The Swedish Parliament Corpus

LREC-COLING 2024 Presentation

Väinö Yrjänäinen
Johan Jarlbrink
Pelle Snickars
Fredrik Mohammadi Norén
Lotta Åberg Brorsson
Måns Magnusson
Robert Borges
Anders P. Olsson

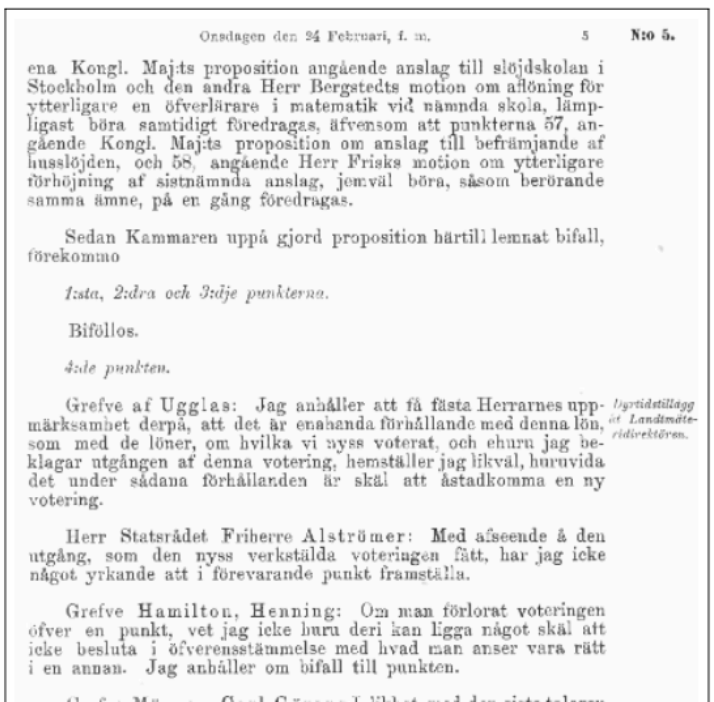
2024-05-02

Background

- The Swedish Parliament (*Riksdagen*)
  - Bicameral parliament 1867-1971
  - Unicameral parliament 1971->
  - Currently 349 members
- All of the debates are written down in the parliamentary records (*protokoll*)
- We work on the subset of 1867-2023

Source Material

- Riksdagen records
  - Printed and scanned 1867-1989
  - Digital records in different formats since 1990
- Different sources for metadata
  - Biograpgy books
  - Wikidata
- Motions etc. other documents in the future



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 hushållsljden, och 58, angående Herr Frisks 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.

Bifölios.

4:de punkten.

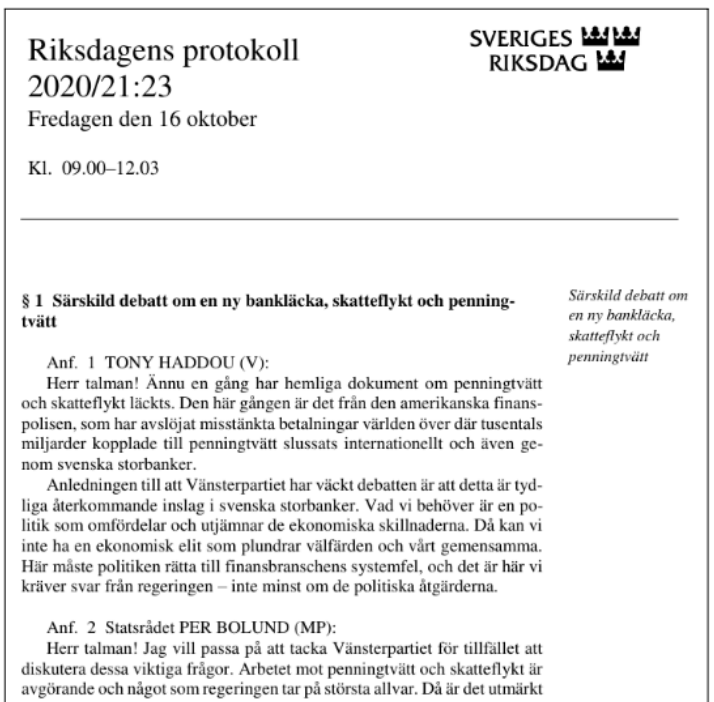
Grefve af Ugglas: Jag anhåller att få fästa Herrarnes uppmärksamhet derpå, att det är enahanda förhållande med denna lön, som med de löner, om hvilka vi nyss voterat, och ehuru jag beklagar utgången af denna votering, hemställer jag likväl, huruvida det under sådana förhållanden är skäl att åstadkomma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

Dygtidstillägg åt Landtmäteridirektören.

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

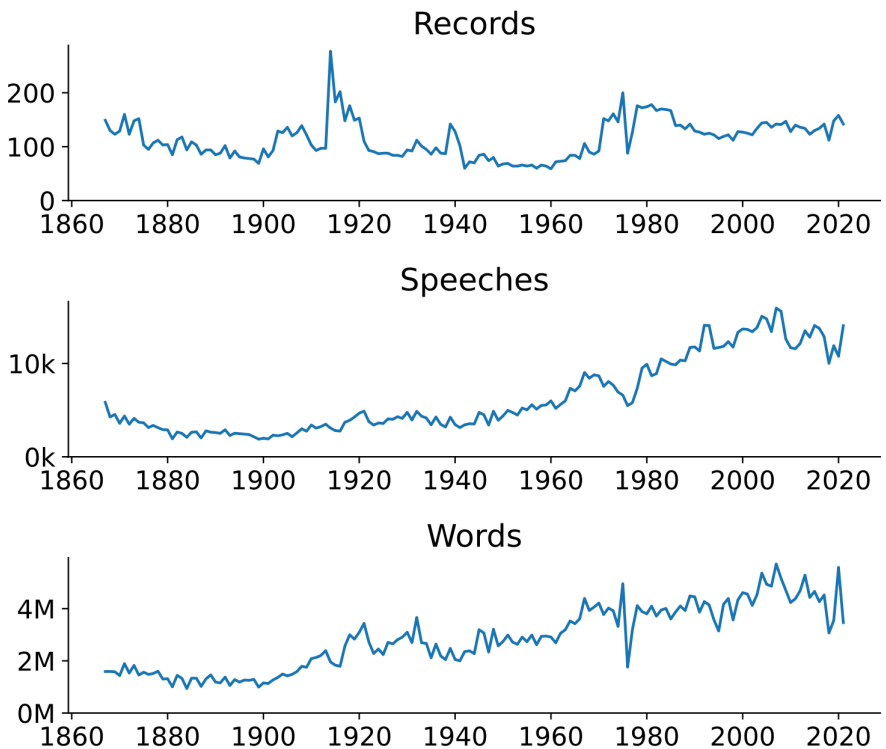
Särskild debatt om en ny bankläcka, skatteflykt och penningtvätt

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



Individuals	6161
Names	14993
i-ort	5184
Party affiliations	6531
Mandate periods	13172

Corpus and format

Corpus API

- Records in Parla-Clarin XML format
- Metadata in tabular form (CSV)
- Full history is stored in git repositories
- Python and R packages to easier access the corpus

```
<!--    [...]    -->
<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>
<!--    [...]    -->
```

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

Biföllos.

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 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.

1867/	1894/	1921/	1948/	1975/	200001/
1868/	1895/	1922/	1949/	197576/	200102/
1869/	1896/	1923/	1950/	197677/	200203/
1870/	1897/	1924/	1951/	197778/	200304/
1871/	1898/	1925/	1952/	197879/	200405/
1872/	1899/	1926/	1953/	197980/	200506/
1873/	1900/	1927/	1954/	1980/	200607/
1874/	1901/	1928/	1955/	198081/	200708/
1875/	1902/	1929/	1956/	198182/	200809/
1876/	1903/	1930/	1957/	198283/	200910/
1877/	1904/	1931/	1958/	198384/	201011/
1878/	1905/	1932/	1959/	198485/	201112/
1879/	1906/	1933/	1960/	198586/	201213/
1880/	1907/	1934/	1961/	198687/	201314/
1881/	1908/	1935/	1962/	198788/	201415/
1882/	1909/	1936/	1963/	198889/	201516/
1883/	1910/	1937/	1964/	198990/	201617/
1884/	1911/	1938/	1965/	199091/	201718/
1885/	1912/	1939/	1966/	199192/	201819/
1886/	1913/	1940/	1967/	199293/	201920/
1887/	1914/	1941/	1968/	199394/	202021/
1888/	1915/	1942/	1969/	199495/	202122/
1889/	1916/	1943/	1970/	199596/	prot-ak.xml
1890/	1917/	1944/	1971/	199697/	prot-ek.xml
1891/	1918/	1945/	1972/	199798/	prot-fk.xml
1892/	1919/	1946/	1973/	199899/	
1893/	1920/	1947/	1974/	19992000/	

person_id	start	end	district	role
i-PFAPNmRqeUAaxDzNRTG1x1	1867	1867	Eskilstuna och Strängnäs valkrets	andrakammarledamot
i-QSYHiJ6G54WwZYYDpVnD4u	1867	1867	Västra Götalands läns västra valkrets	förstakammarledamot
i-Ddmtm1uG9esPH37c8XjUXZ	1867	1867	Torna härads valkrets	andrakammarledamot
i-65rmwEXkkhA1kxSrD4oMUw	1867	1867	Värmlands läns valkrets	förstakammarledamot
i-AvGpsGJvs5PXTcEG4DtbFt	1867	1867	Kristianstads läns valkrets	förstakammarledamot
i-7APUEDdtwmAmmVwZyzCQPJ	1867	1867	Västra Götalands läns västra valkrets	förstakammarledamot
i-T33Na3brmtboXa1NjvckiT	1867	1867	Orusts och Tjörns domsagas valkrets	andrakammarledamot
i-M87tPv4oBZ1v745fULLazy	1867	1867	Värmlands läns valkrets	förstakammarledamot
i-KMNhDimDdFZa9YsdoMSHgX	1867	1867		andrakammarledamot
i-HiLh4CgUmPaDRYddeo7w7X	1867	1867	Jämtlands läns valkrets	förstakammarledamot
i-A7oyGrcQCHCGYURTKKXFvN	1867	1867		andrakammarledamot
i-PsEE3VTZZXPnyW93HZcXyd	1867	1867	Älvsborgs läns valkrets	förstakammarledamot
i-3MH73DLbrRLp6rCtZUN7xj	1867	1867	Härnösands, Umeå, Luleå och Piteå valkre	andrakammarledamot
i-FgPMLnSevim92vXDyJ4J7F	1867	1867	Gävleborgs läns valkrets	förstakammarledamot
i-QmwGRrMY63mWh5oomvqDqy	1867	1867	Landskrona valkrets	andrakammarledamot
i-5u9D97HkMoFCMxfWkAsxz3	1867	1867-01-18	Västbo härads valkrets	andrakammarledamot
i-JybGN9aHM3q5M4kgRwLipb	1867-01-15	1867-01-24		andrakammarledamot
i-QCgdixScKoSHeN5oNtiVxs	1867	1867-03-07		andrakammarledamot
i-SHW87cHVMCQaaznoE9ZBmu	1867	1867-04-24	Örebro läns valkrets	förstakammarledamot
i-8YUu6FdSHbMaWxHkRan8Va	1867	1868		andrakammarledamot
i-R3ezo9eNq9173LWfrJPgK8	1867	1868	Visby stads valkrets	andrakammarledamot
i-LuXD6bo7y9QvBg7xBKqETi	1867	1868		andrakammarledamot
i-Pd4h8w97V3gTekGxSvebsQ	1867	1868	Hallands läns valkrets	förstakammarledamot
i-Dojdaqz1Xnckajnx4HsfYk	1867	1868		andrakammarledamot
i-Jrqp3xhubpckqh6NNnYVm4	1867	1868	Kalmar läns norra valkrets	förstakammarledamot
i-JHpaqU2z8URQR83kWF8zv4	1867	1868	Jönköpings läns valkrets	förstakammarledamot

```
chair_mp.csv
chairs.csv
described_by_source.csv
external_identifiers.csv
government.csv
location_specifier.csv
member_of_parliament.csv
minister.csv
name.csv
party_abbreviation.csv
party_affiliation.csv
person.csv
place_of_birth.csv
place_of_death.csv
portraits.csv
references_map.csv
riksdag-year.csv
speaker.csv
twitter.csv
wiki_id.csv
```

Curation methods

Curation Methods: Optical character recognition

- *Convert an image to text*
- Tesseract OCR
  - Take in scanned images
  - Outputs words and paragraph splits

```
tesseract -l swe prot--1985--001-1.jpg prot--1985--001-1 alto
```

Curation Methods: Paragraph classification

- *Is this paragraph a speech or something else?*
- Speech, non-speech, and speaker introductions
- Neural network (BERT) approach

```
>>> classifier = pipeline("text-classification", model="./trained/binary_note_seg_model/")
>>> classifier("1:sta, 2:dra och 3:dje punkterna. Biföllos. 4:de punkten.")
[{'label': 'note', 'score': 0.9890651702880859}]
```

Curation Methods: Metadata scraping

- *What is the date of this record?*
- Filter out long paragraphs
- Find substrings that contain a date using regex
- Use the *dateparser* Python library to convert to a proper format

```
>>> dateparser.parse('den 13.. januari 1997')
datetime.datetime(1997, 1, 13, 0, 0)
```

Curation Methods: Speaker mapping

- *Who gave this speech?*
- Detect speaker introductions using neural networks
- Convert intro to using regex
- Do a heuristic search against the metadata database

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

Curation process & agile methodology

Curation Process

1. Prototype corpus
2. Iterative improvement
    1. Address one issue
    2. Do quality control on the proposal
    3. Do automatic testing of data
    4. Merge if acceptable quality and tests pass
3. Release continuously

Quality control

1. Sample 50 diffs
2. Classify each of them into correct or incorrect
3. If >25 of them are correct, accept revision

Quality control (*Myers difference algorithm*)

N:o 9. Angående höjning af anslaget till Ultuna [-landibruks–]{+landibrulks-+} institut [-m. m.-]{+m.m.+} (Forts.) 16 Lördagen den 23 Februari, f. m. dana föräldrar, som bedrifva jordbruk, dels, då praktisk undervisning ej kan erhållas hemma, på andra egendomar, innan de komma till skolorna, samt sedan eleverna kommit från institutet kunna de få sin praktiska utbildning å egendomar, hvilka på många ställen i landet skötas synnerligen väl; och der skall denna utbildning blifva af större betydelse och nytta än någonsin vid landtbruksinstitut kan blifva fallet, och dertill mycket billigare särskildt för eleverna. Jag har försökt tydliggöra de skäl jag angifvit i min reservation, som finnes tryckt i betänkandet, och herrarne kunna finna, att jag för min del vill till- styrka kammaren att bifalla denna reservation, till hvilken jag nu yrkar bifall. I detta anförande instämde herrar Odell, [-Nilsson-]{+MNilsson+} i Skärhus, Andersson i Baggböle, Andersson i [-Löfhult-]{+Löthult+} och Björkman. Herr Lasse Jönsson yttrade: Herr talman! Inom statsutskottet var jag till en början något villrådig om, huru vida det vore klokast att nu antaga detta förslag eller låta det hvila till framtiden, men då jag märkte, hvarthän ett uppskof skulle leda, nemligen till att från landtbruksläroverken borttaga så väl den lägre landtbrukskursen som äfven den praktiska undervisningen för eleverna, och detta är ett håll, åt hvilket jag icke vill gå, så ansåg jag det bättre att nu få frågan afgjord i sitt föreliggande skick än att vänta på något, som enligt min Åsigt komme att gå 1 rak strid mot det rätta. Det råder nem- ligen nu för tiden en viss strömning att i allt lägga an på den teore- tiska bildningen. Allting skall vara så lärdt och endast teori. Skulle vi nu slå in på den vägen äfven vid de läroverk, som från början äro inrättade för [-utbildning. ar-]{+utbildning af+} praktiska män, vid de [-amdt rn task,-]{+landtbruksinstitut, dit föräldrar, som icke sjelfva kunna meddela sina söner någon prak-+} tisk [-omen-]{+undervisning+} i [-ord örn-]{+jordbruk,+} borde kunna med [-legat ören-]{+trygghet öfverlemna+} dem just för att [-förvärfva-]{+förvärlva+} en sådan utbildning, skulle denna praktiska undervisning nu der tagas bort, är jag rädd för att man kunde lika så gerna [-upphäfva-]{+upphätfva+} hela institutionen, ty den teoretiska undervisningen kunde i så fall lika väl förläggas till universiteten. Denna teoretiska strömning, som jag omnämnt, har tyvärr redan visat sina skadliga verkningar här och der i landet. Så har till exempel hushållningssällskapet i det län, jag tillhör, redan fattat be- slut att indraga den lägre landtbruksskola, som der funnits, hvilket beslut jag anser vara till stor skada för länet. Äfven såg jag i tid- ningarna här om dagen, att något dylikt lär hafva händt i Vermland. Är detta den rätta vägen att utbilda jordbrukare? Nej, enligt min åsigt är den alldeles felaktig. Skola vi nemligen hafva någon slags undervisning i jordbruk, [-Jåtom-]{+låtom+} oss då på alla sätt gå i motsatt rigt- ning och mera framhålla det praktiska än det teoretiska, ty det teo- retiska har enligt mitt förmenande blott föga värde för den man, som skall egna sig åt jordbruket. Det går nu till den grad långt i orätt rigtning, att landtbruksläroverken kommit i sådan misskredit, att ingen vågar taga någon förvaltare derifrån. Man importerar hellre sådana från länder, der den praktiska utbildningen sättes i främsta

Sampled changes

data/199293/prot-199293--010.xml

Diff starting from line 5

```
@@ -5,9 +5,6 @@
       <titleStmt>
         <title>Protokoll</title>
       </titleStmt>
-      <editionStmt>
-        <edition>0.4.2</edition>
-      </editionStmt>
       <publicationStmt>
         <authority>National Library of Sweden and the WESTAC project</authority>
       </publicationStmt>
```

- [ ] Correct
- [ ] Incorrect

Data Integrity testing

- Automatically test things about the revisions that should always hold
- We run 12 different data integrity tests
  - Nobody should be present in the chamber before they were born
  - All text elements should have an ID
  - etc.

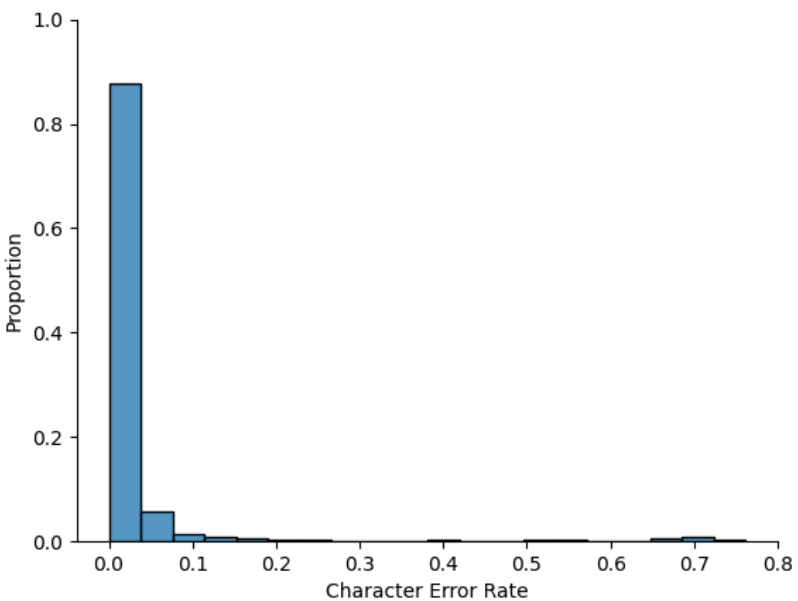
Evaluation

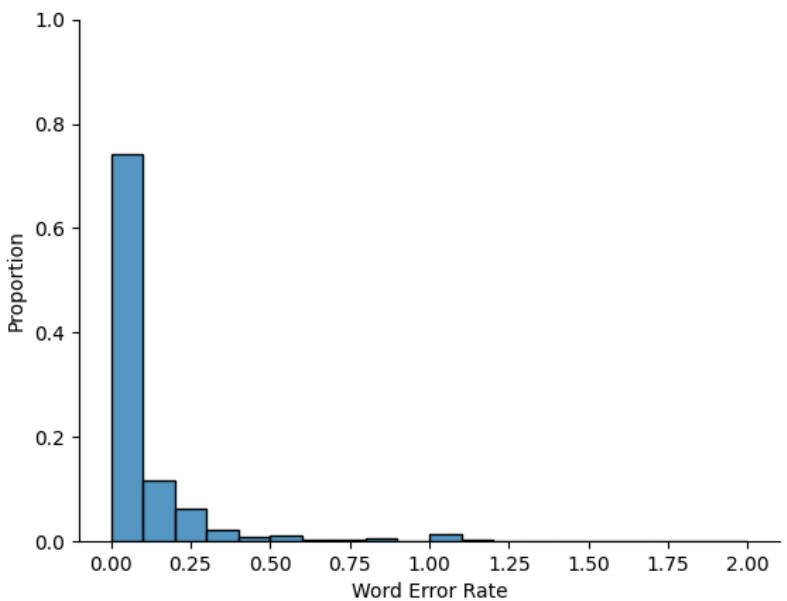
Evaluation / estimation of corpus quality

1. Define *quality dimensions* to assess
   - Eg. OCR, segment classification
2. Take a stratified sample
3. Annotate the sample
4. Calculate quality metric

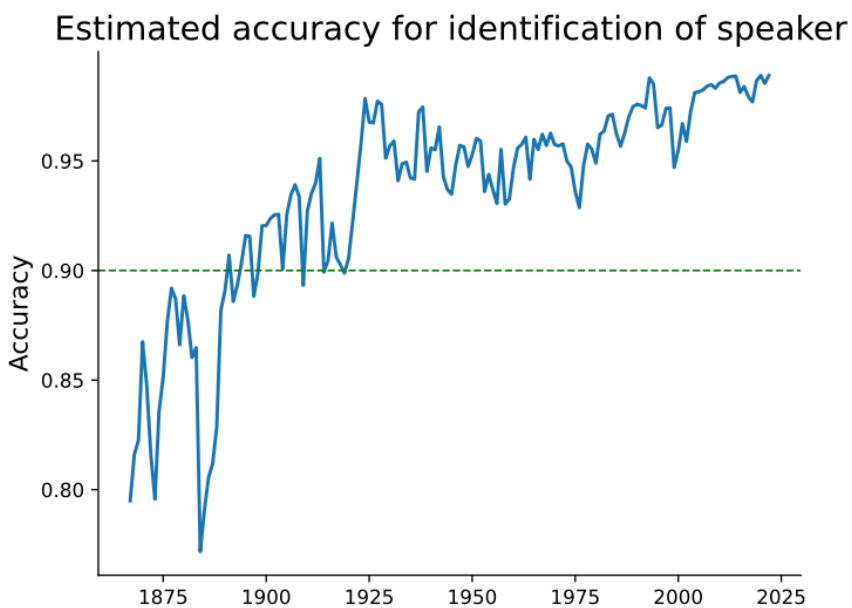
OCR evaluation results

OCR evaluation results

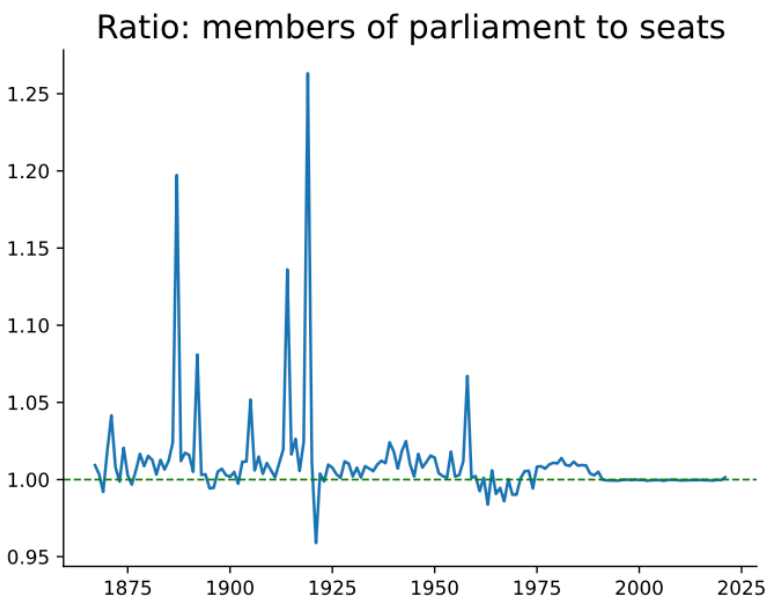




Speaker mapping



MP coverage


Ratio: members of parliament to seats
1.25
1.20
1.15
1.10
1.05
1.00
0.95
1875
1900
1925
1950
1975
2000
2025

Summary

- Data and metadata spanning over 150 years on the Swedish parliament
- The corpus is built iteratively
  - Each change is statistically evaluated
- We provide the corpus in ParlaCLarin XML/CSV formats, along with Python and R modules
  - https://github.com/swerik-project
- Corpus quality is estimated on multiple metrics