From Transcripts to Insights: A Digital Corpus and Interactive Speech Analysis Platform for Turkish Parliamentary Records

Başak Tepe, İrem Nur Yıldırım, Onur Güngör, Suzan Üsküdarlı
Department of Computer Engineering, Bogazici University
Istanbul, Turkey
basaktepe2020@gmail.com, yildirimiremnur42@gmail.com,
onurgu@pt.bogazici.edu.tr, suzan.uskudarli@bogazici.edu.tr

Abstract

Turkish parliamentary transcripts constitute a unique longitudinal record of the country's political, institutional, and linguistic evolution starting from 1920. Yet much of this archive has remained computationally inaccessible due to scanned and analog typewritten transcripts, historical orthography, and heterogeneous formats. We present a unified, machine-readable corpus of the Grand National Assembly of Türkiye (TBMM), comprising 26,648 session transcripts and 1.7 million pages encompassing ten diverse parliamentary entities spanning a century of legislative history. In addition, we introduce an open-access web platform for speech-level analysis of parliamentary debates from 1983 to 2024. The platform integrates named entity recognition, topic modeling, and diachronic semantic shift detection, enabling exploration of discourse patterns across time and parties, including the frequency and thematic focus of speech activities of specific Members of Parliament. By bridging the gap between raw archival scans and modern NLP tools, the dataset and platform support reproducible research in NLP, digital humanities, and computational social science.

Keywords: Turkish parliamentary corpus, corpus digitization, corpus exploration platform, semantic shift detection, topic modeling, speaker-topic attribution

1. Introduction

The Grand National Assembly of Türkiye (TBMM) has been in continuous session since 23 April 1920. Its parliamentary transcripts (Türkiye Büyük Millet Meclisi, 2024) constitute a comprehensive record of the country's major debates, political concerns, and institutional transformations. Over the course of a century, the TBMM has operated under four distinct constitutions (1921, 1924, 1961, and 1982), and most recently under the presidential system introduced by the 2017 constitutional amendments. Each of these transitions redefined the parliament's role and competences, contributing to significant institutional and textual diversity within the archive.

Parliamentary proceedings document these changes across multiple record types: member speeches, proposed bills, written and oral questions, memoranda, and plenary debates. Taken together, they capture the evolving positions of Members of Parliament (MPs, *Milletvekili*) and political parties, shaped by the political issues of their time. Despite their value for political science, history, and computational linguistics, substantial portions of this archive have remained difficult to use computationally, as records have primarily been distributed as scanned documents rather than machine-readable text.

The institutional evolution of the TBMM, spanning constitutional reforms, transitions between single-party and multi-party systems, and periods of military intervention, has produced significant discontinuities in legislative procedures, documentation practices, orthographic conventions, and archival standards. The result is a fragmented archive with heterogeneous formats, organizational structures, and metadata schemas, which has made systematic and longitudinal analysis difficult.

In this work, we present two openly available resources that address this gap. First, we release a unified, research-ready corpus of Turkish parliamentary session transcripts with standardized organization and metadata. Second, we provide a web platform for the analysis of parliamentary speeches, built on a separate processing pipeline. The platform applies NLP methods including named entity recognition, topic modeling, and semantic shift detection. Both the corpus and the platform source code are publicly available.[1] Together, they support both scholarly research and public exploration of Turkish parliamentary discourse.

The main contributions of this work are:

- A comprehensive, machine-readable corpus of Turkish parliamentary transcripts (1920–2024) with standardized metadata for sessions, legislative terms, and document types, addressing the previously fragmented nature of this historically significant archive. The corpus consists of 10 parliamentary bodies, 26,877 total sessions and 1,933,461 pages. The corpus and its processing code are openly available.

[1] Platform can be accessed at this URL. Source code is available at project repository. The corpus is publicly available in Parquet format on Hugging Face at turkish-parliamentary-corpus.

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, pages 44–55
16 May 2026.

- A web platform for the semantic analysis of Grand National Assembly of Türkiye parliamentary speeches (1983–2024), featuring MP-level topic profiling, keyword-based temporal analysis, and a speech browser. The platform uses corpus data from Term 17 onwards, featuring 1,339 total MPs, 27,662 speeches and 1,343 total sessions. The platform source code is openly available.
- Empirical analyses of Turkish parliamentary speeches, including topic modeling, semantic shift detection, and speaker-level analysis, demonstrating the utility of the released resources for computational studies of Turkish political language.

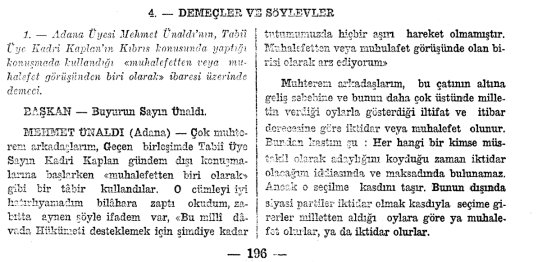
The remainder of this paper is organized as follows. Section 2 reviews related work on parliamentary corpora and computational analysis of political discourse. Section 3 describes the data collection and preprocessing steps, and the structure of the released corpus. Section 5 details the analysis platform, including MP speech extraction, the construction of a search index with enriched metadata, and the applied NLP analyses. Section 7 discusses key findings and situates them within the context of Turkish language studies, digital humanities, and NLP research, while also addressing the limitations of the presented resources.

2. Related Work

Parliamentary corpora and their computational analysis are studied internationally across many countries. Existing efforts range from large-scale comparable corpora to national archives and to tools for diachronic and semantic analysis. This section reviews these lines of work and situates the resource and platform presented in this paper.

International comparable corpora. Erjavec et al. (2025) present ParlaMint 5.0, a set of comparable corpora containing transcriptions of parliamentary debates from 29 European countries and autonomous regions. The Turkish Parliament is included from 2011 to 2022; however, the Turkish subcorpus offers neither page-level granularity nor full coverage of the TBMM archive, leaving room for a dedicated, fine-grained language resource for Turkish parliamentary research.

Turkish Parliament. For Turkish Parliament studies, the corpus of Gungor et al. (2018) is a valuable prior resource. It covers transcripts of the Grand National Assembly of Türkiye but was transcribed using the Tesseract OCR engine and was available only until 2015. It provides session-level data only, without page-level granularity, and thus



4. — DEMEÇLER VE SÖYLEVLER

1. — Adana Üyesi Mehmet Ünaldı'nın, Tabii Üye Kadri Kaplan'ın Kıbrıs konusunda yaptığı konuşmada kullandığı «muhalefetten veya muhalefet görüşünden biri olarak» ibaresi üzerinde demeci.

BAŞKAN — Buyurun Sayın Ünaldı.

MEHMET ÜNALDI (Adana) — Çok muhterem arkadaşlarım, Geçen birleşimde Tabii üye Sayın Kadri Kaplan gündem dışı konuşmalarına başlarken «muhalefetten biri olarak» gibi bir tâbir kullandılar. O cümleyi iyi hatırlıyamadım bilâhara zaptı okudum, zabıtta aynen şöyle ifadem var, «Bu millî davada Hükümeti desteklemek için şimdiye kadar tutumumuzda hiçbir aşırı hareket olmamıştır. Muhalefetten veya muhalefet görüşünde olan birisi olarak arz ediyorum»

Muhterem arkadaşlarım, bu çatının altına geliş sebebine ve bunun daha çok üstünde millette verdiği oylarla gösterdiği iltifat ve itibar derecesine göre iktidar veya muhalefet olunur. Burdan kastım şu : Her hangi bir kimse müstakil olarak adaylığını koyduğu zaman iktidar olacağını iddiasında ve maksadında bulunamaz. Ancak o seçilme kasdını taşır. Bunun dışında siyasi partiler iktidar olmak kasdıyla seçime girerler milletten aldığı oylara göre ya muhalefet olurlar, ya da iktidar olurlar.

— 196 —

Figure 1: Example of historical TBMM scan (two-column layout, low-contrast artifacts) illustrating OCR challenges for pre-1996 material.

serves as a benchmark for subsequent work. Furthermore, data before 1996 is not digitized on the official TBMM website and requires robust OCR; historical scans often exhibit two-column layout and low-contrast artifacts (see Figure 1), which complicate traditional OCR pipelines. The dataset and platform we present address these limitations by providing a unified corpus from 1920 to 2024 with page-level granularity, an OCR pipeline suited to historical scans, and an analysis platform with speech-level and MP-level access.

Other national parliamentary resources. The Italian Parliamentary Corpus (IPSA) (Frasnelli and Palmero Aprosio, 2024) shares characteristics with the Turkish setting: a long-span dataset (over 175 years), with pre-1996 material available mainly in PDF form and thus requiring OCR. The Austrian National Council is supported by tools such as those at Somes Project Team (2024), which allow tracking MP activity over time and browsing amendments, laws, resolutions, and voting records by party. That work builds on a pre-existing, fine-grained open government data (OGD) Cooperation OGD Austria (2024) corpus. Hyvönen et al. (2023) take a data-driven approach to model parliamentary work as an ontology on the Semantic Web, using the Parliament of Finland as a case study. The MP-centric design of that approach informed the structure of our own analysis platform.

Diachronic and semantic analysis. For semantic shift analysis—one of the components of our platform—the goal is to capture how the meaning of a word changes over time. Early contextual-embedding approaches such as Giulianelli et al. (2020) operationalize semantic change by comparing word representations across *two* time periods; however, such pairwise frameworks require post-hoc cluster alignment and do not scale to fine-grained, year-by-year tracking of meaning evolution. This limitation is particularly relevant in political discourse, where terminology is subject to rapid and continuous reinterpretation across legislative peri-

ods. Periti et al. (2022) address these shortcomings with WiDiD (What is Done is Done), a memory-based clustering method that incrementally accumulates word-sense clusters along a diachronic timeline without requiring inter-period alignment. Because our corpus is partitioned by year and spans over a century of parliamentary debate, WiDiD's alignment-free, scalable design is a natural fit: we adopt it in our pipeline to analyze lexical semantic change in Turkish parliamentary discourse.

Positioning of this work. Despite these advances, no single resource has so far provided a comprehensive, page-level, machine-readable corpus of TBMM transcripts from 1920 to 2024, together with a reproducible OCR pipeline and an open platform for speech-level and MP-level analysis. Our contribution fills this gap by releasing (1) a unified dataset with standardized metadata and page- and session-level granularity, including the 1920–1928 Ottoman–Modern Turkish transition period, and (2) an interactive analysis platform that applies named entity recognition, topic modeling, and semantic shift detection to parliamentary speeches, thereby supporting reproducible research in NLP, digital humanities, and computational social science.

3. Turkish Parliamentary Data

All parliamentary transcripts are publicly available online on the Grand National Assembly of Türkiye (TBMM) website (Türkiye Büyük Millet Meclisi, 2024). This website is the origin of the data.

The organizational structure of TBMM records reflects the formal structure of the Turkish legislature. A *legislative term* (*dönem*) corresponds to a parliamentary period of typically four to five years, beginning after a general election. Within each term, parliamentary activity is divided into *session years* (*yasama yılı*). Each sitting of the assembly produces a *session transcript* (*tutanak*), which is the primary documentary unit of the corpus.

Throughout the republic's history, different types of parliamentary bodies were formed (see Figure 2). All combined, they correspond to 26,648 session transcripts between 1920-2024, consisting of 1.7 million pages. There are 10 different types of parliamentary entities that contribute to this corpus.

A format discontinuity exists in the digital archive. Proceedings published before 1996 are available only as scanned, print-layout PDFs. These documents typically follow a two-column page design and contain scan artifacts, marginal noise, degraded typography, and inconsistent print quality. As they are image-based rather than digitally typed, they are not directly machine-readable and are therefore prone to OCR errors. In contrast, post-1996 proceedings are digitally born and publicly available in structured formats (HTML and Word), in addition to PDF.

To process the material, we employed a three-step pipeline (see Figure 3). First, we obtained all historical data from the Grand National Assembly of Türkiye (TBMM) website (Türkiye Büyük Millet Meclisi, 2024), preserving the original institutional ordering of parliamentary terms and sessions. PDFs were converted to page-level PNG images at 300 DPI, with a standardized directory layout to preserve traceability to source documents. Optical character recognition was performed using the DeepSeek-OCR vision–language model (Wei et al., 2025). The conversion stage consumed approximately 500 GB of RAM and required seven days of wall time; OCR inference on a single NVIDIA H100 GPU completed in nine days.

4. Data Preparation

4.1. Information Flow

The pipeline transforms raw parliamentary documents into the released dataset and analysis platform. Raw PDFs are converted to page images (PNGs), then to text via OCR. The resulting text files are made available on Hugging Face in Apache Arrow format. The same data is then fed into speech extraction pipeline, whose output is stored in a secondary search index. The indexed speeches support three downstream analyses: topic modeling, named entity extraction, and semantic shift analysis.

4.2. Optical Character Recognition

To automate the large-scale processing of Turkish parliamentary records, we developed a pipeline that orchestrates the entire data acquisition and OCR workflow on a per-term basis.

The pipeline consists of four major stages, each designed with fault tolerance and resumability:

1. *PDF Retrieval:* We automatically scrape the official TBMM web archives to identify and download 26,648 session PDFs across all legislative terms, storing valid records in structured directories.

2. *PDF to Image Conversion:* Each PDF file is converted into page-level images at 300 DPI by default, resulting in a repository of 1.7 million PNG images. This ensures a balance between visual clarity and memory usage.

3. *OCR and Corpus Formation:* The repository of images is processed using the *DeepSeek* model. This produces 1.7 million individual page text files, which are then merged back

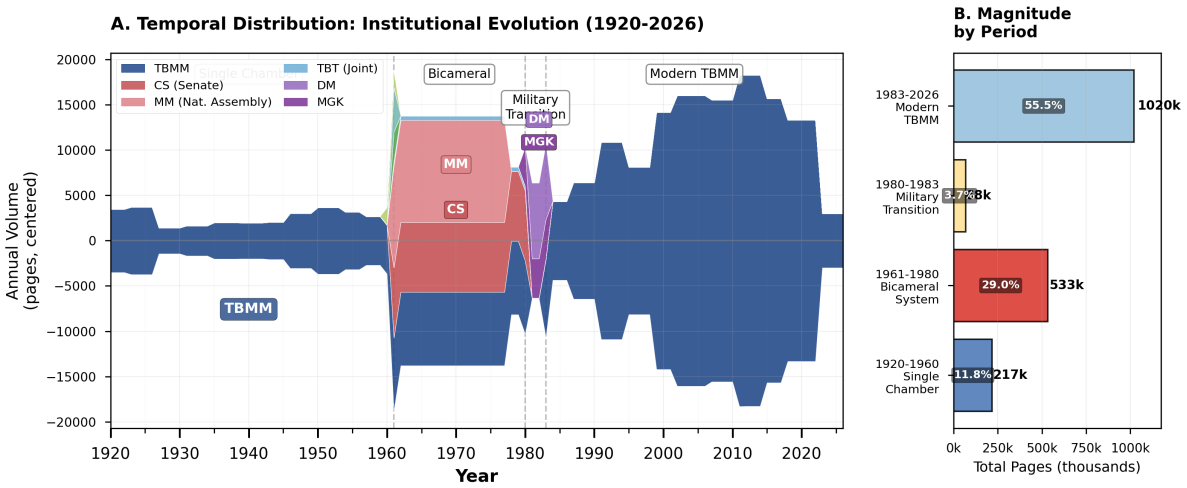


Figure 2: Corpus overview representing different parliamentary bodies' relative contribution.

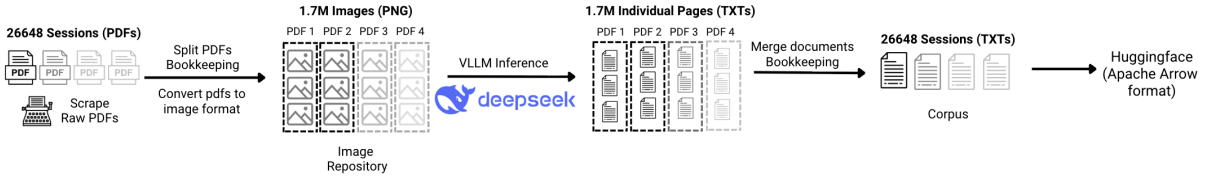


Figure 3: OCR Processing Pipeline

into session-level transcripts to form the finalized TBMM Corpus (26,648 sessions).

4. *Indexing and Storage:* The resulting transcripts and extracted metadata are indexed into *ElasticSearch*. This enables high-performance full-text search and serves as the primary data source for the analysis platform and downstream NLP tasks.

4.3. Parliamentary Speech Extraction

4.3.1. Corpus Scope and Rationale

Parliamentary speeches represent MPs' positions on legislative and societal issues. While the formed corpus spans over a century of records, we apply automated speech extraction specifically from Term 17 (1983) onward when plenary transcripts offer a consistent, comparable record of individual MP speeches. In pre-1983 structure, speeches were spontaneous and unstructured, making systematic extraction infeasible; formalized sections (e.g., "Gündem Dışı Konuşmalar" under "Başkanlığın Genel Kurula Sunuşları") emerged from Term 17, enabling rule based extraction. Many pre-1983 sittings were dominated by roll calls and failed quorums, so transcripts contain little sustained, attributable speech; much debate occurred in committees and informal settings rather than in plenary. MP-level contributions in plenary sessions were mostly motions of censure (gensoru) and formal proposals (önerge), with little direct MP–speech linkage; speeches were rare in this period. In addition, pre-1983 records are split across the National Assembly (Millet Meclisi), Joint Sessions (Birleşik Toplantı), and the Senate of the Republic (Cumhuriyet Senatosu), hindering a unified, comparable speech-level corpus.

Speeches given by MPs in plenary sessions are classified into two types: *Off-the-agenda remarks* (*gündem dışı konuşma*) are statements delivered outside the formal agenda, typically at the opening of a sitting. *Formal statements* (*açıklama*) are declarations made in response to agenda items or on behalf of a party group *(see Appendix)*. For the purposes of analysis, we define an MP's *speech* as the aggregate of their off-the-agenda remarks and formal statements within a given session.

4.4. Text Normalization Framework

4.4.1. Core Principle: Error vs. Variation

We systematically distinguish between digitization artifacts and authentic document variations. OCR-induced errors (i.e., text not present in the original documents) are corrected through normaliza-

tion. Linguistic variations present in the original transcripts are preserved to maintain fidelity to the source material. This approach ensures historical authenticity while enabling reliable text extraction.

4.4.2. OCR Error Correction

Turkish diacritic confusions: Turkish diacritic confusions frequently affect the phrase "gündem dışı" which is systematically misrecognized as gündemdeği, gündemdeş, gündemişi, or gündemiş due to character substitutions (e.g., ş→ğ, ı→e, d→i) and truncation. These variants are non-words in Turkish, do not appear in pre-digitization documents, and are not attested in manually verified original transcripts. Accordingly, all such forms are normalized to "gündem dışı" (see Table 1).

Punctuation standardization: Unicode variants such as em dash (U+2014), en dash (U+2013), and apostrophes (U+2019) are normalized to ASCII equivalents (U+002D, U+0027), as they carry no semantic distinction in the Turkish parliamentary context.

Whitespace normalization: Multiple consecutive spaces resulting from PDF extraction artifacts are collapsed into a single space.

4.4.3. Preserved Linguistic Variations

Topic prepositions: Three semantically equivalent but stylistically distinct forms introduce speech topics: *ilişkin* (formal), *hakkında* (general), and *konusunda* (common). Term 17 shows a balanced distribution (42% / 28% / 30%), while Term 22 reflects a formalization trend (58% / 30% / 12%). These variations are preserved to enable diachronic stylistic analysis.

Vowel harmony suffixes: Turkish possessive suffixes (*'nın, 'nin, 'nun, 'nün*) are determined by phonological rules. Normalization would introduce grammatical errors and is therefore avoided.

Section header formats: Historical variations such as *A)*, *A—*, and *A–)* reflect evolving style guides and are preserved for historical authenticity.

5. Analysis Platform

We developed a platform that allows researchers to browse and analyze Turkish parliamentary speeches. The platform applies a range of NLP methods, including named entity recognition, topic modeling, and semantic shift detection to the corpus. It supports browsing and searching speeches by keywords, political parties, topics, entities, speakers, and dates. It is also possible to see the thematic evolution of a given word within the parliamentary debates — how it evolved, in which different contexts it is used within each year, and when new contexts emerged or disappeared (see Table 3 and Figure 6a). Each MP has a profile page that shows a distribution of their interested topics through time. In addition, the platform profiles each member's topical emphasis relative to their political party (see Figure 6b).

Stage	Content
Input	A) GÜNDEMDEŞİ — KONUŞMALAR ... 1. - Mehmet Vedat Melik'in, Ceylanpınar Tarım İşletmeleri sınırları içerisinde yaşayan göçerlerin sorunlarına ilişkin gündemdeş konuşması...
Changes	1) GÜNDEMDEŞİ → GÜNDEM DIŞI 2) Punctuation: en/em-dash, curly quotes → ASCII (-, ') 3) Whitespace normalization (collapse doubles, fix line joins)
Preserved	Named entities and morphology kept intact: "Mehmet Vedat Melik", "Ceylanpınar", grammatical tokens like "ilişkin" and possessive "'in".

Table 1: Normalization example: preserves names and morphology while fixing OCR noise critical for section and speech detection.

5.1. Topic Modeling

The topic modeling pipeline involves:

Keyword Extraction: The Aya Expanse (Dang et al., 2024) model was used for enhanced semantic representation. Given that the model would run over more than $25{,}000$ speeches, resource and cost-efficiency was prioritized. The purpose of keyword extraction was to perform a normalization strategy before making these speeches subject to topic analysis. There is significant heterogeneity in the length and content of speeches in parliamentary texts: they range from brief speeches to multi-page discourses. There are cases where speeches are to the point, or they have rhetorical filler content. Speeches are even interrupted by automatic microphone cut-offs. By distilling these speeches into a fixed set of keywords, we achieved uniform semantic density, ensuring that the topic in each speaker's turn was fairly represented in the model regardless of the original word count or rhetorical content. The high coefficient of variation in speech length motivated the use of a fixed $k = 10$ keywords per speech.[2]

Embedding TR-MTEB turkish-embedding-model-fine-tuned Baysan and Gungor (2025) (728 dimensions) was used to embed the extracted keywords. Given the limited cardinality of the

[2]Word count statistics for statements (trimmed: bottom and top 3% excluded): $n = 25{,}611$, mean 270.9 words, standard deviation 357.5, median 140, CV 132%.

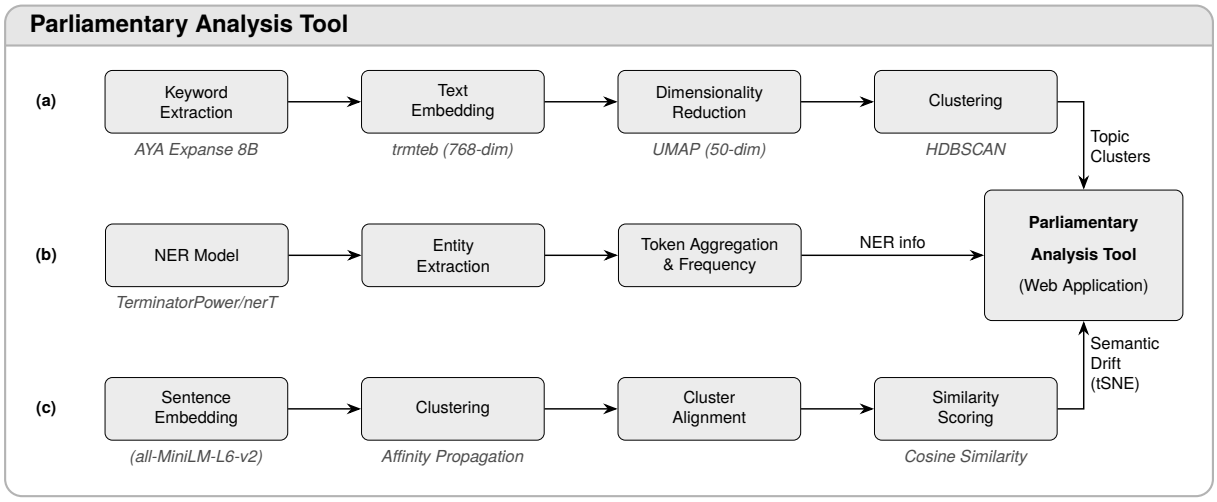


Figure 4: Analysis pipelines within the Parliamentary Analysis Tool: (a) topic modeling, (b) named entity extraction, and (c) semantic shift detection (tSNE). Process boxes list the operation name; the technology used is shown in italics below each box.

keyword set, the resulting high-dimensional representations exhibited sparsity. To mitigate this issue, Uniform Manifold Approximation and Projection (UMAP) McInnes et al. (2020) was applied to reduce the embeddings to 50 dimensions, increasing representational density while preserving the essential topological structure of the latent semantic space.

Clustering HDBSCAN was used due to its ability to perform unsupervised discovery of clusters without a pre-defined number of topics. An alternative approach would be to predefine the topics; however, this would risk missing time-specific topics related to emerging events and newly introduced terms over time (e.g., COVID-19).

Cluster refinement In this corpus, the medium is inherently political; terms such as democracy, institution names (e.g., Meclis, TBMM), ruling party abbreviations, and recurrent conflict-related vocabulary appear across a large number of speeches. Consequently, one cluster (distinct from the outlier label -1) accumulated more than one third of all speeches, with topic labels reflecting shared political rhetoric rather than substantive differences in content. For practical use of the topic assignments, a single refinement step was applied: a stop list of domain-common terms (institution and party names, frequently occurring MP names from the index) was identified and removed from the keyword strings of speeches in that cluster only. The filtered keywords were then re-embedded using the same model, and speeches were either reassigned to existing topics based on cosine similarity to topic centroids or re-clustered under the previous settings.

5.2. Named Entity Recognition

Named entity recognition (NER) was conducted using the `TerminatorPower/nerT` model (Bayraktar Ezel, 2024), targeting three entity categories: PERSON, LOCATION, and ORGANIZATION (Figure 5). As the model employs BERT-style WordPiece tokenization, subword tokens were reassembled into complete entity spans via post-processing. Resulting annotations were aligned to speech-level segments for use in downstream analyses.

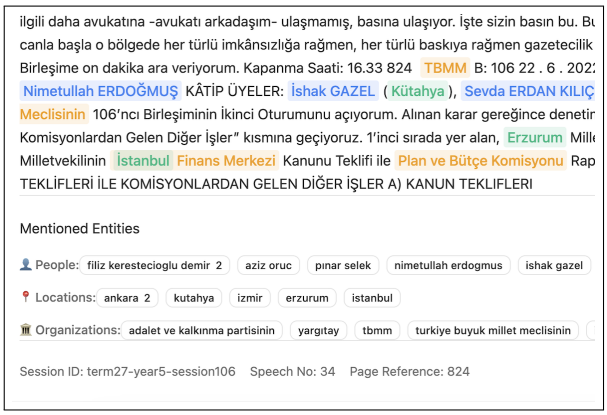


Figure 5: The persons (blue), locations (green), and organizations (yellow) in a speech.

5.3. Semantic Shift

To detect semantic shift in the parliamentary corpus, context windows of 20 tokens were extracted around each target word, capturing the local distributional context contributing to lexical meaning. Departing from the exact-match retrieval of the original WiDiD implementation (Periti et al., 2022), a

regex-based morphological expansion (root $+ \omega^*$) was employed (see Table 2), ensuring coverage of all inflected and derived forms of target keywords.

Target Word	Captured Morphological Variations
iklim *(Climate)*	iklimler (plural) iklimsel (derivational) iklimimiz (possessive) iklimin (genitive)
emekli *(Retiree)*	emeklilik (noun-forming) emeklilerimiz (plural-possessive) emekliye (dative)
döviz *(Exchange)*	dövizdeki (relational) dövizlerin (plural-genitive) dövizle (instrumental)

Table 2: Examples of Turkish morphological variations captured via root $+ \omega^*$ regex expansion, mitigating data sparsity in the embedding phase.

ID	Year	Context (Summarized)
0	1988	Migrant worker population in Council of Europe member states
0	1988	Permanent residence rights for migrant workers in Europe
0	1988	Turkish workers among 15 million migrant workers in Europe
2	1988	Discourse surrounding Palestinian refugees at the United Nations since 1940s
4	1989	Bulgarian Turks migrating to Türkiye (1949–1950)
13	2016	Refugee tragedy and migrant smuggling in Çanakkale
13	2019	Irregular migrants and Syrian refugees framed as a security issue (Aegean/Mediterranean)
13	2022	Allegations of mistreatment of irregular migrants in Greece (Lesbos)

Table 3: Topic clusters for the term *göç* across selected years, illustrating the semantic drift from labour migration discourse in the 1980s to irregular migration and border security framing in the 2010s.

The original WiDiD framework represents a target word w in corpus slice C_j via pseudo-contextual embeddings derived from Doc2Vec (Le and Mikolov, 2014), which produces static embeddings for sequences observed during training (Řehůřek and Sojka, 2010). In the present study, context windows are instead encoded using the pretrained `all-MiniLM-L6-v2` SentenceTransformer (Reimers and Gurevych, 2019), following Yin and Zhang (2024), who demonstrate its efficacy on sentence pairs that are semantically similar but structurally distinct, and vice versa. Applied without corpus-specific fine-tuning, this model yields contextualized representations that capture sense distinctions from the surrounding textual context.

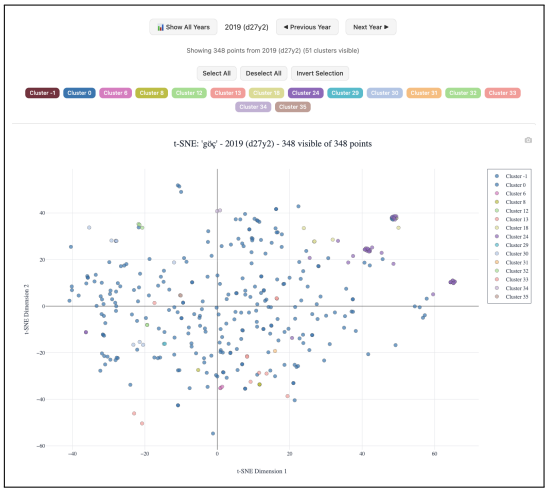
Clustering and alignment: For each term–year, affinity propagation is applied to that year's context embeddings to obtain local cluster labels, followed by a cluster-limiting step in which only the largest N clusters by size are retained and the remainder mapped to an outlier label. Centroids are computed for the retained clusters. A cluster aligner maintains a global list of centroids and assigns stable global IDs: as each new term–year is processed, its centroids are compared to the stored ones via cosine similarity; if the similarity exceeds a threshold, the cluster receives the same global ID, otherwise a new ID is created. This incremental alignment constitutes the core of WiDiD and enables consistent sense identities across time. A single t-SNE projection is then computed on the concatenated embeddings from all term–years, ensuring that spatial coordinates remain comparable across years.

5.4. Illustrative Example

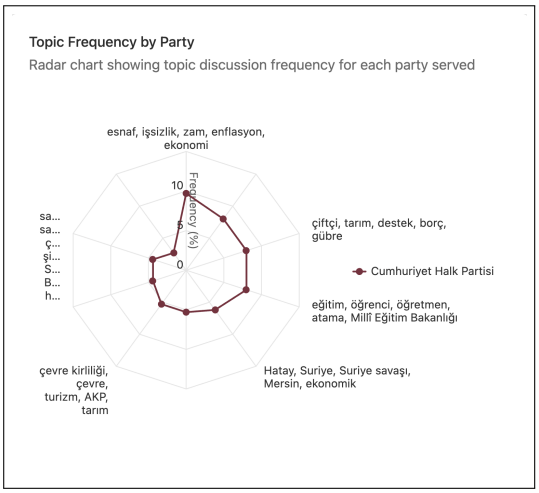
To illustrate the platform's capabilities in practice, we trace the semantic trajectory of the term *göç* (migration) across Turkish parliamentary discourse.

Figure 6(a) presents the contextual topic map associated with *göç* across time. In the late 1980s, the term is distributed across two distinct thematic clusters. Figure 6(b) presents the party-relative topical profile of a CHP member of parliament. Each axis represents a topic to which this MP makes a notable contribution within their party's discourse, with the value indicating the proportion of the party's speeches on that topic attributable to this member. The first cluster concerns labour migration, encompassing Turkish workers among the estimated 15 million migrant workers in Council of Europe member states, their rights to permanent residence, and related policy discussions (Cluster 0). The second associates *göç* with forced displacement in an international context, such as the discourse surrounding Palestinian refugees debated at the United Nations (Cluster 2). Representative contexts for each cluster and year are summarized in Table 3.

By the 2010s, the semantic environment of *göç* shifts considerably. A new cluster associated with irregular migration, refugee flows, and border security emerges (Cluster 13). Speeches in this period address the refugee tragedy in Çanakkale, migrant smuggling in the Aegean and Mediterranean, and the framing of Syrian migrants as a security concern. This transition reflects a broader reorientation of migration discourse in the Turkish parliament fol-



(a) Contextual topic map for *göç* (migration), term 27, year 2. Each point represents a context; colours denote cluster membership. The spatial layout is produced via t-SNE dimensionality reduction of the topic distributions.



(b) Party-relative topical profile of a CHP member of parliament. Each axis represents a topic to which this MP makes a notable contribution within their party's discourse.

Figure 6: Platform visualizations for member-level and semantic analysis.

lowing the onset of the Syrian migration.

The platform further enables attribution of this discourse at the member level. A pronounced contribution to the topic *Suriye* (Syria) in Figure 6(b) indicates that a given MP accounts for a disproportionate share of their party's discourse on Syrian migration, directly linking the observed semantic shift to specific parliamentary actors.

6. Conclusions

We presented a unified, machine-readable corpus of Turkish parliamentary transcripts (1920–2024) together with an interactive platform for speech-level analysis covering 1983–2024. The corpus consolidates 26,648 session transcripts and 1.7 million pages across ten parliamentary entity types into a standardized and research-ready format. The dataset will be publicly released on Hugging Face, facilitating transparent access, reuse, and integration into existing NLP workflows.

The modular structure of the data collection and indexing process makes it straightforward to incorporate newly published sessions, allowing the corpus to remain up to date as parliamentary records continue to expand. The accompanying web platform demonstrates how the resource can support topic analysis, named entity exploration, and semantic shift studies across time, parties, and individual MPs.

We see this work primarily as an enabling resource. Future directions include adding committee-level debates, proposals, bills and memoranda to broaden coverage of legislative activity and aligning the corpus with international parliamentary datasets to support comparative research. Methodologically, future developments involve building a medium-agnostic topic-modeling pipeline tuned for political discourse and refining semantic-shift clustering to automatically identify dominant discourse(s) within specific historical periods. We also aim to improve tooling for auto-updating the corpus and platform as the source data are revised. In addition, future iterations may integrate entity linking to support the construction of knowledge graphs that capture interconnections, mentions, and references among political actors, institutions, and concepts over time. We hope that this release encourages collaborative extensions and reproducible research on Turkish political language.

7. Discussion of Limitations

We acknowledge several limitations. First, the current release of the corpus extends only to 2024 and depends on the availability of data from the official parliamentary website; consequently, it is not continuously updated and requires periodic refreshes to remain current. Second, the extraction step only included MP speeches (for practical and time-related reasons); incorporating bills, committee reports, and other legislative documents would broaden coverage and analytical perspective. Third, automated speech extraction and OCR may fail on documents with exceptional formatting or substantial noise, potentially introducing gaps or transcription errors in specific sessions. Fourth, coverage for the 2011–2022 period remains imperfect and could be further improved through integration with ParlaMint Turkish corpora.

Despite these limitations, the dataset and tools presented here provide a foundation for future re-

search and have the potential to enhance public understanding and democratic accountability by making parliamentary language more accessible, searchable, and systematically analyzable.

8. Code and Data Availability

Speech Analysis Platform can be accessed at this URL. Source code is available at project repository. The corpus is publicly available in Parquet format on Hugging Face at turkish-parliamentary-corpus with different configurations for page-level and session-level transcripts as well as a tbmm-only option for eliminating other parliamentary bodies that emerges and disappears over time.

9. Acknowledgements

We are grateful to Melikşah Türker for their technical assistance and VNGRS for funding the high-performance computational resources and infrastructure required for the data processing and OCR stages of this research.

Bayraktar Ezel. 2024. Terminatorpower/nert: Turkish named entity recognition model. https://huggingface.co/TerminatorPower/nerT. Accessed: 2026-01-15.

Mehmet Selman Baysan and Tunga Gungor. 2025. TR-MTEB: A comprehensive benchmark and embedding model suite for Turkish sentence representations. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 8867–8887, Suzhou, China. Association for Computational Linguistics.

Cooperation OGD Austria. 2024. data.gv.at – The Austrian Open Government Data Portal. https://www.data.gv.at. Accessed: 2026-02-21.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier.

Tomaž Erjavec, Matyáš Kopp, Taja Kuzman Pungeršek, Nikola Ljubešić, Maciej Ogrodniczuk, Petya Osenova, Manex Agirrezabal, Tommaso Agnoloni, José Aires, Monica Albini, Jon Alkorta, Iván Antiba-Cartazo, Ekain Arrieta, Mario Barcala, Daniel Bardanca, Starkaður Barkarson, Roberto Bartolini, Roberto Battistoni, Nuria Bel, Maria del Mar Bonet Ramos, María Calzada Pérez, Aida Cardoso, Çağrı Çöltekin, Matthew Coole, Roberts Darģis, Ruben de Libano, Griet Depoorter, Sascha Diwersy, Réka Dodé, Kike Fernandez, Elisa Fernández Rei, Francesca Frontini, Marcos Garcia, Noelia García Díaz, Pedro García Louzao, Maria Gavriilidou, Dimitris Gkoumas, Ilko Grigorov, Vladislava Grigorova, Dorte Haltrup Hansen, Mikel Iruskieta, Johan Jarlbrink, Kinga Jelencsik-Mátyus, Bart Jongejan, Neeme Kahusk, Martin Kirnbauer, Anna Kryvenko, Noémi Ligeti-Nagy, Giancarlo Luxardo, Carmen Magariños, Måns Magnusson, Carlo Marchetti, Maarten Marx, Katja Meden, Amália Mendes, Michal Mochtak, Martin Mölder, Simonetta Montemagni, Costanza Navarretta, Bartłomiej Nitoń, Fredrik Mohammadi Norén, Amanda Nwadukwe, Mihael Ojsteršek, Andrej Pančur, Vassilis Papavassiliou, Rui Pereira, María Pérez Lago, Stelios Piperidis, Hannes Pirker, Marilina Pisani, Henk van der Pol, Prokopis Prokopidis, Valeria Quochi, Paul Rayson, Xosé Luís Regueira, Andriana Rii, Michał Rudolf, Manuela Ruisi, Peter Rupnik, Daniel Schopper, Kiril Simov, Laura Sinikallio, Jure Skubic, Lars Magne Tungland, Jouni Tuominen, Ruben van Heusden, Zsófia Varga, Marta Vázquez Abuín, Giulia Venturi, Adrián Vidal Miguéns, Kadri Vider, Ainhoa Vivel Couso, Adina Ioana Vladu, Tanja Wissik, Väinö Yrjänäinen, Rodolfo Zevallos, and Darja Fišer. 2025. Multilingual comparable corpora of parliamentary debates ParlaMint 5.0. Slovenian language resource repository CLARIN.SI.

Valentino Frasnelli and Alessio Palmero Aprosio. 2024. There's something new about the Italian parliament: The IPSA corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16037–16046, Torino, Italia. ELRA and ICCL.

Mario Giulianelli, Marco Del Tredici, and Raquel Fernández. 2020. Analysing lexical semantic change with contextualised word representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages

3960–3973, Online. Association for Computational Linguistics.

Eero Hyvönen, Petri Leskinen, and Jouni Tuominen. 2023. A data-driven approach to create an ontology of parliamentary work: Case parliament of finland on the semantic web. In *Proceedings of the International Workshop on Semantic Web and Ontology Design for Cultural Heritage (SWODCH'23)*, volume 3540 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

Francesco Periti, Alfio Ferrara, Stefano Montanelli, and Martin Ruskov. 2022. What is done is done: an incremental approach to semantic shift detection. In *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, pages 33–43, Dublin, Ireland. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Somes Project Team. 2024. SOMES – Social Frames: Platform for Political Transparency. https://www.netidee.at/somes. Accessed: 2026-02-21.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

Chen Yin and Zixuan Zhang. 2024. A study of sentence similarity based on the all-minilm-l6-v2 model with "same semantics, different structure" after fine tuning. In *Proceedings of the 2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)*, pages 677–684. Atlantis Press.

10. Appendix

- *off-the-agenda (gündem dışı):*
  - Terms 17–22 (1983–2007): Speeches appear as subsections under "BAŞKANLIĞIN GENEL KURULA SUNUŞLARI".
  - Terms 23–now (2007–): Speeches appear as main sections with the subsection "A) Milletvekillerinin Gündem Dışı Konuşmaları".
- *statement (açıklama):*
  - Terms 23–now (2007–): Introduction of the "Açıklamalar" (Statements) section.

The structural composition of the corpus is detailed in the following tables. Table 4 provides a breakdown of parliamentary bodies that existed and contributed to the corpus during the century, reflecting constitutional transitions such as the bicameral system and military interventions to current day unicameral system.

Table 5, on the other hand, presents a comprehensive profile of the constructed dataset's scale, categorized by legislative term and document type. This table quantifies the distribution of raw files and page counts formed by the institutions in Table 4: across the three primary data categories: indices, agendas, and full transcripts.

Comparative OCR Performance: A Sample Page Analysis

Figure 7 provides a comparative visualization of a representative archival transcript (Term 10, Year 1, Session 3, 24 May 1954) in three forms: the original document scan, the Tesseract-based OCR results from Gungor et al. (2018), and the output produced by the DeepSeek OCR pipeline. The historical scans frequently exhibit a two-column layout (see Figure 1). Although OCR engines can use layout-aware segmentation, the Tesseract-based transcripts from Gungor et al. (2018) still exhibit frequent character-level errors and fragmented line structure on such pages, as illustrated in Figure 7(b). The DeepSeek OCR pipeline recovers a cleaner transcript with consistent left-to-right reading order and structured formatting, as shown in Figure 7(c).

11. Language Resource References

Onur Gungor, Mert Tiftikci, and Çağıl Sönmez. 2018. A corpus of grand national assembly of turkish parliament's transcripts. In *Proceedings*

English Description	Turkish Name	Abbr.	Period
Grand National Assembly of Türkiye	Türkiye Büyük Millet Meclisi	TBMM	1920 – Curr.
GNAT Joint Session / Joint Sitting Sessions	TBMM Birleşik Toplantı	TBT	1961 – 1980
Senate of the Republic	Cumhuriyet Senatosu	CS	1961 – 1980
National Assembly Sessions	Millet Meclisi	MM	1961 – 1977
Secret Sessions	Gizli Celse	GC	1920 – 2011
Consultative Assembly	Danışma Meclisi	DM	1981 – 1983
National Security Council	Milli Güvenlik Konseyi	MGK	1980 – 1983
Constituent Assembly	Kurucu Meclis	KM	1961
Assembly of Representatives	Temsilciler Meclisi	TM	1961
National Unity Committee	Milli Birlik Komitesi	MBK	1960 – 1961

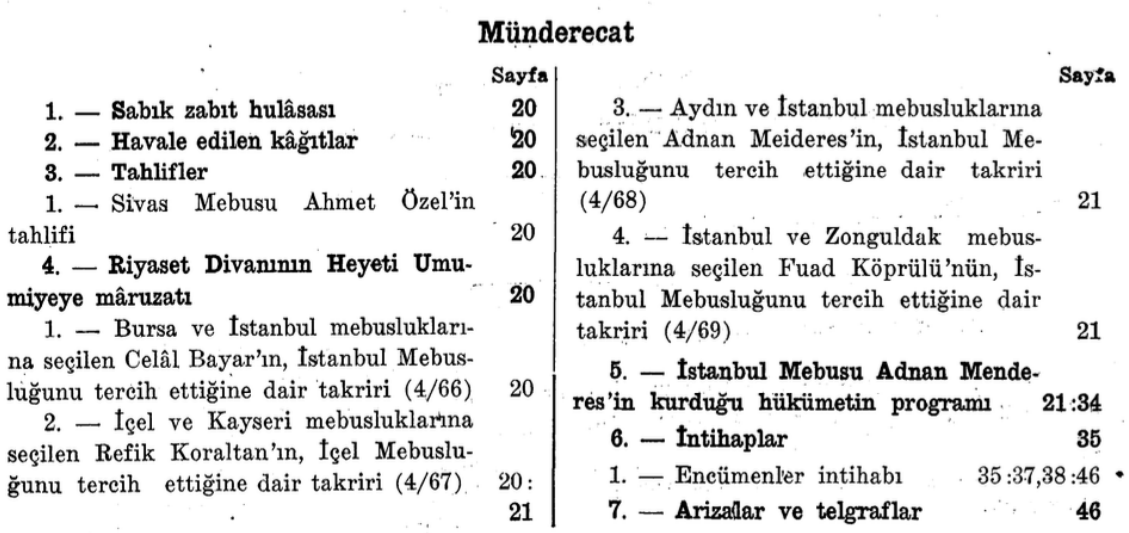
Table 4: Legislative bodies and session types with Turkish nomenclature (açılım) and corresponding date ranges. *Grand National Assembly of Türkiye.

Parliamentary Entity	Term	Index		Agenda		Transcript		Total	
		Files	Pages	Files	Pages	Files	Pages	Files	Pages
TBMM	D01	29	921	0	0	1,104	27,009	1,133	27,930
	D02	33	980	0	0	948	36,236	981	37,216
	D03	26	360	0	0	372	13,906	398	14,266
	D04	25	436	0	0	294	16,148	319	16,584
	D05	29	543	0	0	318	19,496	347	20,039
	D06	30	531	0	0	313	19,350	343	19,881
	D07	24	406	0	0	312	16,022	336	16,428
	D08	25	765	0	0	367	29,565	392	30,330
	D09	29	1,241	0	0	404	35,520	433	36,761
	D10	20	760	0	0	292	24,576	312	25,336
	D11	14	681	0	0	240	20,852	254	21,533
	D17	44	1,650	448	1,401	448	40,354	940	43,405
	D18	63	1,788	448	7,793	448	54,725	959	64,306
	D19	98	3,044	550	16,928	550	88,888	1,198	108,860
	D20	71	2,414	423	11,493	423	67,483	917	81,390
	D21	103	2,987	446	10,416	446	100,150	995	113,553
	D22	161	4,448	617	18,942	617	169,171	1,395	192,561
	D23	100	3,586	492	14,222	492	137,833	1,084	155,641
	D24	113	4,727	511	20,452	511	157,679	1,135	182,858
	D25	2	44	10	36	10	1,944	22	2,024
	D26	73	3,002	337	6,920	337	115,571	747	125,493
	D27	121	5,300	496	17,015	496	137,616	1,113	159,931
	D28	30	0	131	862	131	23,097	292	23,959
Cumhuriyet Senatosu (CS)		114	3,132	1,655	3,421	1,752	147,842	3,521	154,395
Danışma Meclisi (DM)		23	552	335	673	335	23,862	693	25,087
Gizli Celse (GC)		0	0	0	0	213	3,625	213	3,625
Kurucu Meclis (KM)		2	46	0	0	26	3,206	28	3,252
Millet Meclisi (MM)		159	6,371	2,374	18,097	2,515	167,035	5,048	191,503
Milli Birlik Komitesi (MBK)		6	222	0	0	103	3,626	109	3,848
Milli Güvenlik Konseyi (MGK)		11	372	189	220	189	16,808	389	17,400
TBMM Birleşik Toplantı (TBT)		19	370	337	593	358	8,221	714	9,184
Temsilciler Meclisi (TM)		7	170	0	0	110	4,712	117	4,882
Total		1,604	51,849	9,799	149,484	15,474	1,732,128	26,877	1,933,461

Table 5: Granular breakdown of the corpus by legislative term and document type.

of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France. European Language Resources Association (ELRA).

Türkiye Büyük Millet Meclisi. 2024. Tutanak dergisi PDFler – meclis dönemleri. Official source of parliamentary session transcripts (scanned PDFs) by legislative period. Accessed: 2026-02-23.



Münderecat

(a) Original scan

```
Münderecat
Sayfa
SayJa
1. - Sabık zabıt hulâsası
20
3. - Aydın ve istanbul mebusluklarına
2. - Havale edilen kâğıtlar
!20 seçilen Adnan Meideres'in, istanbul Me
3. - Tahlifler
20
busluğunu tercih ettiğine dair takriri
1. - Sivas Mebusu Ahmet Özel'in
(4/68)
21
tahlifi
20
4. - İstanbul ve Zonguldak mebus
4. - Riyaset Divanının Heyeti Umuluklarına seçilen Fuad Köprülü'nün, is
miyeye mâruzâtı
20
tanbul Mebusluğunu tercih ettiğine dair
```

(b) Tesseract OCR (Gungor et al., 2018)

```
Münde recat

| Sayfa | Sayfa |
|--------|--------|
| 1. - Sabık zabıt hulâsası | 20 |
| 2. - Havale edilen kâğıtlar | 20 |
| 3. - Tahlifler | 20 |
| 1. - Sivas Mebusu Ahmet Özel'in tahlifi | 20 |
| 4. - Riyaset Divanının Heyeti Umumiyeye mâruzat | 20 |
| 1. - Bursa ve İstanbul mebusluklarına seçilen Celâl Bayar'ın, İstanbul Mebusluğunu tercih ettiğine dair takriri (4/66) | 20 |
| 2. - İçel ve Kayseri mebusluklarına seçilen Refik Koraltan'ın, İçel Mebusluğunu tercih ettiğine dair takriri (4/67) | 20 |
| 1. - Enümeler intihabı | 35 |
| 2. - Arizalar ve telgraflar | 46 |

3. - Aydın ve İstanbul mebusluklarına seçilen Adnan Meideres'in, İstanbul Mebusluğunu tercih ettiğine dair takriri (4/68) | 21 |
4. - İstanbul ve Zonguldak mebusluklarına seçilen Fuad Köprülü'nün, İstanbul Mebusluğunu tercih ettiğine dair takriri (4/69) | 21 |
5. - İstanbul Mebusu Adnan Menderes'in kurduğu hükümetin programı | 21:34 |
6. - İntihaplar | 35 |
1. - Encümeler intihabı | 35:37,38:46 |
7. - Arizalar ve telgraflar | 46 |
```

(c) DeepSeek OCR pipeline

Figure 7: OCR quality comparison for a representative page (Term 10, Year 1, Session 3, 24 May 1954). Original scan (a); Tesseract-based output with recognition errors and fragmented line structure (b); DeepSeek OCR output with structured formatting (c).