A multi-scale, traceable analysis of Slovenian parliamentary discourse 1992–2022

Every claim in this thesis can be walked back to the speeches it rests on, on CLARIN.SI infrastructure, end to end.

David Bordon
CLARIN.SI PhD representative
david.bordon@ff.uni-lj.si
davidbordon.com/academia
↑ Poster and abstract, to download

CLARIN
UNIVERSITY OF LJUBLJANA | FF Faculty of Arts

CLARIN IN / OUT

The chain is built to re-run on any ParlaMint corpus.

USES
siParl 4.0 deposit (hdl 11356/1936). Annotation layer joined by Speech-ID. CLARIN.SI concordancer for CQL, collocation and concordance. Slovenian LT Overview. ParlaCAP classifier. ParlaMint comparability frame.

TO RETURN
To the siParl maintainers once the thesis is finished: interpellation session metadata, corrections to mislabelled speeches, and body taxonomy behaviour with the exclusion-based CQL workaround.

PLANNED DEPOSIT
Embeddings and the topic-modelling pipeline, documented for re-running on ParlaMint corpora. Plenary speech table, session- and date-ordered (CSV and XLSX) under CC BY 4.0. CQL query register and term lists. ParlaCAP labels extended back to 1990. Traceability reader: topic → terms → query → hits, on one screen.

NEGATIVE FINDING
No monolingual Slovene sentence encoder exists. `xlm-r-parla` was tested as an embedding model and was not suitable. BERTopic's default encoder reads a median 29 % of a speech (due to truncation).

RQ1
MACRO
213 372
PLENARY SPEECHES

01 / TOPIC MODELLING
siParl 4.0 CoNLL-U
LT Overview
ParlaCAP

RQ1 – what the Assembly discusses

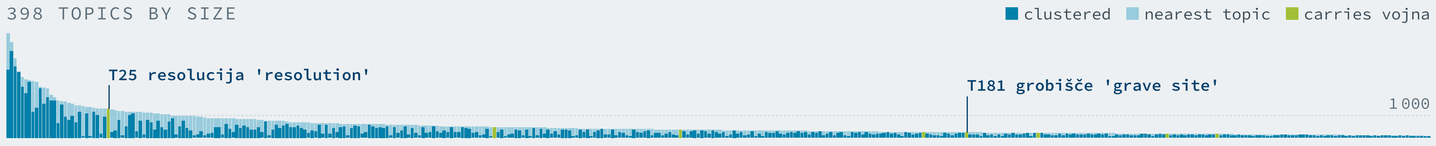
213 372 SPEECHES → 398 TOPICS
FINAL RUN

QUESTION
What does parliament discuss, how does its attention shift, and how do small, niche parties engage relative to mainstream ones?

METHOD
BERTopic over all 213 372 speeches in one topic space; parties and terms are read from it. Qwen3-Embedding-0.6B reads each speech whole. A minimum cluster size of 75 gives 398 topics.

WHAT IT YIELDS
A ranked term list per topic – an inspectable object, not a score. External check against the ParlaCAP topic classifier (CLASSLA) on 10 049 held-out speeches: 70 % agreement where the clustering placed the speech, 50 % where it was placed afterwards; the majority baseline is 14 %.

398 TOPICS BY SIZE



WHAT IS NOT CLAIMED
The 398 topics are not yet coded to CAP, so no attention share by party or term is shown. Two speeches in five are placed by nearest topic, and every share will be read with and without them.

RQ2
MESO
2 059
CONCORDANCE LINES

02–04 / CORPUS QUERYING
CLARIN.SI concordancer
collocation · CQL
concordance export

RQ2 – how the contested past is mobilised, traced term by term

398 TOPICS → ONE TERM → 2 059 LINES
WORKED EXAMPLE BELOW

01 TOPIC MODELLING ▲
siParl 4.0 CoNLL-U, annotation as distributed · the ParlaCAP classifier as the external check

A topic surfaces a term list. Nothing is claimed from it.

Established method. The chain starts where topic modelling usually stops.

02 TERM LIST
CLARIN.SI concordancer (NoSketchEngine) – collocation on siParl 4.0

Domain-salient seeds expanded by collocation snowball. Windows ±1 and ±5; LogDice primary – ≥7 strong, ≥6 retained; MI only at frequency ≥10; T-score discarded.

Keyness and raw frequency are not used to choose terms – they surface procedural boilerplate. Frequency traces a term once chosen.

03 CQL QUERY
CLARIN.SI concordancer – CQL

Plenary-only by *exclusion* of working bodies. Prompted / unprompted split by session date, verified against corpus dates rather than PISRS milestones.

The query is printed, not described. Register: 66 excluded dates, 72 excluded bodies.

04 CONCORDANCE LINES
CLARIN.SI concordancer – concordance export, joined by Speech-ID

Concordance export with date, party and speaker role. *Prompted* – clustering around a legislative event. *Unprompted* – independent of any trigger.

Lines by party and period, each one addressable.

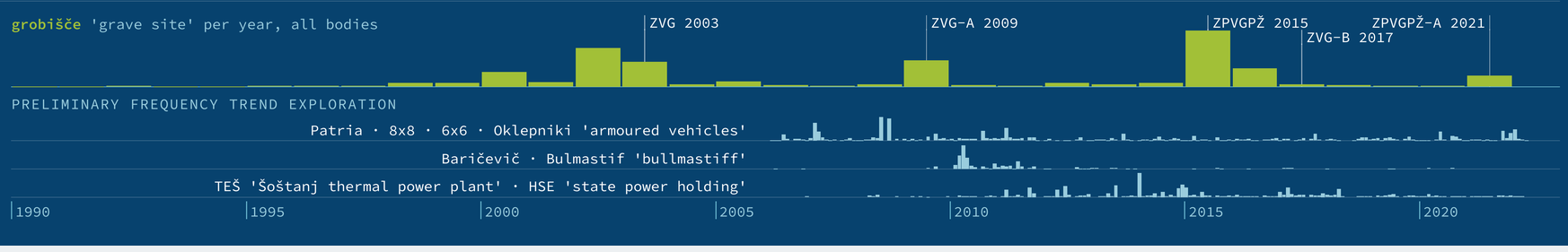
05 CLOSE READING ▼
Method, not a CLARIN component – the thesis apparatus, applied to CLARIN output

Critical discourse analysis, two modes of one apparatus. **Ideational** – nomination, predication, argumentation. **Interpersonal** – mood, modality, stance.

Every passage carries its chain reference back to 01.

Validation loop – 05 is checked back against 01 before any claim rests on a topic

PERSISTENCE AND SPIKE
A second strand traces scandal terms as monthly frequencies, to set a spike profile against the persistence of the contested past.



RQ3
MICRO

05 / CLOSE READING
Interpellation subcorpus, tagged by genre

RQ3 – how accountability is enacted under pressure

INTERPELLATIONS → FEATURE RATES → EXTREMES READ IN FULL
DESIGN

QUESTION
Interpellations are the genre in which parliament calls a minister to account, and they rarely end in dismissal. What does adversarial language look like when accountability is performed rather than delivered?

INTERPERSONAL CATEGORIES
Mood, modality, evaluation, stance. Acceptance, deflection or reattribution of blame. Face-work between the one who demands answers and the one who must give them.

METHOD
Every interpellation speech is tagged by genre and set against ordinary debate on the same topic. CQL rates per 1 000 words for deontic modality, negation, address and blame pronouns. The highest- and lowest-scoring speeches on each feature are read in full, so a measure picks the passages.

STATUS
Design. The list of interpellations and the metadata table (dates, addressee, proposer, outcome) are built next, from the National Assembly record. **No results are claimed.**

BACK UP TO RQ1 ▲
Within one interpellation debate, count the RQ1 topics its speeches spread across, against ordinary debate on the same topic. A wide spread means the debate carries other agendas.

WORKED EXAMPLE

grobišče
"mass grave / war grave" – the contested-memory term of the Slovenian Assembly. Anchors ZVG 2003, ZPVGPŽ 2015.

One seed, all five links, nothing hidden.

01 / TOPIC MODELLING

T181 239 speeches, 210 placed by the clustering

TOP-10 LEMMAS
grobišče 'grave site' · vojen 'war (adj.)' · pokop 'burial' · žrtev 'victim' · prikrit 'concealed' · posmrten 'mortal (remains)' · vojna 'war' · grob 'grave' · ostanek 'remains' · morišče 'killing site'

Eight of the 398 topics carry *vojna* 'war' among their 15 lemmas. T181, the war graves, is the one taken, from domain knowledge.

T25	resolution	T220	disabled person
T107	veteran	T286	Iraq
T108	war (adj.)	T291	injustice
T181	**grave site**	T292	army

Topic membership is not evidence. Every claim below is made from lines, not from the model.

02 / TERM LIST

	vojna +	'war' +	7 = strong
1	svetovni	world	11.23
2	svetovne	world	11.00
3	svetovno	world	10.12
4	II.	Second	9.36
5	žrtev	victims	9.00
6	Slovenijo	Slovenia	8.95
7	**grobišča**	grave sites	**8.92**
8	osamosvojitveni	independence	8.87
9	veteranov	veterans	8.79
10	1991		8.62
11	vojni	war	8.61
12	vojno	war	8.57

collocates of vojna 'war', word forms, ±5, logDice; plenary
collocate frequencies counted over the whole corpus

03 / THE QUERY, PRINTED

```
[lemma="grobišče"]
  within <speech date!="1998-10-08|…|2019-03-20"/>
    ↳ 66 sitting days, each written out
  within <speech body!="Odbor …|Zbor občin|…"/>
    ↳ 72 bodies, each written out
```

The full entry writes out all 66 excluded sitting days and all 72 excluded bodies: 3 683 characters in one line of CQL: 18 for the lemma, 751 for the dates, 2 914 for the bodies.

Positive selection on body returns the whole corpus – the attribute is multivalued with hierarchical inheritance. Exclusion is the only reliable filter. Flip `date!=` to `date=` and the same query returns the prompted set.

2 059	1 613	446
hits, total	prompted	unprompted

Counts confirmed on the CLARIN.SI concordancer.

04 / ONE CONCORDANCE LINE

2004-06-16 · NSi · MP · Jožef Bernik
unprompted · own topic T9 'corruption'

… and corruption! Then, the shameful law on **GRAVE SITES**. Is that not a unique case of corruption? …

05 / CLOSE READING

"[…] the commission for concealed grave sites was a farce […]. And those in power hold on to a falsified history […]. Is that not corruption?"

Translated from Slovene.
chain ref: T181 → grobišče → CQL (66 dates · 72 bodies) → **Redna-37-2004-06-16.u48**

CORPUS — siParl 4.0 – Pančur et al. (2024), hdl 11356/1936, CC BY 4.0. Parla-CLARIN encoded; annotated with Ljubešić & Dobrovoljc (2019), consumed as distributed. National Assembly plenary sessions, terms 1–8.

NODE — CLARIN.SI, the Slovenian node of CLARIN ERIC – repository, concordancer and Slovenian Language Technologies Overview used throughout.

SCALE — **13 233** TEXTS · **1.22 M** UTTERANCES · **239 M** WORDS · **213 372** PLENARY SPEECHES

REFERENCES — Pančur et al. (2024) *siParl 4.0* · Ljubešić & Dobrovoljc (2019) neural morphosyntactic model for Slovene · Grootendorst (2022) BERTopic · Zhang et al. (2025) Qwen3 Embedding · Kuzman Pungeršek et al. (2026) ParlaCAP · Pegan & Krašovec (2021) Slovenian Parliament · ParlaMint / Parla-CLARIN encoding