesevi, Muhammed, İbrahim, and Atatürk. The other high-frequency names include Hüseyin, Süleyman, İsmail, Hasan, Yusuf, Abay, Mahmud, Abdullah, Köroğlu, Nasreddin, Stalin, Mahmut, and Cengiz. It is seen that the corpus under scrutiny includes historical and literary figures from different periods and countries. The tool developed by Savaşan [21] for access to daily news on the Internet made it possible to access to the news on the same topic published in different newspapers from a single point with the aid of an automatic tag cloud. Such studies can be used for analyzing documents that constitute a large volume of data particularly in the context of language and literary studies as a part of social sciences. In addition, Varol's [25] classification study for discovering the poet of an anonymous poem through text mining can inspire researchers to focus on literature.

As another group of words searched in this study, nationality words demonstrated that *Türk* is the most frequently used item. It is followed by the nationality words such as *Osmanlı, Rus, Kazak, Türkmen, Sovyet, Arap, Kırgız, Uygur, Rum,* and *Selçuklu*. Similarly, as a language name, *Türkçe* was noted as the most recurrent term in *bilig* corpus. The other outstanding items in this regard are *Farsça, Arapça, İngilizce,* and *Rusça.*

As the last step in this study, a list of academic words was prepared on the basis of *bilig*. The list can be useful for undergraduate and graduate students because academic word lists are considered as an important material for increasing academic Turkish skills of particularly foreign students studying in Turkey.

5. Recommendations

In the light of our study findings, following recommendations were brought for future research: Our study was conducted on a body of *4,850,817* words. Further studies can be done on more comprehensive texts leading to different explorations on Turkish vocabulary.

In the future, it would be useful to study compilations in different disciplines to exploit the words used most often in the academic context so that research and curriculum development studies can be carried out to this end.

Subsequently, social network analysis could be performed on the names (of city, country, region, persons, and so on) derived from the social sciences corpus in our study.

Finally, the academic word list for social sciences derived from *bilig* corpus could be replicated in journals specialized in other areas.

Acknowledgements

1. I would like to thank Dr. Burak Aydın, one of my colleagues at RTE University. He helped me to shape the idea and provided supervision on how to utilize the programming language R to carry out this study. However, later some additional measures had to be taken since there is no single R package that can perform root-suffix decomposition for Turkish words. Therefore, the resources regarding a project called Zemberek were examined via GitHub and the project manager was contacted. I am grateful to Mr. Ahmet Afşın Akın for his great contribution to the completion of the work. I would like to thank them very much.
2. Figures 2 and 3 were drawn with the ggplot2 [27] package.

REFERENCES

[1] Akın, A., A., & Akın, M., D., (2007). Zemberek an open source NLP framework for Turkic languages. Structure, 10. Available at: https://github.com/ahmetaa/zemberek-nlp

[2] Akküçük, U. (2011). *Veri madenciliği: kümeleme ve sınıflama algoritmaları*. İstanbul: Yalın Yay.

[3] Akpınar, H. (2000). Veri tabanlarında bilgi keşfi ve veri madenciliği. *İ.Ü. İşletme Fakültesi Dergisi, 29,*1, s: 1-22.

[4] Al, U., Soydal, İ. & Haydar, Y. (2010). Bibliyometrik Özellikleri Açısından bilig'in Değerlendirilmesi. bilig, *Türk Dünyası Sosyal Bilimler Dergisi 55*: 1-20.

[5] Bayer, H. (2011). *Veri madenciliğinde bir metin madenciliği uygulaması*. Unpublished master's thesis, Beykent Üniversitesi, İstanbul.

[6] Büyük, E. (2016). *Conflict analysis for Turkish debates using text mining and text segmentation techniques*. Master Thesis. İstanbul: Bahçeşehir University.

[7] Çelikyay, E., K. (2010). *Metin madenciliği yöntemiyle Türkçede en sık kullanılan ve birbirini takip eden harflerin analizi ve birliktelik kuralları*. Unpublished master's thesis, Beykent Üniversitesi, İstanbul.

[8] Doğrusöz, A. (2007). *Makine öğrenmesi teknikleri ile metinlerin otomatik olarak sınıflandırılması*. Unpublished master's thesis, Yıldız Teknik Üniversitesi, İstanbul.

[9] Fan, W., Wallace, L., Rich, S., & Zhang, Z. (2006). Tapping the power of text mining. *Communications of the ACM, 49*(9), 76-82.

[10] Fayyad, U., Piatestsky-Sharpiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine, 17*(3), 37-54.

[11] Güven, A. (2007). *Türkçe belgelerin anlam tabanlı yöntemlerle madenciliği*. Unpublished doctoral dissertation, Yıldız Teknik Üniversitesi, İstanbul.

[12] İlhan, U. (2001). *Application of k-NN and FPTC based text categorization algorithms to Turkish news reports*. Master Thesis. Ankara: Bilkent University.

[13] Karaca, M., F. (2012). *Metin madenciliği yöntemi ile haber sitelerindeki köşe yazılarının sınıflandırılması*. Unpublished master's thesis, Karabük Üniversitesi, Karabük.

[14] Longman Dictionary, Retrieved June 15, 2017, from http://www.ldoceonline.com/dictionary/data-mining

[15] Oğuzlar, A. (2011). *Temel metin madenciliği*. Bursa: Dora Yay.

[16] Ooms, J. (2017). pdftools: Text Extraction and Rendering of PDF Documents. R package version 1.2. https://CRAN.R-project.org/package=pdftools

[17] Oxford Dictionary, Retrieved June 17, 2017, from https://en.oxforddictionaries.com/definition/data_mining

[18] Özkan, Y. (2013). *Veri madenciliği yöntemleri*. İstanbul: Papatya Yay.

[19] Pilavcılar, İ. F. (2007). *Metin madenciliği ile metin sınıflandırma*. Unpublished master's thesis, Yıldız Teknik Üniversitesi, İstanbul.

[20] R Core Team, (2017). R: A language and environment for statistical computing. Vieanna, Austria: R Foundation for

Statistical Computing.

[21] Savaşan, S. (2011). *Türkçe içeriklerden otomatik etiket bulutu oluşturma*. Unpublished master's thesis, Yıldız Teknik Üniversitesi, İstanbul.

[22] Silahtaroğlu, G. (2013). *Veri madenciliği kavram ve algoritmaları*. İstanbul: Papatya Yay.

[23] Şimşek Gürsoy, U., T. (2012). *Uygulamalı veri madenciliği sektörel analizler*. Ankara: Pegem Akademi Yay.

[24] Uzun, V. (2014). *Semantic text mining and an application in Turkish documents*. Master Thesis. İzmir: Dokuz Eylül University.

[25] Varol, M. (2011). *Metin madenciliği yöntemlerini kullanarak Türkçe dokümanlarda tür ve yazar tanıma*. Unpublished master's thesis, Süleyman Demirel Üniversitesi, Isparta.

[26] Visa, A. (2001). Technology of Text Mining, *In International Workshop on Machine Learning and Data Mining in Pattern Recognition* (p. 1-11). Berlin: Springer.

[27] Wickham, H. (2009). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York.