ISBN: 978-84-1091-017-1 (PDF)
DOI: https://doi.org/10.14201/0AQ0373559569

Exploring Semantic-Thematic Fields and Lexical Patterns in Parliamentary Debates — Topic Modeling Across Comparable Corpora

Exploración de campos semántico-temáticos y patrones léxicos en los debates parlamentarios: topic modeling en corpora comparables

Katrin HERGET
University of Aveiro, CLLC
kherget@ua.pt

Jean-Ezra YEUNG
Columbia Startup Lab, Columbia University
jy2343@caa.columbia.edu

Michelle RUFRANO
Fordham University
mrufrano3@fordham.edu

Teresa ALEGRE
University of Aveiro, CLLC
teresaalegre@ua.pt

ABSTRACT: The collection and analysis of multilingual text corpora assumes a fundamental role in an ever-digitizing world. The access to large-volume corpora allows the study and description of various linguistic items and patterns. Recently, extensive

research has been conducted on topic modeling for analysing political debates (Guldi 2019; Ristilä and Elo 2023). Our study is based on the exploration of comparable corpora of parliamentary debates (Erjavec et al. 2023) in three languages. The analysed corpora were collected from the language resource repository CLARIN ERIC for the Austrian, British, and Portuguese datasets. Our study employs a topic modeling approach to investigate semantic-thematic fields in political speeches. The extracted data from the three countries will be compared to draw conclusions regarding the analysis of potential thematic convergences or divergences.

KEYWORDS: multilingual text corpora; topic modeling; artificial intelligence; parliamentary debates; semantic-thematic fields.

RESUMEN: La recopilación y el análisis de corpus multilingües desempeñan un papel fundamental en un mundo cada vez más digital. El acceso a corpus de gran volumen permite estudiar y describir una gran variedad de elementos y patrones lingüísticos. Recientemente, se ha llevado a cabo una amplia investigación sobre el uso del *topic modeling* para el análisis de debates políticos (Guldi 2019; Ristilä y Elo 2023). Nuestro estudio se basa en la exploración de corpus comparables de debates parlamentarios (Erjavec et al. 2023) en tres idiomas. Los corpus austriaco, británico y portugués analizados se recopilaron del repositorio de recursos lingüísticos CLARIN ERIC. En nuestro estudio, empleamos un enfoque basado en *topic modeling* para investigar los campos semántico-temáticos en los discursos políticos. Los datos extraídos de los tres países se compararán para alcanzar conclusiones relativas al análisis de posibles convergencias o divergencias temáticas.

PALABRAS CLAVE: corpus textuales multilingües; *topic modeling*; inteligencia artificial; debates parlamentarios; campos semántico-temáticos.

1. INTRODUCTION

In a data-driven age, governments are challenged to address the overabundance of both accurate and false information in the context of increasing human disasters, i.e. infodemics related to climate change, pandemics, and conflict (Gostin 2022; Wells and Scheibein 2022). Infodemics necessitate the use of technological tools in addressing data voids that can be weaponized as misinformation (National Center for Immunization and Respiratory Diseases 2021; Wilhelm 2023). For text data, lexical equivalence may provide insights into cultural distinctions between countries that may also help us understand differences in access to information (Eysenbach 2009). Knowledge translation is vital; therefore, the collection and analysis of multilingual text corpora play a crucial role in characterizing the information landscape.

Access to large-volume corpora enables the study and description of diverse linguistic items and patterns. Our preliminary study explores comparable corpora of parliamentary debates (Erjavec et al. 2023) in three countries (Austria, Great Britain, and Portugal). The corpora were sourced from the language resource repository CLARIN ERIC, containing parliamentary speeches from each country.

In this study, we combine topic modeling with qualitative research analysis to explore themes and sub-themes within political speeches. In a subsequent phase, we aim

to compare the extracted data to draw conclusions regarding the analysis of potential thematic convergences or divergences.

1.1. *Comparable Corpora for Data Analysis*

The exploration of multilingual datasets plays an important role in understanding and comparing thematic fields in different language-cultural communities. Topic modeling can be used to automatically identify abstract topics or themes within large-scale corpora. With the advancements in Natural Language Processing (NLP), the collection of corpora is becoming increasingly prevalent and significant. Comparable corpora are an essential resource for various NLP tasks, such as machine translation, cross-lingual information retrieval, and contrastive linguistics studies. Unlike parallel corpora, comparable corpora are not influenced by the structure of the source text. According to McEnery and Hardie (2012, 20), a comparable corpus consists of «components that are collected using the same sampling method, e.g., the same proportions of texts of the same genres in the same domains in a range of different languages within the same sampling period». Similarly, Mikhailov and Cooper (2016, 217) define comparable corpora as collections of texts compiled «based on the same principles (size of collections, size of samples, covered topics, chronological period, etc.) in different languages or different variants of the same language, such as texts on atomic energy in French and Spanish, or texts in German from Germany, Austria, and Switzerland».

1.2. *Study Motivation*

In times of unlimited access to information through large corpora collections, our study aims to identify and analyse thematic fields from parliamentary debates in three European countries: Austria, Great Britain, and Portugal. Exploring these topics provides insights into political, economic, and social issues, allowing us to observe convergences and divergences across the three language corpora. Identifying and comparing these topics is particularly relevant for cross-linguistic and cross-cultural analyses, especially in the fields of Languages for Specific Purposes and Specialized Translation.

2. METHODOLOGY

To conduct an exploratory analysis, we used both LDA (Latent Dirichlet Allocation), a probabilistic model, and ATLAS.ti, a qualitative data analysis tool, to identify and analyse recurrent themes in our selected corpora. By automatically assigning each utterance to one or more topics based on the distribution of words, LDA provided an initial overview of the primary themes present in our data. However, LDA does not capture all the nuances within the text. We also required ATLAS.ti to facilitate a deeper exploration and coding of qualitative data within topics, allowing us to identify and analyse themes, patterns, and relationships among the analysed corpora that might not have been immediately apparent from the initial topic modeling results.

2.1. *LDA: a Brief Overview*

LDA overcomes challenges found in prior alternatives, such as (probabilistic) latent semantic indexing, which suffers from assigning probabilities to words previously unseen, and overfitting, i.e. not as generalizable to new data (Blei 2003). While there are many newer alternatives to topic modeling from unsupervised learning techniques, LDA has an open-source implementation, has been widely used, and generates probabilities for each topic, enabling utterances to be assigned to more than one topic, which is a required flexibility to anticipate multi-topic utterances and degrees to which an utterance is associated with a topic. The key limitation is that LDA assumes that the order of words does not matter because it models topics based on the mixture of words; a «bag-of-words» model.

To enable the technical feasibility of LDA, we dramatically reduced the vocabulary size, so that it did not exceed the sample size (number of utterances). The vocabulary size was ultimately reduced to approximately a factor of 10 compared to the sample size to enable several topics to be potentially elicited. Reduction of vocabulary size started with removal of stop words, and then frequently occurring words greater than the 95th percentile of frequently occurring words and low frequency words.

Other pre-processing steps included using the term frequency, and removal of short utterances (30 characters or less), which was a conservative cut-off from exploring the histogram to ensure removal of transitory statements that were a matter of parliamentary formality.

For each language corpora, we took the latest year of data, should there be sufficient data to run the LDA algorithm successfully. For Portuguese, there was insufficient data in the latest year, 2022, so 2021 was used instead.

We provided the option to re-run the algorithm on some topic segments of the corpora to identify any sub-topics if the topics were large. The model output assigns each utterance a probability score and the corpora is accompanied with a list of word tokens sorted in descending frequency, reflecting the top sampled mixture of words for each topic.

2.2. *Qualitative Research with ATLAS.ti*

According to Gupta (2024, 127), ATLAS.ti is a software for qualitative data analysis of multimodal texts. It provides tools for managing, extracting, comparing, and exploring meaningful information segments, allowing for a deeper exploration of compiled data. Through visualization, integration, discovery, and exploration features, ATLAS.ti facilitates the development of new knowledge. The software allows the extraction, categorization, filtering, and interlinking of data segments, enabling the identification of patterns and themes within a diverse range of source documents. ATLAS.ti is used across various domains, such as social sciences, engineering, business administration, media, psychology, linguistics (2024, 127).

Recently, ATLAS.ti introduced AI assistance for open and descriptive coding, which is based on OpenAI’s GPT model (ATLAS.ti 2024).

> ATLAS.ti uses artificial intelligence to quickly retrieve and display words, concepts, and phrases that the researcher searches in the data. ATLAS.ti 23 has AI (artificial intelligence) coding features that researchers can use to code their data with much precision and organize codes based on significance. Further, the auto-coding function can filter, link groups, and enable researchers to perform various tasks such as comparing, relating, interpreting, and building theory from the data. (Lewins and Silver 2007; Gupta 2024, 120-21)

AI-generated codes and sub-codes required human validation with corresponding utterances (revision and modification of AI Codes), and interpretation. The utterance analysis was based on LDA's previously assigned topics within each language corpus. From each topic within a language, a random sample of 200 utterances was drawn for analysis due to time and effort constraints.

2.3. *Corpus Presentation – Parliamentary Debates Corpora (ParlaMint)*

ParlaMint 4.0 (Clarin 2024) presents a comprehensive collection of comparable corpora comprising transcriptions of parliamentary debates from 29 European countries and regions, mainly spanning from 2015 to mid-2022, totaling over 1.1 billion words. The corpora are meticulously structured, with daily divisions including details on terms, sessions, and meetings, alongside speeches marked with speaker information and transcriber comments. Rich metadata encompassing speaker details, political affiliations, and additional information is provided, with some corpora featuring further metadata, such as speaker birth years and Wikipedia links. The corpora include linguistic annotations covering tokenization, sentence segmentation, lemmatization, part-of-speech tagging, morphological features, syntactic dependencies, and named entities (ParlaMint.ana 4.0).

3. RESULTS

The date range, utterance size and vocabulary size of the corpora datasets is presented in Table 1.

Data (after short utterance removal)	**Date Range**	**Parliamentary Session Frequency**	**Utterances (Model Size)**	**Vocabulary Size (Model Size)**
Great Britain (Commons)	January – July 2022	Daily	37 867 (34 080)	42 212 (3 000)
Austria (German)	January – October 2022	~Monthly	5 635 (5 071)	53 230 (600)
Portugal (Portuguese)	January – December 2021	~Monthly	16 648 (14 983)	42 533 (1 800)

Table 1. Corpora datasets

3.1. *Austrian Corpus*

For the Austrian corpus, we analysed samples from three topics (AT0, AT1, AT2) to illustrate thematic occurrences. From these three topics, we chose AT0 to exemplify our analysis approach within the scope of this paper. Figure 1 illustrates the human-validated codes and their respective sub-codes for sample AT0.

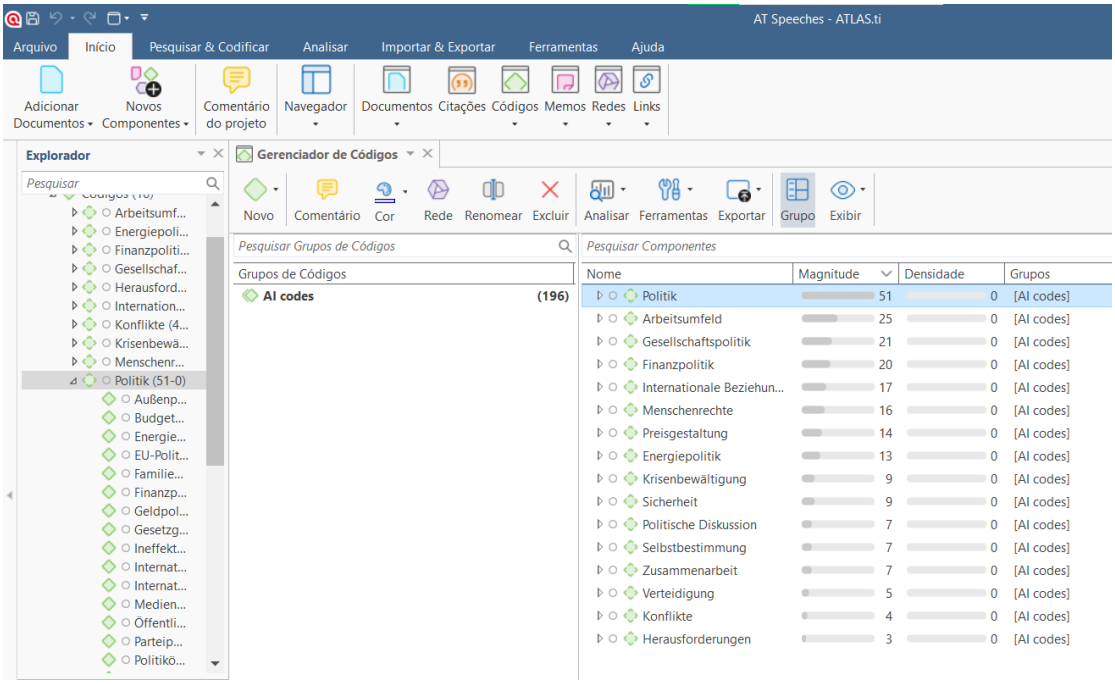

Figure 1. Codes and sub-codes for sample AT0

The range of topics suggests a comprehensive analysis of various aspects, including politics, society, economics, international relations, human rights, energy policy, and crisis management. The inclusion of such diverse topics underscores the complexity of contemporary challenges. Many topics overlap or intersect; for instance, energy policy intersects with environmental issues, economic policy influences social welfare, and international relations impact security concerns.

The significant focus on social topics – such as tolerance, equality, women's empowerment, and social justice – reveals a fundamental concern for addressing pressing social issues and fostering inclusivity and equity.

The inclusion of topics related to crisis management, such as poverty alleviation, conflict resolution, and pandemic management, indicates a recognition of the prevalence and severity of crises in contemporary society.

The codes and sub-codes can be organized into tables, with their frequency in the dataset shown on the left. On the right, the occurrence of specific topics in quotations is displayed. In the example below (Figure 2), the topic of «energy policy» appears 13 times in the selected text sample, which discusses recent trends, such as the energy crisis and strategies to mitigate it.

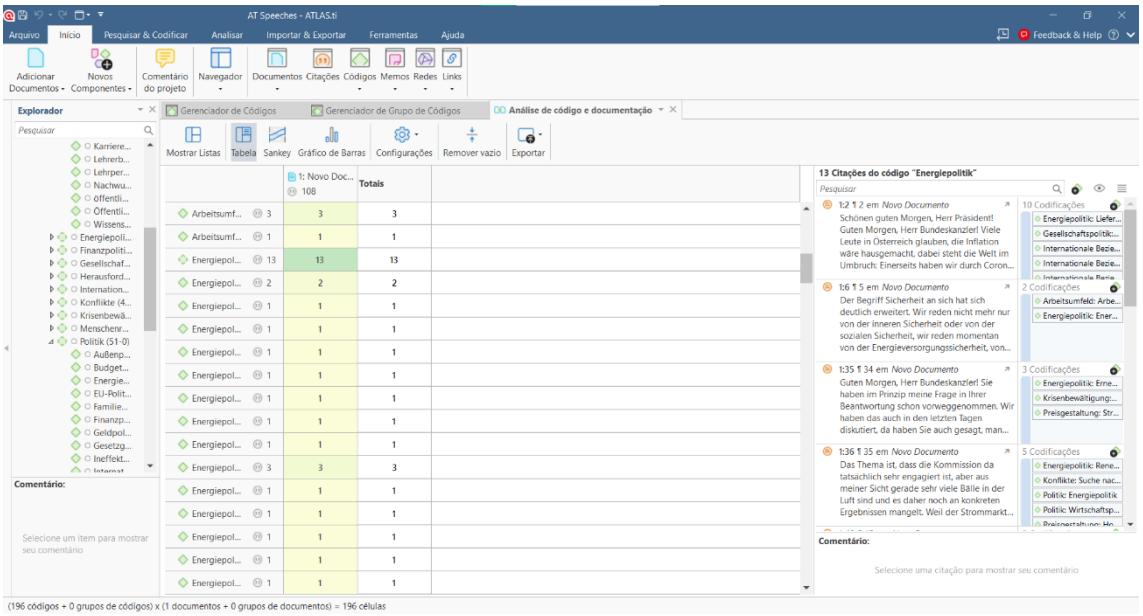

Figure 2. Frequency and occurrence of «Energy Policy» in sample AT0

3.2. *Portuguese Corpus*

For the Portuguese corpus, we analysed samples from three topics (PT0, PT1, PT2) to illustrate thematic occurrences. The provided data (Figure 3) presents themes and their respective magnitudes, indicating the frequency of each theme within the dataset.

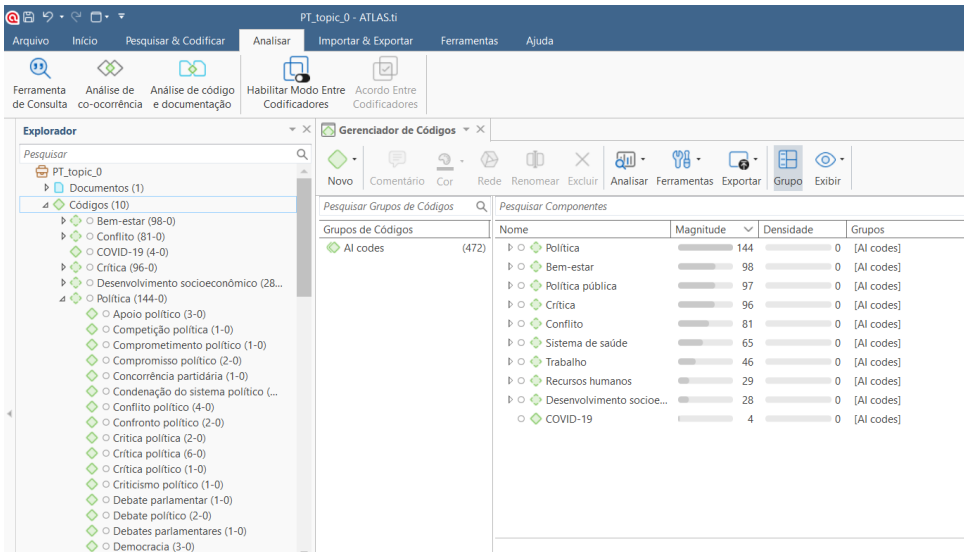

Figure 3. Themes and magnitudes in sample PT0

Politics emerges as a substantial theme encompassing various aspects, such as political criticism, democracy, government, legislation, public policy, and accountability. Public policy is a notable subset, focusing specifically on policies related to health care, social welfare, finance, justice, and transparency. This underscores the importance of policy discussions within the dataset.

Secondly, the code «well-being» includes a variety of sub-themes, such as benefits of proximity, trust, quality of life, respect, transparency, and solidarity. The high magnitude of this code suggests that issues related to well-being are both prevalent and diverse within the dataset.

Thirdly, conflict emerges as another significant theme, with sub-themes including corruption and social injustice. Criticism is notable, as there are sub-themes covering a wide range of topics, such as environmentalism, ethics, economy, health care, and social justice. This indicates that critical viewpoints are expressed across various domains within the dataset. Healthcare is highlighted as a significant theme, addressing sub-topics such as access to care, quality of services, workforce shortages, and public health concerns. Work-related issues are also prominent, including conditions of work, productivity, workforce reinforcement, and dignity of labour.

Overall, the analysis reveals a diverse range of themes, encompassing political, social, economic, and health-related issues.

3.3. *British Corpus*

For the analysis of the British corpus, we exclusively examined the utterances from the House of Commons, ensuring the comparability of the datasets. To enable the compilation of comparable corpora, we did not analyse the utterances from the House of Lords so that the corpora structure is similar across the three countries.

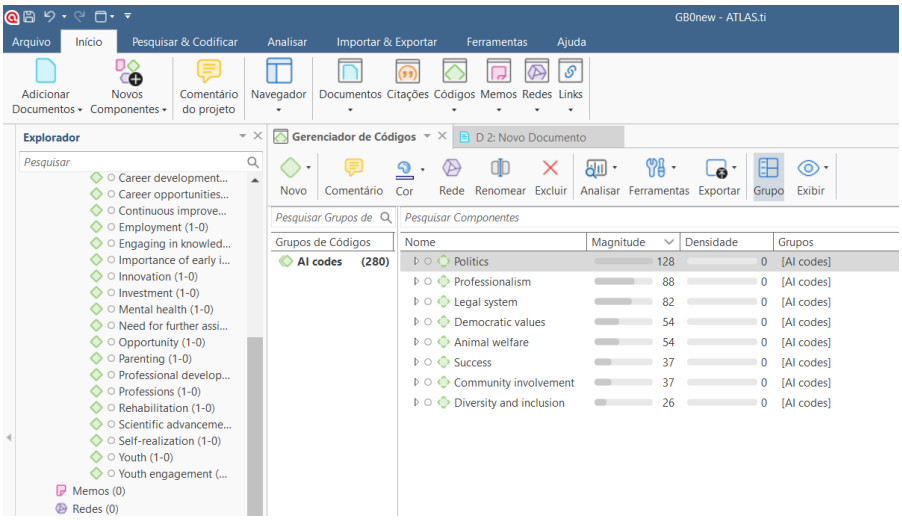

Figure 4. Themes and magnitudes in sample GB0

Within this corpus, several dominant sub-topics emerge. Government policies, political discourse, power dynamics, and public opinion are prominently represented. The legal system is another significant topic, including various legal aspects such as legislation, law enforcement, and criminal justice. Animal welfare is addressed through sub-topics such as animal rights, cruelty, experimentation, and testing. Diversity and inclusion are also prominent, highlighting issues related to activism, cultural sensitivity, social justice, and inclusivity. This reflects a growing awareness of diversity and inclusion matters.

In summary, the data reveals a rich variety of thematic content, with significant emphasis placed on political, legal, and professional discourse, as well as issues related to diversity, inclusion, and community engagement.

4. DISCUSSION

4.1. Conclusion

The three countries converge on politics and government, such as international relations, diplomacy, and global affairs, indicating a shared focus on governmental issues and political processes. Social issues, such as community involvement, diversity, inclusion, and societal concerns, are also prevalent, reflecting a shared concern for societal welfare. Legal frameworks and ethical standards feature other common topics in parliamentary speeches.

Despite these commonalities, the frequency distribution of topics varies significantly, and instances of the topics diverge across countries. For instance, the Austrian corpus emphasizes energy policy and crisis management, particularly regarding Europe's dependency on Russian gas during the Russian invasion of Ukraine. In contrast, the British corpus focuses more on international relations and government processes. Additionally, entries on formal processes, parliamentary procedures, and bureaucratic structures are more prominent in the Austrian corpus than in the British and Portuguese corpora.

4.2. *Limitations*

This study explored the use of topic modeling (LDA) and qualitative research analysis with ATLAS.ti in three multilingual text corpora of parliamentary speeches from Austria, Great Britain, and Portugal. To validate topic modeling and ATLAS.ti, the researchers came from various domains of expertise in sociology, applied linguistics, and public health. The limitations of this study parallel the limitations of LDA and ATLAS.ti's application of OpenAI's model, which can capture complex patterns in large datasets but cannot identify novel and unique patterns. Important data points out that, in the digital landscape, information-seeking behaviours, data voids, and hierarchy within data go undetected with these tools (Eysenbach 2009; Golebiewski and Boyd 2019). Any information that is considered novel and unique is easy to miss using traditional machine learning (LDA) and deep learning (ATLAS.ti) techniques because of the need for large datasets and regular updates to large language models, such as OpenAI's model; and limitations in linear algebraic methods and cluster analysis (OpenAI API; Nicolau et al. 2011; Carlsson 2020).

Consequently, this potentially mystifies instances where governments can impede their own ability to respond to misinformation (Pomeranz and Schwid 2021). If there is little quality content for any search engine to return, much is left to the imagination of both well-meaning governments and adversarial actors to weaponize misinformation

(Golebiewski and Boyd 2019). Misinformation as a threat to democracy was identified through sociological analyses in our preliminary research (Bayer et al. 2021; House of Lords Media Notices 2020). However, topic modeling and ATLAS.ti did not readily extract nor provide remnants to identify it as a topic, sub-topic, or thematic field. Furthermore, a general search of utterances revealed the theme of misinformation as a social problem across corpora. In Great Britain, there were utterances concerning misinformation, disinformation, deepfake and fake news; for Austria, it was *Desinformation(s)(kampagne),* deepfake, fake news; and for Portugal, it was *desinformação*, deepfake and fake news. We suspect that the pattern is too unique and novel to be detected by traditional machine learning and deep learning techniques.

As a result, we believe that future research will need to ascertain to what extent such a mounting issue in society can escape analytic tools, and whether this is an alignment with the general problem of algorithmic bias and then being intertwined with social and linguistic biases. The strength of this preliminary study is that it aligns with the main research objective of identifying relevant topics and subtopics within comparable corpora. These identified topics will be subjected to a more comprehensive analysis in subsequent stages of the research. In sum, this study sheds light on how topic modeling using LDA and AI coding using ATLAS.ti can facilitate the identification of topics.

REFERENCES

ATLAS.ti. AI Coding Powered by OpenAI. Accessed June 7, 2024. https://atlasti.com/ai-coding-powered-by-openai.

Bayer, Judit, Irin Katsirea, Olga Batura, Bernd Holznagel, Sarah Hartmann, and Katarzyna Lubianiec. 2021. «The Fight Against Disinformation and the Right to Freedom of Expression». Accessed June 12, 2024. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/695445/IPOL_STU(2021)695445_EN.pdf.

Blei, David. M., Andrew Y. Ng, and Michael I. Jordan. 2003. «Latent Dirichlet Allocation». *Journal of Machine Learning Research* 3 (1): 993-1022.

Carlsson, Gunnar. 2020. «Topological Methods for Data Modelling». *Nature Reviews Physics* 2 (12): 697-708.

Erjavec, Tomaž, et al. 2023. *Linguistically Annotated Multilingual Comparable Corpora of Parliamentary Debates ParlaMint.ana 3.0. Slovenian Language Resource Repository.* Accessed February 20, 2024. http://hdl.handle.net/11356/1488.

Eysenbach, Gunther. 2009. «Infodemiology and Infoveillance: Framework for an Emerging Set of Public Health Informatics Methods to Analyze Search, Communication and Publication Behavior on the Internet». *Journal of medical Internet research* 11 (1): e11.

Gostin, Lawrence O. 2022. «Living in an Age of Pandemics—From COVID-19 to Monkeypox, Polio, and Disease X». *JAMA Health Forum* 3 (9): e224062. Accessed February 20, 2024. https://jamanetwork.com/journals/jama-health-forum/fullarticle/2796824.

Golebiewski, Michael, and Danah Boyd. 2019. «Data Voids: Where Missing Data Can Easily Be Exploited». Accessed June 12, 2024. https://datasociety.net/wp-content/uploads/2018/05/Data_Society_Data_Voids_Final_3.pdf.

Guldi, Jo. 2019. «Parliament's Debates about Infrastructure: An Exercise in Using Dynamic Topic Models to Synthesize Historical Change». *Technology and Culture* 60: 1-33.

Gupta, Ajay. 2024. *Qualitative Methods and Data Analysis Using ATLAS.ti. A Comprehensive Researchers' Manual*. Berlin: Springer.

House of Lords Media Notices. 2020. *Democracy under Threat from «Pandemic of Misinformation» Online– Lords Democracy and Digital Technologies Committee. UK Parliament House of Lords Media Notices, 20 (June).* UK Parliament. Accessed on June 12, 2024. https://www.parliament.uk/business/lords/media-centre/house-of-lords-media-notices/2020/jun-20/democracy-under-threat-from-pandemic-of-misinformation-online-lords-democracy-and-digital-technologies-committee/.

Kuzman, Taja, et al. 2023. «Linguistically Annotated Multilingual Comparable Corpora of Parliamentary Debates in English ParlaMint-en.ana 3.0». *Slovenian Language Resource Repository*. Accessed February 20, 2024. http://hdl.handle.net/11356/1810.

McEnery, Tony, and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

Mikhailov, Mikhail, and Robert Cooper. 2016. *Corpus Linguistics for Translation and Contrastive Studies. A Guide for Research*. London, New York: Routledge.

National Center for Immunization and Respiratory Diseases. 2021. *How to Address COVID-19 Vaccine Misinformation. Centers for Disease Control and Prevention.* Centers for Disease Control and Prevention. Accessed February 20, 2024. https://www.cdc.gov/vaccines/covid-19/health-departments/addressing-vaccine-misinformation.html.

Nicolau, Monica, Arnold J. Levine, and Gunnar Carlsson. 2011. «Topology Based Data Analysis Identifies a Subgroup of Breast Cancers with a Unique Mutational Profile and Excellent Survival». *Proceedings of the National Academy of Sciences* 108 (17): 7265-70.

OpenAI API. OpenAI. Accessed June 12, 2024. https://platform.openai.com/docs/models.

Pomeranz, Jennifer L., and Aaron R. Schwid. 2021. «Governmental Actions to Address COVID-19 Misinformation». *Journal of public health policy* 42: 201-10.

Ristilä, Anna, and Kimmo Elo. 2023. «Observing Political and Societal Changes in Finnish Parliamentary Speech Data, 1980-2010, with Topic Modelling». *Parliaments, Estates and Representation* 43 (2): 149-76.

Wells, John S., and Florian Scheibein. 2022. «Global Pandemics, Conflict and Networks–the Dynamics of International Instability, Infodemics and Health Care in the 21st Century». *Journal of Research in Nursing* 27 (3): 291-300.

Wilhelm, Elisabeth et al. 2023. «Measuring the Burden of Infodemics: Summary of the Methods and Results of the Fifth WHO Infodemic Management Conference». *JMIR infodemiology* 3 (1): e44207.