The Swedish Parliamentary Motions Corpus 1867-2024

Robert Borges[*], **Fredrik Mohammadi Norén**[†], **Lotta Åberg Brorsson**[‡], **Väinö Yrjänäinen**[*], **Hanna Bäck**[§], **Robert Klemmensen**[§], **Måns Magnusson**[*]
[*]Department of Statistics, Uppsala University, Uppsala, Sweden.
{robert.borges, vaino.yrjanainen, mans.magnusson}@statistik.uu.se
[†]School of Arts and Communication, Malmö University, Malmö, Sweden. fredrik.noren@mau.se
[‡]The Riksdag Library, Stockholm, Sweden. {lotta.aberg.brorsson}@riksdagen.se
[§]Department of Political Science, Lund University, Lund, Sweden.
{hanna.back, robert.klemmensen}@svet.lu.se

Abstract

Motions submitted to the Swedish Parliament are important data for social science and humanities researchers. We introduce a new research corpus, the Swedish Parliamentary Motions Corpus, which is larger and more developed than previously available research corpora for the Swedish motions. The corpus contains annotated and structured parliamentary motions over more than 150 years, through the bicameral parliament (1867–1970) and Sweden's current unicameral parliament (1971–). Along with the corpus, we describe procedures to measure and ensure transparency around issues related to the data quality of the corpus. In addition, we link motions' authors to a rich metadata set, ensuring the corpus's utility in various research applications.

Keywords: parliamentary data, parliamentary motions, Swedish legislative process, data curation, language resources

1. Introduction

Parliamentary motions, or private member bills, are an essential part of legislative activity. They are introduced by individual MPs, providing backbenchers and opposition MPs with a formal channel to raise issues and shape debate (Bruinsma et al., 2023). Although much less likely to become law than government bills (Cavari et al., 2023), they reveal the priorities and strategies of legislators and parties, making them a valuable resource for research across corpus linguistics (Virk et al., 2024), gender studies (Gilchrist and Brooks, 2023), history (Palonen, 2016), political science (Williams and Indridason, 2018), and sociology (Popping and Wittek, 2015).

Political scientists have long examined parliamentary motions to understand party formation, polarization, and legislative signaling. A central line of research studies motion cosponsorship, i.e. endorsement of motions by multiple legislators, often using network analysis to map relationships within and between parties (see, e.g. Fowler 2006). Studies have examined whether the number of cosponsors affects success rates (Wilson and Young, 1997; Volden and Wiseman, 2014) and whether legislators from minority groups differ in cosponsorship behavior (Rocca and Sanchez, 2008). Motions thus serve as empirical material for exploring collaboration and influence in parliamentary systems.

Beyond institutional research, motions form a rich resource for computational linguistics and historical inquiry. Large-scale, metadata-rich corpora such as ParlaMint demonstrate how parliamentary texts can be interoperably annotated and compared across countries (Erjavec et al., 2022), supporting studies of ideology, framing, and linguistic change. Motion-speech pairs have been used for sentiment labeling and stance detection (the HanDeSeT dataset, Abercrombie and Batista-Navarro, 2018), and the Swedish dataset of motions (1971–2015) enables longitudinal topic modeling (Bruinsma et al., 2023). More broadly, such corpora underpin work in NLP and machine learning (Abercrombie et al., 2019; Mochtak et al., 2024; Stammbach et al., 2024), while historians use motions to trace procedural reform (Hoetink, 2024; Palonen, 2024), study policy evolution (e.g. Wiell, 2018), and examine the changing nature of parliamentary documents and actors (e.g. Hägglund, 2023; Norrbin, 2008; Stjernquist, 1996).

1.1. The Swedish Parliament Motions

The Swedish Parliament (*Riksdag*) traces its origins to the sixteenth-century Riksdag of the Estates, replaced in 1867 by a bicameral parliament with an upper (First) chamber and a lower (Second) chamber (Bengtsson et al., 1985). The bicameral system, retained through the democratic reforms and introduction of universal suffrage in the 1920s, was replaced by a unicameral parliament in 1971 (Stjernquist, 1996). The First Chamber had about 150 indirectly elected members serving eight-year terms, and the Second about 230 directly elected members serving three-year terms (Stjernquist, 1996). After 1971, MPs were elected every third year until 1994 and every fourth year thereafter (Möller,

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6772–6782
11–16 May 2026.

2015). Today, the Riksdag has 349 directly elected MPs (The Swedish Parliament, 2023).

Legislation may be introduced either through a government bill (*proposition*) or a private member's motion (*motion*). The latter, dating back to the seventeenth century, was formalized in the eighteenth century (Petersson and Wahlgren, 1999). The Riksdag Act (*Riksdagsordningen*) regulates how and when motions may be submitted. Today, MPs may submit motions during the general motions period (*allmänna motionstiden*), a short period following the government's budget bill, or in response to specific government bills (författningssamling, 2014).

There are three types of motions: independent motions (*fristående motioner*) submitted during the general period; motions responding to government bills (*följdmotioner*); and motions arising from events of major significance (*händelsemotioner* or *motioner med anledning av en händelse av större vikt*). A motion can also be categorized by the entity behind its submission: individual motions (*ensklilda motioner*) motions are submitted by one or more members of parliament from one or more parties; party motions (*partimotioner*) are submitted with backing of one or more parties; committee motions (*kommittémotioner*) are submitted by a group of MPs from one or more parties with backing of one or more committees (*utskott*) (see e.g. Riksdagen, 2012). All are referred to the relevant parliamentary committee for consideration (författningssamling, 2014).

Since 1867, about 251,000 motions have been submitted, most after the Second World War (see Figure 4). In the first 50 years of the bicameral era, around 295 motions were submitted annually; after 1971, the average rose to about 3,200 per year. Yet only 0.1–1 % of motions reach a parliamentary vote, depending on the balance of power between government and legislature (Mattson, 2016).

Despite low success rates, motions serve key political purposes. Sponsorship and co-sponsorship act as signalling mechanisms, allowing MPs and parties to communicate policy preferences to colleagues and constituents (Volden and Wiseman, 2014). Repeated submissions may help build policy traction, while motions also serve for political profiling and demonstrating activity to voters (Mattson, 2016).

The format of motions has remained relatively stable since 1867. Each includes a motion number, heading, sponsoring MP, body text, proposal points, date, and signatures (see Figure 1). During the bicameral era, it was common for one chamber to mirror the other chamber's motion through an identical submission from the same party, referenced with the phrase *lika lydande*, 'same as'.

1.2. Our Contributions

In this paper, we present the Swedish Parliamentary Motions Corpus and outline its construction, curation, and research potential. We can summarize the contributions as follows.

1. We unify and structure the Swedish parliamentary motions from 1867 to 2024 into a single corpus with a consistent data format.
2. We connect each motion with the signing members of parliament, enabling rich contextual metadata for individual motions.
3. We provide statistical estimates of important quality dimensions of the corpus, such as annotation of text containing key metadata (motion authors, titles, and dates) and optical character recognition (OCR) quality.

2. The Riksdagen Motion Corpus

The Riksdagen Motions Corpus consists of written motions from all chambers from 1867 until 2024. It can be seen as a sister corpus of the Swedish Parliament Corpus of Yrjänäinen et al. (2024), which contains records of business conducted in the main parliamentary chamber including speeches and debates, in that roughly the same period is covered and the motions corpus connects to the same member of parliament metadata to facilitate co-usability of the two corpora.

Similar to the Swedish Parliament Corpus, the motions corpus is designed as a general research infrastructure for scholars working with parliamentary proceedings data. Key characteristics include: (a) continuous growth as new parliamentary sessions are added, (b) data volumes too large for manual curation, and (c) use by a wide range of researchers with diverse questions. Hence, we also follow the same design choices as Yrjänäinen et al. (2024), i.e. preserving data authenticity, applying an iterative process for data improvement, and treating the data as a dynamic state. Descriptive statistics of the motions corpus are provided in Table 1.

Statistic	Size
Size of corpus	3.5GB
N documents	251,146
N tokens	213,801,637
N signatories	994,766

Table 1: Summary description of the Riksdagen motions corpus 1867–2024.

2.1. The Motion Corpus Content

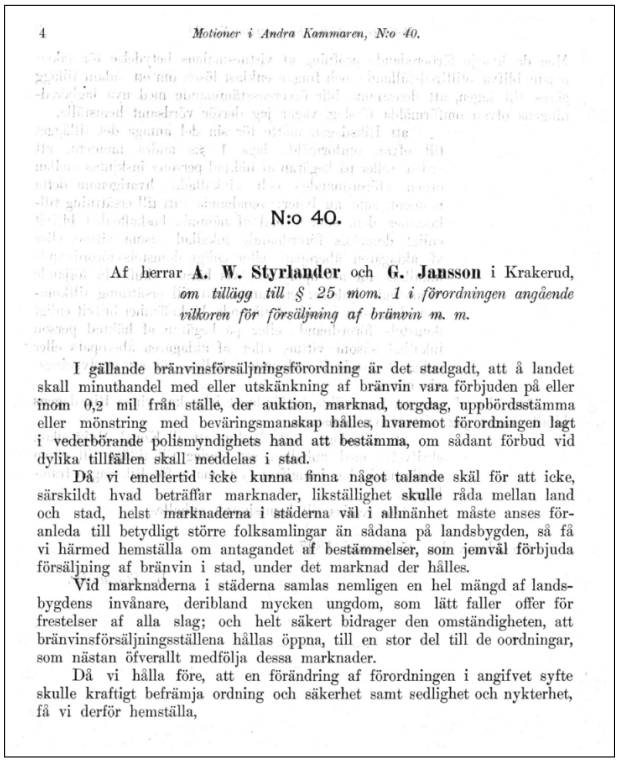
The motion corpus is encoded in a modified TEI (Text Encoding Initiative) schema designed to preserve Riksdag motions' linguistic content and for-

mal structure. Each motion is represented as structured text with four main components reflecting parliamentary drafting conventions. The submission information section contains the Riksdag document identifier, main author, and declared topic. Proposals for parliamentary decisions, expressed as operative clauses (*att-satser*) or under the heading 'Proposed decision of the Riksdag' (*Förslag till riksdagsbeslut*), form a distinct text division, positioned according to historical practice: near the end in two-chamber and early unicameral motions, but at the beginning in the later unicameral period (see Figure 1). The main body contains the substantive argumentation, while the signature block records signatories and submission date. Each signatory is linked to the Member of Parliament database in Yrjänäinen et al. (2024), enabling cross-referencing of legislators across documents and sessions. See Figure 2 for an example of the TEI structure corresponding to Figure Figure 1a.

The TEI header of each motion encodes unified metadata describing the document and its corpus relations. In Version 1.0.0, the header provides a minimal yet consistent record, including links between signatories and the MP database, the motion's title or topic, submission date, and external identifiers from the Riksdag's open data API. It also cross-references alternative digital versions, such as the Riksdag's official publication, our in-house PDF, and ALTO XML files from OCR processing. This ensures provenance and version transparency across archival environments (see Figure 3).

The header–body structure mirrors the division between metadata and historical text. The body reproduces the motion as printed, preserving layout conventions and embedded cues like submission dates or author attributions. The header provides a normalized, comprehensive metadata layer that supplements or corrects missing information. For example, modern motions omit submission dates, which are supplied via the Riksdag's API. Recent documents also include richer metadata in general, e.g. on committee referrals and administrative handling, which is completely absent from older materials. This dual design secures both historical fidelity and interoperability with current metadata standards.

Standardized encoding of text and metadata enables longitudinal analyses of parliamentary language and behavior. Because each motion is linked to metadata on authorship, topic, and date, patterns over time can be traced, such as trends in motion volume, length, cross-referencing (see Figure 4), textual complexity (subsection 4.1), or cosponsorship (subsection 4.2). Examination of such data trends is an analytically valuable basis for research, as recurrent textual patterns often expose implicit norms and changing understandings of legislative



4 *Motioner i Andra Kammaren, N:o 40.*

N:o 40.

Af herrar **A. W. Styrlander** och **G. Jansson** i Krakerud, *om tillägg till § 25 mom. 1 i förordningen angående vilkoren för försäljning af brännvin m. m.*

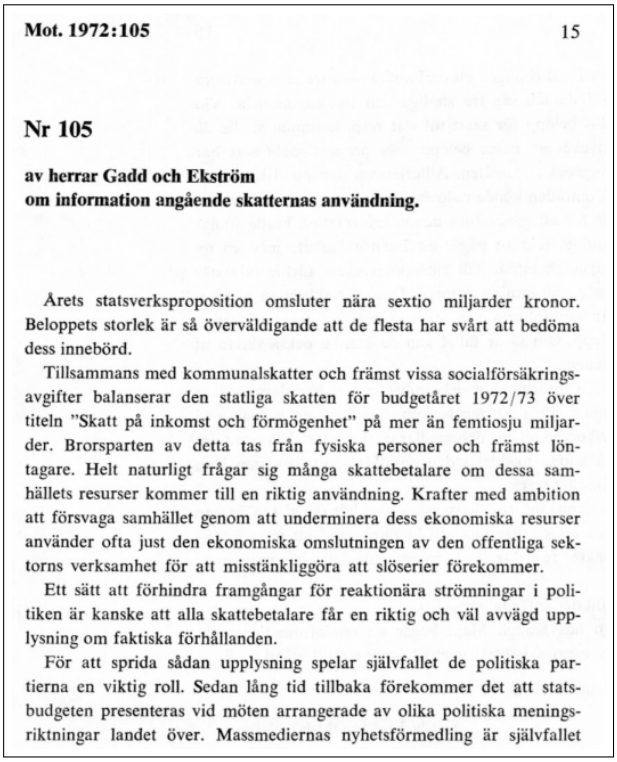
I gällande brännvinsförsäljningsförordning är det stadgadt, att å landet skall minuthandel med eller utskänkning af brännvin vara förbjuden på eller inom 0,2 mil från ställe, der auktion, marknad, torgdag, uppbördsstämma eller mönstring med beväringsmanskap hålles, hvaremot förordningen lagt i vederbörande polismyndighets hand att bestämma, om sådant förbud vid dylika tillfällen skall meddelas i stad.

Då vi emellertid icke kunna finna något talande skäl för att icke, särskildt hvad beträffar marknader, likställighet skulle råda mellan land och stad, helst marknaderna i städerna väl i allmänhet måste anses föranleda till betydligt större folksamlingar än sådana på landsbygden, så få vi härmed hemställa om antagandet af bestämmelser, som jemväl förbjuda försäljning af brännvin i stad, under det marknad der hålles.

Vid marknaderna i städerna samlas nemligen en hel mängd af landsbygdens invånare, deribland mycken ungdom, som lätt faller offer för frestelser af alla slag; och helt säkert bidrager den omständigheten, att brännvinsförsäljningsställena hållas öppna, till en stor del till de oordningar, som nästan öfverallt medfölja dessa marknader.

Då vi hålla före, att en förändring af förordningen i angifvet syfte skulle kraftigt befrämja ordning och säkerhet samt sedlighet och nykterhet, få vi derför hemställa,

(a) Motion 1897:40 (second chamber)



Mot. 1972:105 15

Nr 105

av herrar Gadd och Ekström
om information angående skatternas användning.

Årets statsverksproposition omsluter nära sextio miljarder kronor. Beloppets storlek är så överväldigande att de flesta har svårt att bedöma dess innebörd.

Tillsammans med kommunalskatter och främst vissa socialförsäkringsavgifter balanserar den statliga skatten för budgetåret 1972/73 över titeln "Skatt på inkomst och förmögenhet" på mer än femtiosju miljarder. Brorsparten av detta tas från fysiska personer och främst löntagare. Helt naturligt frågar sig många skattebetalare om dessa samhällets resurser kommer till en riktig användning. Krafter med ambition att försvaga samhället genom att underminera dess ekonomiska resurser använder ofta just den ekonomiska omslutningen av den offentliga sektorns verksamhet för att misstänkliggöra att slöserier förekommer.

Ett sätt att förhindra framgångar för reaktionära strömningar i politiken är kanske att alla skattebetalare får en riktig och väl avvägd upplysning om faktiska förhållanden.

För att sprida sådan upplysning spelar självfallet de politiska partierna en viktig roll. Sedan lång tid tillbaka förekommer det att statsbudgeten presenteras vid möten arrangerade av olika politiska meningsriktningar landet över. Massmediernas nyhetsförmedling är självfallet

(b) Motion 1972:105.

Figure 1: Example motions from parliamentary motions corpus.

```xml
<body>
  <pb facs="riksdagen-motions-pdf/data/1897/mot_1897__ak__40/
  mot_1897__ak__40-001.pdf"/>
  <fw xml:id="i-2FEFQZszu9rmRobJuWysQ7">
  4 Motioner i Andra Kammaren, N:o 40.
  </fw>
  <div xml:id="i-QHY1c76BXP3JMvKx654YhM" type="motBody">
    <div xml:id="i-K8mwrUExZaNvQ5tLT1jAwB">
      <p xml:id="i-LBhyogamvsSzBDoXn946tF">
        N:o 40.
      </p>
      <p xml:id="i-6Bw8ExcUdnQAnFvP4HUuB7" type="titleString">
        Af blertar A.: W. Styrlander och G. Jansson i Krakerud, om
        tillägg. till: $ 25 mom. 1 i förordningen angående vilkoren
        för försäljning af bränvin m. m.
      </p>
      <p xml:id="i-HQweTRhu6bBXpHs3SKCSLW">
        I gällande bränvinsförsäljningsförordning är det stadgadt,
        att å landet skall minuthandel med eller utskänkning af
        bränvin: vara förbjuden på eller inom 0,2 mil från: ställe,
        der auktion, marknad, torgdag, uppbördsstämma eller mönstring
         med beväringsmanskap hålles, hvaremot förordningen lagt i
        vederbörande polismyndighets hand att bestämma, om sådant
        förbud vid dylika tillfällen ”skall meddelas i stad.
      </p>
      ...
    </div>
  </div>
</body>
```

Figure 2: TEI excerpt for 1897:40 (for the same motion as in Figure 1a).

```xml
<teiHeader>
  <fileDesc>
    <titleStmt>[...]</titleStmt>
    <publicationStmt>[...]</publicationStmt>
    <sourceDesc>
      <bibl>
        <title>
          om information angående skatternas användning.
        </title>
        <idno type="rdwebb">FV02105</idno>
        <rs type="number">105</rs>
        <ref type="dokument_url_text">
          http://data.riksdagen.se/dokument/FV02105/text
        </ref>
        [...]
      </bibl>
    </sourceDesc>
  </fileDesc>
  <profileDesc>
    <particDesc>
      <listPerson>
        <person gender="man">
          <idno>i-AnrRW8P2FoZPwmfbQGBp8u</idno>
          <name>ARNE GADD</name>
          <state type="partyAffiliation" ref="Q105112">
            <desc>s</desc>
          </state>
        </person>
        [...]
      </listPerson>
    </particDesc>
    <correspDesc>
     <correspAction xml:id="i-5UY4AJY8Nmcxu9xJbAkSd1" type="signed"/>
      <correspContext corresp="i-5UY4AJY8Nmcxu9xJbAkSd1">
        <note type="signatory" corresp="i-AnrRW8P2FoZPwmfbQGBp8u"/>
        [...]
      </correspContext>
    </correspDesc>
    [...]
  </profileDesc>
</teiHeader>
```

Figure 3: TEI header for 1972:105 (for the same motion as in Figure 1b).

procedure. Notice, for example, how in Figure 5 motion length declines while cross-referencing increases between 1920 and 1970, peaking in the 1960s when MPs began linking broadly to related motions beyond mirrored versions across chambers.

2.2. The Creation of the Corpus

The motions corpus was constructed through two parallel data-processing pipelines reflecting parliamentary data's changing formats and availability. For the born-digital period (2006–present), data were obtained from the Riksdag's open data service. Each motion was retrieved via automated web queries, with body text in HTML and metadata in JSON. A custom Python pipeline converted this material into normalized TEI, merging metadata into the header and aligning document structure with textual content. Meaningful HTML markup was retained to guide TEI encoding: `<h1>` elements typically marked submission information, `<h2>`–`<h3>` became `<head>` elements within divisions, and headings such as *Förslag till riksdagsbeslut* delimited sections summarizing proposed actions.

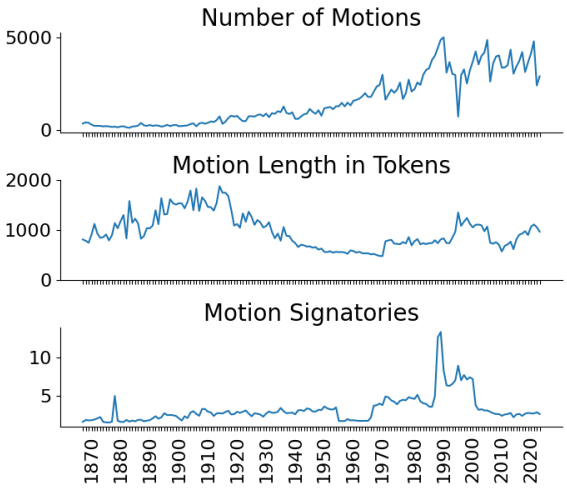


Figure 4: Motions by year: Number of motions submitted (top), average length of motions in tokens (middle), and average number of MP signatures per motion (bottom).

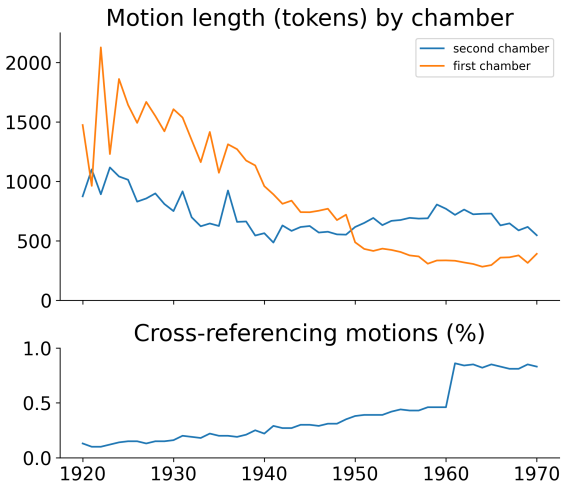


Figure 5: Average number of tokens per chamber, 1920 through 1970 (top). Shortening of motions from the 1950s is due, in part, to the increasing use of cross-references to existing motions (bottom).

For earlier motions lacking structured text, documents were obtained from the Riksdag's archive in PDF form and processed with Tesseract OCR to ALTO XML, later simplified to minimal XML capturing page, block, and line layout. Motions from the transitional years 1995/96–2005/06, mostly in legacy `.doc` format, were converted to PDF and routed through the same OCR pipeline. Each PDF page was recorded with `<pb>` elements, while printed headers and page numbers were encoded as `<fw>` to preserve original pagination.

After the creation of an initial TEI structure, all corpus documents underwent a unified post-processing phase added persistent `xml:id` attributes and launched an iterative enrichment loop. Successive beta versions incorporated structural and semantic annotations—identifying signature blocks, linking signatories to MP identifiers, and feeding extracted metadata back into the `<teiHeader>`. The full processing environment, including transformation scripts and utilities, is publicly documented and released.[1]

2.3. The Corpus Curation Process

The curation process of the motion corpus follows the same process as the Swedish Parliament corpus. For a detailed description, we refer to Yrjänäinen et al. (2024) and Yrjänäinen et al. (2025). In summary, curation consists of the following steps:

1. Small iterative improvement
   E.g., correcting identified errors in the corpus, adding new motions, more metadata, etc.
2. Quality control
   (a) Data integrity testing
   (b) Error measurement
   (c) Revision control
3. Release of a new version

Steps 2a and 2b are, in large part, automated using git and continuous integration (CI). Revision control (Step 2c) involves thorough manual evaluation of sampled changes. Recent theoretical work suggests that manual assessment of a random sample of 20 edits is sufficient to strike a balance between cost and quality control; give the sample size, if 15 edits are correct, i.e. accurately represent the true underlying documents, then the revision can be accepted with high probability of continuous improvement to the corpus (see Yrjänäinen et al., 2025, for details).

2.4. The Corpus Research Interface

A central component of the corpus is its formal interface, which defines how users interact with the data. Conceptually similar to an application programming interface (API), this "corpus API" provides a stable access layer even as the underlying content evolves. Following general software-engineering principles for long-term usability (de Souza et al., 2004), stability means avoiding disruptive changes, such as relocating files, altering formats, or modifying metadata columns, which potentially break downstream workflows. The corpus employs a semantic-style versioning system (see Preston-Werner, 2013; Yrjänäinen et al., 2024), where new major versions signal changes to the corpus API.

The repository structure supports sustainability and reproducibility (see Figure 6). The `data/` directory contains TEI-XML motion files organized by parliamentary year (e.g., `data/1867/`, `data/202324/`), mirroring the chronological organization of source material and enabling targeted analyses by session or period. Project documentation resides under `docs/`, and is automatically generated from embedded READMEs and docstrings during each release via CI. Updated documentation is published with every release to reflect the current corpus and processing pipelines.

Error measurement and validation are implemented through the `quality/` and `test/` modules, forming a transparent evaluation framework. Within `quality/`, Python scripts compute quality metrics from manually annotated samples (`quality/data/`), with sampling documentation in `quality/docs/`, and results stored under `quality/estimates/`. The `test/` directory hosts automated integrity tests, with supporting data in `test/data/`, documentation in `test/docs/`, and version-controlled results in `test/results/`. This integrated structure ensures that all stages, from data ingestion to validation and documentation, remain reproducible, traceable, and openly verifiable.

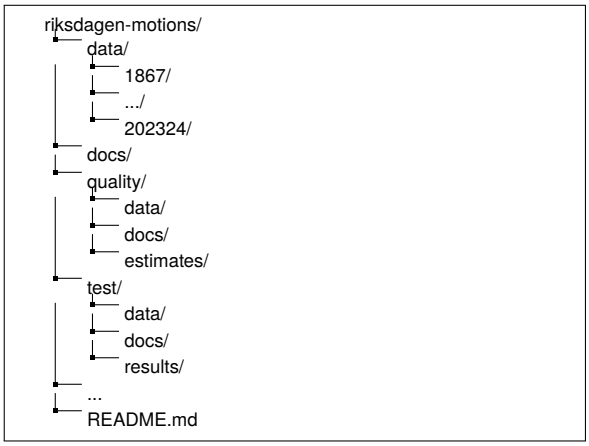


Figure 6: Directory structure of the corpus.

The dataset is publicly available in an open online

[1] https://github.com/swerik-project/scripts

repository, currently hosted on GitHub,[2] to ensure transparent access and version control. We plan to provide regular, versioned releases of the dataset, with each release archived and assigned a DOI to facilitate citation and long-term reproducibility. Regarding licensing, the exact license governing the parliamentary proceedings is currently under deliberation by the administration of the Swedish Parliament. However, the data will certainly be released under an open license once this process is finalized, ensuring that researchers can freely access and reuse the material in accordance with the applicable terms.

3. The Corpus Data Quality

Because the corpus has undergone extensive automated and manual processing, users should know its overall quality. Here, quality refers primarily to textual representation accuracy, i.e. how faithfully the digital text reproduces the original document (Bodell et al., 2022). We distinguish between data integrity testing and quality estimation (Yrjänäinen et al., 2024; Yrjänäinen et al., 2025): the former applies systematic quality checks to the corpus as a whole, while the latter measures the prevalence of representation errors against manually annotated reference samples. The results described here refer to Version 1.0.0, which already meets the reliability needs of most research applications. Ongoing curation following the process outlined in Section 2.3 aims to further improve data quality in subsequent releases.

3.1. Data-Integrity Tests

The corpus undergoes 12 distinct data integrity tests, each designed to evaluate a specific dimension of data quality. These tests encompass, among other aspects, verification that all record files conform to the TEI XML schema, detection of potential duplicate motions, and confirmation that each motion was submitted within the corresponding mandate period of the signing members of parliament. Furthermore, the tests assess the completeness of the metadata and textual annotations associated with each motion. Specifically, elements that are expected to be present in all motions are systematically verified, e.g. date of submission, the title, and the signature (see Figure 7 for the current result). The complete list of integrity tests is provided in Appendix A.

[2] https://github.com/swerik-project/riksdagen-motions

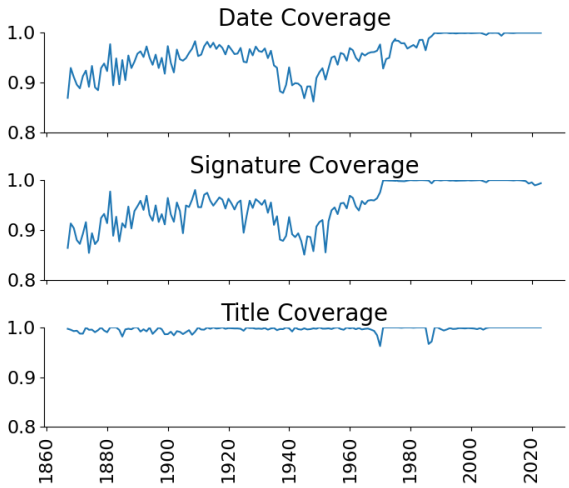


Figure 7: Data integrity testing: all motions should have key metadata annotated in the textual representation and / or listed in the teiHeader. Date, signature, and title coverage across the corpus.

3.2. Data-Quality Estimation

Data quality is evaluated against a random sample gold-standard reference set, stratified by parliamentary year and legislative chamber, which is manually annotated by domain experts, such as Swedish parliamentary historians and political scientists with relevant background. In this process, multiple dimensions of corpus quality are systematically assessed through a comparative analysis between the corpus's digital representation and the corresponding source material.

Optical character recognition (OCR) quality

Evaluating OCR accuracy is essential for corpora derived from scanned documents, as recognition errors can compromise data reliability and distort downstream analyses such as information retrieval, text mining, and linguistic annotation. To ensure the digitized text faithfully represents the original material, we systematically assess OCR quality by sampling three pages per chamber and year and manually annotating three to six lines per page. Performance is benchmarked using three complementary metrics: Levenshtein distance, word error rate (WER), and character error rate (CER). These capture character-level similarity, token-level recognition accuracy, and overall typographical precision. Together, these measures provide a robust and interpretable assessment of OCR quality across linguistic and analytical dimensions (see Figure 8).

Titles, signature blocks, and dates

Text-division quality in the corpus is assessed through a stratified sampling procedure ensuring

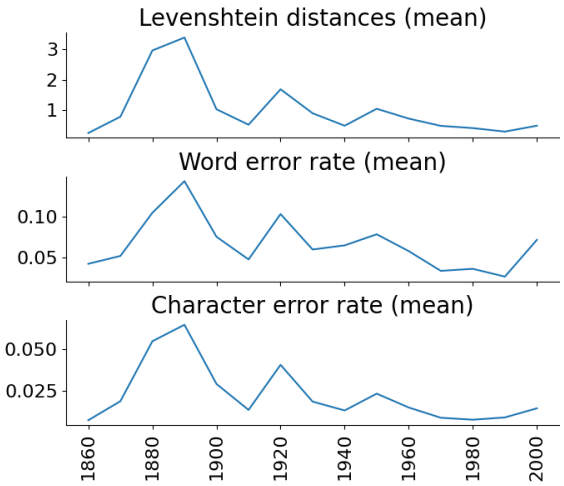


Figure 8: OCR quality metrics by decade.

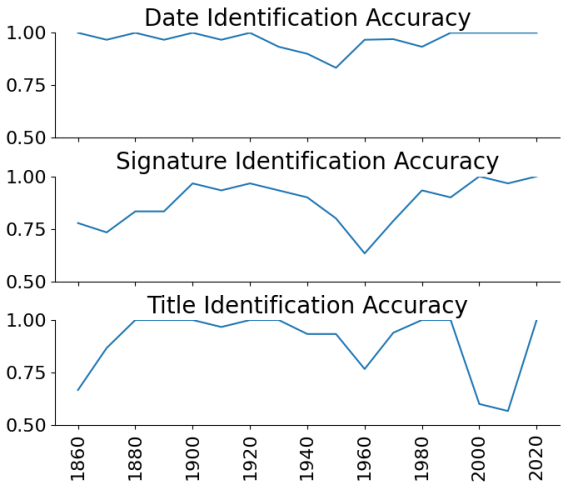


Figure 9: Estimated accuracy for identification of key textual divisions over time.

temporal representativeness. Each parliamentary year contributes a random sample of three motions per chamber in which key structural components, submission date, title, and signature block (see Figure 9), are manually annotated by domain experts. These annotations serve as a benchmark for evaluating the automated segmentation algorithms that detect corresponding divisions across the corpus. Systematic comparison of manual and automated results enables estimation of precision and overall reliability, providing both a diagnostic tool for improvement and an empirical basis for assessing markup consistency.

Overall, text-division accuracy is high. The temporary decrease in title and signature accuracy during the 1960s reflects minor formatting variations in the source material. Reduced title accuracy in the 2000s results from titles appearing only in metadata rather than in the document text. From 1970 onward, however, all titles are consistently available through metadata extracted from the Riksdag's open data service.

Connecting motions to members of parliament

Signatures in the motions are mapped to the database of members of parliament introduced by Yrjänäinen et al. (2024), enabling full interoperability with the Swedish Parliament Corpus and the complete set of metadata associated with each MP. The algorithm takes a name string extracted from the document body or `teiHeader` along with a small dictionary of metadata that can be reliably extracted from (a) the document content, including a party abbreviation, location specifier of the MP or the document date, and (b) the file name itself, including the parliament year and the chamber; name strings are matched against names in the MP database and strings matching multiple persons are disambiguated with the associated metadata. The mapping performs exceptionally well for motions from the unicameral period (see Figure 10). Despite a well-developed algorithm for linking name strings to the MP database, the bicameral period presents greater challenges, primarily due to the complexity of parsing blocks of multiple signatures (e.g. the dip in coverage in the 1940s and 50s). These difficulties arise from variation in the number of given names and surnames, inconsistent use of initials, and shifting conventions in the spatial layout and punctuation used to delimit list items.

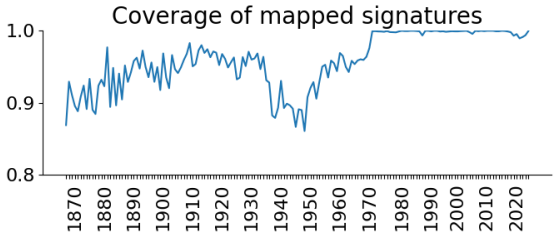


Figure 10: Coverage of signatures mapped to entries in the MP database.

4. Case Studies

We illustrate the use of the corpus with two minor use cases. The use cases are not intended to be full-fledged scientific inquiries, but instead point to the potential example use of the corpus for researchers.

4.1. Textual Complexity

To capture how Swedish texts vary in linguistic sophistication over time, six complementary met-

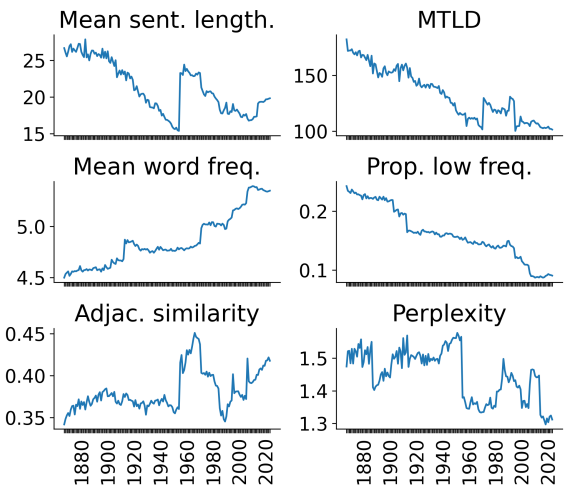


Figure 11: Complexity metrics

rics were computed that jointly describe different dimensions of textual complexity (see Figure 11). Mean sentence length reflects syntactic complexity: longer sentences typically involve more embedded clauses and coordination, offering a rough but reliable indicator of grammatical elaboration. Measure of Textual Lexical Diversity (MTLD McCarthy and Jarvis, 2010) quantifies how quickly vocabulary repeats, providing a length-independent measure of lexical variety and stylistic richness. Average word frequency and the proportion of low-frequency words gauge lexical sophistication, indicating the balance between common, easily accessible vocabulary and rarer, domain-specific terms that require more advanced linguistic competence (cf. Chen and Meurers, 2016).

To extend the analysis beyond surface form, adjacent-sentence similarity is computed using multilingual sentence embeddings, estimating how semantically cohesive a text is; high similarity suggests tight topical focus and repetition, while lower similarity implies greater conceptual progression or narrative dynamism (Foltz et al., 1998; Reimers and Gurevych, 2019). Finally, pseudo-perplexity derived from a pretrained Swedish BERT language model approximates the overall predictability of the text: lower perplexity indicates more conventional and predictable phrasing, whereas higher values suggest syntactically or semantically complex constructions (Griciūtė et al., 2022; Malmsten et al., 2020).

Together, these measures provide a multidimensional yet computationally efficient profile of textual complexity, integrating structural (morphosyntactic), lexical, and semantic perspectives without requiring manual annotation or language-specific training. From 1867 to 2024, the six complexity measures reveal distinct yet coherent patterns. Average word frequency increases steadily, while both lexical diversity (MTLD) and the proportion of low-frequency words decline, indicating a gradual shift toward more common and uniform vocabulary. Mean sentence length decreases from the late nineteenth century to the 1960s, rises sharply mid-century, and then falls again in recent decades. Adjacent-sentence similarity remains relatively stable until the 1960s, followed by alternating increases and decreases, whereas pseudo-perplexity fluctuates irregularly between 1.3 and 1.7 without a consistent trend. Overall, the results show a long-term reduction in lexical and syntactic complexity, with largely stable cohesion and linguistic predictability across the 157 years.

4.2. Motion Cosponsorship

Numerous scholars have studied bill cosponsorship patterns, primarily in the US Congress. Fowler's (2006) seminal work introduced a measure of 'connectedness' based on the frequency of cosponsorship to analyze the social distance between legislators. An important strand of research analyzes cosponsorship networks to measure political polarization, focusing on bipartisan patterns where polarization occurs when same-party members collaborate through cosponsorship while cross-party members do not (Neal, 2020).

Although most cosponsorship research focuses on the US Congress, some studies examine non-US legislatures, including, for example, Italy (Papavero and Zucchini, 2018) and the Netherlands (Wijnen and van der Brug, 2025). The most comprehensive study is the one presented by Briatte 2016, which analyzes cosponsorship by MPs within and across party lines in 19 European countries over 10–30 years, including the Swedish Riksdag (1988–2015).

We show the share of motions signed by MPs from more than one party, capturing cross-partisan cosponsorship (see Figure 12). The pattern reveals a slight downward trend, suggesting that such collaboration has become less common from the 1990s to today. This implies that MPs from different parties are increasingly unlikely to draft motions

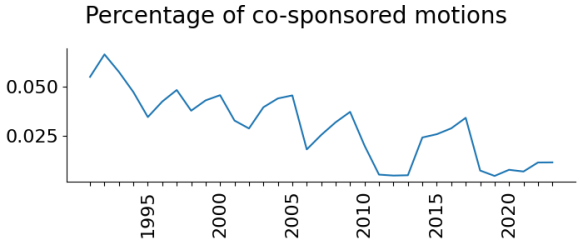


Figure 12: Percentage of co-sponsored motions with signatories from multiple parties.

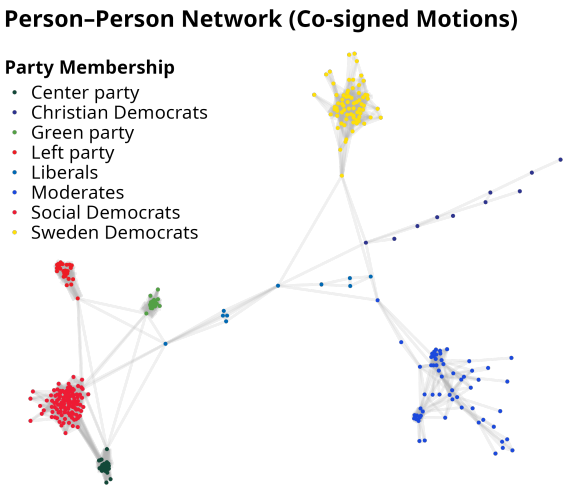


Figure 13: Person-to-person network of motion co-sponsorship in the 2023/24 parliament year.

together, possibly reflecting rising political polarization over the period.

Co-sponsorship patterns are further illustrated through a network graph of motions from the most recent parliament year (2023/24) (Figure 13). As expected, MPs within the same party most frequently cosponsor bills with one another. Collaboration also tends to occur among parties belonging to the same political bloc. In this period, the governing bloc comprised the Moderates, Christian Democrats, Liberals, and the Sweden Democrats. While at the same time, opposition parties such as the Social Democrats, Center Party, Greens, and Left Party often cosponsor with each other. Cross-bloc collaboration is rare, indicating polarization between the blocs, though a few Liberal MPs co-sponsored motions with opposition parties.

5. Conclusion

We have compiled and standardized Swedish parliamentary motions from 1867 to 2024 into a unified corpus with a consistent data format. In addition, the corpus includes metadata on each motion covering the same period. To continuously improve the resource, we apply computational methods and machine learning in an iterative process. Version control is managed according to semantic versioning principles, and each proposed revision undergoes data integrity tests, quality estimation, and manual quality control to ensure quality. Combining this corpus with the material of Yrjänäinen et al. 2024 will further enable researchers to study parliamentary processes and its language.

References

Gavin Abercrombie and Riza Batista-Navarro. 2018. Handeset: Hansard debates with sentiment tags. Mendeley Data, V2.

Gavin Abercrombie, Federico Nanni, Riza Batista-Navarro, and Simone Paolo Ponzetto. 2019. Policy preference detection in parliamentary debate motions. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 249–259, Hong Kong, China. Association for Computational Linguistics.

Ingemund Bengtsson, Herman Schück, and Nils Stjernquist, editors. 1985. *Riksdagen genom tiderna*. Sveriges riksdag, Stockholm.

Miriam Hurtado Bodell, Måns Magnusson, and Sophie Mützel. 2022. From documents to data: A framework for total corpus quality. *Socius*, 8:23780231221135523.

François Briatte. 2016. Network patterns of legislative collaboration in twenty parliaments. *Network science*, 4(2):266–271.

Rinse Bruinsma, Simon Davidsson, and Johan Lindvall. 2023. Legislative activity and political polarization in sweden, 1971–2022. *Quality & Quantity*.

A. Cavari, M. Rosenthal, and I. Shpaizman. 2023. Reevaluating the policy success of private members bills. *Research & Politics*, 10(2). Original work published 2023.

Xiaobin Chen and Detmar Meurers. 2016. Characterizing text difficulty with word frequencies. In *Proceedings of the 11th workshop on innovative use of nlp for building educational applications*, pages 84–94.

Cleidson RB de Souza, David Redmiles, Li-Te Cheng, David Millen, and John Patterson. 2004. Sometimes you need to see through walls: a field study of application programming interfaces. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 63–71.

Tomaž Erjavec, Darja Fišer, et al. 2022. The parlamint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 56(3):853–876.

Peter W. Foltz, Walter Kintsch, and Thomas K. Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2–3):285–307.

James H. Fowler. 2006. Connecting the congress: A study of cosponsorship networks. *Political analysis*, 14(4):456–487.

Svensk författningssamling. 2014. Riksdagsordningen. SFS 2014:802. Riksdagsordning.

D. J. Gilchrist and G. Brooks. 2023. The history and impact of women in the parliament of western australia: From golden age to disappointment. *Australian Feminist Studies*, 38(117):339–357.

Bernadeta Griciūtė, Marc Tanti, and Lucia Donatelli. 2022. On the cusp of comprehensibility: Can language models distinguish between metaphors and nonsense? In *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, pages 173–177, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Frederik Hoetink. 2024. Efficiency and the modernization of parliamentary procedure. *Parliaments, Estates and Representation*.

Josefin Hägglund. 2023. *Demokratins stridslinjer: Carl Lindhagen och politikens omvandling, 1896–1923*. Södertörn University, Huddinge.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of sweden – making a swedish bert.

Ingvar Mattson. 2016. Parliamentary committees: A ground for compromise and conflict. In Jon Pierre, editor, *The Oxford handbook of Swedish politics*, pages 670–690. Oxford University Press.

Philip M. McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2024. The parlasent multilingual training dataset for sentiment identification in parliamentary proceedings.

Tommy Möller. 2015. The parliamentary system. In Jon Pierre, editor, *The Oxford handbook of Swedish politics*, pages 115–129. Oxford University Press Oxford.

Zachary P. Neal. 2020. A sign of the times? weak and strong polarization in the us congress, 1973–2016. *Social networks*, 60:103–112.

Camilla Norrbin. 2008. *Från isolering till integrering: En kollektivbiografisk studie över de kvinnliga riksdagsledamöterna under tvåkammarriksdagens tid 1922–1970*. Ph.D. thesis, Umeå University, Umeå.

Kari Palonen. 2016. *The Politics of Parliamentary Procedure: The Formation of the Westminster Procedure as a Parliamentary Ideal Type*. Barbara Budrich Publishers, Opladen.

Kari Palonen. 2024. Parliamentary procedure and the conceptualization of politics. *Parliaments, Estates and Representation*.

Licia C. Papavero and Francesco Zucchini. 2018. Gender and party cohesion in the italian parliament: a spatial analysis. *Italian political science review/Rivista italiana di scienza politica*, 48(2):243–264.

Olof Petersson and Anna Wahlgren, editors. 1999. *Sveriges konstitutionella urkunder*. SNS Förlag, Stockholm. 1. uppl., serien SNS författningsprojekt.

R. Popping and R. Wittek. 2015. Success and failure of parliamentary motions: A social dilemma approach. *PLOS ONE*, 10(8):e0133510.

Tom Preston-Werner. 2013. Semantic versioning 2.0. https://semver.org/.

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.

Sverige. Riksdagen. 2012. *Betänkandehandboken: mars 2012*. Sveriges riksdag.

Michael S. Rocca and Gabriel R. Sanchez. 2008. The effect of race and ethnicity on bill sponsorship and cosponsorship in congress. *American Politics Research*, 36(1):130–152.

Dominik Stammbach, Philine Widmer, Eunjung Cho, Caglar Gulcehre, and Elliott Ash. 2024. Aligning large language models with diverse political viewpoints. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7257–7267, Miami, Florida, USA. Association for Computational Linguistics.

Nils Stjernquist. 1996. *Tvåkammartiden: Sveriges riksdag 1867–1970*. Sveriges riksdag, Stockholm.

The Swedish Parliament. 2023. Members and parties. Accessed: 2023-09-30.

Shafqat Mumtaz Virk, Claes Ohlsson, Nina Tahmasebi, Henrik Björck, and Leif Runefelt. 2024. Enhancing swedish parliamentary data: Annotation, accessibility, and application in digital humanities. In *Proceedings of the 4th International*

Conference on Natural Language Processing for Digital Humanities, pages 280–288, Miami, USA. Association for Computational Linguistics.

Craig Volden and Alan E. Wiseman. 2014. *Legislative effectiveness in the United States congress*. university press, Cambridge.

Karolina Wiell. 2018. *Bad mot lort och sjukdom: Den privathygieniska utvecklingen i Sverige 1880–1949*. Ph.D. thesis, Uppsala University, Uppsala.

Koen T. H. Wijnen and Wouter van der Brug. 2025. Populist parties and parliamentary collaboration: patterns of co-sponsorship. *Acta Politica*, 2025:1–28.

B. D. Williams and I. H. Indridason. 2018. Luck of the draw? private members' bills and the electoral connection. *Political Science Research and Methods*, 6(2):211–227.

Rick K. Wilson and Cheryl D. Young. 1997. Cosponsorship in the us congress. *Legislative Studies Quarterly*, 22(1):25–43.

Väinö Aleksi Yrjänäinen, Fredrik Mohammadi Norén, Robert Borges, Johan Jarlbrink, Lotta Åberg Brorsson, Anders P. Olsson, Pelle Snickars, and Måns Magnusson. 2024. The Swedish parliament corpus 1867 – 2022. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16100–16112, Torino, Italia. ELRA and ICCL.

Väinö Yrjänäinen, Johan Jonasson, and Måns Magnusson. 2025. Iterative data curation with theoretical guarantees.

Appendix

A. Comprehensive list of data integrity tests

For each new proposed revision to the corpus, we do multiple data integrity tests to control the quality of the corpus. These tests are continuously developed and increased as new errors are found or when new manual controls based on the sources are made. Below are the data integrity checks currently included in the corpus.

A.1. Corpus File Integrity

1. Test that there are no duplicate files in the corpus.
2. Test that no file name contains the string "None".

A.2. Document Content Integrity

1. Test that every motion has a `<teiHeader>` element.
2. Test that every motion's `<author>` element under `<titleStmt>` in the `teiHeader` is not empty.
3. Test that every motion's `<title>` element under `<titleStmt>` in the `teiHeader` is not empty.
4. Test that every motion has a `<body>` element.
5. Test that every motion's `<body>` element is not empty.
6. Test that every motion is associated with its date of submission, either in the `teiHeader` metadata or annotated in its `body`.
7. Test that every motion has a title either in the metadata in the `teiHeader` or annotated in its `body`.
8. Test that every motion is associated with at least one MP signatory.
9. Test that MPs only sign motions during their mandate period.
10. Test that XML in all motion files is schema compliant