Domonkos Sik,* Renáta Németh,** Ildikó Barna,***
Theresa Gessler**** & Hanna Orsolya Vincze*****

# 'Text as data': Eastern and Central European political discourses from the perspective of computational social science

Intersections. EEJSP
10(4): 1–5.
https://doi.org/10.17356/ieejsp.v10i4.1462
https://intersections.tk.hu

* [sik.domonkos@tatk.elte.hu] (Eötvös Loránd University, Faculty of Social Sciences, Budapest)
** [nemeth.renata@tatk.elte.hu] (Eötvös Loránd University, Faculty of Social Sciences, Budapest)
*** [barna.ildiko@tatk.elte.hu] (Eötvös Loránd University, Faculty of Social Sciences, Budapest)
**** [gessler@europa-uni.de] (European University Viadrina, Frankfurt/Oder)
***** [vincze.orsolya@fspac.ro] (Babeş-Bolyai University, Cluj-Napoca)

## Textual data in the analysis of the political public sphere

Eastern and Central European countries have followed many different paths in the last thirty years since the fall of communism. We find examples of the strengthening of democratic institutions, the dismantling of the public sphere and civil society, right-wing and left-wing populisms, digital transformations, and disruptions brought about by the rise of online platforms. Many attempts have been made to analyse these processes: local and global dynamics of capital and field relations, historical changes in institutions, values and norms, and the challenges and ambivalences of European integration and globalisation have all offered theoretical frameworks for explaining the heterogeneity of Eastern and Central European regimes.

Studying public discourses can yield important insights into the various trends in the region. Although political discourses are always embedded in the broader context of local and international structures, institutions, values, and behaviour patterns, they also represent a distinct level of analysis. The different layers of the political public sphere (for example, parliament and political actors' communication, the media, online forums and social media platforms) are key arenas for the social construction of reality. These discursive levels are crucial for understanding the functioning of democratic and non-democratic electoral regimes; their complex structure and the interactions within them have the potential to explain political institutions, praxes, and civic culture.

All layers of the political public sphere are producing textual data in vast quantities, and the birth of computational social science opens up new perspectives for their analysis. The aim of this thematic issue is to bring together papers that profit from this new potential. The thematic issue takes a fresh look at the transforming political discourses of Eastern and Central European countries by relying on the tools of large-scale textual data analysis in the broad sense.

## NLP as a social research tool

The articles in this thematic issue all approach social research problems through text analysis, using computational techniques, typically natural language processing (NLP). Although language is a crucial tool for social interaction, quantitative social research largely overlooked it for decades, mainly due to a lack of data collection and processing tools. The situation has changed radically in the last decade, with the use of textual data as an empirical social research base spreading at an exponential rate (see the 'text-as-data' movement, Gentzkow et al., 2019; Benoit, 2020). In fact, specific textual representations of all subsystems of society are being created.

These texts allow their authors to express nuanced opinions, reflect observed behaviour, and are not burdened by, for example, recollection bias, thus allowing for more valid conclusions. Online-generated data allow us to follow human behaviour in real context and real-time, which goes far beyond the traditional research methods of social scientists. While social researchers have previously had to make trade-offs between data size and depth, digitisation has made it possible to abandon such constraints.

From a social research perspective, an important feature of the new textual data asset is that penetration is widening, so digital platforms offer a way for anyone to express themselves. This is an important change, as previously the texts that were made available to the public were almost exclusively written by the elite.

In parallel with the revolution in using textual data to describe society, the last decade has seen an explosion in computing power and, in parallel, in text analytics technologies for analysing data, with new technologies providing a relevant depth of text processing. This explosion has spread from computer science and computational linguistics to computational social science, with the development of broader than domain-specific tools and models for social research.

This thematic issue provides an insight into the applications of NLP. NLP is an exciting and promising area for social research located at the intersection of computer science, artificial intelligence research and linguistics. The last decade has seen huge growth in the scientific application of NLP. It has been used in ambitious projects in health, business applications, marketing and national defence. In the last few years, NLP has also started to gain ground in the social sciences, from political science to economics and sociology.

Although NLP is a relatively new interdisciplinary field, quantitative and qualitative text analysis itself has a decades-long tradition in social research. However, traditional quantitative text analysis typically requires researchers to actually read and understand the text they are analysing. This approach started to change after the turn of the millennium. From then on, the text was not presented as an object to be read and understood but rather as an input for automated methods, without any need for actually reading it. The use of NLP in social research is therefore related to this more recent 'text-as-data' approach, where text is treated as an ordered, well-structured, numerical database that provides input for computer algorithms. Traditional quantitative text analysis tended to quantify only the appearance of certain terms or codes in texts. In comparison, the NLP toolbox is a major advance, automating tasks such as identifying the emotional load of texts, measuring the distance between texts, creating text clusters, identifying latent thematic structures or latent semantic relations, and other discursive patterns identifiable on a large scale.

The text-as-data approach in social research is primarily an alternative to opinion polls. The use of digital textual data also has methodological and epistemological advantages over deploying traditional opinion polls. The methodological advantage lies in the fact that the digital revolution has channelled opinion polling to the internet, and computational methods provide access to these vast amounts of data. The epistemological advantage stems from the fact that digital textual data are 'found data' in the sense that they are usually produced for some purpose other than scientific analysis. Therefore, unlike opinion poll data, there is no possibility of non-response. Furthermore, since opinions and interactions can be observed in their natural environment, these data reflect observed behaviour, as opposed to self-reported responses, so the views derived from them are likely to have greater internal validity, for example, as they are not subject to recall bias or social desirability bias. Other advantages include real-time availability, the potential for longitudinal analysis, and their coverage of virtually the entire population, making them useful for studying rare phenomena and hard-to-reach subpopulations. It is precisely these possibilities that are exploited in the analysis of depression forums.

We hope that this thematic issue will demonstrate that NLP represents a unique opportunity to explore regularities in texts, which can also serve as explanations for theory building. At the same time, NLP is associated with methodological pitfalls. The primary concern is that the sheer volume of data and the complexity of the methods that are used can create a false sense of reliability. It is important to stress that in most sociological applications of NLP, the use of qualitative methods is inevitable at some point in the analysis, most notably at the validation and interpretation stages. Similarly, qualitative approaches are often used to support model interpretation, as complex NLP models are difficult to interpret without going back to the original texts. Without domain-specific knowledge, NLP is a technique with little scientific value for the social sciences. Some papers in this thematic issue also illustrate this statement.

## Content of the thematic issue

Our thematic issue, published in two parts (issues 2024/4 and 2025/1), consists of 11 research papers and 6 shorter contributions. From a thematic perspective, several clusters can be outlined.

Based on parliamentary debates in Hungary between 2009 and 2019, Theresa Gessler analyses the extent to which opposition parties draw on the concept of (liberal) democracy when criticising the government's policies and how democratic backsliding affects these debates.

A distinct set of articles complements these approaches by analysing similar questions through the notion of populism. Elena Cossu analyses ten Central European countries from the perspective of the populist rhetoric in electoral manifestos, creating a semantic and sentiment-based map applicable to the discursive strategies of left- and right-wing parties. Jogilė Ulinskaitė and Lukas Pukelis analyse the populist discourse of Lithuanian political parties over a 30-year period (1990–2020). Their article seeks to identify populist content in a corpus of political party manifestos, websites, and columns written by party members. Zsófia Rakovics and Ildikó Barna elaborate a case study of 'main-

streaming the extreme': the example of Jobbik, a Hungarian radical party that transformed into a centre-right party, focusing on its rhetoric and networking strategies. Through the lens of these analyses, a complex picture of Central European democratic culture emerges: while there are differences at the national and party levels, overall, it seems that thirty years after the transition, Central European democracies are in a vulnerable position.

The diagnoses concerning the vulnerability of Central European democracies can be further refined by reflecting on ongoing discursive mechanisms. One of the most dangerous potential outcomes of political polarisation and the resulting intensification of social conflict is various forms of scapegoating. Kata Knauz, Attila Varga, Zsolt Szabó and Sára Bigazzi analyse Hungarian political discourses about the Roma minority. Besides the explicit stereotypes and hostile prejudices, they attempt to detect the more subtly biased, paternalistic discourses present in the communication of parties. Rok Smrdelj, Roman Kuhar and Monika Kalin Golob analyse another hotspot in contemporary identity politics, namely the debates surrounding gender. Based on Twitter posts, they attempt to map filter bubbles of anti-gender discourses. Emese Túry-Angyal and László Lőrincz also analyse similar discursive mechanisms, partly related to the technological infrastructure of algorithmic public spheres. They explore how echo chambers, homophily, and network type affect the spread of information on Facebook.

Besides ethnic minorities and gender, one of the main issues implying scapegoating is the semantics of nationhood. Several articles explore how geopolitical discourses construct national identity in opposition to external others (i.e., foreign countries). Áron Szalay and Zsófia Rakovics analyse the enemy images appearing in the speeches of Viktor Orbán, the prime minister of Hungary. Besides mapping the range of abstract (e.g., migrants) and concrete (e.g., Soros, Brussels) enemies, they explore discourses of fearmongering in detail. According to Radu M. Meza and Andreea Mogoș, generating fear and loathing is not a rare tendency in Central European political discourses. In their analysis of the headlines of *Sputnik News* (distributed in Poland, Romania, and the Czech Republic), they mapped the discursive strategies and the affective framing that appeared in the communication of a Russian government-financed news agency. In a similar spirit, Ilya Sulzhytski and Varvara Kulhayeva analysed Belorussian Telegram channels from the perspective of their interpretation of the Ukrainian invasion. Besides mapping the main discursive panels, they also demonstrate that local pro-government activists are central channels for creating and disseminating hatred towards Ukrainians in Belarus. Although the impact of war is the most tangible in the present, collective traumas also linger in collective memory. Renáta Németh, Eszter Katona, Péter Balogh, Zsófia Rakovics, and Anna Unger explore, through parliamentary debates, the discourses surrounding the Carpathian Basin, a central metaphor for a collective identity anchored in the narratives of historical Hungary.

Besides the original research articles, the thematic issue also includes some shorter reflections introducing specific challenges of natural language processing research. Miklós Sebők, Csaba Molnár, and Anna Takács address the difficulties of building an appropriate dataset, discussing the construction of a dataset of parliamentary speeches, bills, and laws for Czechia, Hungary, Poland and Slovakia from the early 1990s to the 2020s. Zsófia Rakovics and Márton Rakovics provide a critical methodological analysis of large language models, exploring how they can be used to extract information about different

groups of society and utilised as data providers by acting as virtual respondents. In a similarly critical manner, Renáta Németh and Domonkos Sik summarise some methodological conclusions from natural language processing research: they argue that topic models need to be complemented with hermeneutic tools during the interpretative process. Tamás Varga reviews a book that can serve as a guide to using computational text analysis to learn about the social world, *Text as Data: A New Framework for Machine Learning and the Social Sciences* (by Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart). In another book review, Mohammad Ashraful Alam writes about the book *Natural Language Processing in the Real World: Text Processing, Analytics, and Classification*, written by Jyotika Singh, which is a practical guide for building natural language processing solutions. In addition to the articles above, the thematic issue also contains a short data note concerning a comparative analysis of information society in Central and Eastern Europe by Árpád Rab, Tamás Szikora and Bernát Török. Findings reveal both regional commonalities and distinct national attitudes towards online manipulation, social media usage, and the impact of internet communication on personal relationships.

Although the various contributions focus on different fields and rely on different methods, the authors share a common ambition. Computational social science opens up new possibilities for research on Central European public spheres. However, in order to fully tap the inherent potential of the emerging datasets and methods, there is a need to link the existing insights of the social sciences (originating from non-NLP analyses) with the new ones. Only in this way can the trap of 'methodological fetishism' be avoided so computational social science can be integrated into the broader process of exploring discursive dynamics in a meaningful manner. Our thematic issue is a step in this direction.

## References

Benoit, K. (2020). Text as data: An overview. In L. Curini & R. Franzese (Eds.), *The SAGE handbook of research methods in political science and international relations* (pp. 461–497). SAGE. https://doi.org/10.4135/9781526486387.n29

Gentzkow, M., Kelly, B. & Taddy, M. (2019). Text as data. *Journal of Economic Literature, 57*(3), 535–574. https://doi.org/10.1257/jel.20181020