FACULTY
OF INFORMATICS
Masaryk University

# Data gathered with automatic tools from European parliamentary chambers

Autor: Ota Mikušek

# Outline

1. Motivation and objective
2. Results
3. Toolset maintenance

# Motivation

# Motivation

- Easier and uniform access to debates in the parliaments of EU countries.

# Motivation

- Easier and uniform access to debates in the parliaments of EU countries.
- Text source without errors.

# Motivation

- Easier and uniform access to debates in the parliaments of EU countries.
- Text source without errors.
- Texts suitable for machine learning.

# Objective

# Objective

- Creation of a toolset for the automatic continuous download of stenoprotocols of the parliaments of EU countries.

# Objective

- Creation of a toolset for the automatic continuous download of stenoprotocols of the parliaments of EU countries.
- Provide:
  - Reliable sources.

# Objective

- Creation of a toolset for the automatic continuous download of stenoprotocols of the parliaments of EU countries.
- Provide:
  - Reliable sources.
  - Creating programs for downloading and formatting stenoprotocols.

# Objective

- Creation of a toolset for the automatic continuous download of stenoprotocols of the parliaments of EU countries.
- Provide:
  - Reliable sources.
  - Creating programs for downloading and formatting stenoprotocols.
  - Detecting errors in running programs.

# Reliable sources

# Reliable sources

- Located 30 reliable sources out of 38 possible chambers.

# Reliable sources

- Located 30 reliable sources out of 38 possible chambers.
- Static websites / API

# Reliable sources

- Located 30 reliable sources out of 38 possible chambers.
- Static websites / API
- Format: HTML, JSON, CSV, XML, XLSX, and DOCX.

# Created toolset

# Created toolset

- One tool per chamber.

# Created toolset

- One tool per chamber.
- 29 tools created. (Ireland lower and upper chamber is on same website.)

# Created toolset

- One tool per chamber.
- 29 tools created. (Ireland lower and upper chamber is on same website.)
- Each tool consist of
  - Shared code

# Created toolset

- One tool per chamber.
- 29 tools created. (Ireland lower and upper chamber is on same website.)
- Each tool consist of
  - Shared code
  - Downloaded

# Created toolset

- One tool per chamber.
- 29 tools created. (Ireland lower and upper chamber is on same website.)
- Each tool consist of
  - Shared code
  - Downloaded
  - Prevertbuilder

# Example of a prevertical

```
<doc source_url="https://www.psp.cz/eknih/2017ps/stenprot/zip/
012schuz.zip" url_access_time="2023-05-09 20:20:47 UTC"
filename="2017_012schuz.prevert" date="2022-02-22" date_day="22"
date_month="2" date_year="2022">
<p>
(Schůze zahájena ve 14.01 hodin.)
</p>
<speaker name="Předseda PSP Radek Vondráček">
<p>
Vážené paní poslankyně, vážení páni poslanci, vážení členové vlády,
zahajuji 12. schůzi Poslanecké sněmovny a všechny vás tu vítám.
</p>
<p>
Organizační výbor Poslanecké sněmovny stanovil návrh pořadu 12.
schůze dne 29. března 2018. Pozvánka vám byla rozeslána tentýž den.
</p>
```

# Resulting corpora

# Resulting corpora

- Creation of 29 corpora.

# Resulting corpora

- Creation of 29 corpora.
- Partial or complete download of the parliaments of 22 of the 27 EU countries.

# Resulting corpora

- Creation of 29 corpora.
- Partial or complete download of the parliaments of 22 of the 27 EU countries.
- Tools available at `https://gitlab.com/Atom194/european-parliamentary-protocols`

# Comparison with the ParlaMint project (PM)[1]

# Comparison with the ParlaMint project (PM)[1]

- Toolset
- 22 nations of EU
- 1 329,06 M tokens
- 1 146,25 M words
- All data at least since 2022

- ParlaMint Corpora
- 17 corpora
- 555,63 M tokens
- 472,44 M words
- From 2015 to 2020

# Change in past six manths

| tool name | words | words now | change | from year |
|---|---|---|---|---|
| bg_deputies | 5.40M | 5.82M | +0.42M | 2022 |
| cz_deputies | 18.41M | 20.71M | +2.30M | 2018 |
| cz_senate | 11.32M | 11.51M | +0.19M | 2010 |
| dk_deputies | 79.00M | 79.55M | +0.55M | 2007 |
| nl_deputies | 71.20M | 80.20M | +9.00M | 2013 |
| nl_senate | 9.99M | 11.01M | +0.02M | 2019 |
| ir_deputies | 40.70M | 87.28M | +46.58M | 2022 |
| ee_deputies | 9.04M | 10.47M | +1.43M | 2020 |
| fi_deputies | 21.09M | 21.09M | 0 | 2015 |
| be_deputies | 54.94M | 56.70M | +1.76M | 2007 |
| be_senate | 0.06M | 0.69M | +0.63M | 2019 |
| fr_deputies | 21.09M | 59.55M | +38.46M | 2015 |
| fr_senate | 169.08M | 173.52M | +4.44M | 2004 |

# Change in past six manths

| tool name | words | words now | change | from year |
|---|---|---|---|---|
| at_deputies | 6.94M | 7.19M | +0.25M | 2022 |
| at_senate | 2.73M | 2.87M | +0.14M | 2019 |
| de_deputies | 125.03M | 125.53M | +0.50M | 1950 |
| gr_deputies | 58.31M | 59.47M | +1.16M | 2015 |
| hu_deputies | 3.08M | 3.93M | +0.85M | 2022 |
| it_deputies | 3.32M | 5.15M | +1.83M | 2022 |
| it_senate | 13.31M | 14.61M | +1.30M | 2018 |
| pl_senate | 20.08M | 20.25M | +0.17M | 2011 |
| pt_deputies | 141.10M | 154.36M | +13.26M | 1976 |
| ro_deputies | 14.02M | 14.86M | +0.84M | 2016 |
| ro_senate | 26.36M | 26.88M | +0.52M | 2001 |
| sk_deputies | 6.76M | 8.73M | +1.97M | 2022 |
| si_deputies | 15.49M | 23.69M | +8.20M | 2018 |

# Change in past six manths

| tool name | words | words now | change | from year |
|---|---|---|---|---|
| es_deputies | 66.66M | 68.73M | +2.07M | 2019 |
| se_deputies | 131.74M | 131.74M | 0 | 1994 |
| sum | 1,146.25M | 1,286.09M | +139.84M | - |

# Toolset maintenance

# Parliament of Slovenia

(27. marca 2007) -- (27. marec 2023)
(28. aprila 1999) -- (28. april 2023)
(3. decembra 1999) -- (18. december 2018)

(16. JUNIJ 2010) -- (15. junij 2020)

# Parliament of Bulgaria

- HEADERS = ’User-Agent’: ’curl/7.82.0’
- TIME_SLEEP_SECONDS = 16

# Connection errors

- 72 errors out of 299 total (in past nine months).
- Error is solved the next day during the next download attempt.

# Conclusion

- Proof that automatic and autonomous parliamentary corpora creation is possible.

# Conclusion

- Proof that automatic and autonomous parliamentary corpora creation is possible.
- Only a few sheared metadata.

# Conclusion

- Proof that automatic and autonomous parliamentary corpora creation is possible.
- Only a few sheared metadata.
- Maintenance costs are low.

# Bibliography

[1] CLARIN.SI. *CLARIN.SI - NoSketch Engine*. 1999. URL: https://www.clarin.si/noske/ (visited on 12/07/2023).