Pártélet: A Hungarian Corpus of Propaganda Texts from the Hungarian Socialist Era

Zoltán Kmetty[1], Veronika Vincze[2], Dorottya Demszky[3], Orsolya Ring[1], Balázs Nagy[4], Martina Katalin Szabó[1,4]

[1] Centre for Social Sciences, CSS-RECENS
1097, Budapest, Tóth Kálmán street 4, Hungary
[2] MTA-SZTE Research Group on Artificial Intelligence
6720 Szeged, Tisza Lajos krt. 103., Hungary
[3] Stanford University, Linguistics Department
450 Jane Stanford Way, Stanford, CA 94305
[4] University of Szeged, Institute of Informatics
6720 Szeged, Dugonics tér 13., Hungary
kmetty.Zoltan@tk.mta.hu, vinczev@inf.u-szeged.hu, ddemszky@stanford.edu
ring.Orsolya@tk.mta.hu, bnagy@inf.u-szeged.hu, martina@inf.u-szeged.hu

Abstract

In this paper, we present *Pártélet*, a digitized Hungarian corpus of Communist propaganda texts. *Pártélet* was the official journal of the governing party during the Hungarian socialism from 1956 to 1989, hence it represents the direct political agitation and propaganda of the dictatorial system in question. The paper has a dual purpose: first, to present a general review of the corpus compilation process and the basic statistical data of the corpus, and second, to demonstrate through two case studies what the dataset can be used for. We show that our corpus provides a unique opportunity for conducting research on Hungarian propaganda discourse, as well as analyzing changes of this discourse over a 35-year period of time with computer-assisted methods.

Keywords: historical corpus, propaganda, communism, Hungarian language, discourse analysis, computational history, conputational semantics

1. Introduction

The active and decisive period of Hungarian history from 1956 to 1989 (also referred to as the Kádár era after the eponymous General Secretary of the Hungarian Socialist Workers' Party) is a widely examined topic in sociology, social history, political history, and history in a broader sense. At the same time, most of the studies used traditional methods for analyzing these phenomena (Szabó, 2007; Pap, 2017; Gyáni, 2016).

Until now, the biggest barrier for the analysis of a huge amount of textual data with NLP methods was the lack of a reasonably-sized digitized corpus. We address this issue by compiling a large Hungarian database that contains all the articles of the *Pártélet* (literal translation: "life of the party") journal, which was published from 1956 to 1989.

In this paper, we present *Pártélet*, a digitized Hungarian corpus of Communist propaganda texts. *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. The structure of our dataset provides a unique opportunity for the corpus-based examination of the temporal dynamics of several elements of the political discourse of the given era. In addition, since digitized corpora of propaganda texts written in Hungarian have not been compiled so far, it also enables the corpus-based examination of manipulative language use. Outcomes of the detailed analysis of propaganda techniques, for instance, can be potentially useful for propaganda detection in contemporary texts.

We illustrate the potential usage of our corpus in two case studies. In the first case study, we conduct statistical analysis of frequency distributions of some of the key concepts of the era in question. We analyze the change in the frequency of words over time as a proxy for understanding changes in the importance of these concepts over time. In the second case study, we construct lexicons of propaganda language and find a decrease in the prominence of propagandistic language over time.

2. Motivation

2.1. Lack of Language Resources

There are only a few corpora in Hungarian language that can be used for historical research purposes.

There are some specifically historical corpora, namely *Magyar Történeti Szövegtár* "Hungarian Historical Corpus" [1], *Ómagyar Korpusz* "Old Hungarian Corpus" [2], *Történeti Magánéleti Korpusz* "Historical Corpus of Personal Texts"[3]. We briefly review these corpora below.

The Old Hungarian Corpus contains digitized and annotated codices, minor texts, letters and Bible translations from the Old and Middle Hungarian period (Simon, 2014). It contains more than 3.2 million tokens. The aim of the creators of the corpus was to construct an annotated corpus comprising all existent texts from the Old Hungarian period (896–1526) and several texts from the Middle Hungarian period (1526–1772).

The Historical Corpus of Personal Texts contains texts belonging to two genres that are supposed to best represent

[1] http://www.nytud.hu/hhc/
[2] http://omagyarkorpusz.nytud.hu
[3] http://tmk.nytud.hu/

Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2374–2381
Marseille, 11–16 May 2020

informal language use: private letters and testimonies of witnesses in trials (Novák et al., 2018). All the sources predate 1772, the symbolic end of the Middle Hungarian period. It is normalized and morphologically annotated. Its current size is approximately 6M characters (850 thousand tokens).

The Hungarian Historical Corpus was originally a collection of texts of various genres and styles, produced between 1772 and 2000. In 2015 it was extended to about 3 millions of text words, produced between 2001 and 2010. The corpus can be queried by year or time period and by text genres (prose, poetry, theatre). However, the corpus consists of raw texts and it does not contain any additional information about the language features of the data, such as stemming or morphological analysis. Only a part of the Hungarian Historical Corpus can be used to examine the Hungarian Socialist era. However, texts of press releases during the Kádár era are underrepresented in the database.

Lastly, there is one more database that contains documents produced during the Hungarian Socialism, the so called *Magyar Nemzeti Szövegtár* "Hungarian Gigaword Corpus"[4]. It is an extended new edition of the Hungarian National Corpus, with upgraded and redesigned linguistic annotation and an increased size of 1.5 billion tokens (Oravecz et al., 2014). Texts of the corpus were produced between the end of the 20th and the beginning of the 21st century. However, despite the large amount of data in the corpus, the Kádár era is still underrepresented in this database.

The underrepresentation of the era in question in the abovementioned large corpora is probably connected to the challenges of the digitization procedure. In the case of solely printed or written texts, time-consuming and costly digitization is just the first step. Moreover, digitized texts have to be further processed most of the time to make them suitable for the automated analysis (e.g. the correction the texts after the application of Optical Character Recognition (OCR) (Hoekstra and Koolen, 2019)).

2.2. Research Gap

As mentioned above, the active and decisive period of Hungarian history from 1956 to 1989 is a widely examined topic in sociology and well as political-historical sciences. However, the analysis of large datasets, especially from a quantitative perspective is an unusual approach in Hungarian historical sciences. However, the application of natural language processing and especially the analysis of the temporal dynamics of semantic features of specific concepts in a large corpus may prove to be beneficial as such a study might be able to identify notable changes in the political-social discourse of this era (Xu and Kemp, 2015; Jatowt and Duh, 2014; Kulkarni et al., 2014; Hamilton et al., 2016b; Hamilton et al., 2016a; Garg et al., 2018).

As for propaganda texts, despite the fact that the quantitative and qualitative analyses of propaganda discourse are important from an NLP and a political-historical point of view, very few studies seem to provide a systematic analysis of a huge amount of propaganda texts (Propaganda Analysis, 1938; Rashkin et al., 2017; Barrón-Cedeño et al., 2019; Vincze et al., 2019). This is a quite significant research gap concerning the Hungarian language. We are aware of only one paper, focusing specifically on the features of a Hungarian totalitarian discourse by using NLP methods (Vincze et al., 2019).

3. Dataset

3.1. Source of the Data

The *Pártélet* journal was the official journal of the governing party, the Central Leadership of the Magyar Szocialista Munkáspárt or MSZMP (in English: Hungarian Socialist Workers' Party, 1956-1989). It consists of 33 volumes with 12 issues every year (one issue per month). We note here, however, that the November and December issues in 1956 were never published and therefore are missing from our compilation. The last edition of *Pártélet* was published in April 1989.

Propagating the political ideology with a special focus on practical aspects, *Pártélet* was an important tool for direct political agitation and propaganda. Articles published in the journal were primarily intended for the party leaders and functionaries (not for the average citizen). At the same time, it was published in 54,150 copies, and this large number suggests that the journal was read not only by the top executives of the party but by people at the lower levels of the political hierarchy as well.

The complete selection of issues of the *Pártélet* journal covering the time period 1956-89 is available online at the Arcanum Digitheca[5].

3.2. Corpus Compilation

The scanned pdf pages of *Pártélet* were downloaded and processed using our own scripts as well as open-source tools. We applied Optical Character Recognition (OCR) engine to convert the scanned journal pages into analyzable text. As the OCR engine works with image files, the individual pdf pages were converted to portable pixmap format, i.e. png image files, using the *pdftoppm* converter. Then, png files were binarized, that is, converted to black-and-white images with ImageMagick[6]. A threshold value of 50% was applied to enhance the efficiency of the OCR process by increasing the contrast between the actual text and its background (Szabó et al., 2019). These binary png images were then processed by *tesseract*, an open-source OCR engine.[7] The output of the OCR procedure, i.e. the raw texts were further processed by removing the page numbers, whitespaces as well as hyphenations. We used our own Linux shell scripts and Python routines for this step.

The individual steps in the workflow are shown in Figure 1. The resulting text was then processed with *magyarlanc*[8] (Zsibrita et al., 2013), a toolkit written in JAVA for the linguistic processing of Hungarian texts. With this tool, the

[4] http://clara.nytud.hu/mnsz2-dev/

[5] https://adtplus.arcanum.hu/en/collection/Partelet/

[6] https://imagemagick.org

[7] https://github.com/tesseract-ocr

[8] http://www.inf.u-szeged.hu/rgai/magyarlanc

Figure 1: The individual steps of the corpus compilation procedure

text was split into sentences, tokenized and lemmatized. We also removed punctuation and stopwords (Szabó et al., 2019).

Due to possible OCR errors, the *hunspell* spellchecker was unable to recognize a significant amount of tokens, more precisely, approximately 22000 words. To address this issue, we took the following steps. First, we collected "unknown" words that occurred at least twenty times in the corpus and consisted of at least three characters. Second, we corrected all these words manually. Then we replaced words with "unknown" label in the corpus with the corrected versions. We corrected 64% of the unknown words with this method.

Our final processed corpus contains a total of 13,185,200 tokens. The distribution of tokens in the corpus is well balanced among the years, that is, the number of tokens are roughly the same for each year.[9]

3.3. Usability of the Corpus

The corpus of *Pártélet* provides a unique opportunity to analyze the official discourse in Hungary between 1956 and 1989. In this paper we present two short examples how this corpus could be used to examine the period. In the first example we map the semantic change of some central concepts of the era, namely the changing role of control and decision in the given time period. In the second example we present the change of propaganda related text over time.

[9]At this stage of the work, the corpus is only available for the researchers within the research project. But upon request we can share the corpus with interested researchers.

4. The Changing Role of Decision and Control

One of the goal of our research work is to map out the semantic change of selected concepts over time. For this purpose we have applied word embedding linguistic modeling methods from the field of Natural Language Processing (NLP).

4.1. Methods

In order to allow for an analysis using temporal approach (Kozlowski et al., 2019), we divided the initial corpus into six different overlapping time periods. It is proven that the quality of word embedding heavily depends on the quality and size of the corpus. Even though the initial corpus is quite large, the size of the resulted six sub-corpora would be obviously smaller. In order to address this issue, our time periods overlap with each other to a certain extent. At the same time, despite the overlap, each time period has its own vector space.

As a next step, we calculate a unique word vector representation for each time period. For this we train 300-dimensional embeddings for unigrams using the `Glove` method (Pennington et al., 2014).

We also test the reproducibility of the embeddings by repeating the training process several times for a selected subcorpus, and only slight differences among these test results was observed. However, to certainly get a robust and statistically more reliable correct outcome, we train 20 embedding models for each time period. For each of the 20 distinct vector spaces, we then calculate the distance between the selected concepts in terms of cosine similarity of their representing vectors. The final distance is obtained by averaging the 20 individual similarities (Antoniak and Mimno, 2018). According to our tests, this approach produces a stable and reliable solution for mapping the distance between our concepts.

4.2. Statistical Analysis of Word Frequency Distributions

We use changes in word frequency as a proxy for studying changes in the importance of the concepts. In order to analyze the word frequency distribution, we calculate each word's occurrence per 1000 words. We use unigrams for this analysis. Data are visualized in Figure 2.

Based on the results we can conclude that there is a significant increase in the frequency of the word "decision", which moves from the 19th place in the first period to the 11th place in the sixth period, while the "comrade" dropped from the 6th place in the first period to the 19th place in the sixth period.

If we examine the top ten expressions of the first and last periods, we can see exactly the same expressions in both cases (except for one word), but the frequency of these ten words are not the same in the two periods in question. The only exception is the word "comrade" that is replaced by the concept "society" in the end of the era. This modification can be considered as a good indicator of a change in the direction of the discourse.

4.3. Results of the Word Embedding Analysis

After the descriptive analysis, we focus on the semantic relatedness of the concepts examined. Figure 3 presents the cosine similarity of some of the given concepts in the 6 time periods.

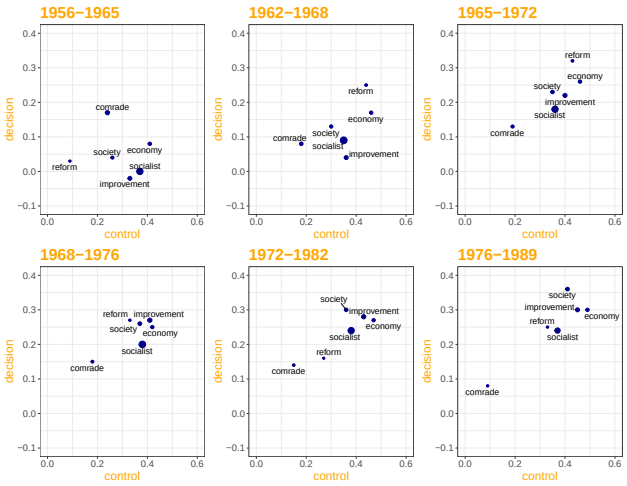

Figure 3: Changes in the cosine similarity of the concepts examined over time.

Our results illustrate that the relationship between *döntés* "decision" and *irányítás* "control" with the tested concepts has undergone a significant change. In the first two periods, our concepts are closer to the word *irányítás* "control", which indicates that discourse contains references to decision processes that represent directives rather than alternatives. From 1965, expressions are getting closer to *döntés* "decision", indicating alternatives that appear in the discourse. As a result of the reform of 1968, the "new economic mechanism" (Kornai, 1980), the independence of individual companies increased, and companies were able to invest part of their own profits.

The term *társadalom* "society", for example, has a rather weak link with the expression of decision in the first two periods (0.04–0.13) but it has a strong connection with *irányítás* "control" (0.26–0.30). At the same time, in the next four periods, though it has a strong association with *irányítás* "control" (0.35–0.41) it also has the same strong semantic connection with *döntés* "decision" (0.23–0.36).

An interesting tendency can be observed in case of the term *reform* "reform", whose cosine similarity to both the concepts *döntés* "decision" (0.03) and *irányítás* "control" (0.09) was low during the first period, but high in all the other periods, only a slight weakening of the *reform* "reform" and *döntés* "decision" in the 1970s (0.16), which can be explained by the cease of the 1968 economic reform at this time. In the case of reform governance, it is above 0.2 in every periods, wich can be expalined also by economic measures.

Lastly, our analysis revealed high cosine similarity score (0.41–0.49) in each time periods between *gazdaság* "economy" and *irányítás* "control", while the *gazdaság* "economy" and *döntés* "decision" only from the second half of the 1960s.

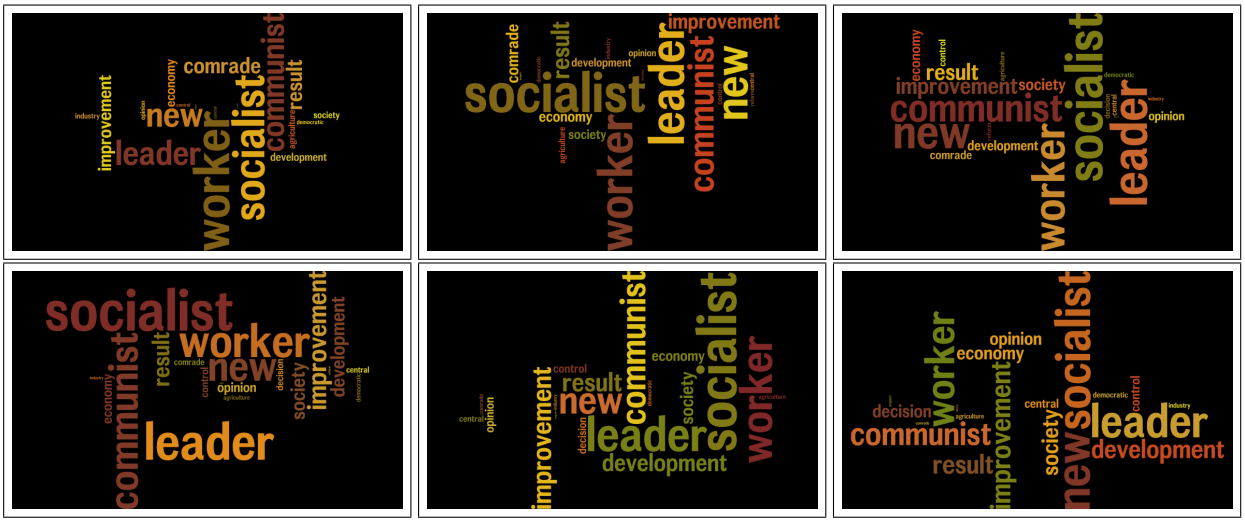

Figure 2: Frequency of selected words in each period.

5. Trends of Communist Propaganda Discourse

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 provide a systematic analysis of a huge amount of propaganda texts (Barrón-Cedeño et al., 2019; Rashkin et al., 2017). In this case study, we demonstrate how the dynamic change of propaganda can be studied using the *Pártélet* corpus.

This analysis is not the first one pertaining to this topic and corpus. Vincze et al. (2019) analyze this corpus and find remarkable differences between the linguistic characteristic of the journal in 1957 and 1988-89. They found more imperatives in the beginning of the era and lower percentages of conditional verb mood. High number of superlative adjectives and adverbs were also the feature of the beginning period which might be a function of propaganda techniques. They found that positive sentiment was more dominant in the first period and negative sentiment in the second period. The set of frequent words also changed between the two periods. Nouns linked to the system were used in the first period (e.g. *terv* "project", *teljesítés* "fulfilment"), while in the second period nouns like *döntés* "decision" and *vélemény* "opinion" were more common.

Vincze et al. (2019) also found that mostly positive propaganda was used in this journal. However, we decided to search for both negative and positive propaganda. The study also showed that in the last period words linked to technology and development become more frequent. Based on this result, we decided to study the dynamic change of this phenomenon as well.

5.1. Creating Propaganda Lexicons

To create a lexicon for positive and negative communist propaganda terms, we use a similar method to that of Rice and Zorn (2013). Specifically, we manually select a list of seed terms (see Table 5.1.) that we identify with positive and negative propaganda, based on our intuition and previous literature Vincze et al. (2019). We then train 300-dimensional embeddings for unigrams and bigrams (henceforth, tokens) based on the entire corpus, using the `word2vec` method (Mikolov et al., 2013) with negative sampling and window size of 5.

We then sum the embeddings of the seed tokens, yielding two vectors v^P and v^N for positive and negative propaganda, respectively. To capture the meaning of propaganda tokens of only one valence (positive or negative), we subtract v^N from v^P and v^P from v^N, which results in two vectors, v^{P-N} and v^{N-P}. We expand the positive and negative propaganda lexicons by selecting the 20 closest tokens in the embedding space to v^{P-N} and v^{N-P}, respectively.

We also develop a lexicon of tokens associated with technology and devevelopment. We sum the vectors of these tokens to yield v^D, and we expand the lexicon by taking the 20 closest embeddings to v^D.

5.2. Results

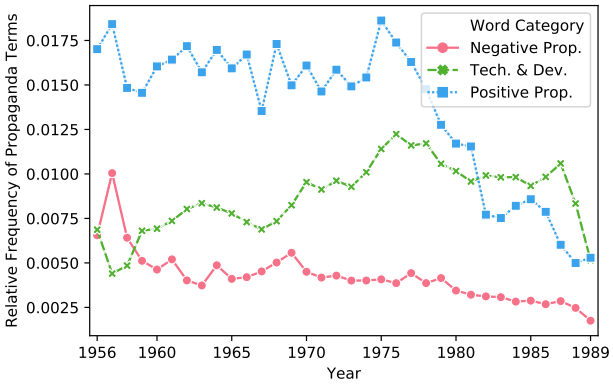

Figure 4: Relative frequencies of positive and negative propaganda terms as well as terms associated with technology and development, calculated over time in our corpus.

We compute the relative frequency of terms belonging to each lexicon, separately for each year in our corpus. As we expected, positive propaganda was more common during most of the period. The prominence of positive propaganda terms remains steadily high until 1975 after which

Lexicon	Tokens
Positive Propaganda	***szocialista brigád*** "socialist brigade", *10. Pártkongresszus* "10th Party Congress", *kongresszus tisztelet* "congressional reverence", ***kommunista*** "communist", *kongresszusi munkaverseny* "congressional labor competition", ***tömegszervezet*** "organization of the masses", ***munkaverseny*** "labor competition", *11. Pártkongresszus* "11th Party Congress", *brigád vállalás* "brigade commitment", *alapszervezet munkaterv* "basic work plan", *szocialista versenymozgalom* "socialist competition movement", *versenymozgalom* "competition movement", ***ötéves terv*** "five-year plan", *munkaverseny-mozgalom* "labor competition movement", *kongresszusi felszabadulás* "congressional liberation", *pártcsoportmunka* "party group-work", *KISZ-szervezetek* "communist youth organizations", ***elvtárs*** "comrade", *jubileumi munkaverseny* "anniversary labor competition", *eredményesen teljesít* "perform effectively", *8. Pártkongresszus* "8th Party Congress", ***munkásosztály*** "working class", *vállalás teljesítés* "fulfill commitment", *Csepel Vas* "Csepel Iron", ***szocialista*** "socialist", ***osztályharc*** "class struggle", *Háry Béla* "Béla Háry", ***proletárdiktatúra*** "dictatorship of the ploretariat", *szocialista munkaverseny-mozgalom* "socialist labor competition movement", ***pártszervezet*** "party organization"
Negative Propaganda	***burzsoá*** "burgeois", *gyarmatosító* "colonis", ***monopólium*** "monopoly", *imperialista ország* "imperialist country", ***kulák*** "kulak", ***imperialista*** "imperialist", ***szektás*** "sect member", *uszítás* "incitement", *fenyegető* "threatening", *esztelen* "brainless", ***reakciós*** "reactionary", *józanabb* "more sober", *zsarolás* "blackmailing", *reakciósabb* "more reactionary", *nagybirtokos* "landowner", *faji* "racial", *veszedelmes* "dangerous", *feudális* "feudal", ***ellenforradalom*** "counter-revolution", *amerikai imperialista* "American imperialist", ***tőkés*** "capitalist", *totális* "absolute", ***kapitalista*** "capitalist", *lepel* "cover", *szít* "incite", ***kispolgár*** "petty burgeois", ***ellenség*** "enemy", *kaland* "adventure", *makacs* "stubborn"
Technology & Development	*teljesítményorientáltság* "performance-orientedness", *hatékony* "effective", ***fejlődés*** "development", *lényegi változás* "substantive change", *energia felszabadítás* "energy release", ***vállalkozás*** "enterprise", ***kezdeményezés*** "initiative", *erőteljes kibontakozás* "powerful evolution", *szemlélet erősödés* "perspective strengthening", *előrehaladás kulcskérdése* "key to progress", ***megújítás*** "renewal", ***eredmény*** "result", *demokratizmus továbbfejlesztése* "development of democracy", *újjáalakítás* "reconstruction", *alkotókepesség kibontakoztatás* "evolvement of creative ability", ***korszerűség*** "modernity", ***szemléletváltás*** "change of perspective", *kormányzati irányátás* "governmental guidance", *fejlesztést célzó* "targeting development", *mechanizmus korszerűsítése* "modernizing the mechanism", ***fejtesztés*** "development", *kibontakoztatás* "evolvement", ***hatékonyság*** "effectiveness", ***önállóság*** "autonomy", *racionalitás* "rationality"

Table 1: List of tokens belonging to the three lexicons we use to analyze trends in propaganda discourse. **Bold** indicates seed terms and the rest of the words are acquired via our word embedding based method.

it shows a sudden decline. In the early 80's, the decline is stopped for few years, but from 85 it continues. It is interesting to highlight how the word *osztályharc* "class struggle" changes during the period. As class struggle is an essence of communism, it has a clear propagandic function. The relative frequency of class struggle was around 0.2–0.3 (per 1000 words) between 1956 and 1973. Then it declined and fluctuated around 0.1 until 1981. In the final few years it hardly reached the 0.05 value. This shows clearly how key concepts of communism faded away.

Negative propaganda is less prominent, at least in the official journal of the regime. The baseline of the negative propaganda is much lower compared with the positive one, and it starts to further decrease at the beginning of the period. 1957 is an exception, but as that year was right after the revolution against the system in 1956, this result is not surprising. The relative frequency of words like enemy and counter-revolution was high in this period. The negative propaganda reaches a small local peak at 1969. This short-term increase was led by words linked to foreign policy such as *imperialista* "imperialist". USSR invaded Czechoslovakia in August 1968, and Hungary was part of this military movement. Most of the negative propaganda linked to this event, based on the qualitative analysis of the journal.

If "class struggle" is a typical function of positive propaganda, *polgár* "burgeois" is a typical negative one. The temporal change of this word is a representative of the decline in negative propaganda. The relative frequency (per 1000 words) of the word *polgár* "burgeois", was around 0.4–0.6 until 1961. Then it decreased to 0.2–0.3 (with a small peak at the end of 60's). And in the 80's it went below 0.1. One of the most typical negative words nearly disappeared by the end of the period.

But what was in the journal instead of the propaganda? The previous study of Vincze et al. (2019) shows that words linked to technology and development were more common in 1989 than in 1956–57. We use our technology and development lexicon to study the trends in the prominence of these terms. The results confirm our previous hypothesis. Terms related to technology and development show a steady increase over time, and they become more prominent than positive propaganda terms in the 1980's. Some words like *eredmény* "result" and *fejlődés* "development" were popular in the whole period. But new terms also emerged, such as *kezdeményezés* "initiative" and *önállóság* "autonomy".

As Centeno (1993) states, there could be a certain natural relatedness between the phenomena of technocracy and market capitalism. Namely, capitalism attempts to legit-

imize itself partly by reference to the productive efficiency of its economic mechanism. In connection with this feature, the increasing number of technocratic elements in our corpus over time may be considered as a sign of the upcoming free-market capitalism.

6. Conclusions

Until now, the biggest barrier for the analysis of a huge amount of textual data regarding the Kádár era with NLP methods was the lack of a reasonably-sized digitized corpus. We address this issue by compiling a large Hungarian database that contains all the articles of *Pártélet*, which was published from 1956 to 1989. We illustrated the potential usage of this corpus in two case studies.

In the first case study we mapped out the semantic change of some selected concepts over time. We focused on the role of decision and control and their link with six key concepts. Based on the results we can conclude that there is a significant increase in the frequency of the word "decision", which moved from the 19th place in the first period to the 11th place in the sixth period. Results based on word embeddings also showed that "decision" became closer to central concepts like "economy" and "reform".

In the second case study, we constructed lexicons of propaganda language to analyze the change of the official language in the period. Positive propaganda was more common during most of the period than negative propaganda. However, both positive and negative propaganda declined over time. On the other hand, terms related to technology and development show a steady increase over time, and they become more prominent than positive propaganda terms in the 1980's, which may be considered as a sign of the upcoming free-market capitalism.

Both case studies clearly show the potential of our corpus to complement traditional research methods in the social sciences. With this unique dataset we will be able to better understand how official discourse changed over time in the Kádár era, which we plan to investigate in the near future. Moreover, the corpus enables a detailed analysis of propaganda techniques, which can also be potentially useful for propaganda detection in contemporary Hungarian texts. Last but not least, we intend to analyse how certain political events were reflected in the official communication of the Communist party, hence linguistic tools for framing can also be identified in the data.

7. Acknowledgements

This work was supported by the Hungarian Research Fund (NKFIH / OTKA, grant number FK 131826). Travelling of the presenter (Martina Katalin Szabó) was supported by Centre for Social Sciences, CSS-RECENS, Hungary and University of Szeged, Institute of Informatics, Department of Software Engineering, Hungary. The work of Zoltán Kmetty was funded by the Premium Postdoctoral Grant of the Hungarian Academy of Sciences.

8. Bibliographical References

Antoniak, M. and Mimno, D. (2018). Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics*, 6:107–119.

Barrón-Cedeño, A., Jaradat, I., Martino, G. D. S., and Nakov, P. (2019). Proppy: Organizing the news based on their propagandistic content. *Information Processing Management*, 56(5):1849–1864.

Centeno, M. (1993). The new leviathan: The dynamics and limits of technocracy. *Theory and Society*, 22:307–335, 06.

Garg, N., Schiebinger, L., Jurafsky, D., and Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. 115 16:E3635–E3644.

Gyáni, G. (2016). *A történelem mint emlék(mű)*. Kalligram, Budapest.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016a). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany, August. Association for Computational Linguistics.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016b). Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, pages 2116–2121.

Hoekstra, R. and Koolen, M. (2019). Data scopes for digital history research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 52(2):79–94.

Jatowt, A. and Duh, K. (2014). A framework for analyzing semantic change of words across time. In *IEEE/ACM Joint Conference on Digital Libraries*, pages 229–238.

Kornai, J. (1980). The dilemmas of a socialist economy: the Hungarian experience. *Cambridge Journal of Economics*, 4(2):147–157.

Kozlowski, A. C., Taddy, M., and Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, 84(5):905–949.

Kulkarni, V., Al-Rfou', R., Perozzi, B., and Skiena, S. (2014). Statistically significant detection of linguistic change. In *WWW*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Novák, A., Gugán, K., Varga, M., and Dömötör, A. (2018). Creation of an annotated corpus of Old and Middle Hungarian court records and private correspondence. *Language Resources and Evaluation*, 52(1):1–28.

Oravecz, Cs., Váradi, T., and Sass, B. (2014). The Hungarian Gigaword corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1719–1723, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Pap, M. (2017). A népitől a szocialista demokráciáig. A korai Kádár-korszak demokráciafogalma a pártfolyóiratok tükrében. *Múltunk*, 1:202–226.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Propaganda Analysis, I. f. (1938). How to Detect Propaganda. In *Propaganda Analysis. Publications of the Institute for Propaganda Analysis*, volume I, pages 210–218.

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., and Choi, Y. (2017). 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*, pages 2931–2937, Copenhagen, Denmark, September. Association for Computational Linguistics.

Rice, D. R. and Zorn, C. (2013). Corpus-based dictionaries for sentiment analysis of specialized vocabularies. *Political Science Research and Methods*, pages 1–16.

Simon, E. (2014). Corpus building from Old Hungarian codices. *The evolution of functional left peripheries in Hungarian syntax*, pages 224–236.

Szabó, M. K., Ring, O., Nagy, B., Kiss, L., Koltai, J., Berend, G., Vidács, L., and Kmetty, Z. (2019). Mapping the dynamic change of the concept "industry" and "agriculture" in the Hungarian Socialist era using a word embedding model. *Historical Methods: A Journal of Quantitative and Interdisciplinary History. Submitted for review.*

Szabó, M. (2007). A dolgozó mint állampolgár. Fogalomtörténeti tanulmány a magyar szocializmus három korszakáról. *Korall*, 27:151–171.

Vincze, V., Szabó, M. K., and Ring, O. (2019). Automatic analysis of linguistic features in Communist propaganda texts. `https://propaganda.qcri.org/bias-misinformation-workshop-socinfo19/paper4_final_Vincze_et_al_propaganda.pdf`.

Xu, Y. and Kemp, C. (2015). A computational evaluation of two laws of semantic change. In *CogSci*.

Zsibrita, J., Vincze, V., and Farkas, R. (2013). magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In *Proceedings of RANLP*, pages 763–771.