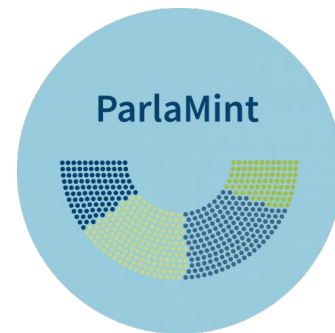


Searching the ParlaMint corpus

Class #12b, May 3 2022

Barbora Hladká hladka@ufal.mff.cuni.cz



ParlaMint

is a project of compiling parliamentary debates into uniformly annotated multilingual corpora. It is financially supported by [CLARIN](#).

<https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora>

ParlaMint v. 2.1

= corpora of 17
European
parliaments

Table 1 Basic information about the ParlaMint corpora including the corpus ID, the covered language(s), the houses and number of terms included, from and to months of included transcriptions, the number of years covered, the number of millions of words per year and in total

ID	Lang	Houses	Ts	From	To	Yrs	Mw/Yr	Mw
BE	nl+fr	Lower	2	2015–11	2020–08	4.8	6.50	31.37
BG	bg	Unicameral	2	2014–10	2020–07	5.8	3.42	20.02
CZ	cs	Lower	2	2013–11	2021–04	7.5	3.03	22.56
DK	da	Unicameral	–	2014–10	2020–09	6.1	4.85	29.40
ES	es	Lower	5	2015–01	2020–12	6.0	2.19	13.10
FR	fr	Lower	1	2017–07	2020–07	3.0	10.75	32.73
GB	en	Lower+Upper	4	2015–01	2021–03	6.3	17.25	109.30
HR	hr	Unicameral	1	2016–11	2020–05	3.6	5.81	20.65
HU	hu	Unicameral	2	2014–05	2020–12	6.7	0.13	0.87
IS	is	Unicameral	3	2015–01	2020–09	5.8	4.06	23.66
IT	it	Upper	2	2013–03	2020–11	7.8	3.46	26.94
LT	lt	Unicameral	2	2012–11	2020–11	8.1	1.82	14.78
LV	lv	Unicameral	2	2014–11	2021–02	6.3	1.02	6.48
NL	nl	Lower+Upper	5	2014–04	2020–11	6.6	7.74	51.45
PL	pl	Lower+Upper	4	2015–11	2020–08	4.9	5.66	27.45
SI	sl	Lower	2	2014–08	2020–07	6.0	3.34	20.19
TR	tr	Unicameral	4	2009–04	2021–02	12.0	3.65	43.99

Source: (Erjavec, T., Ogrodniczuk, M., Osenova, P. et al., 2022)

ParlaMint v. 2.1

- The complete corpora hdl.handle.net/11356/1432
- The complete corpora with added linguistic annotations
hdl.handle.net/11356/1431
- Available on Github: Samples of the corpora, the XML schema and various processing and validation scripts
<https://github.com/clarin-eric/ParlaMint>
- Corpora in concordancers
 - NoSketch Engine clarin.si/noske
 - KonText <https://www.clarin.si/kontext>

Use case :: Leaving EU in ParlaMint

1. How many times did the speakers say leave|leaves|left|leaving the European Union in their speeches?

Examples

- As we leave the European Union, changes to regulations might be required and ...
- ...so that we have a smooth transition from where we are today to leaving the European Union
- to be able to have its own free trade policy once we have left the European Union

2. How did the overall frequency of the text change over time?

Use case :: Leaving EU in ParlaMint

Data

- ParlaMint-GB 2.1 (British parliament)

Tools

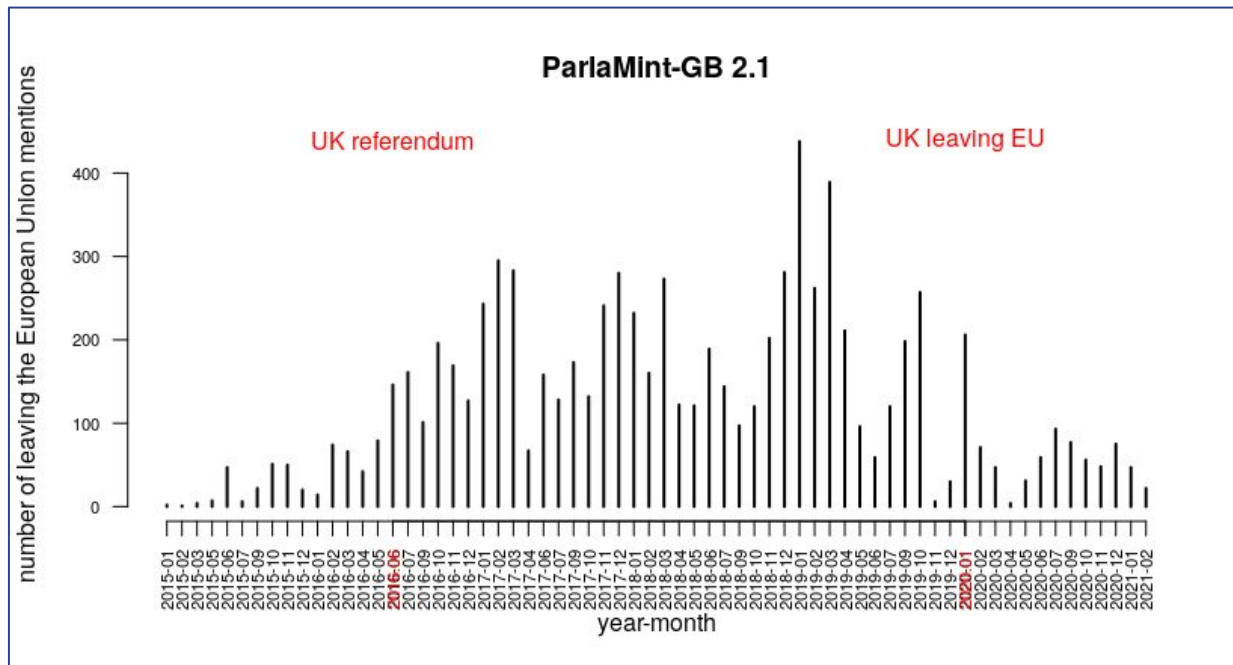
- combination of KonText search and R programming

1. How many times did the speakers say leave|leaves|left|leaving the European Union in their speeches?

	A	B
1	speaker	frequency
2	May, Theresa Mary	736
3	Gove, Michael Andrew	121
4	Leadsom, Andrea Jacqueline	116
5	Smith, Julie	113
6	Callanan, Martin	110
7	Davis, David Michael	101
8	Fox, Liam	96
9	Forsyth, Michael Bruce	84
10	Bridges, James	78
11	Cairns, Alun Hugh	78
12	Ahmad, Tariq	70
13	Soubry, Anna Mary	69
14	Cash, William Nigel Paul	68
15	Eustice, Charles George	67

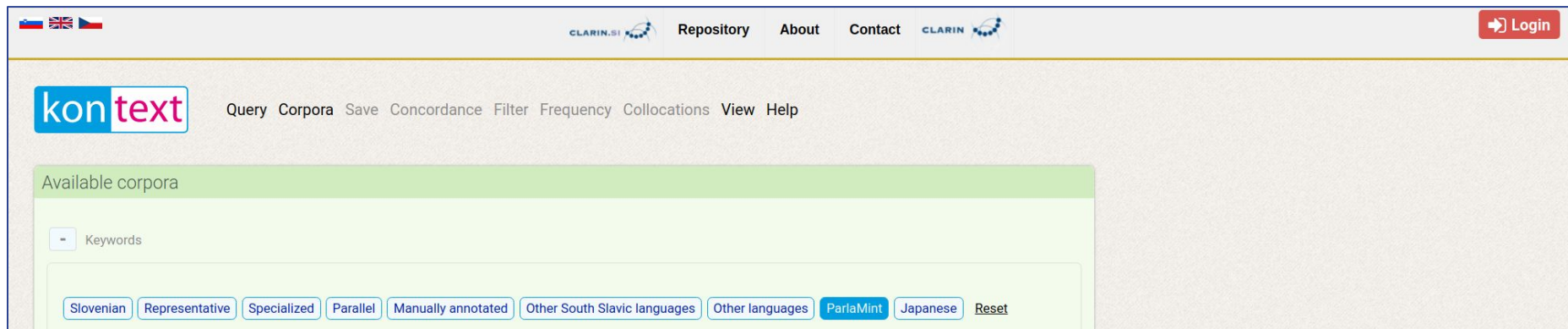
<https://ufal.mff.cuni.cz/~hladka/2022/docs/s-fr.csv>

2. How did the overall frequency of the text change over time?



ParlaMint 2.1 in KonText

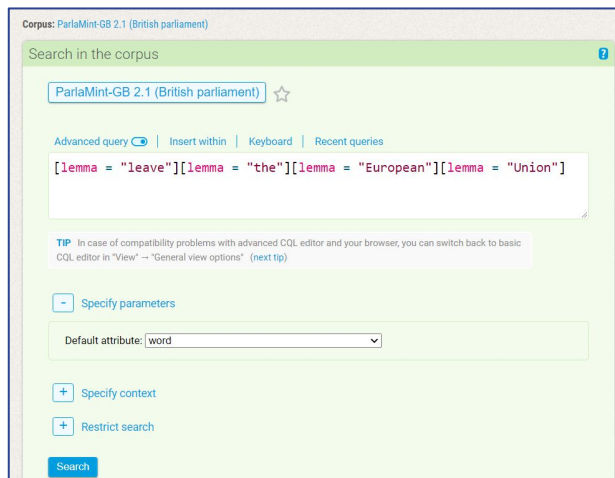
<https://www.clarin.si/kontext/corpora/corplist>



The screenshot shows the KonText web interface. At the top, there is a navigation bar with flags on the left, the CLARIN.SI logo, and links for Repository, About, and Contact. A red Login button is on the right. Below the navigation bar, the KonText logo is on the left, and a menu with links for Query, Corpora, Save, Concordance, Filter, Frequency, Collocations, View, and Help is on the right. The main content area is titled "Available corpora" and features a search box labeled "Keywords". Below the search box, there is a row of buttons for filtering corpora: Slovenian, Representative, Specialized, Parallel, Manually annotated, Other South Slavic languages, Other languages, ParlaMint (highlighted in blue), and Japanese. A Reset button is located at the end of this row.

KonText search in ParlaMint-GB 2.1

https://www.clarin.si/kontext/query?corpname=parlamint21_gb



Corpus: ParlaMint-GB 2.1 (British parliament)

Search in the corpus

ParlaMint-GB 2.1 (British parliament) ☆

Advanced query ☒ | Insert within | Keyboard | Recent queries

[lemma = "leave"][lemma = "the"][lemma = "European"][lemma = "Union"]

TIP In case of compatibility problems with advanced CQL editor and your browser, you can switch back to basic CQL editor in "View" → "General view options" (next tip)

☐ Specify parameters

Default attribute: word

☐ Specify context

☐ Restrict search

Search

- Switch Advanced query ON
 - Enter query in CQL (see Lecture #5 [video](#) from 00:24:37)
- [lemma = "leave"][lemma = "the"][lemma = "European"][lemma = "Union"]

KonText search in ParlaMint-GB 2.1

kon text

Query Corpora Save Concordance Filter Frequency Collocations View Help

Corpus: ParlaMint-GB 2.1 (British parliament) | Query: leave, the, European, Union (8,228 hits)

Hits: 8,228 | i.p.m.: 72.8 (related to the whole corpus) | ARF: 2,829.79 | Result is sorted

1 / 206 ►►

Line selection: simple ▾

☐	2019-04-08 + Brokenshire, James Peter	been preparing for a range of issues . As we	leave the European Union	, changes to regulations might be required and training and
☐	2019-04-08 + Brokenshire, James Peter	have a smooth transition from where we are today to	leaving the European Union	. The hon. Member for North Wiltshire (James Gray
☐	2019-04-08 + Leadsom, Andrea Jacqueline	House can support as a way to ensure that we	leave the European Union	in very short order . However , if the talks
☐	2019-04-08 + Leadsom, Andrea Jacqueline	slightly different way forward . It remains our intention to	leave the European Union	with a deal that both means we leave in line
☐	2019-04-08 + Leadsom, Andrea Jacqueline	was held in 2016 and enable the United Kingdom to	leave the European Union	in a way that would ensure that we met the
☐	2019-04-08 + Leadsom, Andrea Jacqueline	Friend is correct : the legal date for us to	leave the European Union	is indeed this Friday , 12 April . However ,

Settings

kon text Query Corpora Save Concordance Filter Frequency Collocations **View** Help

Corpus: [ParlaMint-GB 2.1 \(British parliament\)](#) | Query: [leave, the, European, Union](#) (8,228 hits)

Hits: **8,228** | [i.p.m.:](#) **72.8** (related to the whole corpus) | [ARF:](#) **2,829.79** | Result is sorted

Line selection: [simple](#) ▾

☐	2019-04-08 + Brokenshire, James Peter	been preparing for a range of issues . As we
☐	2019-04-08 + Brokenshire, James Peter	have a smooth transition from where we are today to
☐	2019-04-08 + Leadsom, Andrea Jacqueline	House can support as a way to ensure that we
☐	2019-04-08 + Leadsom, Andrea Jacqueline	slightly different way forward . It remains our intention to
☐	2019-04-08 + Leadsom, Andrea Jacqueline	was held in 2016 and enable the United Kingdom to
☐	2019-04-08 + Leadsom, Andrea Jacqueline	Friend is correct : the legal date for us to

Corpus-specific settings for ParlaMint-GB 2.1 (British parliament)

[Positional attributes](#) [Structures](#) [References](#) [Additional functions](#)

☐ <#> ☐ Token number	☒ <speech> ☐ Document number ☐ Text ID ☐ speech.title ☐ speech.subcorpus ☐ speech.house ☐ speech.term ☐ speech.session ☐ speech.meeting ☐ speech.sitting ☐ speech.agenda ☒ speech.from ☐ speech.to ☐ speech.wordcount ☐ Speech ID ☐ speech.speaker_id ☒ speech.speaker_name ☐ speech.speaker_role ☐ speech.speaker_type ☐ speech.speaker_party ☐ speech.speaker_party_name ☐ speech.party_status ☐ speech.speaker_gender ☐ speech.speaker_birth	☐ <p> ☐ p.id ☐ p.lang	☐ <s> ☐ Sentence ID	☐ <name> ☐ name.type
☐ Select all	☐ Select all	☐ Select all	☐ Select all	☐ Select all

KonText search in ParlaMint-GB 2.1


Query Corpora Save Concordance Filter Frequency Collocations View Help

Corpus: ParlaMint-GB 2.1 (British parliament) | Query: leave, the, European, Union (8,228 hits)

Hits: 8,228 | i.p.m.: 72.8 (related to the whole corpus) | ARF: 2,829.79 | Result is sorted

Line selection: simple ▾

☐	2019-04-08 + Brokenshire, James Peter	been preparing for a range of issues . As we leave the European Union , changes to regulations might be required and training and
☐	2019-04-08 + Brokenshire, James Peter	have a smooth transition from where we are today to leaving the European Union . The hon. Member for North Wiltshire (James Gray
☐	2019-04-08 + Leadsom, Andrea Jacqueline	House can support as a way to ensure that we leave the European Union in very short order . However , if the talks
☐	2019-04-08 + Leadsom, Andrea Jacqueline	slightly different way forward . It remains our intention to leave the European Union with a deal that both means we leave in line
☐	2019-04-08 + Leadsom, Andrea Jacqueline	was held in 2016 and enable the United Kingdom to leave the European Union in a way that would ensure that we met the
☐	2019-04-08 + Leadsom, Andrea Jacqueline	is indeed this Friday , 12 April . However ,

◀ and why we still have £ 10 million available for any immediate pressures that may emerge in the forthcoming year . What is all this money for ? Councils have been preparing for a range of issues . As we leave the European Union , changes to regulations might be required and training and support might be needed , as well as contingency planning so that we have a smooth transition from where we are today to leaving the European Union . The hon. ▶

KonText search in ParlaMint-GB 2.1


Query Corpora Save Concordance Filter Frequency Collocations View Help

Corpus: ParlaMint-GB 2.1 (British parliament) | Query: leave, the, European, Union (8,228 hits)

Hits: 8,228 | i.p.m.: 72.8 (related to the whole corpus) | ARF: 2,829.79 | Result is sorted

Line selection: simple

Speech ID	ParlaMint-GB_2019-04-08-commons.u6	speech.title	Minutes of the House of Commons, Daily Session 2019-04-08
speech.from	2019-04-08	speech.to	2019-04-08
speech.subcorpus	Reference	speech.house	Lower house
speech.term	57	speech.speaker_name	Brokenshire, James Peter
speech.speaker_gender	M	speech.speaker_role	Regular
speech.speaker_type	MP	speech.speaker_party	CON
speech.speaker_party_name	Conservative		

been preparing for a range of issues . As we **leave the European Union** , changes to regulations might be required and training and

for North Wiltshire (James Gray

However , if the talks

means we leave in line

ensure that we met the

, 12 April . However ,

Save the results <https://ufal.mff.cuni.cz/~hladka/2022/docs/gb.csv>

 Query Corpora **Save** Concordance Filter Frequency Collocations View Help

Corpus: ParlaMint-GB 2.1 (British parliament) | Query: leave, the, European, Union (8,228 hits)

Hits: 8,228 | i.p.m.: 72.8 (related to the whole corpus) | ARF: 2,829.79 | Result is sorted

Line selection: simple ▾

☐	2019-04-08 + Brokenshire, James Peter	been preparing for a range of issues . As we	leave the European Union	, changes to regulations might be required and training and
☐	2019-04-08 + Brokenshire, James Peter	have a smooth transition from where we are today to	leaving the European Union	. The hon. Member for North Wiltshire (James Gray
☐	2019-04-08 + Leadsom, Andrea Jacqueline	House can support as a way to ensure that we	leave the European Union	in very short order . However , if the talks
☐	2019-04-08 + Leadsom, Andrea Jacqueline	slightly different way forward . It remains our intention to	leave the European Union	with a deal that both means we leave in line
☐	2019-04-08 + Leadsom, Andrea Jacqueline	was held in 2016 and enable the United Kingdom to	leave the European Union	in a way that would ensure that we met the
☐	2019-04-08 + Leadsom, Andrea Jacqueline	Friend is correct : the legal date for us to	leave the European Union	is indeed this Friday , 12 April . However ,

Web www.clarin.si/fiká
One-click saving of data to the csv format saves at most 10,000 lines.
OK?

OK Zrušit

Processing using the R system

- install R <https://www.r-project.org/>
- install RStudio <https://www.rstudio.com/>
- download <https://ufal.mff.cuni.cz/~hladka/2022/docs/gbeu.R>, run it

Further reading

- Erjavec T. et al. *Multilingual comparable corpora of parliamentary debates ParlaMint 2.1*. Slovenian language resource repository CLARIN.SI. 2021. <http://hdl.handle.net/11356/1432>
- Erjavec, T., Ogrodniczuk, M., Osenova, P. et al. The ParlaMint corpora of parliamentary proceedings. *Lang Resources & Evaluation*. 2022. <https://doi.org/10.1007/s10579-021-09574-0>
- Hladká, B., Kopp, M., & Straňák, P. Compiling Czech parliamentary stenographic protocols into a corpus. In *Proceedings of the LREC 2020 Workshop on Creating, Using and Linking of Parliamentary Corpora with Other Types of Political Discourse (ParlaCLARIN II)* (pp. 18–22). European Language Resources Association (ELRA). 2020. [url](#).