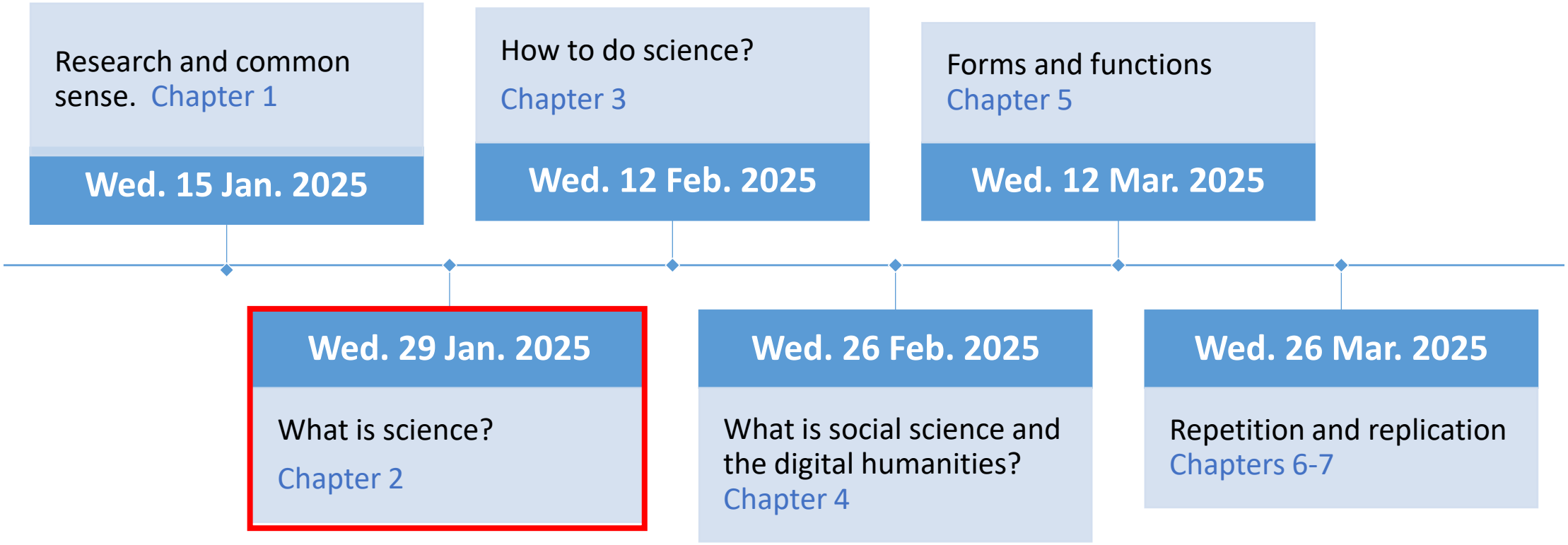
Open research group

ESRC Centre for Corpus Approaches to Social Science

Lancaster University

Camera on, sound off if possible

Topics: spring term



Data segmentation in corpus-assisted parliamentary discourse studies

Anna Kryvenko

INZ (Slovenia), NISS (Ukraine)

Data segmentation as a corpus design and research design problem

A dependency between the level of aggregation of the data under analysis and the results produced

Temporal segmentation in MD-CADS, or modern-diachronic corpus-assisted discourse studies (Marchi, 2018):

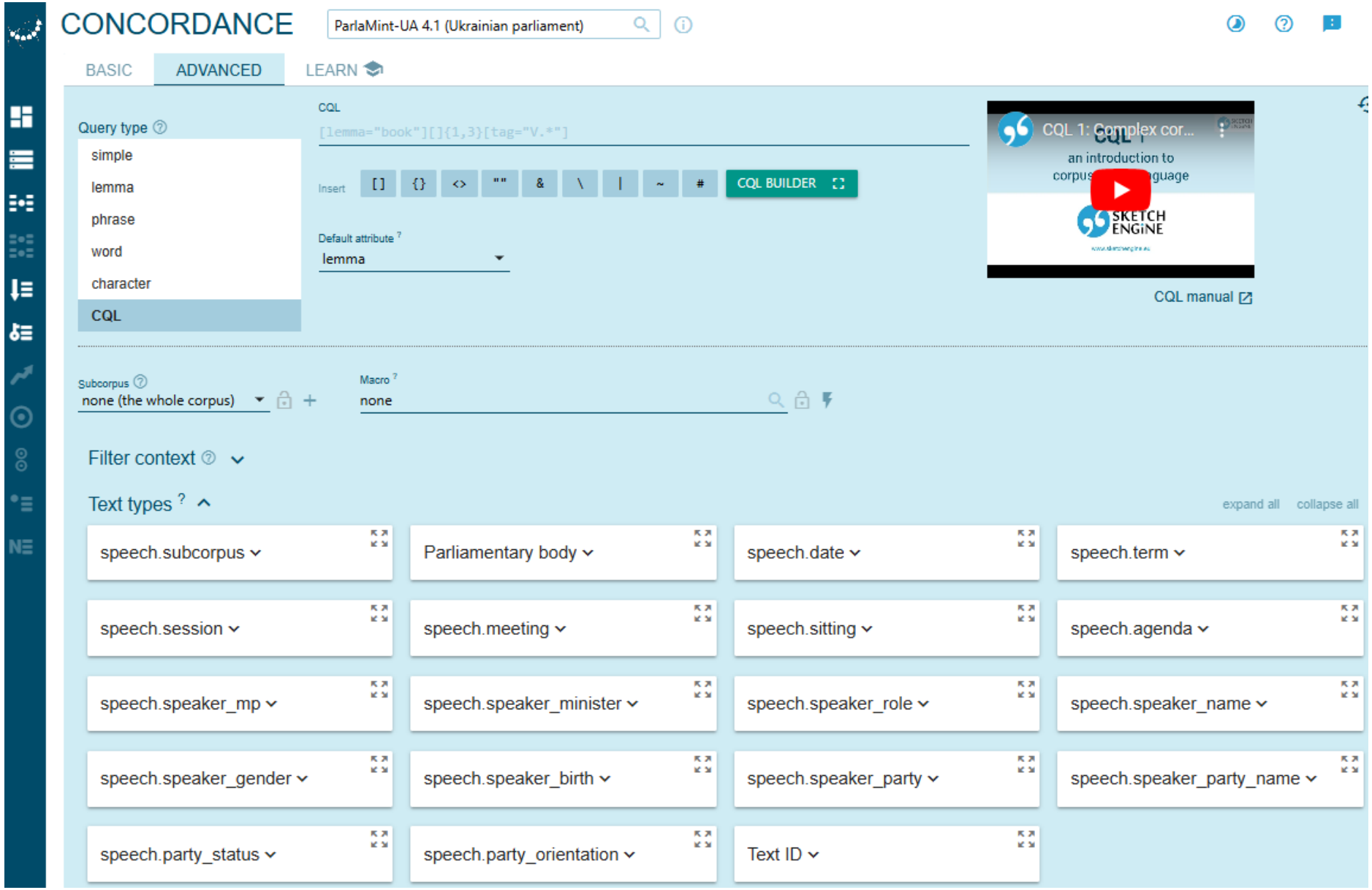
- tracking a phenomenon or process over continuous data
- comparing between two or more distinct moments
- comparing before and after defined “turning points”

Speaker-related data segmentation

- personal (name, age, gender)
- institutional (parliamentary body, party affiliation, role in the parliament, power position, political leaning)

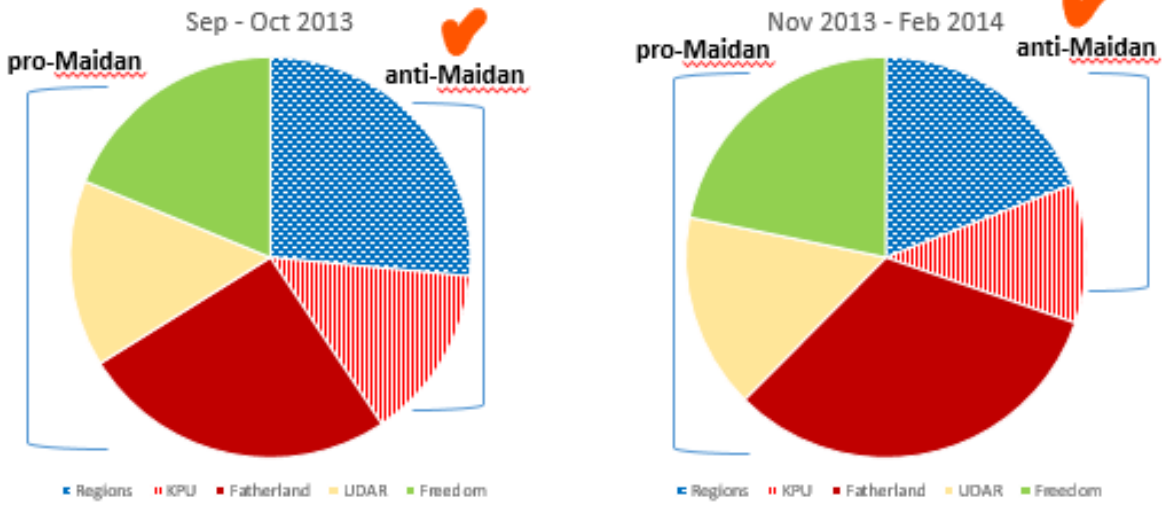
ParlaMint project:

- 29 European countries and autonomous regions
- at least the period from 2015 to 2022
- over 1 billion words
- original speeches and MT to English

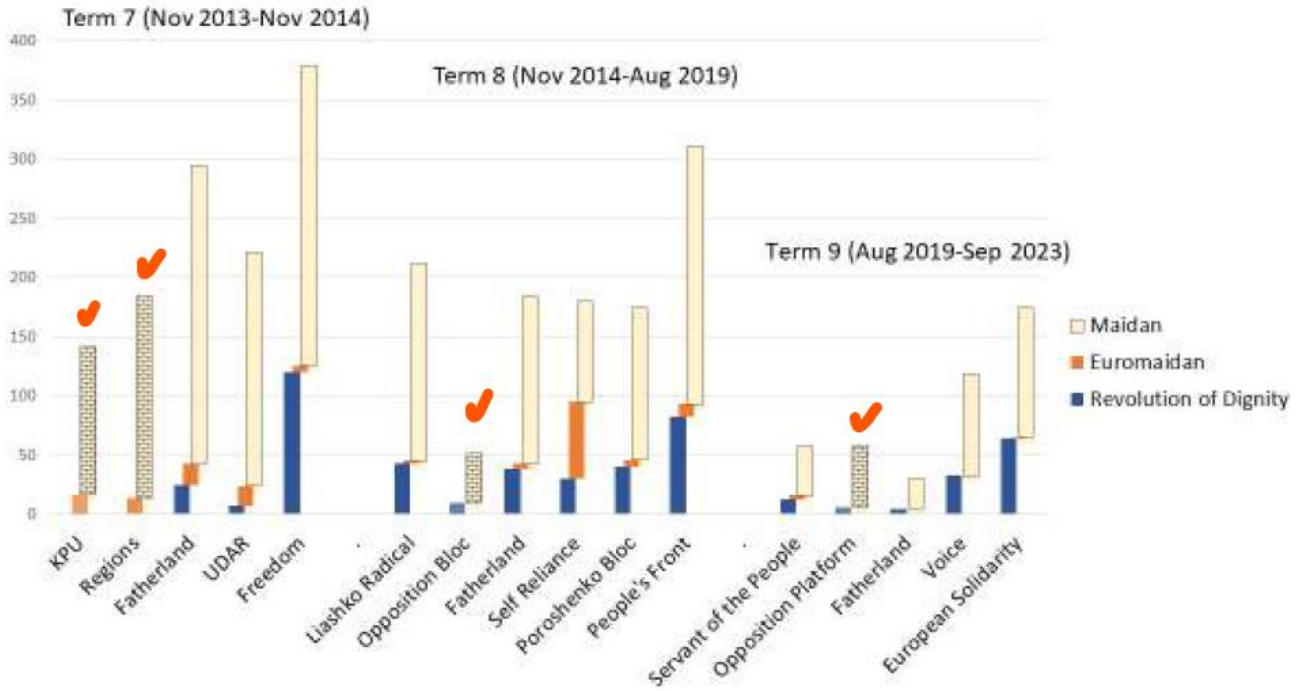


CLARIN.SI & LEXICAL COMPUTING

Case study: pro- and anti-Maidan parties in the Ukrainian parliament



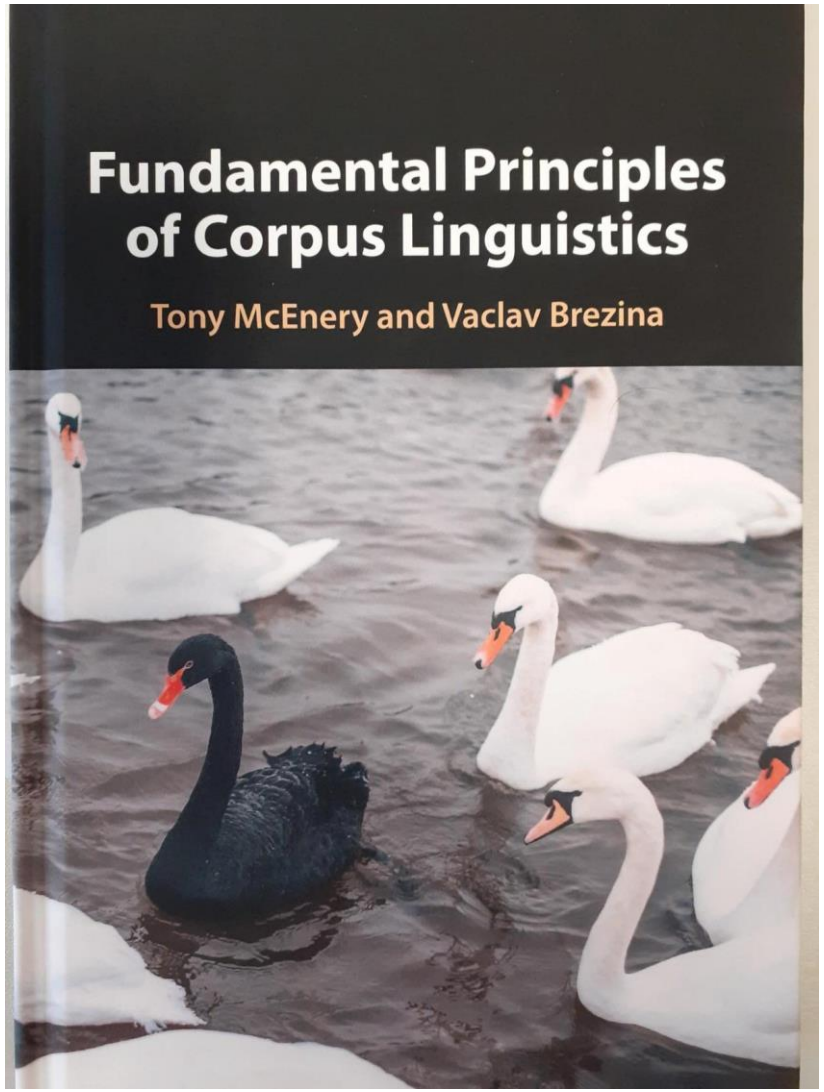
Distribution of tokens produced before and during the protests per faction



Naming choices for the 2013-14 protests per faction (per million tokens in text type)

References

- Erjavec, T., Kopp, M., Ljubešić, N. *et al. 2024.* ParlaMint II: advancing comparable parliamentary corpora across Europe. *Lang Resources & Evaluation*.
- Kryvenko, A. 2025. ‘Maidan has become part of Ukrainian identity’: The dynamics of naming and framing civil resistance in parliamentary discourse [Manuscript accepted for publication]. Corpora 20 (3).
- Marchi, A. 2018. ‘Dividing up the data: epistemological, methodological and practical impact of diachronic segmentation’ in C. Taylor and A. Marchi (eds) Corpus Approaches to Discourse: A Critical Review, pp. 174–196. London and New York: Routledge.



“

These scientists seem entirely blind to the problem of induction. Similarly, they are blind to the substantial human intervention hidden behind machine learning – the coding of the algorithm, the setting of parameters, the choice of data, the encoding of data, the development of ontologies – all these are really about the choice of hypothesis to be examined, at least indirectly. They are also silent on the role that humans play in explaining the results produced by such analyses. What appears to be a statistical magic wand is, in fact, a screen drawn across a lot of human choice and, crucially, a screen drawn across the problem of induction. In fact, this is a good example of an issue with induction that Popper calls *a priorism*. One has to assume that some things are true, or given, to avoid a problem of inductive recursion – something outside of the system has to set the system so that it does not loop back on itself, running inferences to

[23] Though see Talbot (2015: 20–21) for a more critical account of Popper's quasi-inductive approach and its relationship to the quasi-inductive framework of Hempel (1945).

What time is it?

What assumptions do we make when telling the time? What assumptions do we make when do our research?

What assumptions do we make when telling the time?

- **Clockwise Movement:** We assume that clock moves in a clockwise direction.
- **Numerical Arrangement:** Typically, clock numbers increase sequentially in a clockwise manner.
- **Trust in Timekeeping:** We trust that clocks provide accurate time based on conventional movement.

Opportunity for next week:
5 min. presentation

What scientific processes (methods, protocols, assumptions etc.) do you use in your research?