**Polıtıcal Dıscourse in the Kazakh Language: A Revıew of the Lıterature and the Creatıon of a Text Corpus**

***Ayaulym SAIRANBEKOVA1,[[1]](#footnote-1)\*[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0000-0002-0814-1532), Laura ORYNBAY1[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0009-0006-4787-7944), Banu YERGESH2[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0000-0002-8967-2625), Gulmira BEKMANOVA2[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0000-0001-8554-7627), Altynbek SHARIPBAY1[C:\Users\Abdullah\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ORCID-iD_icon-16x16.gif](https://orcid.org/0000-0001-5334-1253)***

*1Faculty of Information Technology, Department of Artificial Intelligence Technologies, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan*

*2Digital Development and Distance Learning Department, L. N. Gumilyov Eurasian National University, Astana, Kazakhstan*

|  |
| --- |
| **Abstract**  The corpus of political discourse texts is a collection of text documents containing texts related to political topics, debates and discussions. Such texts can be political speeches, laws, political programs, articles, discussions on social networks, news materials, interviews with political figures and much more.  In this article, the results of a study on the review of literature at the world and local level are written, the work of various researchers is analyzed and examples of their work are given.  This article also provides an overview of the analysis of Internet resources devoted to political discourse in the Kazakh language, as well as describes the initial steps to create a corpus of texts for research in this area. The study's significance lies in supporting the concept of a “hearing state” and contributing to scientific and technological development in Kazakhstan. |
| Keywords: political discourse, a text corpus, analysis, Kazakh language, Internet source |

1. **Introduction**

Political discourse, or the language of power and governance, is critical to understanding the complexities of politics, ideology, and public opinion. In the age of the digital revolution, the Internet has become a valuable repository of political discourse. It allows us to understand the beliefs, aspirations and debates that influence countries and communities.

This article is devoted to a comprehensive study of political discourse analysis in the Kazakh language and this research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP19679847).

The first part of the article reviews relevant studies and methodologies related to the creation of structured semantic databases for analyzing political discourse.

The second section is devoted to the study of online resources devoted to political discourse. The issues of categorization and statistical analysis of the huge array of Internet resources collected for the analysis of political discourse in Kazakhstan are considered.

The third part then examines the process of creating a corpus of political texts in Kazakh. The methodology of creating a corpus of political discourse texts is revealed, and the use of a synonymizer containing typical examples of synonymous words in socio-political discourse is demonstrated.

1. **Literature Review**

Political discourse analysis (PDA) is an interdisciplinary field that plays a key role in understanding political communication, ideology and public opinion [1]. Creating structured semantic databases for political discourse in a specific language, such as Kazakh, involves various linguistic, computational and socio-political challenges. This literature review provides an overview of relevant studies and methodologies related to the creation of structured semantic databases for political discourse analysis.

*Political Discourse Analysis and Semantics*. An overview of key theoretical and methodological approaches to analyzing political discourse was presented by Chilton [2]. This book provides detailed case studies of how political discourse shapes public opinion and policy outcomes. For example, he analyzes how the discourse of European integration has evolved over time, from the early days of the European Union to the present day, and how it has been shaped by different actors and ideologies.

Regarding key studies on semantics in the PDA, one study examined the argument that in political contexts metaphor can be used and is often used for ideological purposes because it activates unconscious emotional associations and thereby contributes to the creation of a myth: politicians use metaphor to tell the right story [3].

Semantic approaches have also been applied to analyze political speeches in the context of emotional expression [4]. The study used an affective computing approach to automatically detect emotional content in political speeches and identify rhetorical strategies through semantic analysis. The results showed the importance of emotional language in political communication and emphasized the value of using semantic approaches for PDA.

*Language Resources and NLP*. Fundamental knowledge about language processing methods and technologies needed to analyze, categorize and extract information from political texts, speeches and debates can be found in the book by Jurafsky and Martin [5]. The comprehensive coverage of speech and language processing concepts provides a solid foundation for the development of tools and systems that meet the specific requirements of analyzing political discourse in the Kazakh language.

*Sentiment Analysis in Political Discourse*. Sentiment analysis, especially in the context of political discourse, is a crucial area of research. Bekmanova et al. [6] discuss sentiment processing in socio-political discourse and public speeches, emphasizing the importance of understanding public and social media sentiment. They propose methods and tools for analyzing emotional states and opinions expressed through textual data.

*Analyzing emotional states in social networks*. The work of Yergesh and Kenjina [7] deals with analyzing the emotional states of users on social media. The study focuses on identifying and understanding the emotions expressed by users on social platforms. This study is in line with the broader task of analyzing sentiments in socio-political contexts, as it aims to identify sentiments in society.

*Text Corpus and Corpus Linguistics*. A foundational text that discusses the methodology, theory, and practical applications of corpus linguistics can be found in McEnery and Hardie [8], whereas Baker and Egbert [9] provide valuable insights into the methods and techniques used in corpus linguistic research. These works serve as foundational texts for researchers and linguists interested in studying language patterns, usage, and discourse within large datasets called corpora. Although these studies do not specifically focus on political discourse, they provide important insights into corpus linguistics, which is highly relevant in the context of creating a structured semantic database of political discourse in Kazakhs.

*Information Sources for Political Discourse in Kazakhstan*. Valuable resources for collecting data on political discourse in Kazakh are various online news agencies, political party websites and social media platforms. These sources are analyzed in more detail in the next section.

1. **Examination of Online Resources regarding Political Discourse**

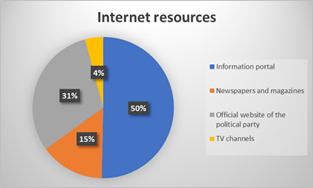
A variety of sources can be used to create a text corpus of political discourse via the Internet. For example, official government websites can provide access to laws, regulations, official statements, and other policy documents; News websites and agencies provide news, analysis, and commentary on political events and topics; Websites of political parties and organizations contain political programs, speeches of leaders and other materials; Publications and comments on social networks where users discuss political topics; Political blogs and forums may contain opinions and discussions about current political issues; Digital libraries and databases that provide access to academic articles and research in political science and sociology; Videos of political speeches, interviews and discussions on platforms such as YouTube; Online archives containing historical political documents and materials; Recordings of public speeches by political figures available online; Many newspapers and magazines have online versions where articles on political topics can be found.

In Kazakhstan, political texts can be found in such Internet resources as information portals, newspapers and magazines, the official website of the political party, and TV channels (Table 1).

**Table 1.** Categories of Internet resources and the Number of Resources Collected.

|  |  |
| --- | --- |
| Type of Internet resource | Аmount of resources collected |
| Information portal | 439 |
| Newspapers and magazines | 128 |
| Official website of the political party | 269 |
| TV channels | 36 |
| Total | 872 |

In this article, not only the results of the text corpus are published, but also the statistics of Internet resources for creating the corpus are described in detail. Of these, 50% are information portals, such as news sites, district blogs, and official and unofficial sites of cities and 31% are official sites of political parties, the remaining 15% are data from newspapers and 4% are TV channels (Figure 1).



**Figure 1**. Categories of Internet resources and the Number of Resources Collected (in percentage).

Analyzing political discourse from online sources requires a discerning approach, considering factors like source evaluation, bias and objectivity, relevance, audience and intent, and cross-referencing. Reliable sources of political discourse in the Kazakh language include online publications like “Informburo.kz” [10], “Tengrinews” [11], “Kazpravda.kz” [12], “Sputnik” [13], “BAQ.KZ” [14], “Kazinform” [15], “Egemen Qazaqstan” [16] and more. Bias, objectivity, and the relevance of information must be assessed when evaluating these sources.

Political discourse occurs on the web pages of political parties in Kazakhstan, such as “AMANAT” [17], “Nation's Party of Kazakhstan” [18], “Ak Zhol” [19], “Auyl” [20], and “Baitaq” [21]. Additionally, political science journals and newspapers serve distinct audiences and purposes in shaping political discussions.

1. **Establishing a Political Text Corpus in the Kazakh Language**

To analyze political processes and opinions in Kazakhstan, a critical analysis of Internet resources related to political discourse was conducted, emphasizing the importance of evaluating sources, identifying bias, and using critical thinking [22]. Based on the analyzed sources, the work on creating a text corpus of political discourse in the Kazakh language was initiated. For this purpose, the previously created synonymizer of standard samples of synonymic words in socio-political discourse and public speech was used [6]. The synonymizer contains 1,000 dictionary entries on the topics “Pre-election advertising”, “Speech of political candidates”, and “Pre-election debates”.

The Synonymizer consists of several columns: “Word”, “Part of Speech”, “Status” (homonym, unambiguous, polysemic), “Meaning”, “Example”, “Synonym”, and “Periphrasis” (Figure 2). It serves as a valuable tool for analyzing and comparing synonyms in texts related to socio-political discourse and public speech.



**Figure 2**. A segment from the synonymizer containing standard examples of synonymous words commonly used in socio-political discourse and public speaking.

1. **Conclusion**

As a result of this research, the sources for the corpus of texts were identified and its formation was successfully carried out. However, since the goal of the project is to develop methods for analyzing political discourse in Kazakh-language social networks to identify both official and unofficial sources of political discourse, as well as to determine the mood of discussions in these sources, the work on this topic will still continue. To achieve this goal, the following tasks will be solved in the future: creation of ontological models on election topics, such as “Pre-election advertising”, “Speeches of political candidates”, and “Pre-election debates”; development of knowledge bases with semantic attributes; formal representation of logical rules for drawing conclusions from these knowledge bases; development of a processor for processing official and unofficial sources of political discourse; creation of a tool (software application) for analyzing the mood of official and unofficial sources of political discourse; creation of a tool (software application) for analyzing the mood of official and unofficial sources of political discourse.

**References**

1. Wodak, R. (2015). Critical discourse analysis, Discourse‐historical approach. *The International Encyclopedia of Language and Social Interaction*, 1–14.
2. Chilton, P. (2004). *Analysing political discourse: theory and practice*. London, United Kingdom: Routledge.
3. Charteris-Black, J. (2011). *Politicians and Rhetoric: The Persuasive Power of Metaphor*. Palgrave Macmillan.
4. Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., & Ungar, L. (2017). Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Stroudsburg, PA, USA: Association for Computational Linguistics.
5. Jurafsky, D., & Martin, J. H. Speech and Language Processing (Draft-Version from Dec 2020).
6. Bekmanova, G., Yergesh, B., Ukenova, A., Omarbekova, A., Mukanova, A., & Ongarbayev, Y. (2023). Sentiment Processing of Socio-political Discourse and Public Speeches. In *Computational Science and Its Applications – ICCSA 2023 Workshops* (pp. 191–205). Cham: Springer Nature Switzerland.
7. Yergesh, B., & Kenzhina, L. (2021). Analysis of the users’ emotional state in social networks. In *ICEMIS'21: The 7th International Conference on Engineering & MIS 2021*. New York, NY, USA: ACM.
8. Hardie, A., & McEnery, T. (2011). *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press.
9. Egbert, J., & Baker, P. (2016). *Triangulating Methodological Approaches in Corpus Linguistic Research*. Taylor & Francis Group.
10. News Informburo.kz - the latest news in Kazakhstan and the world. (n.d.). Retrieved October 25, 2023, from https://informburo.kz/kaz.
11. Tengrinews.kz - The main news portal of Kazakhstan. (n.d.). Retrieved from https://kaz.tengrinews.kz/.
12. Kazakhstan News - current, latest news about everything. (n.d.). Retrieved from https://kazpravda.kz/.
13. Sputnik Kazakhstan. (n.d.). Retrieved from https://sputnik.kz.
14. BAQ.KZ - News. Latest news from Kazakhstan. (n.d.). Retrieved from https://baq.kz.
15. News of Kazakhstan and the world - latest news of the day on inform.kz. (n.d.). Retrieved from https://inform.kz.
16. Sovereign Kazakhstan – Egemen Qazaqstan. (n.d.). Retrieved from https://egemen.kz.
17. AMANAT. (n.d.). Retrieved from https://amanatpartiasy.kz.
18. Official website of the People's Party of Kazakhstan (PPK) (n.d.). Retrieved from https://halykpartiyasy.kz/kz.
19. Democratic Party of Kazakhstan NGO “Ak Zhol”. (n.d.). Retrieved from https://akzhol.kz.
20. Party “Auyl”. (n.d.). Retrieved from https://auyl.kz.
21. «Baytaq» The Green рarty of Qazaqstan. (n.d.). Retrieved from https://baytaq.kz/kk/main-kz/.
22. Sairanbekova, A., Orynbay, L., Ukenova, A., Sharipbay, A., & Razakhova, B. (2023). Database Creation of Political Discourse in The Kazakh Language, The XI International Conference on Computer Processing of Turkic Languages “TurkLang 2023”, Bukhara, Uzbekistan (to be published).

1. \* Corresponding author. *e-mail address: sairanbekova98@gmail.com.* [↑](#footnote-ref-1)