A Corpus of Grand National Assembly of Turkish Parliament's Transcripts

Onur Güngör[1], **Mert Tiftikci**[2], **Çağıl Sönmez**[3]
Department of Computer Engineering,
Bogazici University[1,2,3], Istanbul, Turkey
Huawei R&D Center[1], Istanbul, Turkey
onurgu@boun.edu.tr[1], mert.tiftikci@boun.edu.tr[2], cagil.ulusahin@boun.edu.tr[3]

Abstract

In parliaments throughout the world, decisions that are taken directly or indirectly lead to events that affect the society. Eventually, these decisions affect other societies, countries and the world. Thus, transcriptions of these are important to people who want to understand the world, namely historians, political scientists and social scientists in general. Compiling these transcripts as a corpus and providing a convenient way to query the contents is also important from the point of linguists and NLP researchers. Currently, many parliaments provide these transcriptions as free text in PDF or HTML form. However, it is not easy to obtain these documents and search the interested subject. In this paper, we describe our efforts for compiling the transcripts of Grand National Assembly of Turkish Parliament (TBMM) meetings which span nearly a century between 1920 and 2015. We have processed the documents provided by the parliament to the public and transformed them into a single collection of text in universal character coding. We also offer an easy to use interface for researchers to launch custom queries on the corpus on their own. To demonstrate the potential of the corpus, we present several analyses that give quick insights into some of the linguistic changes in Turkish and in Turkish daily life over the years.

Keywords: parliamentary text, corpus, historical records, transcriptions

1. Introduction

As the technological tools for archiving and disseminating text proliferate, we see an increasing number of parliaments across the world that share the transcripts of legislative meetings with the public (Fišer, 2017). This enables a new line of research for humanities and social sciences (Bayley et al., 2004; Cheng, 2015; Georgalidou, 2017; Pančur and Šorn, 2016) and computational linguistics (Mandravickaite and Krilavičius, 2017; Høyland et al., 2014; Grijzenhout et al., 2014; Rheault et al., 2016).

Although parliamentary data is shared with the public, conducting statistical analysis on them is cumbersome in general. This is mainly because they are usually accessed through a search engine where the common workflow is to search for a specific keyword and use search results to investigate the evidence to the specific research question. If the only way of access is through a search engine, it is not possible to calculate statistics of word usage frequency across time or to employ word clustering algorithms besides others which require access to the whole set of documents at once.

In this work, we present our work to address this issue by crawling, processing and combining the transcripts of Grand National Assembly of Turkish Parliament into a single corpus. Our contributions include easier programmatic access to the corpus and several methods to calculate NLP related statistics over the corpus.

The remainder of this paper is organized as follows: in Section 2. we compiled a summary of related work. We explain the details of the process of building the corpus in Section 3. Then, we present a simple analysis of the corpus in Section 4. Finally we conclude in Section 5.

2. Related Work

Although there have been studies in the literature that employ parliamentary data (Vives-Cases and Casado, 2008), studies that compile and process the transcripts to be accessed in a straightforward manner are relatively scarce and do not follow a single format (Verdonik et al., 2013; Graën et al., 2014; Marx et al., 2010).

A recent workshop organized by CLARIN programme aimed to join forces and motivate research on using NLP technologies to make parliamentary data accessible for humanities and social sciences research. The initiative published a report which summarizes the parliamentary corpora in the CLARIN infrastructure (Fišer, 2017).

3. Building the corpus

In this section, we will give details of data preparation and preprocessing phases.

The members of parliament were elected each five years beginning from 1920 until 2007. After 2007, the elections were made every four years. The time period that a parliament is functional after each election is said to form a "term" and is made up of several "lawmaking year"s depending on the actual duration of the "term" which can change due to unscheduled elections or other unexpected events. Every "lawmaking year" is conducted as a series of meetings which we call "sessions". These "sessions" are transcribed by clerks present in the hall in real time. These transcriptions were traditionally published as a periodical called 'Tutanak Dergisi'[1]. With the introduction of digital media, the transcriptions are published on the web as soon as they are redacted in digitized form by the Library, Documentation and Translation Department and Information Technologies Department[2] of the parliament. However, the transcriptions of the sessions before 20th "term" are not published in digitized form, only as scanned images of 'Tutanak Dergisi'. On the other hand, scanned images of Tutanak Dergisi is available for the first 25 terms (1920-2015).

[1] literally 'Journal of Minutes of the Meeting' in English
[2] https://www.tbmm.gov.tr/kutuphane/

So we have chosen to base our work on these scanned images of Tutanak Dergisi to have a corpus which spans 95 years of transcriptions.

The data preparation process can be summarized as the following: We start by crawling the web pages of TBMM. In this phase, we extract the locations of PDF files which contain the transcriptions. We use a command-line tool to extract text from downloaded PDF files. Then we apply a very simple preprocessing operation in which we only get rid of unprintable characters introduced by the text extraction process. We then tokenize the resulting text and obtain the final version of the transcripts. Finally, we compile every document into a single corpus in a reusable format.

3.1. Crawling

Even though we share the scripts which we used for crawling and downloading, we also give the details of the crawling process here for others to replicate.

We used manual labor to obtain the URLs pointing to the scanned images of 'Tutanak Dergisi'. Our effort started with a single visit to a page which contains pointers to every "term" page. After this, we visited every "lawmaking year" page which is accessed through each "term" page. We used a simple browser extension to extract the URLs found in a 'lawmaking year' page.

3.2. Processing

We used `pdftotext` to extract the text contained in each PDF file. `pdftotext` is a tool which is part of `poppler`[3] PDF rendering library. It produced good results in general but sometimes this approach produced erroneous results or no results at all. This is mainly due to the quality of the scanning done when the parliament publishes these files. We only remove spurious characters at the end of lines to obtain the text in free form. We continue with a simple tokenization and conclude our processing by coding the words using a dictionary. We do not strip out or reorder any word during this process.

Our corpus in its current form only records the date of the session. We did not extract the speaker, the context, the political party the speaker belongs to or try to identify other people during the session as suggested in the literature (Marx et al., 2010). However, we made our file format so simple that it is both human and machine readable. Our corpus file basically contains a single document in each line with words in the document in the order as they appeared in the source documents.

The code that is used to crawl and process the corpus can be found in our Github repository[4].

4. Analyzing the corpus

The resulting corpus contains 208 million tokens in 12645 documents which are derived from transcriptions of general assembly sessions between 1920 and 2015. Each document includes data from a session which usually spans a day. We do not include the transcriptions between 2015 and 2018 in this study as they were provided in HTML format as part of a different mode of distribution.

The total number of unique tokens is 619,505, but if we only consider tokens which are found more than 10 times, this figure decreases to 358,286. The number of unique tokens in a given Turkish corpus is usually more than the expected number for other languages which are not morphologically rich thus do not exhibit extensive inflection and derivation. We tested the coverage of our corpus by looking up these unique tokens in a decent Turkish language dictionary[5] from Turkish Language Institute ("Türk Dil Kurumu"). As a result, we found out that about 70% of all unique tokens can be found in the Turkish dictionary.

The median number of tokens in a document is 9,642. We give figures summarizing the total number of words and sessions held per year in Figure 1a and 1b. The distribution of document lengths follow an exponential pattern as can be seen in Figure 1c.

The total size of PDF files is 3.9 gigabytes. After we process and encode the words with a dictionary, the size of the resulting file decreases to 1.2 gigabytes. We share the corpus in a compatible format with the scientific community in our source code repository. Alternatively, we will share the corpus through the Virtual Language Observatory (VLO) in the CLARIN infrastructure and as a shared LREC resource.

4.1. Access to the corpus

In addition to serving the processed corpus as a downloadable file, we provide an offline interactive interface suited for use by linguists or social scientists[6].

For implementing such an interface, we employed Project Jupyter[7] and created a Jupyter notebook. A Jupyter notebook is a special file which can be used to mix documentation and sample code. This enables the user to run simple queries and write their own exploration scripts.

4.2. Word and Topic Distributions

We employed several analyses on the corpus to demonstrate the potential areas for research. First, we have employed latent Dirichlet allocation model (Blei et al., 2003), to discriminate words into a predefined number of topics. We set the number of topics to 20. Using this allocation, we could interpret the topic distribution of a given transcription. We examine these allocations to interpret the representation quality of the topics. For example, in Figure 2, we plot the average weights of selected topics in a year. We chose to present these topics because topics 4, 9 and 12 contain words that are considered as old in Turkish. This is validated in the figure. On the other hand, topic 15 and 16 are two topics that can be used to mark transcriptions recorded in 1990's and beyond.

A similar observation can be also done by looking at the plot of the word usage frequencies of the Turkish word 'mebus'. This word of Arabic origin is used to refer to a member of the parliament in old Turkish. It was adopted

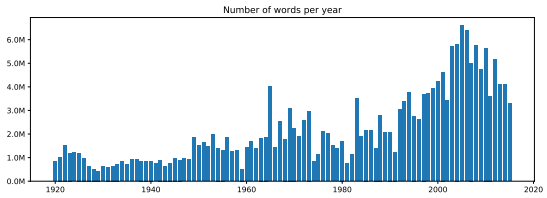
[3] `https://poppler.freedesktop.org/`

[4] `https://github.com/onurgu/turkish-parliament-texts`

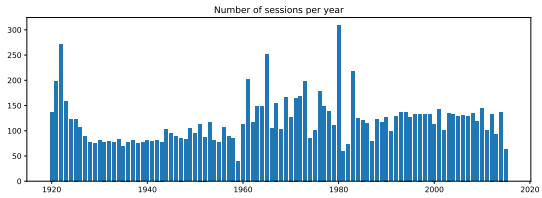
[5] `http://www.tdk.gov.tr/`

[6] Both can be accessed at `https://github.com/onurgu/turkish-parliament-texts/releases`

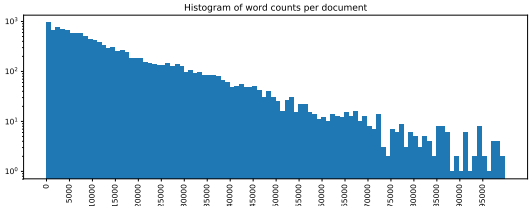
[7] `http://jupyter.org/`



(a) Number of words per year.



(b) Number of sessions per year.



(c) Histogram of word counts per document.

Figure 1: Figures that summarize several statistics about the corpus.

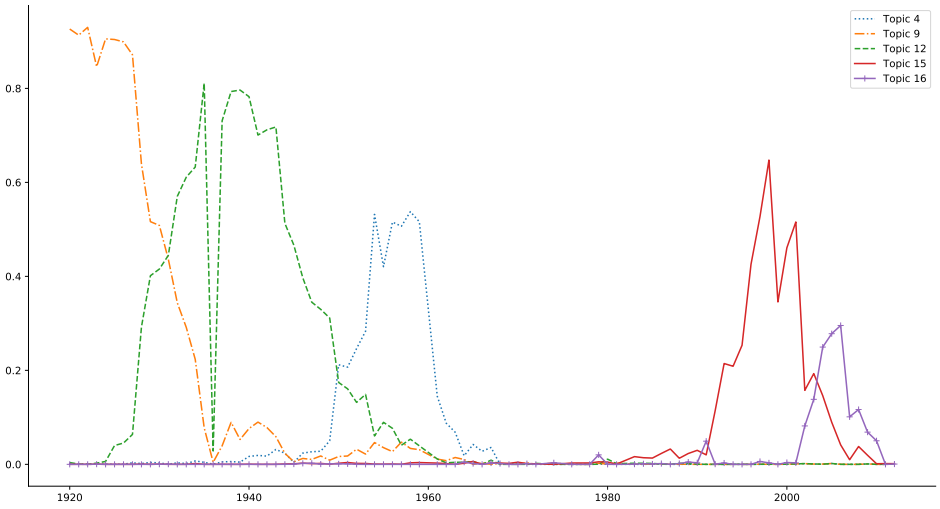


Figure 2: Distribution of selected topics across time.

during the Ottoman era and continued to be used in the republic. It was known that its use decreased through time. We also validate this with a simple query of word counts for the word 'mebus' across the corpus. As can be seen in Figure 3, the usage of 'mebus' is nearly always before 1960. The alternative word 'milletvekili' is mainly used after 1960. This basic plot itself provides an analytical tool for comparing different eras of culture and linguistics. Thus, this shows a good example of the potential of the information contained in the corpus.

In this corpus, there are also traces of introduction of technology in daily life of Turkish citizens. To demonstrate this, we present a comparison between electronic communication devices across the entire corpus ordered by time in Figure 4. The curves in the figure show that television did not become a frequent concept of debate until 1980's. On the other hand, telephone network related issues lost a considerable weight in 1960's. Lastly, we observe the introduction of internet in the parliamentary transcriptions at an increasing pace.

We defer further analytical research to future work. However, we have to note that these analyzes are only scratching the surface. Firstly, due to the high volume of meaningful historical text, we believe that it is possible to conduct comparative linguistics research in Turkish. Second, a wide range of discourse analysis can be done as we can relate every sentence to a specific person belonging to a political party. Moreover, these sentences can be part of a dialogue adding more value to the utterance.

5. Conclusions

In this paper, we present our work on creating a corpus of Grand National Assembly of Turkish Parliament which

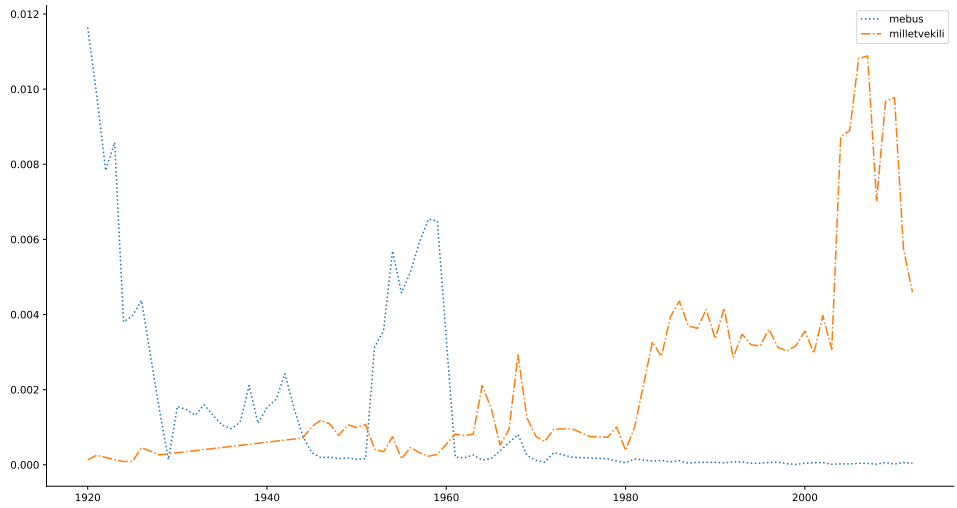


Figure 3: Yearly usage frequencies for ‘mebus’ and ‘milletvekili’ across whole corpus. The yearly usage frequency is normalized over all words in a year.

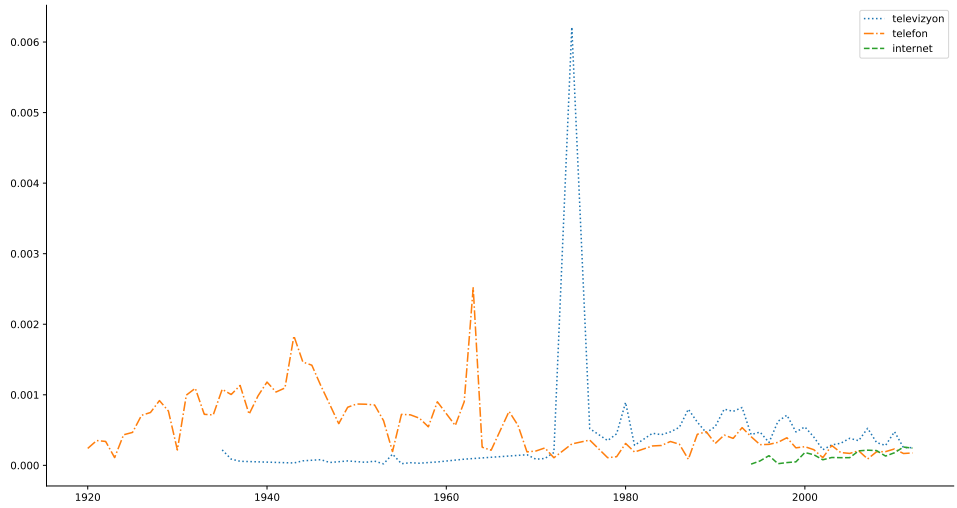


Figure 4: Yearly usage frequencies for ‘televizyon’, ‘telefon’ and ‘internet’ across whole corpus. These words are Turkish translations of ‘television’, ‘telephone’ and ‘internet’ respectively. The yearly usage frequency is normalized over all words in a year.

is intended to be used by social scientists and computational linguists while conducting research on transcriptions of parliamentary sessions. We provide the corpus in digitized form as a single file which can be explored easily with fixed or custom investigative functions.

However, due to the vast amount of work required, we postponed further work such as extensive visualization of documents, extracting person names, political party memberships, mentions of geographical places or buildings and dialogues during the sessions in a structured manner. Also, we omitted the parliamentary sessions between 2015 and today as they were provided in a different format. Future parliamentary sessions will be published in this format. Thus there is work to be done for combining the current version of our corpus with this new source of parliamentary session transcriptions and automatically updating the corpus continuously. Additionally, further spelling correction techniques can be employed to increase the quality of digi-

tization.

6. Acknowledgements

This work is partly supported by the Turkish Ministry of Development under the TAM Project number DPT2007K120610.

7. Bibliographical References

Bayley, P., Bevitori, C., and Zoni, E. (2004). Threat and fear in parliamentary debates in Britain, Germany and Italy. *Cross Cultural Perspectives on Parliamentary Discourse. Amsterdam: John Benjamins*, pages 185–236.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Cheng, J. E. (2015). Islamophobia, Muslimophobia or racism? parliamentary discourses on Islam and Muslims in debates on the minaret ban in Switzerland. *Discourse & Society*, 26(5):562–586.

Fišer, Darja, L. J. (2017). Parliamentary corpora in the CLARIN infrastructure. *CLARIN Annual Conference 2017*.

Georgalidou, M. (2017). Using the Greek parliamentary speech corpus for the study of aggressive political discourse. *CLARIN-PLUS Workshop "Working with Parliamentary Records"*.

Graën, J., Batinic, D., and Volk, M. (2014). Cleaning the Europarl corpus for linguistic applications. In *KONVENS*.

Grijzenhout, S., Marx, M., and Jijkoun, V. (2014). Sentiment analysis in parliamentary proceedings. *From Text to Political Positions: Text analysis across disciplines*, 55:117.

Høyland, B., Godbout, J.-F., Lapponi, E., and Velldal, E. (2014). Predicting party affiliations from European parliament debates. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*, pages 56–60. Association for Computational Linguistics.

Mandravickaite, J. and Krilavičius, T. (2017). Stylometric analysis of parliamentary speeches: Gender dimension. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 102–107, Valencia, Spain, April. Association for Computational Linguistics.

Marx, M., Aders, N., and Schuth, A. (2010). Digital sustainable publication of legacy parliamentary proceedings. In *Proceedings of the 11th Annual International Digital Government Research Conference on Public Administration Online: Challenges and Opportunities*, dg.o '10, pages 99–104. Digital Government Society of North America.

Pančur, A. and Šorn, M. (2016). Smart big data: Use of Slovenian parliamentary papers in digital history. *Prispevki za novejšo zgodovino/Contributions to Contemporary History*, 56(3):130–146.

Rheault, L., Beelen, K., Cochrane, C., and Hirst, G. (2016). Measuring emotion in parliamentary debates with automated textual analysis. *PLOS ONE*, 11(12):1–18, 12.

Verdonik, D., Kosem, I., Vitez, A. Z., Krek, S., and Stabej, M. (2013). Compilation, transcription and usage of a reference speech corpus: the case of the Slovene corpus GOS. *Language Resources and Evaluation*, 47:1031–1048.

Vives-Cases, C. and Casado, D. L. P. (2008). Spanish politicians discourse about the responses to violence against women. *Gaceta sanitaria*, 22 5:451–6.