Automatic analysis of linguistic features in Communist propaganda texts

Veronika Vincze[1]0000-0002-9844-2194, Martina Katalin Szabó[1,2]0000-0002-4192-4352, and Orsolya Ring[2]0000-0002-3710-1118

[1] MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
{vinczev,martina}@inf.u-szeged.hu
[2] MTA TK Computational Social Science - Research Center for Educational and Network Studies, Budapest, Hungary
{Szabo.Martina,Ring.Orsolya}@tk.mta.hu

Abstract. Here we will analyze Hungarian Communist propaganda texts using the digital corpus 'Pártélet', the official journal of the Central Leadership of the Hungarian Socialist Workers' Party between 1956 and 1989. We will focus on the first and the last two years of the era in question and compare some of the statistical, morphologic, syntactic and semantic features of texts from the two periods. Since these texts unequivocally represent a totalitarian language usage, the research results may be utilized in our subsequent political propaganda detection research and development tasks.

Keywords: Propaganda; Discourse analysis; Hungarian Communism

1 Introduction

The active and decisive period of Hungarian history from 1956 to 1989 is a widely examined topic in historical and sociological sciences. At the same time, the linguistic characteristics of the political discourse of this era has not been analyzed so far. To address this issue, we will examine our 'Pártélet' corpus using NLP methods.

'Pártélet' (1956-1989) was the official journal of the Central Leadership of the Hungarian Socialist Workers' Party. Hence it represents not just the media discourse of the era, but also the official discourse of the government. Propagating the political ideology with a special focus on practical aspects, it was an important tool for direct political agitation and propaganda. Articles published in 'Pártélet' were primarily intended for the party members (not for the average person). What is more, it often received letters from the leadership in connection with the work of the Socialist cooperatives and factories, as well as other issues concerning Hungarian society.

For our recent study we selected single-year periods between January 1957 and December 1958 (17 issues) and between January 1988 and April 1989 (16 issues). We selected these periods on the basis of specific historical and sociological criteria. The first period is the beginning of the Kádár era, directly following

the Hungarian revolution in 1956. This is the active starting period of the system. The second period may be viewed as a counterpoint of the first period, just before the regime change in Hungary.

Our main goal is to analyze the linguistic features of this type of discourse, with special regard to biases and deception in political language usage. We would like to explore how and to what extent political forces can affect linguistic choices [5]. This analysis is a preliminary step for our other research works, which will provide an appropriate base for developing an automatic detector of propaganda in online news in the future.

Despite the fact that the quantitative and qualitative analysis of propaganda discourse is important from an NLP and a political-historical point of view, very few studies seem to address the issue of the systematic analysis of a huge amount of propaganda texts [1, 6, 2]. This is a quite significant research gap concerning the Hungarian language; to the best of our knowledge, no study focusing specifically on the features of a Hungarian totalitarian discourse g NLP methods has been carried out so far.

2 Methodology

We hypothesize that authors of 'Pártélet' use both deception in order to conceal those facts that are undesirable for the political leadership, and persuasion, to manipulate beliefs and intentions of the readers [3, 6]. Since the object of the analysis is an ideological journal, the above-mentioned features were probably decisive in both periods: in the first period these features are mainly related to the revolution, and in the second the regime change. At the same time, we presuppose that the linguistic characteristics of the two subcorpora are different from several perspectives.

To achieve this goal, we preprocessed the texts with magyarlanc, a linguistic toolkit for morphological and syntactic parsing [10], then extracted basic statistical features from the subcorpora, concerning the number of sentences and words, and the average sentence length, as well as morphological and syntactic features. Next, in order to examine the phenomena of bias and deception, uncertain words belonging to several classes were extracted and their number and frequency were compared with the total number of tokens [9], together with sentiment and emotion expressions [4]. We also compared our results to the research findings of the Institute for Propaganda Analysis to ascertain what kinds of linguistic devices characterize propaganda texts [1].

3 Results

Our results reveal that there are significant differences between the two periods. In this section, we will just list some of the most interesting results of the analyses.

As for the frequency of the different verb moods across the journal issues, an interesting difference involves the use of imperatives: again, texts from the

beginning of the era use more imperatives, that is, the need for action – as a tool for expressing implicit orders from the party – can be detected there. In connection with the above-mentioned feature, the frequency of the conditional verb mood is very low in the first time period. However, it becomes frequent in 1988-89. This feature might be viewed as a sign of uncertainty – in other words, a sign of political decline.

The high occurrence of the superlative degree of adjectives and adverbs in the first time period also may function as a type of propaganda techniques, reflecting the power of the given political ideology and the confidence and unquestionability of the political system. Our results correlate with the research findings of [1] concerning the linguistic devices used in propaganda texts. Based on the phenomenon called as "glittering generality" (also called "glowing generality"), highly valued concepts and beliefs attract general acclaim and ask for approval without examination of the reason. In addition, the so-called "cherry picking" principle asserts that suppressing evidence, or the fallacy of incomplete evidence point to individual cases or data that seem to confirm a particular position and ignore related cases or data that may contradict that position.

Results concerning the emotions and sentiments of the corpus texts accord with the features described above. Positive emotions like joy and love are mentioned more frequently at the beginning of the era – again a typical device for propaganda: the results agree with the statements in [1] emotionally appealing phrases closely associated with highly valued concepts and beliefs (such as love of country and home, and desire for peace, freedom, glory, and honor) are frequent in propagandistic texts (cf. "glittering generality "). By way of contrast, negative emotions like anger, fear and anxiety occur more frequently in 1988-89, which might anticipate upcoming changes in the Communist regime. On the basis of the above-mentioned features we may conclude that powerful and confident communication, characterizing the first time period became a journal with a more uncertain and powerless discourse by the end of the political era.

As for the frequency of verb tenses in the two subcorpora, a significantly bigger amount of past tense verbs is used in 1956-57 than in 1988-89. A manual analysis of the data revealed that this feature is related to the fact that the effects of a specific past event, namely the 1956 revolution and the political decisions and steps that led to the revolution were intensively discussed at that time.

We also carried out an analysis of the words or phrases denoting uncertainty of the speaker in the veracity of the information expressed in the texts. The results of the analysis showed that three types of uncertainty at the discourse level, namely weasel, peacock and hedge are quite frequent in the first subcorpus compared to that in the second time period [8]. In contrast to semantic uncertainty [7], in the case of discourse level uncertainty "the missing or intentionally omitted information is not related to the propositional content of the utterance but to other factors", e.g. *some, often, much* etc. Bias evoked by these expressions might be viewed as a characteristic feature of propaganda discourse.

In contrast to discourse level uncertainty, the elements of the so-called semantic uncertainty occur significantly less frequently in the first time period. In this

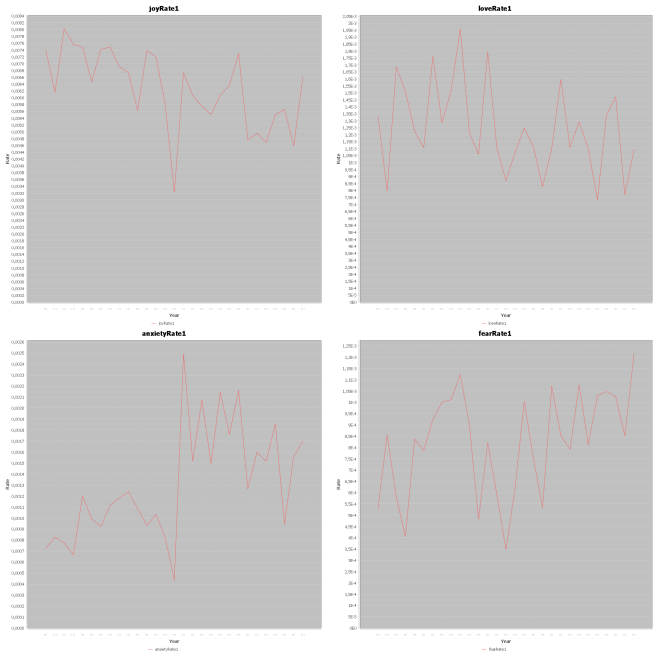

Fig. 1. Frequency distribution of some of the emotions over time

case it is the lexical content (meaning) of the uncertainty marker (cue) that is responsible for uncertainty, e.g. *may, possible, believe* etc. [9]. For instance, epistemic and doxastic types of semantic uncertainty are relatively rare in the first subcorpus. Our results again correlate with the research findings of [1] concerning the linguistic devices in propaganda texts: based on "glittering generality", there is a high accuracy of emotionally appealing phrases that carry conviction without supporting information or reason. From these results we may conclude that the texts of the first subcorpus are implicitly more deceptive compared to the texts of the second subcorpus.

Lastly, the frequency of first person plural verb forms is notably higher in the second subcorpus compared to the first one. A manual analysis of the dataset of the two time periods showed that while in 1956-57 the content of the texts addresses other actors in politics and society (for instance, those groups whose acts and decisions led to the revolution), in 1988-89 the authors discuss the leadership itself.

We also made some lexical analysis of the most frequent words belonging to certain parts-of-speech, i.e. we compared the most common verbs, nouns and adjectives of the two time periods. The results tell us that, for instance, there are a lot of nouns in the first subcorpus semantically related to total authority and power of the leadership (e.g. *project, plan, fullfilment* etc.). At the same time, in the second subcorpus nouns occurring with a higher frequency express the

possibility of divergent opinions and the opportunity of choice between different options (e.g. *possibility, decision, opinion*).

4 Conclusions

In this paper, we provided a linguistic analysis of characteristic features of Hungarian propaganda texts from the 1950s and the 1980s, the beginning and the end of the Communist era. We focused on the morphological, syntactic and semantic features of the texts and also took into account the differences over time.

The results presented in this paper might contribute to the linguistic characterization of the language of propaganda in the Communist era. Moreover, from a historical perspective we can also analyze how these linguistic features change over time. As a next step of the research work we will analyse non-propagandistic texts using the same methods and tools and compare the results with those of our recent analysis. In the long run, we would like to implement an automatic detector of propaganda, for which our current study can serve as a preliminary step.

References

1. How to detect propaganda. Propaganda Analysis. Publications of the Institute for Propaganda Analysis **I**, 210–218 (1938)
2. Barrón-Cedeño, A., Jaradat, I., Martino, G.D.S., Nakov, P.: Proppy: Organizing the news based on their propagandistic content. Information Processing Management **56**(5), 1849–1864 (2019). https://doi.org/10.1016/j.ipm.2019.03.005, https://doi.org/10.1016/j.ipm.2019.03.005
3. Girlea, C., Girju, R., Amir, E.: Psycholinguistic features for deceptive role detection in werewolf. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 417–422. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/N16-1047
4. Horne, B.D., Khedr, S., Adali, S.: Sampling the news producers: A large news and feature data set for the study of the complex media landscape. In: Proceedings of the Twelfth International Conference on Web and Social Media, ICWSM 2018, Stanford, California, USA, June 25-28, 2018. pp. 518–527 (2018), https://aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17796
5. Jalilifar, A.R., Alavi, M.: Power and politics of language use: A survey of hedging devices in political interviews. The Journal of Teaching Language Skills (JTLS) **3(3)**, 43–66 (2011)
6. Rashkin, H., Choi, E., Jang, J.Y., Volkova, S., Choi, Y.: Truth of varying shades: Analyzing language in fake news and political fact-checking. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 2931–2937. Association for Computational Linguistics, Copenhagen, Denmark (Sep 2017). https://doi.org/10.18653/v1/D17-1317
7. Szarvas, Gy., Vincze, V., Farkas, R., Móra, Gy., Gurevych, I.: Cross-genre and cross-domain detection of semantic uncertainty. Computational Linguistics **38**, 335–367 (June 2012)

8. Vincze, V.: Weasels, hedges and peacocks: Discourse-level uncertainty in wikipedia articles. In: Proceedings of the Sixth International Joint Conference on Natural Language Processing, Nagoya, Japan, October 2013. pp. 383–391 (2013), https://pdfs.semanticscholar.org/4a9e/b494a633fb36a6971442bf85be85ea5839fa.pdf
9. Vincze, V.: Uncertainty Detection in Natural Language Texts. Ph.D. thesis, University of Szeged, Szeged, Hungary (2014)
10. Zsibrita, J., Vincze, V., Farkas, R.: magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In: Proceedings of RANLP. pp. 763–771 (2013)