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

This PhD project presents a corpus read at three levels, in a chain in which every step is saved and re-runnable; topic → term list → CQL query → concordance lines → close-read speech. The approach is from distant to close (macro → meso → micro), using a "glass box" approach.

The corpus used is siParl 4.0 (Pančur et al. 2024) from the CLARIN.SI repository, more specifically, its distribution with full linguistic annotation (Ljubešić & Dobrovoljc 2019). The research focuses only on plenary speeches, and chair procedural turns are excluded, netting a total of 213 372 speeches in eight terms and 30 years. The pipeline also relies on the CLARIN.SI concordancer, for corpus analysis, and two CLASSLA resources: the Slovenian Language Technologies Overview (for embedding model candidates) and the ParlaCAP classifier (Kuzman Pungeršek et al. 2026).

The broadest, macro scale (distant reading) employs one BERTopic model over the 213 372 speeches. The default encoder reads a median 29 % of a speech (due to truncation). Five multilingual encoders were compared, each reading whole speeches as raw text, and using a lemma POS filter for the topic labels. The chosen model was Qwen3-Embedding-0.6B (runnable on a free-tier Colab T4), and the production run had a minimum cluster size of 75, where every policy area above 1 % of speeches has at least four topics (set against ParlaCAP labels).

RQ1 asks what parliament discusses, how its attention shifts, and how small, niche parties engage relative to mainstream ones.

The meso (corpus) scale picks terms from the topic model strings and runs a frequency and collocation analysis on a specific issue emerging from RQ1, selected from domain knowledge (in our case 'war'). Collocation parameters are set: logDice ≥ 7 strong, window ±5. From the results, specific cases are chosen for further investigation, in our case "grobišče" ('war grave'), logDice 8.92 (topic T181). A frequency matrix is created to have a broader view of emerging patterns. CQL queries allowed to map the 2 059 plenary hits into "prompted" (1 613 on sitting days of five war-graves acts and related debates) and "unprompted" (446).

At this scale, the salience and dynamics of ideological discourse are compared to scandals (persistence vs spike and fade).

The micro scale (close reading) builds on scandals, focusing on interpellation debates, which call ministers to account but rarely end in dismissal (Pegan & Krašovec 2021). An interpellation subcorpus will be created (enriching the main corpus) and specific adversarial linguistic features will be measured (deontic modality, negation, address forms and blame pronouns) over every interpellation speech, and each feature's

extremes read in full. These speeches' topic labels tie them back to the macro scale; topic dispersion shows whether a debate stays on its policy area.

Upon completion, corpus corrections and various user-based improvement suggestions will be submitted to the siParl authors. The embeddings, topic modeling pipeline, ParlaCAP labels back to 1990, the interpellation metadata, and other deliverables will be deposited in CLARIN.SI, with documentation for re-running it on ParlaMint corpora.

Keywords: siParl; parliamentary discourse; topic modeling; collocation; critical discourse analysis; traceability

References

Kuzman Pungeršek, T., Rupnik, P., Širinić, D., & Ljubešić, N. (2026). Supercharging agenda setting research: The ParlaCAP dataset of 28 European parliaments and a scalable multilingual LLM-based classification. arXiv:2602.16516.

Ljubešić, N., & Dobrovoljc, K. (2019). What does neural bring? Analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing* (pp. 29–34). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-3704

Pančur, A., Meden, K., Erjavec, T., Ojsteršek, M., Šorn, M., & Blaj Hribar, N. (2024). *Slovenian parliamentary corpus (1990–2022) siParl 4.0*. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1936

Pegan, A., & Krašovec, A. (2021). *Slovenian Parliament*. PSA Parliaments briefing. https://psaparliaments.org/wp-content/uploads/2021/03/Slovenia-1.pdf