The Swedish Parliament Corpus 1867–2022

Väinö Yrjänäinen*, Fredrik Mohammadi Norén†, Robert Borges*, Johan Jarlbrink⋆, Lotta Åberg Brorsson‡, Anders P. Olsson‡, Pelle Snickars✶, Måns Magnusson*

*Department of Statistics, Uppsala University, Uppsala, Sweden.
{vaino.yrjanainen, robert.borges, mans.magnusson}@statistik.uu.se
†School of Arts and Communication, Malmö University, Malmö, Sweden. fredrik.noren@mau.se
⋆Department of Culture and Media Studies, Umeå University, Umeå, Sweden. johan.jarlbrink@umu.se
‡The Riksdag Library, Stockholm, Sweden. {lotta.aberg.brorsson, anders.p.olsson}@riksdagen.se
✶Department of Arts and Cultural Sciences, Lund University, Lund, Sweden. pelle.snickars@kultur.lu.se

Abstract

The Swedish parliamentary records are an important source material for social science and humanities researchers. We introduce a new research corpus, *the Swedish Parliament Corpus*, which is larger and more developed than previously available research corpora for the Swedish parliament. The corpus contains annotated and structured parliamentary records over more than 150 years, through the bicameral parliament (1867–1970) and the unicameral parliament (1971–). In addition to the records, which contain all speeches in the parliament, we also provide a database of all members of parliament over the same period. Along with the corpus, we describe procedures to ensure data quality. The corpus facilitates detailed analysis of parliamentary speeches in several research fields.

Keywords: parliamentary data, Sweden, political debate, language resource, politics, data curation

1. Introduction

Parliamentary debates are of great interest to a broad range of academic disciplines. Especially in the humanities and social sciences, the speeches are used as source material for qualitative and quantitative studies in fields such as history (Guldi, 2019; Hägglund, 2023; Jarlbrink and Norén, 2023), sociology (Dolan, 2009; Skubic and Fišer, 2022), linguistics (Gast and Borges, 2023; Korhonen et al., 2023) and political science (Monroe et al., 2008; Perren and Sapsed, 2013; Onursal and Kirkpatrick, 2021).

Thanks to technological advancements and the status of the parliament as a public institution, numerous national parliamentary corpora have been made available in a digital format in recent years. Examples include the Norwegian parliamentary corpus (Lapponi et al., 2018), the Finnish parliamentary corpus (Sinikallio et al., 2021), the UK-Parl (Nanni et al., 2018), the Hansard (Nanni et al., 2019) and the Slovenian parliamentary corpus (Pančur et al., 2018), to name a few recent contributions. Besides the research corpora originating from individual parliaments, collaborative international initiatives such as Europarl (Koehn, 2005) and ParlaMint (Erjavec and Pančur, 2021) have contributed to the creation of additional multilingual and comparable parliamentary corpora, further enabling comparative analysis of parliamentary legislative processes.

Swedish parliamentary data has been made available several times in different ways but never comprehensively regarding annotation and metadata structure for longer periods. For example, Rødven-Eide (2020) developed a small parliamentary corpus containing speeches from 1993 with annotations based on data sourced from the Swedish Parliament's open data platform. A subset of the Swedish parliamentary data covering 2017 and 2022 was also published in the recent 3.0 release of the ParlaMint corpus, along with 26 other languages and 24 other parliaments (Erjavec et al., 2023). Moreover, the Swedish parliament has released different versions of the digitized records back to 1521 (The Swedish National Library, 2023). The data is unstructured for 1867–1993 (The Swedish Parliament, 2023a) and structured for 1993 onwards through the Swedish Parliament's open data platform (The Swedish Parliament, 2023c).

Hence, a full, structured, and unified Swedish parliamentary corpus containing speeches and members of parliament does not exist beyond 1993–2022. For the first time, this paper presents such an annotated parliamentary speech corpus from 1867 to 2022, together with data on the members of parliament (MP) during the same period.

1.1. The Swedish Parliament

With roots in the fifteenth century, the Swedish Riksdag of the Estates (represented by nobility, clergy, burghers, and peasants) was dismantled and replaced by a bicameral parliament with a new representation system in 1867 (Bengtsson et al., 1985). Modelled after other European parliaments, the Swedish parliament now had an upper house (First Chamber) and a lower house chamber (Second Chamber). While the bicameral system sur-

LREC-COLING 2024, pages 16100–16112

vived the democratic breakthrough in the 1920s, it was eventually deemed old-fashioned, and the present unicameral Riksdag was established in 1971 (Stjernquist, 1996).

In the bicameral parliament, the First Chamber had about 150 members on eight-year mandates elected indirectly by local and regional political representatives. The Second Chamber had about 230 members on a three-year mandate, elected directly by citizens entitled to vote (Stjernquist, 1996). After the unicameral reform, which meant that the First Chamber was removed, MPs were elected directly every third year until 1994 and every fourth year since then (Möller, 2015). Today, the Swedish Riksdag has 349 members of parliament (The Swedish Parliament, 2023b). In Sweden, women have been allowed to vote and to be elected as MPs since 1921, but only accounted for about 15 percent in the early 1970s, and then slowly increased to almost 50 percent (see Figure 1 and Freidenvall, 2006).

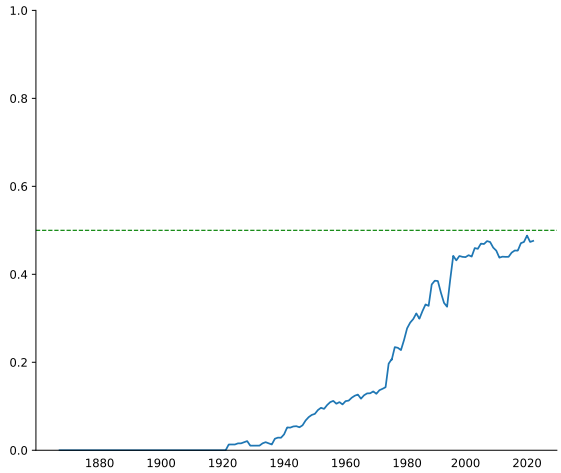


Figure 1: Proportion of female members of parliament over time.

The modern, organized parties started to develop in the late nineteenth century. Before, a political party had a more fluid form (Carlsson, 1992). Between the 1920s and the 1980s, five parties were (and still are) primarily represented in the Swedish parliament: (1) the Social Democratic Party, a social reform-oriented left-wing party often regarded as synonymous with the Swedish welfare state, and has been a dominating power in Swedish politics since the 1920s, (2) the Moderate Party, a conservative and market liberal-oriented party, which has been the main competitor to the Social Democrats since the 1980s, (3) the Centre Party, traditionally an agrarian-oriented party developing liberal liberal tendencies over time (4) the Liberals, a liberal-oriented party but, as all parliamentary parties, has accepted the general principles of the welfare state, and (5) the Left Party, formerly a communist party has reformed into a more traditional European left-wing party. Since the 1980s, a few more parties have entered the parliament that are still represented today: (a) the Christian Democratic Party, a social-conservative party originating from the free church movement, (b) the Green Party, an environmentalist party that entered the parliament in the aftermath of the referendum on nuclear energy and in a context of public concerns about pollution, and (c) the Sweden Democrats, a nationalistic, social-conservative, and right-wing populist party that has positioned itself with an anti-immigrant policy (Aylott, Nicholas, 2016).

1.2. The Swedish Parliament Records

One of the fundamental documents of the Swedish parliament is its records (*Riksdagsprotokoll*), sometimes called proceedings. These parliamentary documents contain all speeches held in the chamber during a parliamentary session (riksmöte) and accounts of various activities in the chamber. From 1867 to the early 1940s, a parliamentary session lasted from the beginning of the year to the summer with a break during the autumn. Due to the Second World War, there was a need to change to a full-year parliamentary session, which is still implemented, today lasting from about September to June. Throughout the whole time period, the parliament convenes a couple of days a week for debates and voting.

There are three types of sub-records: the body record (*stomprotokollet*), the debate record (*debattprotokollet*), and the preliminary record (*snabbprotokollet*). The body record contains descriptions of agendas, vote results, decisions, written questions, etc., and sometimes includes supplemented documents such as committee reports, proposals, and motions. All speeches held by MPs are written down and constitute the debate record (Nygren, 1985). The printed speeches, however, are rarely a verbatim rendering of what was spoken in the chamber. Spoken language is often stylistically polished and meanings interpreted (Harvard, 2011).

Typically, the body and debate records are merged in the order in which events and speeches occurred. These two records constitute the later printed parliamentary record (see Figure 2, for examples of records). The preliminary records only provide an early version of the later printed records and are not included in *the Swedish Parliament Corpus* (Nygren, 1985). Hence, *the Swedish Parliament Corpus* comprises the printed parliamentary records, and not other separated parliamentary document types – government bills (proposi-

tioner), private member's motions (motioner), committee reports (utskottsbetänkanden), et cetera.

1.3. Our contributions

In this paper, we introduce the *Swedish Parliament Corpus*, based on the Swedish parliamentary records from 1867 until today, together with data on all MPs during the same period. We can summarize our contributions in the following way:

1. We unify and structure the Swedish parliamentary records between 1867 and 2022 under one corpus with one data format, along with Python and R modules to easily access the data.
2. We use computational and machine-learning approaches to segment, correct, enrich and annotate the data.
3. We gather and curate metadata on MPs, ministers, speakers of the Riksdag, and governments from multiple sources into a new and structured metadata database.
4. We present a rigorous process to revise the corpus and improve it iteratively over time.
5. We provide statistical estimates of multiple quality dimensions of the corpus.

2. The Swedish Parliament Corpus

The Swedish Parliament Corpus consists of two major parts, (1) the parliamentary records and (2) data on the members of parliament. The corpus consists of the records from all chambers from 1867 until 2022. However, we only focus on the printed records; hence, the preliminary records are not included. In addition, the corpus contains the names of all MPs during this period and additional metadata on each MP.

The corpus is intended to serve as a general research infrastructure for scholars interested in Swedish parliamentary research. Some important properties of the corpus are that (a) it will continue to grow as new parliamentary sessions are held, (b) the data is too large to curate manually, and (c) a large number of different researchers with different research questions will use the material. This has led to three main design decisions to maximize the useability of the corpus. These are summarised as follows.

Preserving Data Authenticity

Since many different researchers are likely use the corpus, we prioritise the structuring of the corpus with minimal analytical interference. We aim to maintain the data as a digital representation of the original documents or what (Hurtado Bodell et al., 2022) refers to as minimizing the "text representation error." Further, the data on MPs tries to represent the information in established biographies as authentically as possible (Norberg et al., 1985; Stjernquist, 1996). As a result, we aim to only store information such as the party affiliation of the parliamentarians by their name at a specific time. Consequently, in cases where a political party undergoes a name change, as has occurred on multiple occasions, we consider the MP to be connected to both these parties during different periods.

Data Curation Through an Iterative Approach

Manual data curation is simply unfeasible, given the size of the corpus. Instead, we adopt a methodology inspired by Voormann and Gut (2008), treating the curation process as an ongoing iterative endeavour. In each iteration, we measure the corpus quality, address issues in a scalable manner and releasing new versions (see Section 3 for details). Furthermore, since the corpus grows with each new parliamentary year, it lacks a definitive endpoint for curation. Consequently, we continually release updated versions of the corpus, addressing identified errors and introducing new data. This iterative process also helps streamline the prioritization of the curation efforts. In each iteration, we focus on the most pressing issue, aligning with the overarching objectives of the corpus.

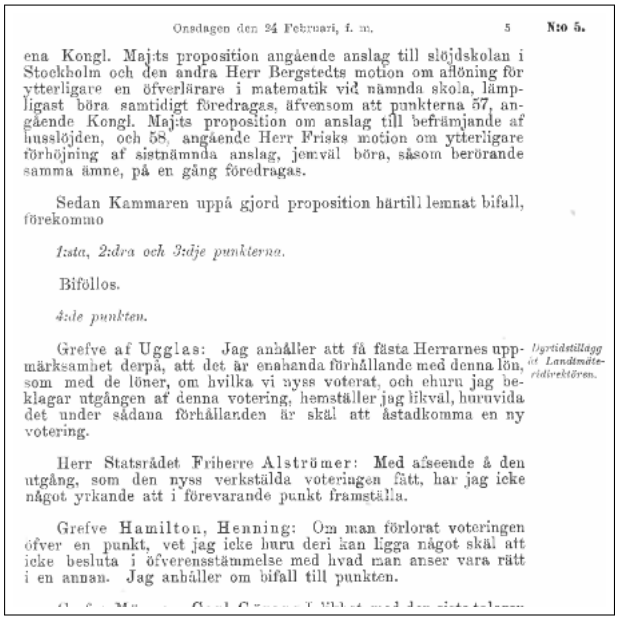
The Data as the State

For most researchers, the data is the main focus and interest, rather than the code processing it. Hence, we version control the data. This shift in perspective allows us to leverage concepts from software engineering to streamline the curation process and simplify the use by researchers and other end users.

2.1. The Corpus Research Interface

A crucial part of the corpus is the formal interface, which defines how the users interact with the corpus. Conceptually, it is similar to an application programming interface (API); thus, we refer to it as the API of the corpus.

Even though the content of the corpus may undergo modifications through curation and expansion, our objective is to maintain the stability of the API as much as possible. This is inspired by APIs in software engineering where stability is an important feature due to similar reasons (de Souza et al.,



Onsdagen den 24 Februari, f. m. 5 N:o 5.

ena Kongl. Maj:ts proposition angående anslag till slöjdskolan i Stockholm och den andra Herr Bergstedts motion om aflöning för ytterligare en öfverlärare i matematik vid nämnda skola, lämpligast böra samtidigt föredragas, äfvensom att punkterna 57, angående Kongl. Maj:ts proposition om anslag till befrämjande af hushållsskolorna, och 58, angående Herr Friska motion om ytterligare förhöjning af sistnämnda anslag, jemväl böra, såsom berörande samma ämne, på en gång föredragas.

Sedan Kammaren uppå gjord proposition härtill lemnat bifall, förekommo

1:sta, 2:dra och 3:dje punkterna.

Bifölles.

4:de punkten.

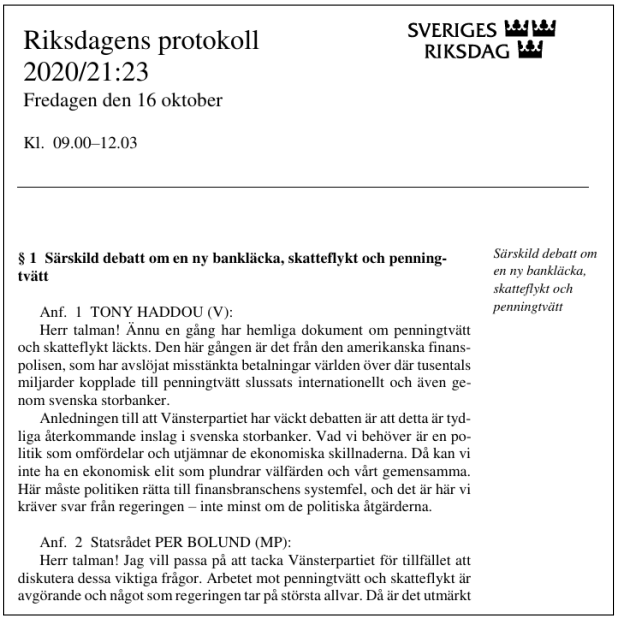
Grefve af Ugglas: Jag anhåller att få fästa Herrarnes uppmärksamhet derpå, att det är enahanda förhållande med denna lön, som med de löner, om hvilka vi nyss voterat, och ehuru jag beklagar utgången af denna votering, hemställer jag likväl, huruvida det under sådana förhållanden är skäl att åstadkomma en ny votering.

Lyrtidstillägg åt Landtmäteriöfverstyrelsen.

Herr Statsrådet Friherre Alströmer: Med afseende å den utgång, som den nyss verkställda voteringen fått, har jag icke något yrkande att i förevarande punkt framställa.

Grefve Hamilton, Henning: Om man förlorat voteringen öfver en punkt, vet jag icke huru deri kan ligga något skäl att icke besluta i öfverensstämmelse med hvad man anser vara rätt i en annan. Jag anhåller om bifall till punkten.

(a) 5th record of the First Chamber's 1875 meeting.



Riksdagens protokoll
2020/21:23
Fredagen den 16 oktober

Kl. 09.00–12.03

SVERIGES RIKSDAG

§ 1 Särskild debatt om en ny bankläcka, skatteflykt och penningtvätt

Särskild debatt om en ny bankläcka, skatteflykt och penningtvätt

Anf. 1 TONY HADDOU (V):
Herr talman! Ännu en gång har hemliga dokument om penningtvätt och skatteflykt läckts. Den här gången är det från den amerikanska finanspolisen, som har avslöjat misstänkta betalningar världen över där tusentals miljarder kopplade till penningtvätt slussats internationellt och även genom svenska storbanker.
Anledningen till att Vänsterpartiet har väckt debatten är att detta är tydliga återkommande inslag i svenska storbanker. Vad vi behöver är en politik som omfördelar och utjämnar de ekonomiska skillnaderna. Då kan vi inte ha en ekonomisk elit som plundrar välfärden och vårt gemensamma. Här måste politiken rätta till finansbranschens systemfel, och det är här vi kräver svar från regeringen – inte minst om de politiska åtgärderna.

Anf. 2 Statsrådet PER BOLUND (MP):
Herr talman! Jag vill passa på att tacka Vänsterpartiet för tillfället att diskutera dessa viktiga frågor. Arbetet mot penningtvätt och skatteflykt är avgörande och något som regeringen tar på största allvar. Då är det utmärkt

(b) 23rd record of the Parliament's 2020/21 meeting.

Figure 2: Example pages from parliamentary records.

2004). Here, stability entails refraining from relocating files or folders, altering formats, changing metadata file columns, or implementing other modifications that might disrupt downstream code and usage.

The API of the Swedish parliamentary corpus comprises the following key elements:

- Records in `data/MEETING` folders (`MEETING` is eg. `1955` or `199798`): contains the parliamentary records as ParlaClarin XML files, stored in folders by the parliamentary year.
- Metadata `data/METADATA.csv` files (`METADATA` is eg. `name` or `government`): The data on members of parliament, speakers of the Riksdag and similar additional metadata is stored here as CSV files.
- Software modules in Python (pyriksdagen) and R (riksdagenr): Software that has been developed to facilitate the work with the corpus further, such as functions for data aggregating, processing and filtering.

The first two are delivered as releases on GitHub[1], while the software modules are available as on PyPi[2] and a GitHub R module[3].

[1] https://github.com/swerik-project/the-swedish-parliament-corpus
[2] https://pypi.org/project/pyriksdagen/
[3] https://github.com/swerik-project/rcr

2.2. The Parliament Records

In the Swedish parliamentary corpus, the speech records are delivered in the ParlaClarin format (Erjavec and Pančur, 2021), which itself is a further specification of the Text Encoding Initiative (TEI) XML format (TEI Consortium, 2007). ParlaClarin is specifically designed for storing parliamentary proceedings.

```
<note xml:id="i-5PtDcaRhvPaqqqaYZfLQYh">
  1:sta, 2:dra och 3:dje punkterna. Biföllos. 4:de punkten.
</note>
<note xml:id="i-6nrbDAuUzk4KNf4ATaAUXm" type="speaker">
  Grefve af Ugglas:
</note>
<u who="unknown" xml:id="i-2L2qwC6so3Hnh6ytNeEE6j">
  <seg xml:id="i-DoRt9GqcPDXbBCQQC6ZhY1">
    Jag anhåller att få fästa Herrarnes uppmärksamhet derpå, att
    det är enahanda förhållande med denna lön, som med de löner,
    om hvilka vi uyss voterat, och ehuru jag beklagar utgången af
    denna votering, hemställer jag likväl, huruvida det under
    sådana förhållanden är skäl att åstadkomma en ny votering.
  </seg>
</u>
```

```
<note xml:id="i-Y32cpnd9wZGz3GqEmMMMms">
  § 1 Särskild debatt om en ny bankläcka, skatteflykt och
      penningtvätt
</note>
<note type="speaker" xml:id="i-125vTa2ZcjDq4jRpcRgdGQ">
  Anf. 1 TONY HADDOU (V):
</note>
<u xml:id="i-daaf4bed8a426df1-0" next="i-daaf4bed8a426df1-1" who="
    i-QhdbBfCQXCqPEr4GTwvK8F">
  <seg xml:id="i-LWtFhTyUMfy43ay4VnAmrE">
    Herr talman! Ännu en gång har hemliga dokument om penningtvätt
    och skatteflykt läckts. Den här gången är det från den
    amerikanska finanspolisen, som har avslöjat misstänkta
    betalningar världen över där tusentals miljarder kopplade
    till penningtvätt slussats internationellt och även genom
    svenska storbanker.
  </seg>
</u>
```

Figure 3: Excerpts of the example records presented in Figure 2 in the ParlaClarin format. 5th record of the First Chamber's 1875 meeting above, 23rd record of the Parliament's 2020/21 meeting below.

The ParlaClarin format starts with record-level metadata. In the Swedish parliamentary corpus, it includes the record identifier , the date or dates of the protocol, licensing information and other technical metadata. This is illustrated in Figure 4.

After the metadata, the document body text starts. It includes a digital version of the original document. All text included in the source material has been retained in the ParlaClarin files: the page body, margins, page number and other visible elements. Segmentation into paragraphs follows the segmentation in the source material, whether digitized or born digital. Page breaks are represented as empty page-break elements (`<pb>`) with a "facs" attribute referring to the original PDF page, as seen in Figure 5 on line 234.

```
5  <titleStmt>
6    <title>Riksdagen records</title>
7  </titleStmt>
8  <publicationStmt>
9    <authority>SWERIK project</authority>
10   <availability>
11    <licence target="https://creativecommons.org/licenses/by/4.0/">
12       Licence: Attribution 4.0 International (CC BY 4.0)
13    </licence>
14   </availability>
15 </publicationStmt>
16 <!-- rest of the metadata -->
```

Figure 4: Metadata in the ParlaClarin XML files. Excerpt of the record 1989/90:11.

Line breaks, text coordinates or typesetting marks are not retained, and dashes are merged. All elements that contain text have a unique identifier stored in the XML `id` attribute.

Transcriptions of speeches in the parliament are a large part of the parliamentary records. They are programmatically detected and tagged as utterances (`<u>`) per the ParlaClarin format. The implementation of this annotation is further explained in Section 3.1. Due to speeches often consisting of multiple paragraphs, all utterances include paragraph segmentation (`<seg>`). Speeches that span multiple pages are connected with the `prev` and `next` attributes in the `<u>` elements. Figure 5 exemplifies the encoding of transcriptions from line 226 to line 233.

Speaker introductions in the records (`<note type="speaker">`) precede transcriptions of speech, as illustrated in Figure 5 on lines 223-225. They are also automatically detected and classified. Using them, speeches are mapped onto person identifiers in the metadata catalogue, further explained in Section 3.1. The speaker is an attribute of utterance elements (`<u who="IDENTIFIER">`), while non-detected utterances have the attribute `who="unknown"`. Other paragraph classifications include margin and transcriber notes (`<note>`) and date notes (`<note type="date">`).

In total, version 1.0.0 of the corpus is made up

```
220 <note xml:id="i-H6hRQtfXpp7iKcvxzoxoE1">
221   3§ Svar på frågorna 1989/90:15 och 72 om koldioxidutsläppen
222 </note>
223 <note type="speaker" xml:id="i-By3GdnFK6aPbETpfJBLVL7">
224   Miljö och energiminister BIRGITTA DAHL:
225 </note>
226 <u xml:id="i-073a81b8bcb98a8d-5" next="i-cef0fe0ee3368cd0-5" who="
        Q4356302">
227   <seg xml:id="i-VDc4MFszrbkSKA9PAXdXvx">
228     Herr talman! Hadar Cars har frågat mig om den regering jag
229     tillhör är obunden av riksdagsbeslut som fattats utan stöd av
230     den socialdemokratiska riksdagsgruppen. Hadar Cars syftar
231     på riksdagsbeslutet om att de svenska utsläppen av koldioxid
232     inte bör öka.
233   </seg>
234 </u>
235 <pb facs="https://betalab.kb.se/prot-198990--11/prot_198990__11
        -003.jp2/_view"/>
```

Figure 5: Data format of the records text in the ParlaClarin XML files. Excerpt of the record 1989/90:11.

of 17,800 records, 1,057,031 pages, 1,022,014 speeches, and 521,404,039 words, 446,349,968 of which are transcriptions of speech. Figure 6 shows the number of records, speeches and words over time. One reason for the growth of words over time could be the increase in plenary session days. The increase in the number of speeches is partly explained by changing parliamentary procedures, for example, the introduction of short replies in 1933 (Stjernquist, 1996). In the 1980s, other debate rules were tried to render livelier debates with short replies and more speeches. In the 1910s, we see a spike in the number of protocols due to extra Riksdag sessions (*urtima*), while the peak in the 1970s is due to shifting from the calendar year to September-June sessions.

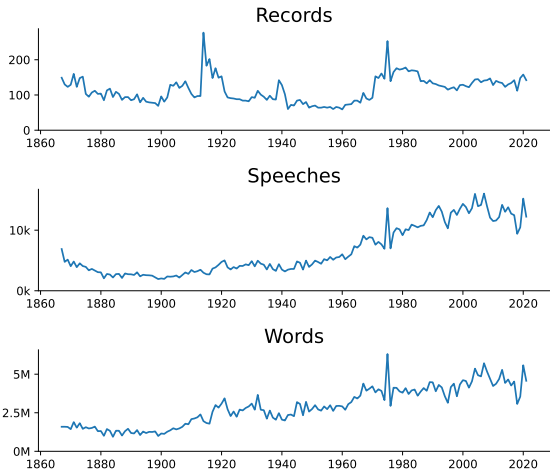


Figure 6: Number of parliamentary records, speeches, and words over time.

2.3. The Members of Parliament

Utterances in the corpus are attributed to their speaker by a unique identifier. This allows linking individual members of parliament to multiple asso-

ciated metadata categories and eliminates potential problems related to identifying individuals by name – e.g. multiple people with the same name, alternative name variants for a single person, or name changes – while also facilitating an iterative expansion of metadata in collaboration with or informed by use by researchers.

The metadata database of individuals with utterances in the parliamentary proceedings is built on previous extensive efforts to compile and link data related to Swedish politicians in the Wikidata repository (Vrandečić, 2012; Wikidata contributors, 2023). Our database is constructed iteratively (see Section 3 below) by executing structured queries to extract individuals who have a defined role as MPs, Speaker of the Riksdag, or minister and associated property attributes likely to be of research interest, such as date of birth and gender.

Most of the data on MPs is built upon Wikidata, and hence extensive quality control is used to assess the correctness of the data with respect to biographical sources such as Norberg et al. (1985) and Stjernquist (1996), see Section 2.4 for details on quality control.

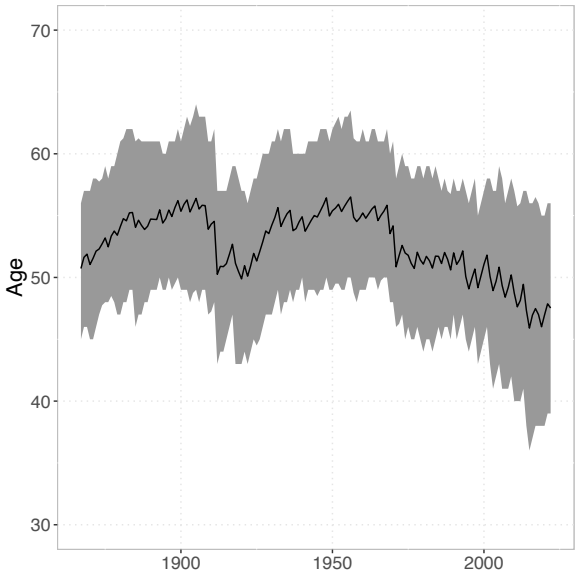


Figure 7: Age distribution among members of parliament, the mean and the 1st and 3rd age quantiles.

Currently, the database consists of 6,161 individuals. Most individuals have basic information such as gender and birth- and death dates. In addition, the database contains (1) 14,993 names, including alternate names, aliases, and spelling variations, (2) 5,184 location specifiers (*i-ort*, a formalized way of geographically disambiguating individuals with similar names), (3) 7,264 party affiliations for 5,911 individuals, including those who were not members of a political party, (4) 13,209 start/end dates associated with mandate period(s) of 5,972 individuals, and 547 twitter/X accounts. Figure 7 shows the development of the age of MPs in the Riksdag. MPs entering after the universal and equal suffrage around 1920 is reflected in a slightly younger parliament, and since the 1970s, the parliament has become both younger and more varied in age. One reason for the decrease in age since the 1970s could be the unicameral reform and the dismantling of the First Chamber, which by tradition had older MPs that also were elected for longer terms than in the Second Chamber (Stjernquist, 1996).

2.4. Corpus Quality

The corpus has undergone extensive automated and manual processing. For this reason, understanding the quality of the corpus is vital for users. To address this need, we evaluate multiple dimensions of the corpus quality. This assessment involves comparing the digital representation of the corpus with the original material. The quality is estimated based on a gold standard, stratified by year and legislative chamber, and partly annotated by domain experts (see below).

OCR Quality

We gauge the optical character recognition (OCR) quality by calculating the ratio of incorrectly parsed characters (character error rate) and words (word error rate). In this context, words were defined as sequences of text delimited by whitespace. Three pages are sampled randomly each year, and three rows are sampled and annotated on each page.

The character error rate for the whole corpus is 0.0311, and the word error rate is 0.0869, i.e., roughly 3 per cent of the characters and 9 per cent of the words are incorrect due to OCR errors. In addition to this, 70.5% of the sampled sequences yielded a fully correct OCR result. On average, the OCR result differs by 1.470 characters from the annotated line. A random sample of the OCR errors is presented in Appendix E.

Paragraph Classification Quality

As discussed in Section 2.2, the parliamentary records are subdivided into various components, including utterances, notes, and speaker introductions. To estimate segment classification errors, we take a stratified sample of three paragraphs per year and chamber during the 1910–2022 period and annotate them manually. We assess the quality of this segmentation by calculating the ratio of accurately categorized paragraphs. This gives a paragraph classification accuracy of 0.9499.

Speech-to-Speaker Mapping Quality

Ideally, every speech should be accurately linked to a corresponding individual in the metadata database. We assess this by calculating the ratio of speeches correctly linked to their respective speakers. To estimate the MP mapping errors, we take a stratified sample of three introductions per year and document type (i.e. three for each chamber) during the period 1867–2022 and annotate them manually. Based on this sample, we estimate the speech-to-speaker accuracy to 0.8589 for the whole corpus.

Furthermore, because the mapping algorithm discussed in Section 2.3 may not be able to identify certain MPs, we also consider the overall proportion of speeches where the speaker remains unidentified as an additional indicator. Figure 8 provides a visual representation of the speech-to-speaker mapping quality, revealing that the accuracy improves over time, with earlier records presenting a more significant challenge. We also see that Figure 8 is biased with respect to the random, manually assessed sample due to errors, rather than known unknowns, in the mapping algorithm.

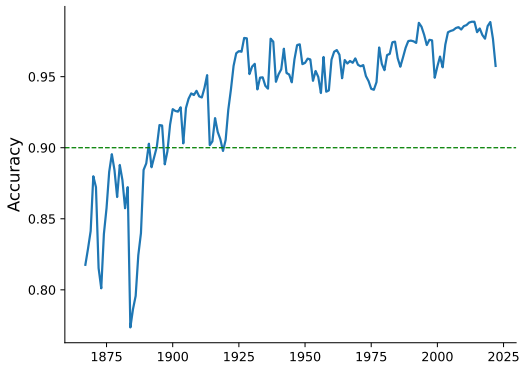


Figure 8: Speech-to-speaker mapping accuracy, according to the algorithm in Section 2.2.

MP Coverage

Lastly, in evaluating the metadata encompassing the parliament members, we summarise the percentage of MPs for each specific parliamentary year, determined by considering the start and end dates of MPs' tenures. Figure 9 illustrates the ratio of the number of MPs vs. the total number of parliamentary seats and the actual count of members of parliament. Due to replacements in the parliament, it is expected to have slightly more MPs per year than the total number of seats. This quality dimension concisely summarises the metadata related to the MPs tenures and coverage. As we can see, there is good, but far from perfect, tenure data coverage, especially before 1920.

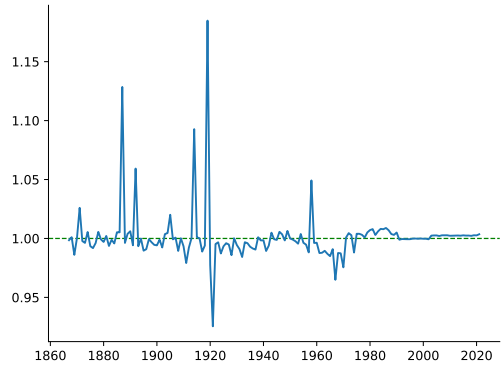


Figure 9: Ratio between the number of MPs and the number of seats in the parliament by year.

3. The Corpus Curation Process

The corpus curation process builds upon ideas of agile corpus creation (Voormann and Gut, 2008) in that the curation is made iteratively to improve the quality and add new data to the corpus. The main workflow can be summarised in the following steps:

1. Small iterative improvement
2. Quality control
   (a) Data integrity testing
   (b) Error measurement
   (c) Revision control
3. Release of a new version

Below, we elaborate on the steps in this curation process.

3.1. Small Iterative Revisions

In our efforts to enhance the corpus, we categorize improvements into two main types of quality improvements. The first is the introduction of new data or features in the corpus. This involves adding various elements, such as identifiers to other sources and new types of data, e.g. details on parliamentary roles of MPs.

The second category focuses on quality improvements. This essentially means reducing textual representation errors (Hurtado Bodell et al., 2022) in the corpus and errors identified in the metadata on MPs, such as speech-to-speaker mapping errors. It also includes actively seeking

and retrieving missing records, ensuring that the corpus maintains a high level of completeness.

To make iterative development possible, features and fixes are *orthogonal* and *modular*, *idempotent* and follow the *data-as-state* approach (see Section 2). This means that you can run all scripts independently from the others (orthogonality and modularity), repeatedly (idempotency), and based on the current data state alone (data as state).

Examples of Data Processing

Images of the printed records are OCRed using Tesseract (Smith et al., 2019). These generate ALTO XML files, which contain information about the text and its location on the page. In addition to the text content resulting from the OCR process, Tesseract's paragraph segmentation is utilized in creating the protocol files. For the born-digital files, a conversion is implemented from the JSON format, along with a fuzzy mapping to the PDF versions of the documents.

We use BERT (Devlin et al., 2018) for segment classification. The model is based on the Swedish pre-trained BERT model (Malmsten et al., 2020), which is subsequently finetuned on specifically annotated training data for the segment classification task. We use two models, one for detecting introductions and another for differentiating speech transcriptions vs. notes. Segment classification is described in detail in Appendix B.

Named entity resolution of the speakers in the speeches is implemented as a hierarchical collection of heuristics. Given an introduction, the name, party, title and similar attributes are detected using regular expressions. These first require a 100% match, and if that fails, the algorithm uses fuzzy matching. If zero or multiple people match the criteria, the speaker is marked as "unknown". Details of the algorithm can be found in Appendix C.

Metadata on the MPs is regularly accessed from Wikidata and processed to match the format of the corpus. If we notice that some available metadata is missing on Wikidata, we manually or programmatically add it to Wikidata and then update the corpus by pulling the data from Wikidata to the corpus. Protocol-level metadata is generated as a combination of the metadata of the source material and metadata scraping from the protocols, described in Appendix D.

Small quality improvements, such as correcting recurring OCR errors or adding missing IDs are addressed by rule-based algorithms. All data processing logic is available on the GitHub repository.

3.2. Quality Control

As we extensively employ computational methods for the curation, such as predictive models, and rely on Wikidata for metadata, ensuring data quality is of paramount importance.

To maintain data quality and integrity, our quality control procedures are implemented at every proposed revision of the corpus. These controls are executed in a three-step process. The first two stages involve checks conducted automatically with each new revision. If these initial automated checks pass successfully, a final manual revision control is carried out by human evaluators.

3.2.1. Data Integrity Testing

The first step of the quality control process is *data integrity testing*. This includes automatic tests to verify properties that should hold for any iteration of the corpus. Currently, we have 12 data integrity tests for different aspects of the corpus. Examples include checking that the record files adhere to the ParlaClarin XML schema, that no MP is present in the parliament before they were born or after they are dead, and that there are no duplicates in the MP database. In addition, when we conduct manual checks and controls of the data, such as a list of start and end dates of individual MPs, this is included as data integrity tests. In this way, we have caught incorrect edits made in Wikidata. All data integrity tests are listed in Appendix A.

3.2.2. Error Measurement

The second automated quality control step is the estimation of the errors for the quality domains of the corpus (see Section 2.4 above). Based on the gold standard created to estimate these errors, we can automatically estimate these error levels and assess if the proposed revision has resulted in an increased error level, which would prompt further investigation into the proposed revision. It also results in a continuous overview of the general quality level of the corpus.

3.2.3. Revision Control

As the corpus is large and many revisions created using predictive models, we cannot manually verify each proposed change in each revision of the corpus. Instead, we use *statistical quality control* to assess the revision quality before accepting the revision. This is done by drawing a random sample of edits in the proposed revision and manually assessing 50 edits. These edits are then manually assessed as correct or incorrect changes compared to the original material. Based on the outcome of this sample, we decide whether the revision should be accepted or rejected.

It's not unusual for a minor percentage of edits within a revision to be incorrect for various reasons. In such cases, our typical approach involves (1)

identifying the underlying cause of these inaccuracies and (2) documenting the source of the error as a separate issue to address in future revisions. Subsequently, we accept the revision while considering this issue for future correction.

3.3. Release of a New Version

After accepting one or more revisions, we release a new corpus version. Striving for multiple monthly releases, we rely on semantic versioning (Preston-Werner, 2013) and release each version with the [MAJOR].[MINOR].[PATCH] formalism. However, since the version concerns data rather than software, we redefine the formalism as follows

- MAJOR version changes when we make incompatible changes that are not backcompatible concerning the corpus interface, i.e. researchers' code based on the interface might break.
- MINOR version changes when we add new functionality, such as new metadata fields or new records in a backwards-compatible manner, and the
- PATCH version changes when we perform fixes or changes that do not introduce new features in the corpus, e.g. fixing errors or adding new data integrity tests.

In this way, we communicate the interoperability and backwards compatibility of the corpus for researchers. When this paper is published, we release the first 1.0.0 stable version of the corpus.

4. Conclusions

We have compiled and structured Swedish parliamentary records from 1867 to 2022 into a unified corpus with a standardized data format. Moreover, we include metadata on the members of parliament during the same period. We use computational methods and machine learning to improve the corpus iteratively. The corpus is version-controlled using semantic versioning principles. We combine data integrity tests, automated error estimation, and statistical quality control of each proposed revision to guarantee quality.

In the future, we will further develop the corpus by adding additional documents, such as private members' motions, government bills, and committee reports, and further connect these documents with the metadata on MPs. Taken together, we believe this will become a unique and useful resource for research.

Ethics statement

Our work will make research of the Swedish parliamentary records and members of parliament more accessible to a broader range of researchers.

We have received a research grant for continued development, extension and refining of the corpus up until the year 2027. We have no conflicts of interest to disclose.

Acknowledgements

The Swedish Parliament Corpus, as this paper, is based on the work conducted in the research infrastructure project "Swedish Riksdag 1867–2022: An Ecosystem of Linked Open Data" (SWERIK), funded by Riksbankens jubileumsfond [IN22-0003] (https://swerik-project.github.io/), as well as in the research project "Welfare State Analytics: Text Mining and Modeling Swedish Politics, Media & Culture, 1945–1989" (WeStAc), funded by The Swedish Research Council [2018-0606] (https://www.westac.se/en/) and "Mining for Meaning" funded by The Swedish Research Council [2018-05170]. We would also like to acknowledge the contribution made by Niklas Bergmark, Lars Brink, Emil Lanzén, Magnus Salgö, Liam Tabibzadeh, Robin Saberi and Mattias Ödevidh. Computational resources and support for this research was provided by KBLab at the National Library of Sweden.

References

Aylott, Nicholas. 2016. The party system. In Jon Pierre, editor, *The Oxford handbook of Swedish politics*, pages 152–168. Oxford University Press, Oxford.

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

Sten Carlsson. 1992. Partiväsendet i den svenska tvåkammarriksdagen 1867–1970. In Anders Norberg and Björn Asker and Andreas Tjerneld, editor, *Tvåkammarriksdagen 1867–1970: Ledamöter och valkretsar*, pages 9–27. Sveriges riksdag, Stockholm.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Paddy Dolan. 2009. Figurational dynamics and parliamentary discourses of living standards in Ireland 1. *The British Journal of Sociology*, 60(4):721–739.

Tomaž Erjavec, Matyáš Kopp, Maciej Ogrodniczuk, Petya Osenova, Darja Fišer, Hannes Pirker, Tanja Wissik, Daniel Schopper, Martin Kirnbauer, Nikola Ljubešić, et al. 2023. Multilingual comparable corpora of parliamentary debates ParlaMint 3.0.

Tomaž Erjavec and Andrej Pančur. 2021. The Parla-CLARIN recommendations for encoding corpora of parliamentary proceedings. *Journal of the Text Encoding Initiative*, (14).

Lenita Freidenvall. 2006. *Vägen till Varannan damernas: Om kvinnorepresentation, kvotering och kandidaturval i svensk politik 1970-2002*. Statsvetenskapliga institutionen, Stockholms universitet, Stockholm.

Volker Gast and Robert Borges. 2023. Nouns, verbs and other parts of speech in translation and interpreting: Evidence from English speeches made in the European parliament and their German translations and interpretations. *Languages*, 8(1).

Jo Guldi. 2019. Parliament's debates about infrastructure: An exercise in using dynamic topic models to synthesize historical change. *Technology and Culture*, 60(1):1–33.

Josefin Hägglund. 2023. *Demokratins stridslinjer: Carl Lindhagen och politikens omvandling, 1896–1923*. Ph.D. thesis, Södertörn University, Stockholm.

Jonas Harvard. 2011. Riksdagsprotokollen som medium. In Staffan Förhammar and Jonas Harvard and Dag Lindström, editor, *Dolt i offentligheten: Nya perspektiv på traditionellt källmaterial*, pages 25–42. Sekel, Lund.

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

Johan Jarlbrink and Fredrik Norén. 2023. The rise and fall of 'propaganda' as a positive concept: A digital reading of Swedish parliamentary records, 1867–2019. *Scandinavian Journal of History*, 48(3):379–399.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Machine Translation Summit*, pages 79–86.

Minna Korhonen, Haidee Kotze, and Jukka Tyrkkö, editors. 2023. *Exploring Language and Society with Big Data: Parliamentary discourse across time and space*. John Benjamins Publishing Company, Amsterdam.

Emanuele Lapponi, Martin G Søyland, Erik Velldal, and Stephan Oepen. 2018. The talk of Norway: A richly annotated corpus of the Norwegian parliament, 1998–2016. *Language Resources and Evaluation*, 52:873–893.

Martin Malmsten, Love Börjeson, and Chris Haffenden. 2020. Playing with words at the national library of Sweden–making a Swedish BERT. *arXiv preprint arXiv:2007.01658*.

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

Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Federico Nanni, Stefano Menini, Sara Tonelli, and Simone Paolo Ponzetto. 2019. Semantifying the UK hansard (1918–2018). In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*,

pages 412–413. Institute of Electrical and Electronics Engineers.

Federico Nanni, M Osman, YR Cheng, SP Ponzetto, and L Dietz. 2018. UKParl: A semantified and topically organized corpus of political speeches. In *Proceedings of LREC*.

Anders Norberg, Björn Asker, and Andreas Tjerneld, editors. 1985. *Tvåkammarriksdagen 1867-1970: ledamöter och valkretsar.* Almqvist & Wiksell International, Stockholm.

Rolf Nygren. 1985. Det svenska riksdagstrycket. In Rolf Nygren, editor, *Handbok i Nordiskt parlamentstryck*, pages 109—137. Gummesson, Falköping.

Recep Onursal and Daniel Kirkpatrick. 2021. Is extremism the 'new' terrorism? The convergence of 'extremism' and 'terrorism' in British parliamentary discourse. *Terrorism and Political Violence*, 33(5):1094–1116.

Andrej Pančur, Mojca Šorn, and Tomaž Erjavec. 2018. Slovparl 2.0: The collection of Slovene parliamentary debates from the period of secession. In *Proceedings of the LREC 2018 Workshop "ParlaCLARIN: LREC2018 workshop on creating and using parliamentary corpora"*, pages 8–14.

Lew Perren and Jonathan Sapsed. 2013. Innovation as politics: The rise and reshaping of innovation in UK parliamentary discourse 1960–2005. *Research Policy*, 42(10):1815–1828.

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

Stian Rødven-Eide. 2020. Anföranden: Annotated and augmented parliamentary debates from Sweden. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 5–10, Marseille, France. European Language Resources Association.

Laura Sinikallio, Senka Drobac, Minna Tamper, Rafael Leal, Mikko Koho, Jouni Tuominen, Matti La Mela, and Eero Hyvönen. 2021. Plenary debates of the parliament of Finland as linked open data and in Parla-CLARIN markup. In *3rd Conference on Language, Data and Knowledge, LDK 2021*. Schloss Dagstuhl-Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing.

Jure Skubic and Darja Fišer. 2022. Parliamentary discourse research in sociology: Literature review. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 81–91.

Ray Smith, Zdenko Podobný, Jim Regan, et al. 2019. Tesseract OCR. Full list of authors: `https://github.com/tesseract-ocr/tesseract/blob/main/AUTHORS`.

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

TEI Consortium. 2007. TEI P5: guidelines for electronic text encoding and interchange.

The Swedish National Library. 2023. Digitaliserat riksdagstryck 1521–1970. Accessed: 2023-09-30.

The Swedish Parliament. 2023a. Dokument och lagar - beställ och ladda ner. Accessed: 2023-09-30.

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

The Swedish Parliament. 2023c. Riksdagens öppna data. Accessed: 2023-09-30.

Holger Voormann and Ulrike Gut. 2008. Agile corpus creation. *Corpus Linguistics and Linguistic Theory*, 4(2):235–251.

Denny Vrandečić. 2012. Wikidata: A new platform for collaborative data collection. In *Proceedings of the 21st international conference on world wide web*, pages 1063–1064.

Wikidata contributors. 2023. Wikidata: Introduction. Accessed: 2023-10-12.

Appendix

A. Data Integrity Tests in the corpus

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. Global data integrity tests

The global data integrity tests are run on the corpus as a whole (or a sample of it) and focus on the overall structure and consistency of the corpus.

A.1.1. Source tests

1. Check a random subset of generated parliamentary records in generated ParlaClarin against the original Alto XML to ensure the difference in content is within an acceptable tolerance.

A.1.2. Metadatabase tests

1. Check that there is no duplicate data in governments, members of parliament, ministers, party affiliations, persons, speakers or Twitter/X tables.

2. Test the integrity of a manually curated list of members of parliament between 1876 and 1994; i.e. that each entry has a Wikidata identifier and birth date.

3. Test that all members of parliament in the manually curated list are present in the database with person, name, location specifier, member of parliament, and party tables.

4. Test that document dates from parliamentary records correspond to known dates of parliamentary sessions; record dates are in the list, and each date in the list appears on a record.

A.1.3. Members of parliament tests

1. Test that all MPs in identified speeches are also found in the metadata database.

2. Test that no speakers are present in parliamentary records before adulthood or after their death.

A.1.4. Technical tests

1. Check that `<u>` element attributes `prev="ELEM-ID"` and `next="ELEM-ID"` actually reference the previous and next `<u>` elements.

2. Check that all XML elements in parliamentary records that contain text (`note`, `u`, `seg`) have an ID attribute (`id`).

3. Validate a controlled sample of parliamentary records against the ParlaClarin XML schema. The sample consists of one randomly selected file from each parliamentary year, which also tests each source type and format, to ensure that different types of records are schematized correctly.

A.2. Revision data integrity tests

The revision data integrity tests focus on the data integrity of the data in the proposed revision.

A.2.1. Unchanged manual edits tests

1. Check that manually curated data, which is considered a gold standard, is not overwritten by edits or deletion of files.

A.2.2. Technical tests

1. Validate any changed or updated documents against the ParlaClarin XML schema.

B. Segment classification

All text in the records is classified into different subcategorias on the segment level. The most important distinctions are between transcriptions of speech, speaker introductions and transcriber and margin notes. The classification process into these classes consists of the following steps

1. Detecting introductions (`<note type="speaker">`) from all other segments

2. Classifying the transcriptions of speech (`<u>`) or transcriber and margin notes (`<note>`)

Both of these steps are implemented using a BERT-based neural network algorithm. The algorithms take in the textual content of the segment, and output class probabilities.

The introduction detection algorithm was initialized from the Swedish KB BERT model. It was then trained on a random sample of the paragraphs, which was subsequently improved with one round of active learning.

The speech vs. note algorithm was also initialized from the Swedish KB BERT model. It was finetuned on paragraphs from a random sample of pages. The best model was selected by a yearly stratified validation set annotated by experts.

C. Speaker Identity Resolution

The speaker entity resolution process in the protocols is a two-step process:

1. Extract metadata from the speaker introduction in the records
2. Match the metadata elements to the metadata database using a hierarchy of increasingly heuristic rules

Step 1 uses a list of regular expressions that match common patterns of presenting the name, party, gender and other metadata in the introductions. It takes the introductions as a string and returns a metadata dictionary, as illustrated in Figure 10.

```
>>> intro_to_dict(``Herr NILSSON i Gävle (k):'')
{`gender': `man', `party': '(k)', `specifier': `Gävle', `name': `
        NILSSON'}
```

Figure 10: Step 1 in the speaker entity resolution: extracting metadata from the introduction.

Step 2 takes in the metadata dictionary and matches it against the metadata database. The metadata folder is filtered by the record year before performing this step.

- If the introduction contains the word minister (minister / stadsråd)
  1. Match by name; and if not found
  2. Match by ministerial role
- If the introduction contains the word speaker (talman)
  1. Match the phrase 1st/2nd/3rd vice speaker (först/andra/tredje vice talman); and if not found
  2. Match the phrase vice speaker (vice talman); and if not found
  3. Default to the speaker of the house (talman)
- If there is no indication of a ministerial or speaker role
  1. Filter by gender if applicable
  2. Match exact name, and full match of other metadata variables
  3. Match exact name with some typos allowed and full match of other metadata variables
  4. In case of three names, match at least two names and have an exact match of other metadata variables

If the process yields 0 or 2 or more matches, it is considered ambiguous, and the speaker is tagged as “unknown”.

D. Document Metadata Extraction

As structured upstream metadata is only available for the 1970–2022 period, some record-level metadata needs to be extracted from the protocols. The dates of the protocols appear in the margin notes for the whole corpus period, so they can be scraped from the documents.

The process consists of the following steps:

1. Exclude transcriptions of speech using a predictive model, as described in Appendix B
2. Exclude paragraphs that are too long to include date metadata ($\geqslant 50$ characters)
3. Match common date formats using regular expressions
4. Convert the textual date format into a Python date `dateparser`[4] Python module

The elements where a date is found are classified as `<note type="date">` according to the ParlaClarin format. All unique dates are also saved in the frontmatter of the XML file as `<docDate>` elements.

E. OCR errors

Annotation	Matched Text in the Corpus
dustri å ett dylikt plan av viss Junkertyp [...]	ndustri å ett dylikt plan av viss Junkertyp [...]
lan kl. 3 och 1/2 4 på e. m. Så krävdes [...]	lan kl. 3 och ½4 på e. m. Så krävdes [...]
hade möjlighet att vara med om den debatten [...]	hade möjlighet att vara med om den debatten [...]
Utförsäljning av Sveriges guldreserv.	Särbeskattning av makars förmögenhet.
rande i vårt land av bedövningstvång vid slakt [...]	örande i vårt land av bedövningstvång vid slakt [...]
företag utgör därvid inget udantag. SJ s service [...]	företag utgör därvid inget undantag. SJ:s service [...]
positionen vara med övervägande ja besvärad [...]	positionen vara med övervägande ja besvarad. [...]
punkter i alla avseenden tillfredsstäl	npunkter i alla avseenden tillfredsstäl
då tekniska fel uppstått på den auto	de synskadade vid tillämpningen av v
men af det uppdrag, nämnda kommittés [...]	ramen af det uppdrag, nämnda kommitté [...]

Table 1: Ten randomly selected OCR errors. On the left, there is the annotated row, on the right there is the closest match on the page in the corpus.

[4] https://dateparser.readthedocs.io/en/latest/