ParlaMint:
Towards Comparable Parliamentary Corpora

Maciej Ogrodniczuk Michał Rudolf	Instytut Podstaw Informatyki PAN
Marta Kołczyńska	Instytut Studiów Politycznych PAN

Seminarium ZIL

7 czerwca 2021

Źródło: Korpus Dyskursu Parlamentarnego

W pigułce:

- dane z Sejmu i Senatu od roku 1919 do chwili obecnej
- różnej jakości (także po OCR-ze i korekcie)
- zbierane w ramach kilku projektów (CESAR, MARCELL, CLARIN-PL 2 i 3, CLARIN-PL-Biz)
- 800 mln segmentów
- trzy typy dokumentów:
  1. stenogramy z posiedzeń plenarnych
  2. sprawozdania z posiedzeń komisji
  3. interpelacje, zapytania i odpowiedzi
- automatycznie anotowane lingwistyczne (ujednoznaczniona morfoskładnia, nazwy własne, rozbiory zależnościowe)
- centrum dowodzenia: `http://clip.ipipan.waw.pl/PPC`

Podobne inicjatywy (z CLARIN-em w tle)

Efekt kuli śnieżnej:

- CLARIN Traveling Campus 'Talk of Europe' (`http://www.talkofeurope.eu/`)
- warsztat CLARIN-PLUS: Working with parliamentary records (`https://www.clarin.eu/event/2017/clarin-plus-workshop-working-parliamentary-records`)
- warsztat ParlaCLARIN na konferencji LREC 2018 i 2020 (`https://www.clarin.eu/ParlaCLARIN`, `https://www.clarin.eu/ParlaCLARIN-II`)
- warsztat ParlaFormat (`https://www.clarin.eu/event/2019/parlaformat-workshop`)
- Parliamentary corpora (`https://www.clarin.eu/content/parliamentary-corpora`)
- wniosek o akcję COST Parlant...

ParlaMint

Niespodziewany miniprojekt:

- **sponsor**: CLARIN ERIC
- **motywacja**: co mówi się w parlamentach narodowych o pandemii?
- **cel**: stworzenie pierwszego wielojęzycznego porównywalnego korpusu parlamentarnego
- **metoda**:
  - dwa podkorpusy: referencyjny (2015–listopad 2019) i „COVID-owy" (listopad 2019–2020)
  - wspólne metadane, struktura korpusu, anotacja lingwistyczna
  - udostępnienie w konkordancerach noSketch Engine i KonText, serwisie Parlameter, stworzenie scenariuszy demonstracyjnych
- **główny link**:
  `https://www.clarin.eu/content/parlamint`

ParlaMint: dwie fazy

Faza 1: lipiec–wrzesień 2020

- seria próbna dla 4 języków: BG, HR, PL, SI
- przetestowanie formatu zapisu
- automatyczna anotacja lingwistyczna (morfoskładnia, lematy, nazwy własne, drzewa zależnościowe)
- udostępnienie korpusów w konkordancerach i serwisie Parlameter

Faza 2: październik 2020 – maj 2021

- mini-granty na dodanie kilku języków
- demonstratory na bazie danych korpusowych

ParlaMint a Parla-CLARIN

Cała infrastruktura:

- Parla-CLARIN: schemat TEI dla posiedzeń parlamentarnych (0.2):
  - utrzymywany na GitHubie: `https://github.com/clarin-eric/parla-clarin/`
  - dość dokładnie udokumentowany: `https://clarin-eric.github.io/parla-clarin/`
- ParlaMint: infrastruktura oparta o format Parla-CLARIN:
  - schematy walidacyjne
  - skrypty XSLT do dodatkowej walidacji, konwersji do formatów obsługiwanych przez konkordancery

Wytyczne ParlaMint

Format reprezentacji danych:

- struktura: kadencje, izby, posiedzenia, wypowiedzi
- mówcy: imię i nazwisko, płeć, data/miejsce urodzenia/śmierci, wykształcenie, przynależność partyjna, linki do zasobów zewnętrznych (Wikipedia)
- partie: nazwa
- wypowiedzi: data, rola mówcy (marszałek, poseł/senator, gość), zdarzenia (aplauz...)

Dane korpusu ParlaMint

Część polska:

Daty	Podkorpus	Segmentów
2015-11-12 – 2019-10-31	referencyjny	29 275 522
2019-11-01 – 2020-08-18	COVID-owy	3 878 873
		33 154 395

https://www.clarin.si/noske/parlamint.cgi/corp_info?corpname=parlamint20_pl

Anotacja lingwistyczna

Przejęta z KDP:

- segmentacja, tokenizacja, lematyzacja, tagowanie (Morfeusz2 + Concraft2)
- nazwy własne: Liner2
  - geogName → LOC
  - orgName → ORG
  - persName → PER, bez podtypów (forename, surname, addName)
  - placeName → LOC, bez podtypów (district, settlement, region, country, bloc)
- struktury zależnościowe: COMBO

Co nowego w stosunku do KDP?

Zdarzenia:

Zdarzenie	Typ	Wskaźnik
note	vote	głosowanie nr
	time	przerwa, początek, koniec, wznowienie
	debate	na posiedzeniu, przewodnictwo, chwila ciszy, chwila przerwy
kinesic	applause	oklaski
	ringing	dzwonek, sygnał telefonu
	laughter	wesołość, śmiech
	signal	uderza laską, pokazuje
	playback	wyświetla, odtwarza, projekcja

Co nowego w stosunku do KDP?

Zdarzenia:

Zdarzenie	Typ	Wskaźnik
vocal	shouting	skanduje
	noise	gwar, rozmowy na sali, poruszenie, uderza w pulpity
gap	reason: inaudible	poza mikrofonem, w tle, zakłócenia wypowiedzi, poza nagraniem, brak nagrania
incident	entering	wchodzi, przybywa, przybycie
	leaving	wychodzą
	action	wstają, włącza, wyłącza, wręcza, otrzymuje, odczytuje, trzyma, prezentuje, podaje, składa gratulacje

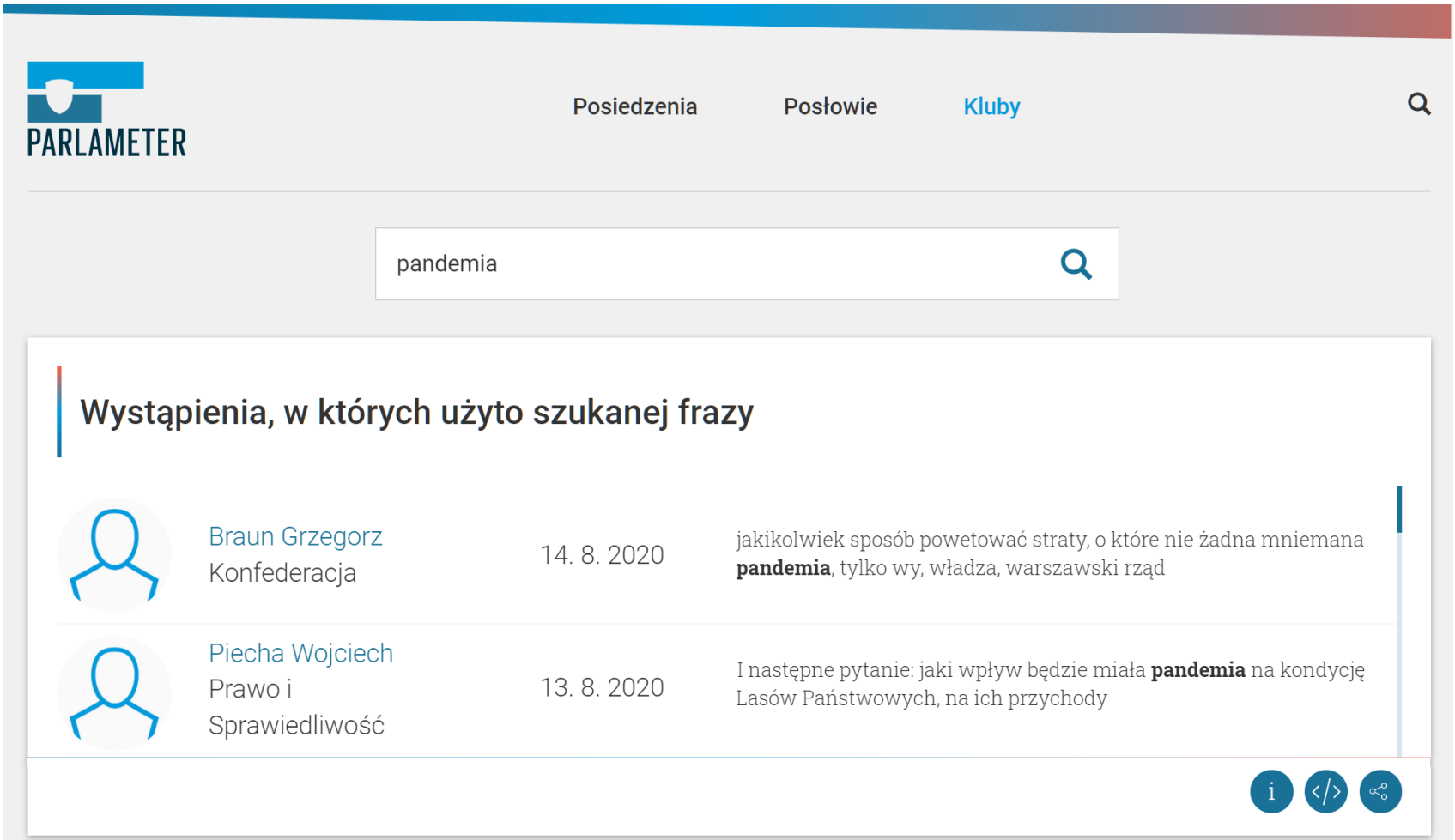
NoSketch Engine

	COVID		*the rest of the corpus*		
lemma_lc	frequency	frequency/mill	frequency	frequency/mill	Score
pandemia	727	187.5	1	0.0	182.3
koronawirusa	543	140.1	0	0.0	141.1
kukiz15	455	117.4	0	0.0	118.4
covid-19	428	110.4	0	0.0	111.4
koronawirusem	408	105.2	0	0.0	106.2
epidemia	1,196	308.5	116	4.0	62.4
maseczka	241	62.2	2	0.1	59.1
kwarantanna	251	64.7	5	0.2	56.1
zdalnie	311	80.2	20	0.7	48.2
tarcza	977	252.0	140	4.8	43.7
covid	163	42.0	0	0.0	43.0
antykryzysowy	324	83.6	29	1.0	42.5
zdalny	292	75.3	24	0.8	41.9

https://www.clarin.si/noske/parlamint.cgi/

https://www.clarin.si/kontext/

ParlaMint: ParlaMeter



https://parlamint.parlameter.org

Faza 2: więcej języków

Kto się zgłosił?

Jesse de Does	Belgia (holenderski/francuski)
Barbora Hladká	Czechy
Costanza Navarretta	Dania
Giancarlo Luxardo	Francja
Ruben van Heusden	Holandia
Steinór Steingrímsson	Islandia
Tomas Krilavičius	Litwa
Robers Dargis	Łotwa
Petru Rebeja	Rumunia
Çağrı Çöltekin	Turcja
Paul Rayson	Wielka Brytania
Miklós Sebők	Węgry
Giulia Venturi	Włochy
María Calzada Pérez	Hiszpania

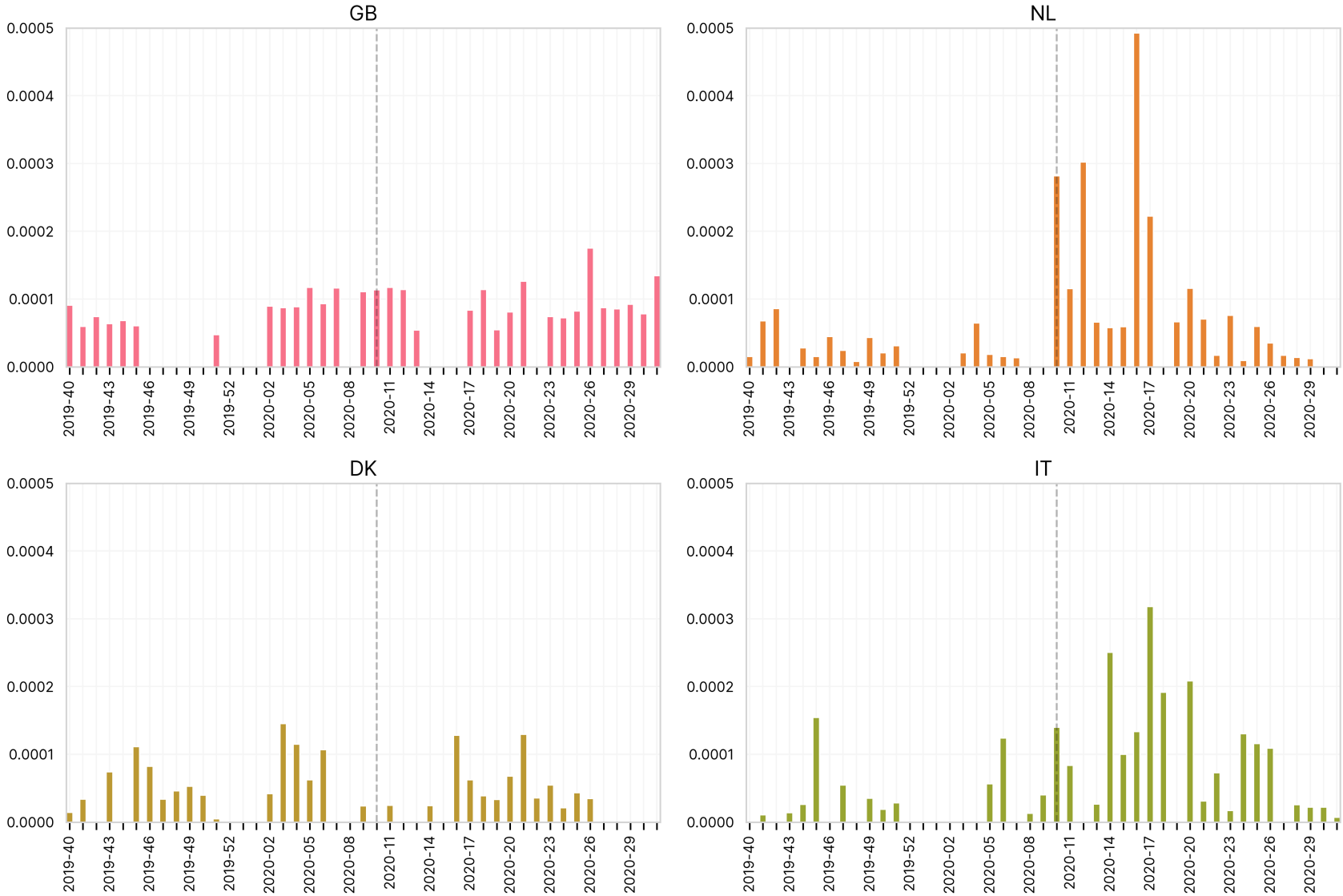
Showcase Rubena Rosa

Pytanie: Czy zaszło zjawisko 'skupienia wokół flagi'?

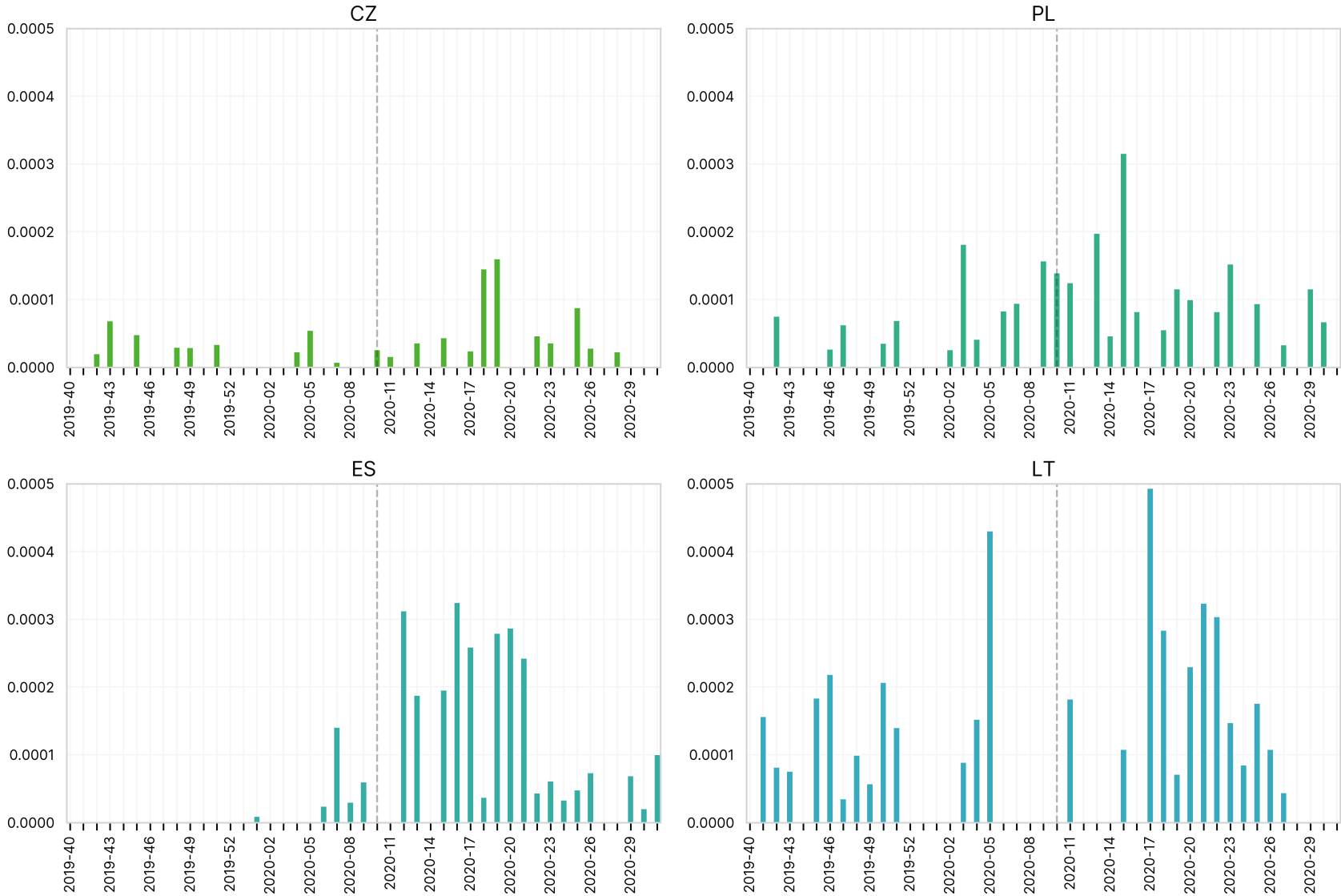
- słowa kluczowe: czy jest więcej odwołań do ekspertów? w czyich wypowiedziach – koalicji czy opozycji?
- TF/IDF: o czym mówiono najczęściej?
- kolokacje: z czym łączyły się interesujące nas hasła?

`https://github.com/rubenros1795/ParlaMintCase/blob/main/blog.md`

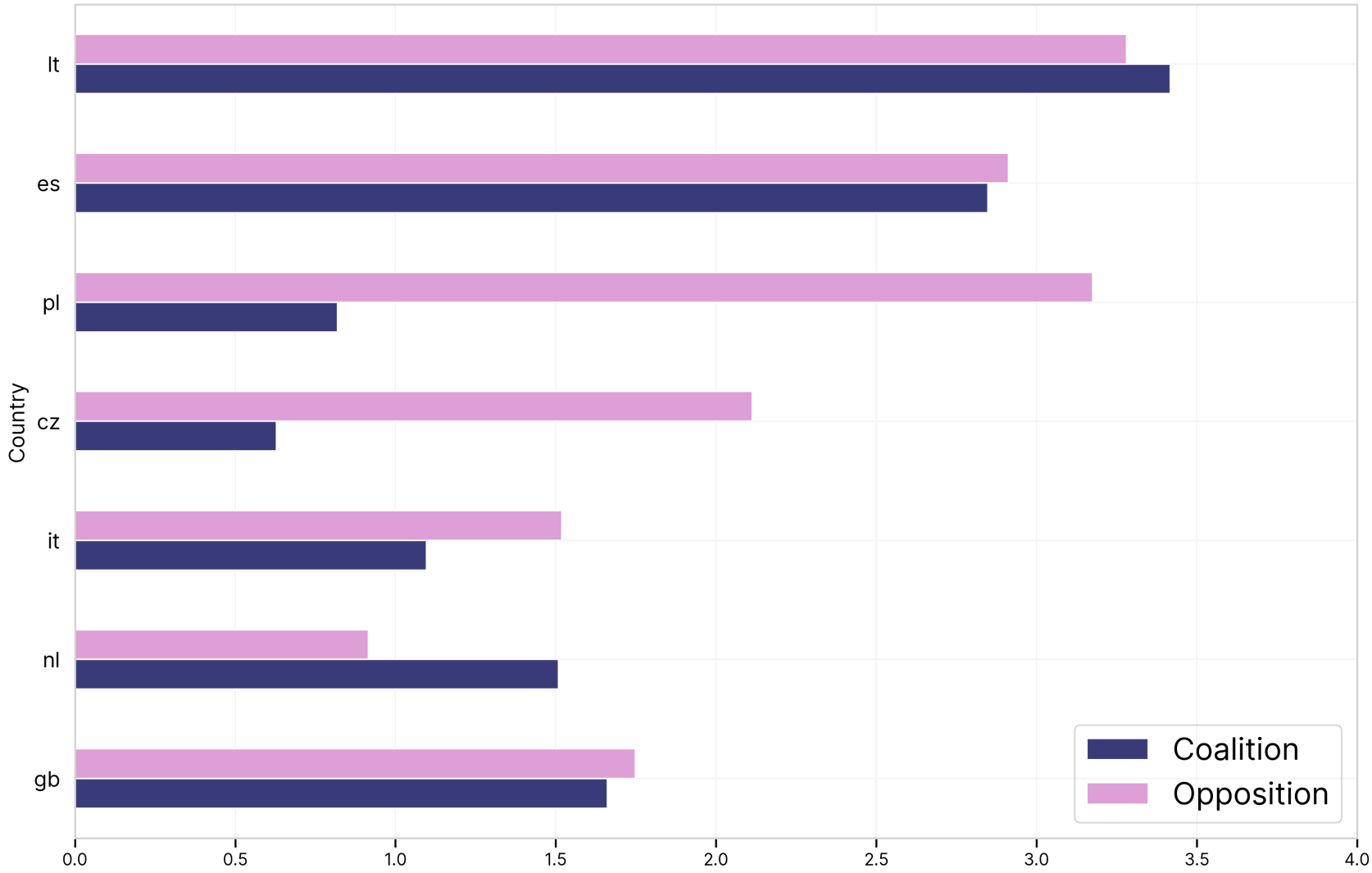
Czy więcej mówimy o ekspertach?



Czy więcej mówimy o ekspertach?



Kto mówi więcej o ekspertach?



Helsiński Hackathon DH 2021

Seria pytań badawczych:

- Czym różnią się wystąpienia na temat COVID-u od pozostałych?
- Jakie tematy pojawiają się w debatach COVID-owych? Które tematy są wspólne dla wszystkich krajów, a które są specyficzne dla danego kraju?
- Czy debaty wskazują na jakieś istotne zmiany w zakresie tematów lub priorytetów w czasie?
- Jaka jest częstotliwość debat związanych z COVID na przestrzeni czasu i czy istnieje jakiś związek między debatami a liczbą przypadków?

Polski akcent: Marta Kołczyńska

`https://dhhackathon.wordpress.com/2021/05/28/parliamentary-debates-in-the-covid-times/`

Co dalej?

- CLARIN Café: ParlaMint Unleashed (28 czerwca)
- ParlaMint 2:
  - więcej metadanych
  - więcej korpusów (austriacki, baskijski, duński, estoński, fiński, grecki, kataloński, niemiecki, norweski, portugalski, szwedzki)
  - tłumaczenie maszynowe na angielski
  - multimodalność
  - DHH 2022
- artykuły: CLARIN Conference, LRE, Gender Studies, ...
- nowy wniosek o akcję COST?
- ParlaMint 3:
  - EuroParl
  - parlamenty lokalne
  - języki pozaeuropejskie
  - zadanie ewaluacyjne (dedukcja afiliacji partyjnej)
  - ...