Thematic Markers and Keywords on the Example of German Political Discourse

Maria Khokhlova
and Mikhail Koryshev

St Petersburg State University, Universitetskaya emb. 7-9-11,
199034 St Petersburg, Russia
m.khokhlova@spbu.ru, m.koryshev@spbu.ru

Abstract. The paper presents the results of keyword extraction and topic modeling based on LDA model from gensim applied to the texts of a German journal "Merkur" collected from 2017 to 2022. The algorithm extracted the most frequent topics that receive attention in the journal and their change over time. The authors also analyze the similarity of the articles with each other.

Keywords: Topic markers, keywords, German language, political discourse

1 Introduction

Topic modeling, like many other applied tasks, was initially carried out on English data. Each text can be represented by several topics, thus it is possible to determine the similarity of the texts. These topics and related keywords allow one to get an idea of the thematic content of texts and reveal latent semantic structures. Applied to the analysis of political discourse, the selection of thematic markers makes it possible to demonstrate which topics are popular, indicate interest in the author's position, and give an idea of the main ideas in the text. The paper [6] is devoted to the study of the materials of the US Senate meetings: the selection of keywords, possible topics, their clustering, as well as their change over time. In [3], the texts of speeches at plenary sessions in the European Parliament were analyzed using a non-negative matrix decomposition (NMF). The dynamics of the discussion of various topics from 1999 to 2014 was traced, and a regression model was built that took into account party membership, the number of speeches, voting for or against a party group, etc. Recently, topic models have also been used to cluster literary texts (see, for example, [8]).

In the German linguistic tradition, special attention is paid to key political issues and their reflection in journalism. Attitudes towards the law on renewable energy sources (Erneuerbare-Energien-Gesetz) were studied by the authors on the basis of German newspapers using structural topic modeling (STM) in [2]. The perceptions of the southern countries (Portugal, Italy, Greece, Spain) were

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023, pp. 165–174, 2023.

analyzed using the same method in the German-language press from 1946 to 2009 on the material of the newspaper "Die Welt" [4].

Our work is part of a project focused on the study of German and Russian political discourse. The working hypothesis is that the extracted markers make it possible to track changes in topics that receive public interest and values shared by groups of people, outline a range of important issues, as well as attitudes towards them over time, taking into account the historical and cultural context that contributes to these changes. The paper presents the results of applying the procedures for topic modeling applied to the texts of a German journal. The most frequent topics that receive attention in the articles were identified, the change in topics over time was traced, and the similarity of the articles with each other was analyzed.

2 Methodology

2.1 Text selection

Founded as a monthly magazine in 1947, "Merkur" [5] follows the idea proclaimed in the subtitle, i.e. "German magazine about European thinking": it publishes leading humanists on politics, aesthetics, social studies, economics, art and literature, where questions are raised, currently perceive attention from the professional German university community. These publications are extremely important, as they prepared the transition of the post-war Germany from German-centric thinking in humanitarian higher education to the European and, let us add, transatlantic vision of the post-war period in the spirit of Robert Schumann. A distinctive feature of the published materials is that the articles do not provide a deepening into the particular problems of the narrowly professional occupations of the authors, but still show how the results of particular subject research allow us to come to conclusions and generalizations of an interdisciplinary and generally significant nature, which are important for the professional humanitarian readership as a whole. At the same time, the conclusions of the authors hit the target set out in the subtitle of this publication, serving the cause of the inclusion of German thought in the pan-European and (earlier, transatlantic) now - globalist context. It should be noted that the editors deliberately make the most interesting texts publicly available, thereby expanding their audience and strengthening their influence - it will not be wrong to say that involvement in the texts of "Merkur" is a kind of pass to the German humanities academic world, and the habit of reading and discussing it publications are perceived as a sign of involvement in the current humanitarian agenda.

We selected articles from 2017 to 2022, access to which is carried out without a subscription (this explains the different amount of data, see Table 1), which allows us to look at those central materials for the editorial board, which, taking into account what has been said above should serve to create a common European humanitarian space through the integration into it of German mentality proper.

Table 1: Text data.

	2017	2018	2019	2020	2021	2022
number of texts	54	18	26	33	22	18
number of tokens	99,906	37,519	68,629	131,266	84,040	73,797

In total, the collection comprises 171 texts (about 500 thousand tokens) that prove to be heterogeneous both in their structures and genres.

2.2 Methods

The purpose of our study is to prove, using automatic procedures, the possibility of identification of topics that were manually identified by experts, as well as to describe and assess thematic components of texts from different time periods.

Preprocessing of texts was carried out using the following procedure: lemmatization was carried out using the HanTa [9] tagger (it shows the highest results for German texts), and stop words and auxiliary parts of speech were removed using the NLTK library. Additionally, rare vocabulary and high-frequency words that can "noise" topics were filtered out: words that occur in less than 5 documents and in more than 75% of documents were not considered. Using Latent Dirichlet Allocation (LDA) [1], implemented in the gensim library, seven models were built, which made it possible to identify the most frequent topics for each year and for the entire corpus as a whole.

To select the optimal number of topics, the coherence measure Cv was used, which demonstrates the most successful results in solving this task [7]. The measure takes values from 0 to 1: the higher the value, the greater the coherence between words, respectively, the better the model or the more interpretable the selected topics are.

3 Results

3.1 Coherence

The interval from 4 to 20 topics was chosen as the range of the number of possible topics. Coherence graphs allow determining the optimal number of selected topics for the considered collections of articles (see an example in Fig. 1).

The measure of coherence showed the highest value for the given sample for 10 topics.

In general, the number of topics does not differ significantly for the examined samples (see Table 2), the standard deviation is 1.13. For corpus 2017–2022, as expected, the largest number of topics was proposed (Cv=0.53).

The number of selected topics needs to be discussed in detail. If there are few of them, then only general topics will be marked, while with an increase in their number, their "fractionality" increases and more intersections appear, which can make it difficult to interpret the thematic components of the texts.

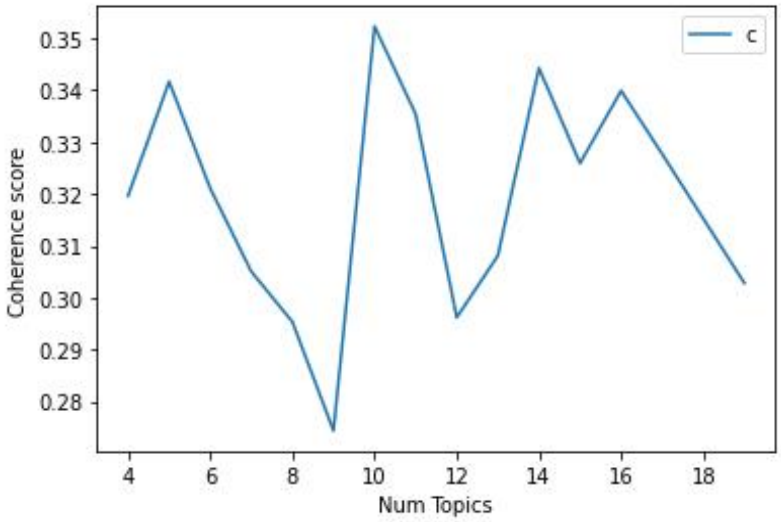

Fig. 1: Coherence for the texts from 2017-2022

Table 2: Number of topics according to the coherence measure.

	2017	2018	2019	2020	2021	2022
number of topics	5	6	6	6	7	8
number of topics ascribed by experts	6	7	4	6	5	5

3.2 Extracted topics

Each topic is presented as a set of keywords with weights assigned to them. In order not to exceed the volume of the paper, we will give an example for one sample of 2017–2022 with the assigned topics:

1) books (book reviews, discussion); 2) sexism; 3) social and political life; 4) literary texts; 5) literature; 6) university life (lectures, freedom of speech); 7) culture and art; 8) theater (theatrical performances).

Below there is a list of keywords extracted for the topic "university life":
(*'Universität'*, 0.00974275),
(*'tun'*, 0.009398366),
(*'Sarrazin'*, 0.009035184),
(*'wissenschaftlich'*, 0.007363963),
(*'Seminar'*, 0.0068394854),
(*'Meinungsfreiheit'*, 0.006065996),
(*'Buch'*, 0.00588211),
(*'deutsch'*, 0.005216127),
(*'ja'*, 0.0046629016),
(*'Du'*, 0.00464889),
(*'Frage'*, 0.0044861315),
(*'Sieg'*, 0.0040239994),
(*'politisch'*, 0.003979097),

('*System*', 0.003978175),
('*Wissenschaft*', 0.0038948406),
('*sagen*', 0.0037951404),
('*Fall*', 0.0037314473),
('*schreiben*', 0.0036307815),
('*solch*', 0.0036227151),
('*Person*', 0.003485797)

Some of the keywords refer to the same topics, showing the intersection between them (Gesellschaft, Kultur). The first topic can be labeled as "writer and creativity" and is the key topic for the June 2019 articles dedicated to the writer Wolfgang Hilbig (*"Den Debilen markieren ... und dann vielleicht klammheimlich schreiben". Ein Porträt des Arbeiters und Schriftstellers Wolfgang Hilbig*) and to the poet Helen Miles (*"Das Ich ist eine sehr bewegliche Angelegenheit" Interview mit Eileen Myles*). The second theme is, in a sense, a continuation of the first one, but it shows more English words. The next topic is formed by the article *"Die Politisierung der Unpolitischen: Moskau, mein Freund Sergej und das Recht auf Stadt"*, published in September, and deals with political events, protests, as well as human rights.

The LDA algorithm shows the probability with which a document belongs to a certain topic. Thus, the article *"Installation einer Freisprechanlage. Ein vorläufiger Bericht in elf Briefen"* (January 2019) belongs to the first cluster with a probability of 0.99, which includes university-related keywords. The content of the article confirms this: it is devoted to a seminar on philosophy and freedom of speech at universities.

In the case of topic modeling, we are talking about fuzzy clustering: a document can refer to several topics. Table 3 presents quantitative data on the distribution of documents by topic. Texts show three macroclusters: "books", "sociopolitical life" and "fiction".

Table 3: Distribution of documents by topics for the corpus 2017–2022.

topic	number of documents
books (book reviews, discussion)	68
sexism	3
social and political life	62
literary texts	21
literature	2
university life (lectures, freedom of speech)	3
culture and art	10
theater (theatrical performances)	2

Figure 2 shows a heat map showing the distribution of the selected topics and corresponding documents. Light colors correspond to a greater likelihood that the document is related to a given topic. The results confirm what was shown

above: it is possible to identify two macro-topics dedicated to fiction and literary studies.

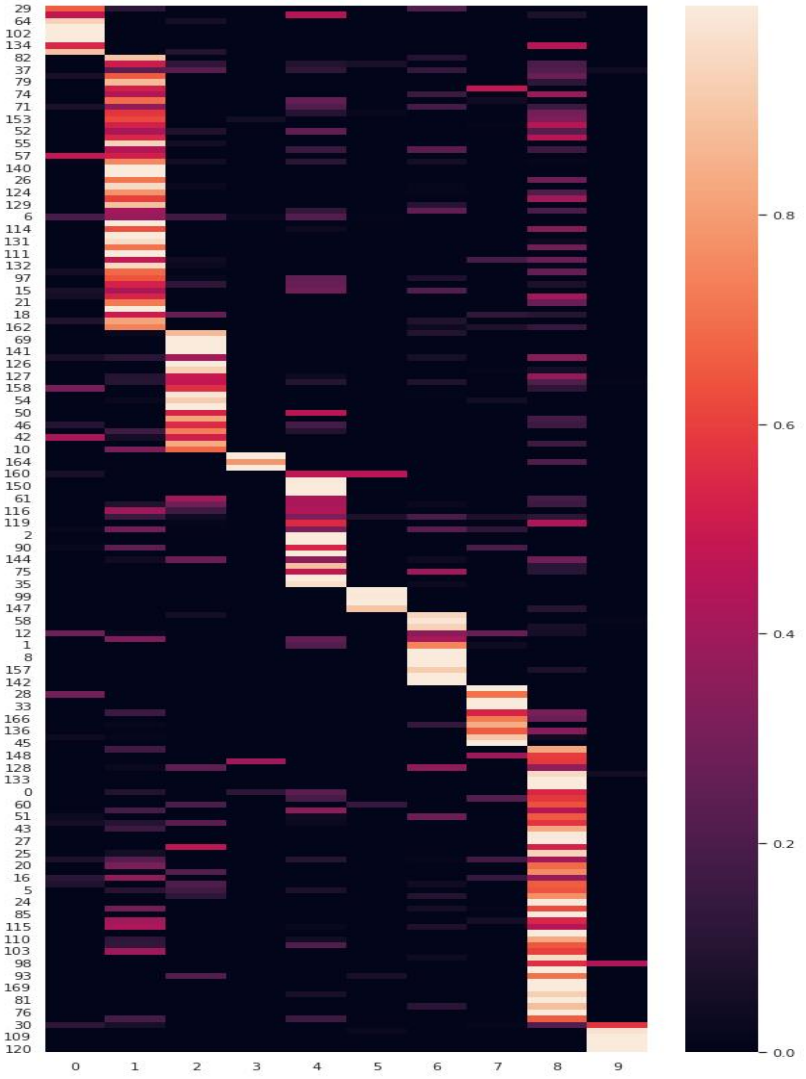

Fig. 2: A heat map for the distribution of documents according to the topics

The last topic, assigned by experts as "fiction" (it reveals general lexis that does not allow building proper clusters for keywords), is controversial, although 18 documents belong to it with a probability of more than 0.4. For example, *"Etc. (Warten; Notizen zur leeren Hand)"* (September 2018) - an article in which the author reflects on the texts themselves and their comprehension - or *"Hausbesuche IV: Bayreuth. Wagner sucht Wagner"* (June 2020) that are notes on the Wagner Festival in Bayreuth of the past year. Although issues related to the literature, theater and university life are given a lot of attention on the pages of the magazine, nevertheless, the algorithm attributed only several documents to these clusters.

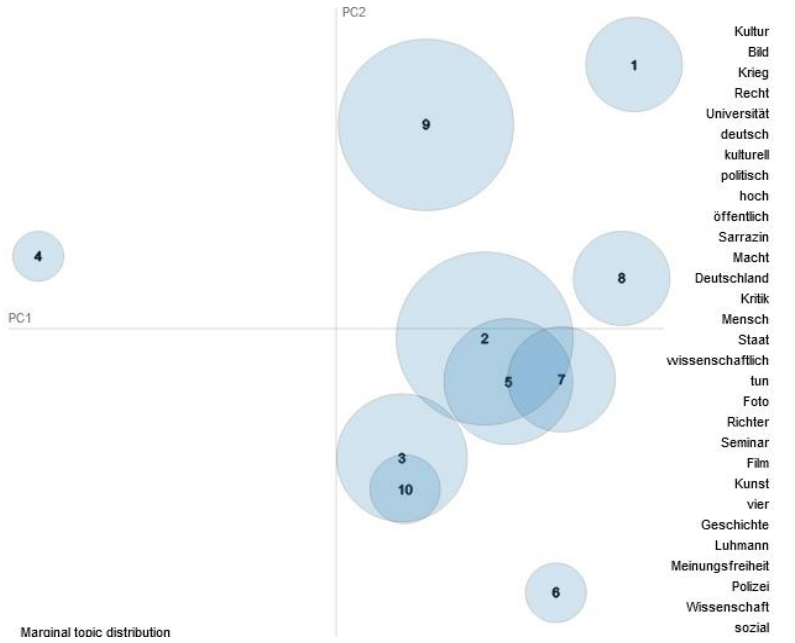

Fig. 3: Intersection of topic clusters

The intersection of topics in clusters is shown in Fig. 3. It should be noted that some topics intersect with each other, which to a certain extent makes it difficult to differentiate them. Topics 1, 4, 8, 6 and 9 are the most distinguishable, as they do not have a common vocabulary among the selected keywords.

The volume of texts in the samples by years is not sufficient: on the example of individual time periods, we noticed that the selected topics are almost identical to the content of some texts and the frequency vocabulary that occurs in them. For example, the article *"Rudolfsheim-Fünfhaus"* (October 2018) with a probability of 0.97 can be assigned a topic related to the description of life in the Rudolfsheim-Fünfhaus area of Vienna, which is very specific.

3.3 Similarity between texts

We calculated the similarity in terms of the cosine measure on tf-idf vectors using the TfidfModel function and the MatrixSimilarity similarity matrix (an example for the articles from 2019 is shown in Fig. 4).

Despite the homogeneous nature of the material, the articles are mostly heterogeneous in their structure and, when compared in pairs, show low similarity. When analyzing texts by years, the greatest similarity (measure value above 0.5) was demonstrated for the following pairs of articles.

1. The topic of elite culture was given attention in a whole series of publications in the magazine, which was reflected in the similarity of the texts:
   (a) *"Priceless. Die hohe Kultur und das Geld (Hohe Kultur 5)"* (April 2017) and *"Hohe Kultur (7)"* (August 2017).

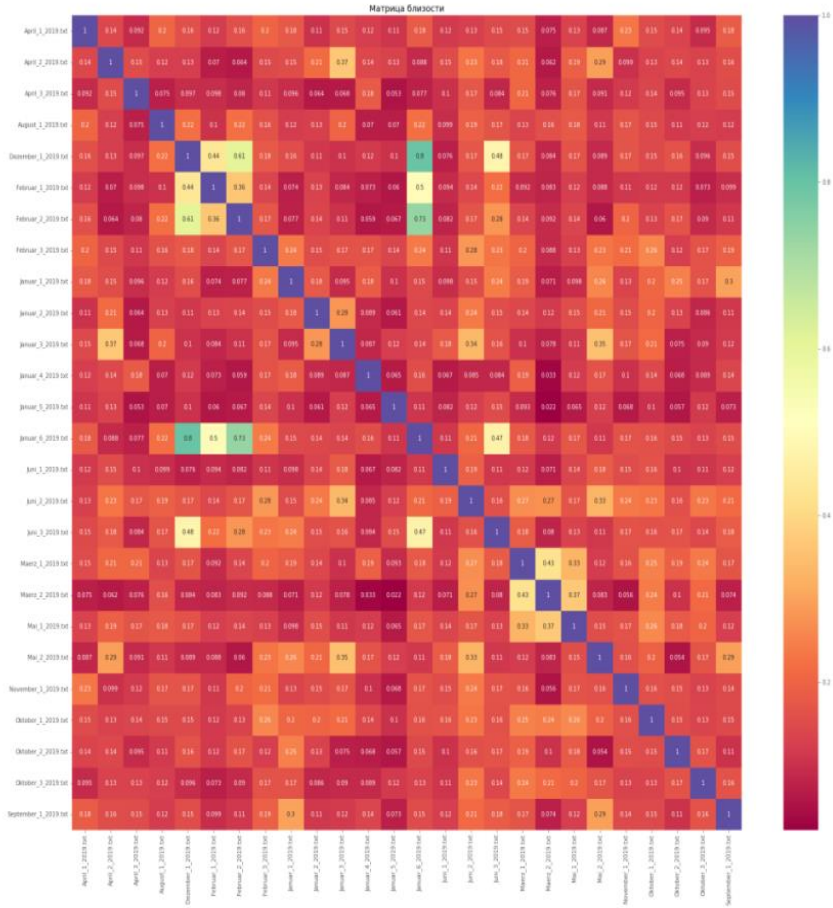

Fig. 4: A heat map for the texts from 2019

(b) *"Hohe Kultur (7)"* (August 2017) and *"Parteiprogramme: Kulturpolitik (Hohe Kultur 8)"* (September 2017).
(c) *"Hohe und niedrige Metaphern (Hohe Kultur 2)"* and *"Kooperation Pop und Merkur"* (February 2017).

2. Freedom of speech and academic freedoms was described in the articles:
(a) *"Installation einer Freisprechanlage. Ein vorläufiger Bericht in elf Briefen"* (January 2019) and *"Fortsetzung und Abschluss des Berichts: Installation einer Freisprechanlage"* (February 2019). Both texts belong to the same author (Erhard Schüttpelz).
(b) *"Installation einer Freisprechanlage. Ein vorläufiger Bericht in elf Briefen"* (January 2019) and *"Wissenschaftsfreiheit und Meinungsfreiheit (Aus Anlass einer Siegener Kontroverse)"* (December 2019). As in the example above, the last article is written by the same author and dwells on the subject of freedom.

3. The lectures of the Swiss writer and journalist K. Kracht:
   (a) *"Der Autor ist anwesend – Ein Abschlussbericht zu Christian Krachts Frankfurter Poetikvorlesungen"* and *"Blitz und Donner – Christian Krachts Frankfurter Poetikvorlesungen als werkbiographische Zäsur"* (May 2018). These articles are written by the same authors (Kevin Kempke; Miriam Zeh) and are focused on the mentioned topic.
4. The topic of politics and mistakes made in foreign policy issues is described in the articles:
   (a) *"Über Fehler in der Politik"* (by Ulrich K. Preuß, June 2022) and *"Fehler in der Politik?"*, (by Franziska Davies, September 2022). A later issue provides commentary on the topics that were raised during the summer.

Other time periods did not demonstrate similarity (the measure takes a low value of less than 0.4, despite the fact that a number of articles in the journal belong to one topic and are a series of publications). While the articles published in the same year turn out to be thematically similar.

4 Conclusion and Future Work

In the paper, we extracted the most frequent topics for the texts of the articles of the "Merkur" magazine published in different years, as well as evaluated the collection as a whole in terms of its thematic content. The LDA algorithm made it possible to identify significant thematic markers, which generally coincide with the expert evaluation and with the content of the articles. The analysis showed a rather low similarity between the texts of different years, however, within the same year samples, similar texts were identified according to the tf * idf measure.

In general, the algorithm has demonstrated successful results, but additional analysis is needed for such tricky linguistic data. More data is required, as well as the evaluation of topics with keywords identified by other algorithms. It is also important to extract longer n-grams, which will allow for a more "elaborate" identification of topics.

Acknowledgements. The presented research was supported by the Russian Science Foundation, project No. 24-28-00937 "Philological regional studies: mindset of German society in an era of instability".

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet Allocation. Journal of Machine Learning Research 3 (4–5), 993–-1022 (2003)
2. Dehler-Holland, J., Schumacher, K., Fichtner, W.: Topic Modeling Uncovers Shifts in Media Framing of the German Renewable Energy Act. Patterns 2, 100–169 (2021)
3. Greene, D., Cross, J.P.: Unveiling the Political Agenda of the European Parliament Plenary: A Topical Analysis. In: Proceedings of the ACM Web Science Conference (WebSci'15), Oxford, UK, pp. 1–10 (2015)

4. Küsters, A., Garrido, E.: Mining PIGS. A structural topic model analysis of Southern Europe based on the German newspaper Die Zeit (1946-2009). Journal of Contemporary European Studies 28 (4), 477–493 (2020)
5. Merkur, `https://www.merkur-zeitschrift.de`. Last accessed 5 Nov 2023
6. Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., Radev, D. R.: How to analyze political attention with minimal assumptions and costs. American Journal of Political Science 54(1), 209–228 (2010)
7. Röder, M., Both, A., Hinneburg, A.: Exploring the Space of Topic Coherence Measures. In: Proceedings of the eight International Conference on Web Search and Data Mining, Shanghai, February 2-6, pp. 399–-408 (2015)
8. Sherstinova, T.Yu., Moskvina, A.D., Kirina, M.A., Karysheva, A.S., Kolpaschikova, A.E.: Thematic modeling of the Russian short story 1900–1930: the most frequent topics and their dynamics [Tematicheskoe modelirovanie russkogo rasskaza 1900–1930: naibolee chastotnye temy i ih dinamika]. In: Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2022", Issue 21, vol. 21. Russian State University for the Humanities, pp. 512–526 (2022)
9. Wartena, Ch.: A probabilistic morphology model for German lemmatization. In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Long Papers, Erlangen, Germany, pp. 40–-49 (2019)