ParlaCLARIN Workshop: Creating and Using Parliamentary Corpora

Miyazaki, 7 May 2018

The Polish Parliamentary Corpus

Maciej Ogrodniczuk

Linguistic Engineering Group
Institute of Computer Science
Polish Academy of Sciences

ParlaCLARIN Worskhop, LREC 2018

Miyazaki, May 7, 2018

The Polish Parliamentary Corpus

In a nutshell:

- a 300M-token collection of linguistically annotated documents from the proceedings of Polish Parliament (Sejm and Senate)
- based on the Polish Sejm Corpus prepared in October 2011 and recently extended in CLARIN-PL
- with new data updated live
- available at: `http://clip.ipipan.waw.pl/PPC`

Corpus data

Data sources:

- 1993–: editable PDFs, other formats available (internally)
- 1989–93: scanned paper transcripts (OCR-ed and manually verified)
- 1919–88: high-quality scans prepared by the Sejm library, manually cleaned, then:
  - human-assisted OCR-ed (area and typo corrections)
  - structure corrected and errors detected using custom scripts

Corpus data

Three political periods:

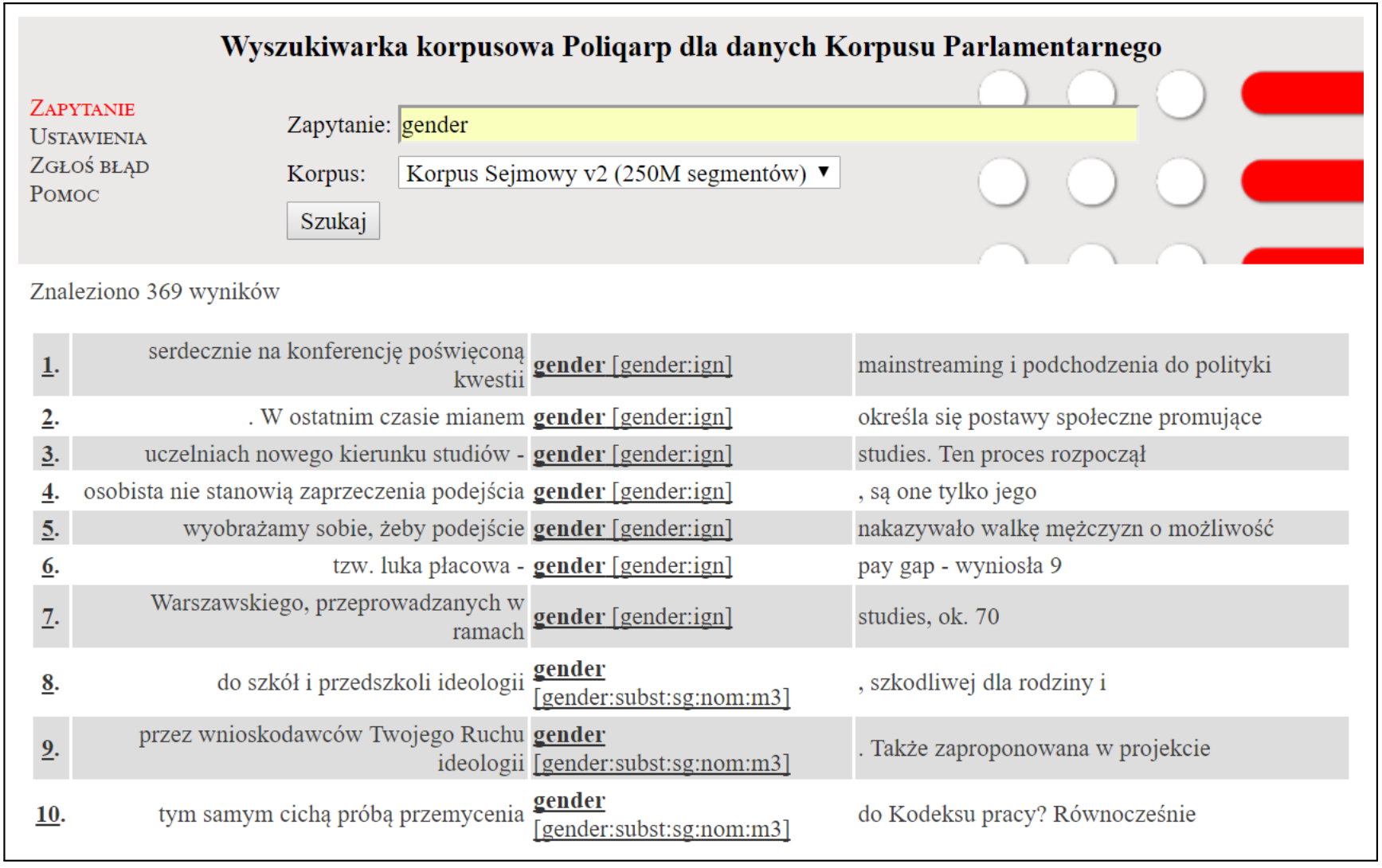
- Second Polish Republic (1919–1939) — 1379 sittings
- People's Poland period (1943–1989) — 481 sittings
  (the Senate abolished by the authorities)
- Third Polish Republic (1989–today) — 1407 sittings.

Corpus data

Corpus format and structure:

- stand-off TEI P5 annotation
- generated with in-house tools:
  - Morfeusz SGJP (text structure, utterance-level segmentation, tokenization, lemmatization)
  - Pantera (disambiguated morphosyntactic description)
  - Spejd (syntactic words and syntactic groups)
  - Nerf (named entities)
  - Corneference (coreference).

Corpus search: Poliqarp



Wyszukiwarka korpusowa Poliqarp dla danych Korpusu Parlamentarnego

ZAPYTANIE
USTAWIENIA
ZGŁOŚ BŁĄD
POMOC

Zapytanie: gender

Korpus: Korpus Sejmowy v2 (250M segmentów)

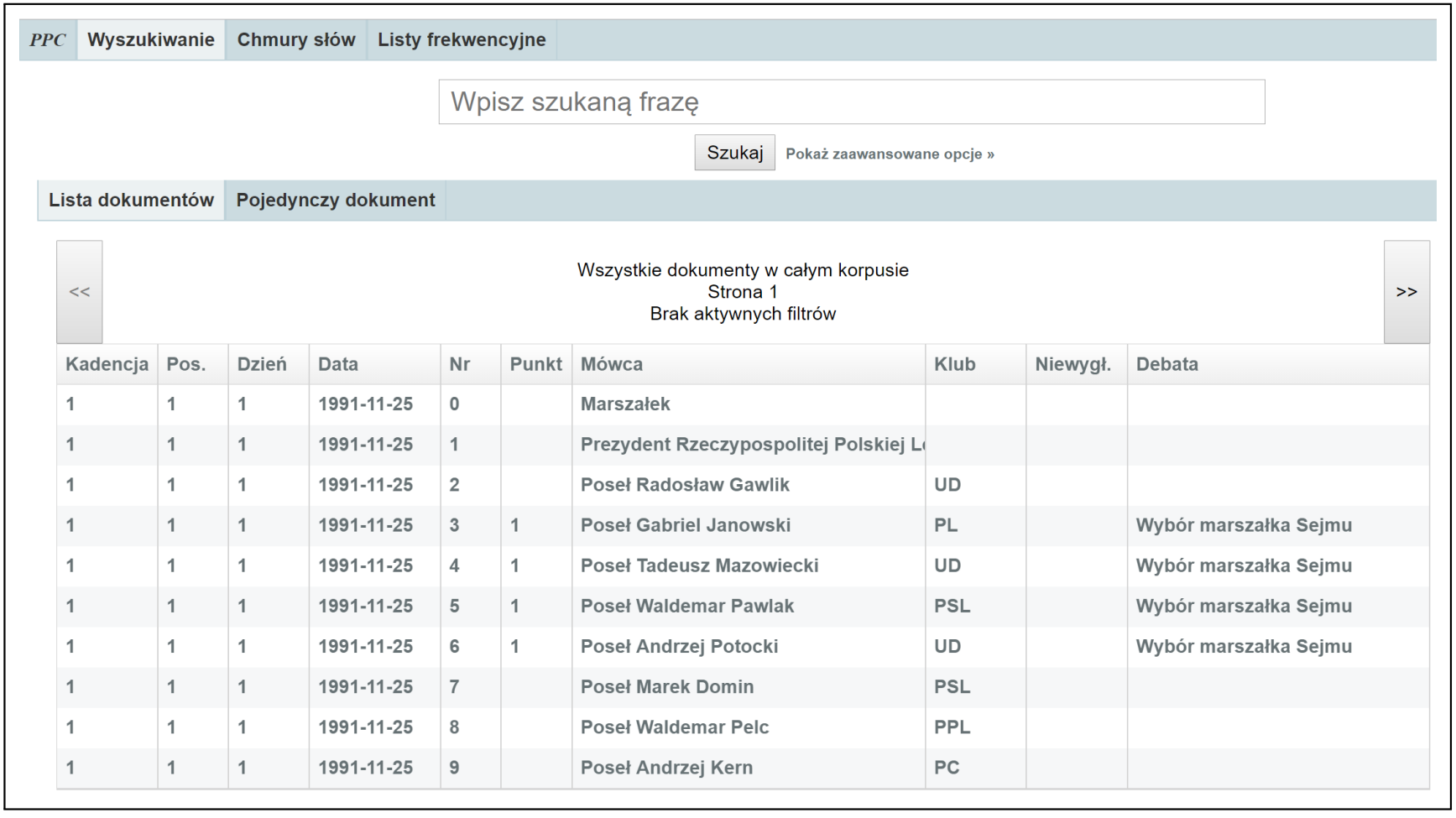
Szukaj

Znaleziono 369 wyników

1.	serdecznie na konferencję poświęconą kwestii	**gender** [gender:ign]	mainstreaming i podchodzenia do polityki
2.	. W ostatnim czasie mianem	**gender** [gender:ign]	określa się postawy społeczne promujące
3.	uczelniach nowego kierunku studiów -	**gender** [gender:ign]	studies. Ten proces rozpoczął
4.	osobista nie stanowią zaprzeczenia podejścia	**gender** [gender:ign]	, są one tylko jego
5.	wyobrażamy sobie, żeby podejście	**gender** [gender:ign]	nakazywało walkę mężczyzn o możliwość
6.	tzw. luka płacowa -	**gender** [gender:ign]	pay gap - wyniosła 9
7.	Warszawskiego, przeprowadzanych w ramach	**gender** [gender:ign]	studies, ok. 70
8.	do szkół i przedszkoli ideologii	**gender** [gender:subst:sg:nom:m3]	, szkodliwej dla rodziny i
9.	przez wnioskodawców Twojego Ruchu ideologii	**gender** [gender:subst:sg:nom:m3]	. Także zaproponowana w projekcie
10.	tym samym cichą próbą przemycenia	**gender** [gender:subst:sg:nom:m3]	do Kodeksu pracy? Równocześnie

http://sejm.nlp.ipipan.waw.pl/

Corpus search: Smyrna



PPC | Wyszukiwanie | Chmury słów | Listy frekwencyjne

Wpisz szukaną frazę

Szukaj Pokaż zaawansowane opcje »

Lista dokumentów | Pojedynczy dokument

<<

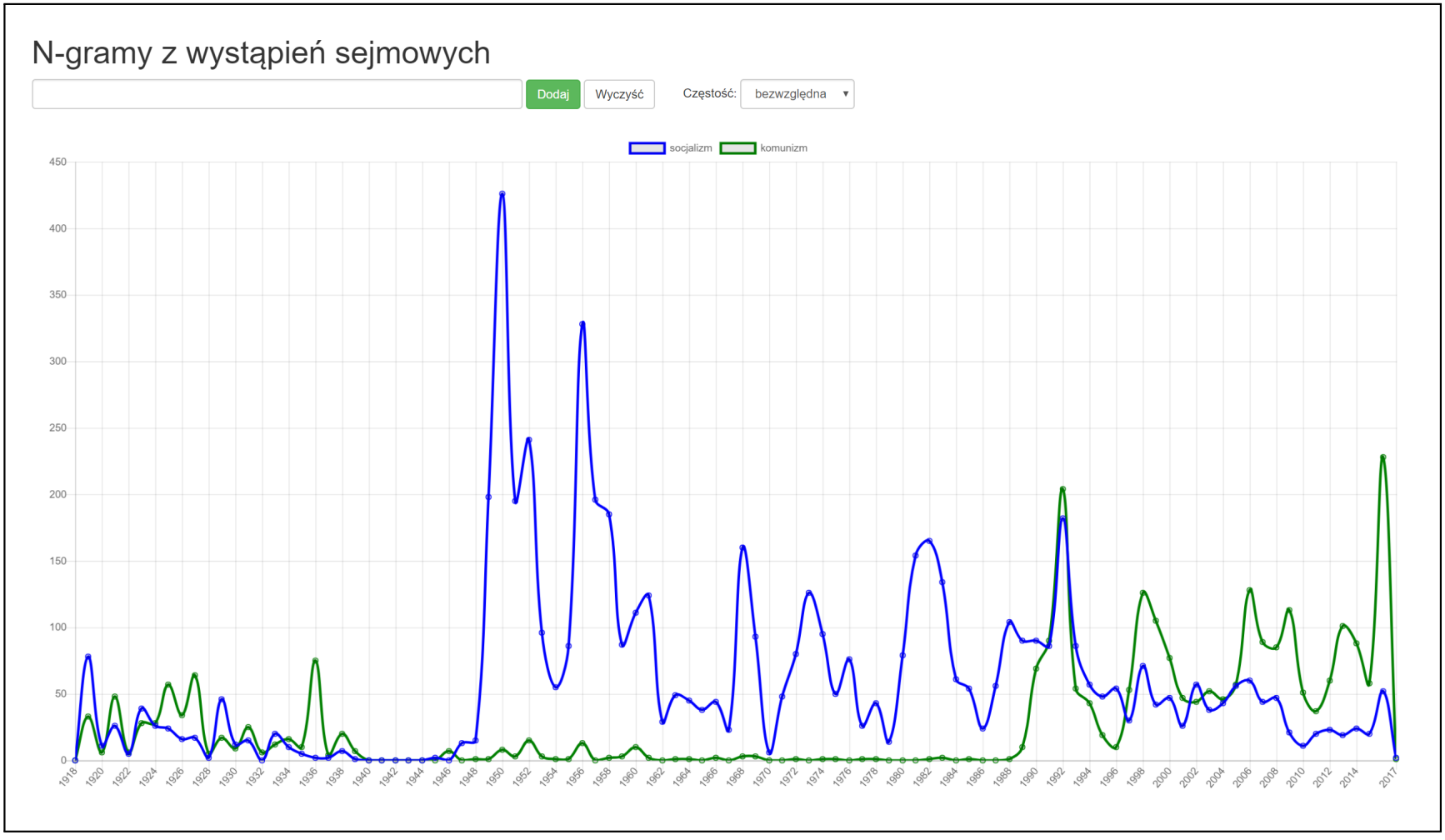
Wszystkie dokumenty w całym korpusie
Strona 1
Brak aktywnych filtrów

>>

Kadencja	Pos.	Dzień	Data	Nr	Punkt	Mówca	Klub	Niewygł.	Debata
1	1	1	1991-11-25	0		Marszałek			
1	1	1	1991-11-25	1		Prezydent Rzeczypospolitej Polskiej L			
1	1	1	1991-11-25	2		Poseł Radosław Gawlik	UD		
1	1	1	1991-11-25	3	1	Poseł Gabriel Janowski	PL		Wybór marszałka Sejmu
1	1	1	1991-11-25	4	1	Poseł Tadeusz Mazowiecki	UD		Wybór marszałka Sejmu
1	1	1	1991-11-25	5	1	Poseł Waldemar Pawlak	PSL		Wybór marszałka Sejmu
1	1	1	1991-11-25	6	1	Poseł Andrzej Potocki	UD		Wybór marszałka Sejmu
1	1	1	1991-11-25	7		Poseł Marek Domin	PSL		
1	1	1	1991-11-25	8		Poseł Waldemar Pelc	PPL		
1	1	1	1991-11-25	9		Poseł Andrzej Kern	PC		

http://smyrna.nlp.ipipan.waw.pl/

Ngram viewer



http://ngram.sejm.nlp.ipipan.waw.pl/

What's currently happening?

CLARIN-PL activities:

- processing data with the newest analytic tools
- better search and presentation (Korpusomat/MTAS), e.g.
  - statistical queries
  - frequency lists
  - graphical summaries

Thank you!

And several funding institutions:

- a European (CIP ICT-PSP) project CESAR: Central and South-East European Resources (grant agreement 271022), part of META-NET
- part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education (DIR/WK/2016/02 and DIR/WK/2018/01)