



Historical Methods: A Journal of Quantitative and Interdisciplinary History

ISSN: (Print) (Online) Journal homepage: <https://www.tandfonline.com/loi/vhim20>

Exploring the dynamic changes of key concepts of the Hungarian socialist era with natural language processing methods

Martina Katalin Szabó, Orsolya Ring, Balázs Nagy, László Kiss, Júlia Koltai, Gábor Berend, László Vidács, Attila Gulyás & Zoltán Kmetty

To cite this article: Martina Katalin Szabó, Orsolya Ring, Balázs Nagy, László Kiss, Júlia Koltai, Gábor Berend, László Vidács, Attila Gulyás & Zoltán Kmetty (2021) Exploring the dynamic changes of key concepts of the Hungarian socialist era with natural language processing methods , Historical Methods: A Journal of Quantitative and Interdisciplinary History, 54:1, 1-13, DOI: [10.1080/01615440.2020.1823289](https://doi.org/10.1080/01615440.2020.1823289)

To link to this article: <https://doi.org/10.1080/01615440.2020.1823289>



© 2020 The Author(s). Published with license by Taylor & Francis Group, LLC



Published online: 23 Sep 2020.



Submit your article to this journal [↗](#)



Article views: 846



View related articles [↗](#)



View Crossmark data [↗](#)

Exploring the dynamic changes of key concepts of the Hungarian socialist era with natural language processing methods

Martina Katalin Szabó^{a,b}, Orsolya Ring^a, Balázs Nagy^b, László Kiss^a, Júlia Koltai^c, Gábor Berend^b, László Vidács^{b,d}, Attila Gulyás^a, and Zoltán Kmetty^e

^aComputational Social Science – Research Center for Educational and Network Studies (CSS-RECENS), Centre for Social Sciences, Budapest, Hungary; ^bDepartment of Software Engineering, University of Szeged, Szeged, Hungary; ^cDepartment of Social Research Methodology, Faculty of Social Sciences, Eötvös Loránd University of Sciences, Budapest, Hungary; ^dResearch Group on Artificial Intelligence (RGAI), Hungarian Academy of Sciences (HAS) and University of Szeged (SZTE), Szeged, Hungary; ^eDepartment of Sociology, Eötvös Loránd University of Sciences, Faculty of Social Sciences, Budapest, Hungary

ABSTRACT

The analysis of social discourses from the perspective of historical changes deserves special attention. Such a study could play a key role in revealing social changes and latent narrative of those in power; and understanding the underlying social dynamic in a given period. Until the recent years, such issues were analyzed mainly in a qualitative approach. In our paper we present a new way of revealing/discovering and interpreting social discourses using an advanced NLP method called word embedding. Based on word similarities we can understand the main structural frames of a given system and using a dynamic approach we can reveal the social changes in a historical period. In our study we created a large corpus from the Hungarian “Pártélet” journal (1956–89). This was the official journal of the governing party, hence it represents not just a media discourse of the era, but the official discourse of the government, too. One of the main focal points of our research is to study the evolution of the semantic content of some of the concepts related to the topics of agriculture and industry, which are two central notions of the examined era.

KEYWORDS

Natural language processing; historical analysis; text mining; socialism; Hungary


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 in Hungary. Analysis of major issues such as social mobility and stratification, economic policy, foreign and domestic policy developments and their social impact provide a comprehensive, clear picture of Hungary's post-World War II history. At the same time, most of the studies used the traditional toolboxes for the analyses of these phenomena.

In recent years, qualitative historical analysis of social discourses related to particular social processes in Hungary has become more common. Among other comprehensive investigations, we can refer to Márton

Szabó's (2007) conceptual history analyses based on discursive political science, Milán Pap (2015, 2017) and Konstantin Medgyesi's text analyses of ideological history (2017), Zsolt Győri's film sociography (2013) and Noémi Herczog's theater discourse analysis (2017). An interesting application of text-based analysis of social life can be found in Melinda Kovai's book (2016), which reconstructs the relationship between politics and psychiatry by processing psychiatric pathologies of the Kádár era. A discourse-based analysis of the representations of cultural memory (Gyáni 2016; Jakab 2012) is also an important research area, which has some interesting results in terms of history and sociology. As for the different social groups, the most widespread discourse analyses were carried out concerning the peasantry (Nagy 2007; Csurgó and Kiss 2007; Csurgó 2007; Kiss 2007).

The analysis of large datasets, especially from a quantitative perspective, is an unusual approach in

CONTACT Martina Katalin Szabó ✉ Szabo.Martina@tk.mta.hu  Computational Social Science – Research Center for Educational and Network Studies (CSS-RECENS), Centre for Social Sciences, 1097 Budapest, Tóth Kálmán str. 4, Hungary.

© 2020 The Author(s). Published with license by Taylor & Francis Group, LLC

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

historical sciences. Automated historical text and discourse analysis, especially applied for the period from 1945 to 1990 is a method that has not been used so far in Hungarian. However, recent developments of historical text analysis suggest that an automated approach could also be promising for mapping the discourse environment of an era (e.g., Miller 2013; Sebők 2016; Cristianini, Lansdall-Welfare, and Dato 2018). The application of Natural Language Processing (NLP) and especially the analysis of the temporal dynamics of semantic features of specific concepts may prove to be beneficial and such a study might be able to identify notable changes in the political-social discourse of this era.

Until now, the biggest obstacle for this analysis was the lack of available digitized textual data of a reasonable size. We addressed this issue by setting up 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. 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.

The corpus gave us the opportunity to analyze the temporal changes of some key concepts of the socialist era. To map the temporal dynamic of selected concepts we applied a word embedding method. This method is a commonly used tool in Natural Language Processing (NLP) and machine learning tasks for capturing the contextual environment of words and expressions; for discovering analogy relationships between any two words; and—in a dynamic perspective—for measuring semantic changes over time (Kim et al. 2014; Xu and Kemp 2015; Jatowt and Duh 2014; Kulkarni et al. 2014; Hamilton, Leskovec, and Jurafsky 2016a, 2016b; Garg et al. 2018; Kutuzov et al. 2018).

Our main research goal is to identify changes in the political discourse related to agriculture and industry, during the years of the Kádár era, between 1956 and 1989 in Hungary. Comparing the vectors from different time periods we will graphically present the temporal dynamics of these decisive concepts in the given discourse. The study will also reveal the temporal changes of the relationships among these words.

The corpus and the historical background

Selected source for the corpus

As we mentioned above, the Hungarian “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, 1956–89). Hence it represents the official discourse of the state-party. 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, it was an important tool for direct political agitation and propaganda. The main topics of the articles were party organization, propaganda and ideology, economic policy and economic analysis. Articles published in “Pártélet” were primarily intended for the party leaders and functionaries (not for the average person). What is more, it often published letters from the leadership in connection with the work of the Socialist cooperatives and factories, as well as other issues concerning Hungarian society. 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.

Historical background

After losing World War II, Hungary came under Soviet military occupation. By the end of the 1940s, a Soviet-style political system had been established in the country. After 1948 the so-called *Rákosi era* began and the heavy industry started to develop along similar lines as the Soviet economy of the 1930s. It was based on the needs of the military industry to “prepare for war”. Economic policy was basically dominated by the central plan of the governing party, not by the country’s actual production ability.

In the 1950s the forced collectivization of the agriculture was implemented. It meant that many individuals belonging to the rich and middle peasantry were labeled as the enemy of the system and private farming was outlawed. Peasant private property was liquidated, the soviet-type kolhoz-system¹ was established. The task of agriculture has been to serve industrial developments and industrial population. As a consequence, agricultural productivity has fallen rapidly and food shortages became constant.

After 1953, the heavy industry developments decreased slightly. At the same time, central support for light industry and agriculture increased. However,

after the fall of Imre Nagy in 1955, the system before 1953 was restored to its previous state.

Following the Hungarian Revolution of 1956, the essence of the consolidation policy of the Kádár era was created. The basis of communist ideology was that the political system rests on economic foundations, so policies primarily aimed at the improvement of efficiency of the centralized planned economy and the raise of people's living standards. Raising the level of consumption also served the purpose of turning people to consumption instead of politics. The basis of communist ideology was that political systems rested on economic foundations, so the policy was primarily aimed at improving the efficiency of the centralized planned economy and raising the living standards of the population. Raising the level of consumption of the population also served the purpose of turning people to consumption instead of politics. More emphasis was placed on raising the level of consumption than at any time before. Investments in the heavy industry continued, but modernization efforts also grew stronger. Export capacity increased.

By 1962, the forced collectivization of agriculture was over. Hungary became a "collectivized country". Financing the investments and, rising living standards at the same time caused serious problems. In order to allocate resources optimally, the party decided to organize the system of trusts. Altogether 15 large plants were established, the role of central management increased, the decision-making power of local plants was narrowed. These actions worsened the supply of the population.

In 1963, the preparation work of a general reform began, which was introduced in 1968. Under the so-called "new economic mechanism", the independence of individual companies has increased, as well as the role of the "second economy" and the traditional private sector. The previously applied, completely artificial price conditions have been replaced by a new price system. Some of the prices were determined by the relationship between supply and demand.

In agriculture, the autonomy of cooperatives increased, including the possibility to set up profitable branches. "Second-economy"—as defined by Gábor (1989, 339)—is "the field of all those economic activities through which the population—legally or illegally—acquires incomes not as employees in the socialist sector." The "second-economy" however was not based on capitalist economic structure—it became an integral part of the Hungarian socialist economy as an important element of the politics of the Kádár era. The employees of the socialist sector using the means

of production of the socialist sector produced for their own benefit.

However, by 1972 the political opponents of these changes overturned the reforms. Central control of the industry increased, and exports were once again centrally regulated. In 1973, the 50 largest industrial companies subjected to direct central supervision. As a result of the oil crisis in 1973, the export capacity of Hungary decreased and this led to higher inflation and the decline of the living standards. The government tried to maintain the level of consumption by taking out foreign loans. This policy led to significant increase in Hungary's national debt.

After the "second oil price explosion" in 1979, it was necessary to reintroduce some elements of the 1968 reforms. Due to unfavorable international economic conditions, the investment rate had to be reduced; the role of the private sector in domestic trade had to be increased. To this end the role of central management was reduced and the "second economic" participation of the population was supported. The reforms of the late 1970s were mostly aimed at preserving the living standards previously achieved. From the beginning of the 1980s, the effects of state debts grew ever stronger.

From 1984, new reforms were introduced, with the aim of establishing a mixed market economy. However, these reforms were no longer effective. The national debt constantly increased, and the productivity of the Hungarian economy declined. The country's economy has plunged into a deepening crisis. The Hungarian state was forced to take out more and more loans, many of which served for the repayment of previous loans. All these tendencies led to a final and definitive crisis of the regime, which terminally ended in 1989 in Hungary.

Methods

Advantages and challenges of automated text analysis

Most of the papers and books that analyzed this era applied a qualitative approach. The qualitative, and, at the same time, manual analysis of contemporary documents is a standard tool used to identify and demonstrate some of the sociological and historical features of this era. However, these studies are time-consuming and costly in most cases as large amount of text has to be read by the researcher: if the dataset is very large, manual annotation may be unfeasible (Desagulier 2018).

In contrast, one of the most important characteristics of automated text analysis is that it can process that large amount of texts, which would be impossible with human capacity. Thus, the innovation is the amount of processed data, which contribute to an important aspect of research, namely the striving for completeness in the analysis of a given historical source (cf. Huistra and Mellink 2016). This eventuate a high level of outer validity as the results arise from a well-shaped and complete textual source.

As for the automatic methods of corpus analysis, one of the basic technologies applied in computer-assisted historical research works is the standard keyword search (cf. Huistra and Mellink 2016). What is more, most digital repositories currently offer the standard keyword search. However, as Huistra and Mellink (2016) explain in detail, it is so difficult for the majority of historical researchers to choose proper keywords even while accessing text sources. The reason for this is that—on the one hand—a topic can be described in many different words, and—on the other hand—a word can have multiple meanings. These features cause notable problems for most of the automatic analyzing methods working with textual sources. To circumvent this difficulty, more advanced search technologies should be applied. Some automated text analysis and NLP methods—such as word embedding method we apply in our analysis—are not only capable of the analysis of the manifest content of a source, but are able to retrieve the latent meaning of a text, which makes the inner validity of the results much higher. These complex methods enable researchers to approach historical corpora in completely new ways, revealing unknown features and interrelations of the data. As the analysis of the large corpus is based on explicitly documented algorithms, the reliability of the research is also high. Because of the above mentioned reasons, automated text analysis can provide results with high inner- and outer validity, and also high reliability and because of the combination of text analysis and algorithms, it offers a jointly quantitative and qualitative method.

However, automated text analysis requires special data formats. From the point of automated text analysis, available historical texts can be divided into three types: initially digital, printed/written but digitalized and not digitalized printed/written ones (Huistra and Mellink 2016). In the case of solely printed or written texts, digitization is just the first step. Digitalized text has to be preprocessed to make it proper for the automated analysis. This includes steps like the correction of Optical Character Recognition

(OCR); concept- or meta tagging or lemmatization, which is also needed in the case of initially digital texts (cf. Hoekstra and Koolen 2019). After the pre-processing phase we can have the corpus, on which it is possible to run the automated text analysis methods.

These corpora have many other advantages for researchers, who are not that familiar with algorithmic analysis. These datasets can be reached and analyzed without geographical limits, opening the physical and temporal boundaries of scientific research. Additionally, simple and more complex searches can be achieved on them, which can spare researchers a lot of time and can help in the selection of relevant sources.

Besides its advantages, there are important pitfalls and tasks that must be taken into account during the application of automated text analysis and more generally, digital historical methods. The historical source, from which the corpus of the analysis is derived, determine the possible analysis. Naturally, source critique is as important here as it is in classical historical analysis. As Huistra and Mellink (2016, 226) stated, “no matter how big the data, scholars still must account for their selection of sources”. For example, in the analysis of such journals as *Pártélet*—which was the official journal of the governing party—will result not the real facts, but the discourse, with which the group in power wanted to form the public opinion. The chosen source and the created corpus also affect the possibly raised research questions (cf. Guldi and Armitage 2014).

Problems during the digitization of the texts, selection of the proper algorithms and the whole data processing and analysis process all contain special risks connected to the computerization of the analysis. To overcome the challenges, cooperation between historians, computational linguists and data scientists should be not just fruitful but necessary in some defined phases of the research project. In the following, we introduce the documentation of the corpus building.

Preprocessing

The complete selection of issues of the “*Pártélet*” journal covering the time period 1956–89 is available online at the Arcanum Digitheca (<https://adtplus.arcanum.hu/en/collection/Partelet/>). Arcanum provides access to the digitized copy of historically important printed resources, such as journals or newspapers. The scanned pdf pages of “*Pártélet*” were

downloaded and further processed using self-written scripts as well as open-source tools as detailed in the following.

We applied Optical Character Recognition (OCR) engine to convert the scanned journal pages into proper text, that can further be processed with NLP tools. First, 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 available in any Linux distribution.

As a second step, png files were binarized, that is, converted to black-and-white images with ImageMagick (<https://imagemagick.org>), also available in any Linux distribution. A threshold value of 50% was applied, meaning that any pixel of the image below this threshold was set to black and the rest to white. This technique was applied to enhance the efficiency of the OCR process by increasing the contrast between the actual text and its background. The threshold of 50% chosen above was an optimum value that provided the most accurate OCR result for a selection of test pages.

These binary png images were then processed by “tesseract,” an open-source OCR engine (<https://github.com/tesseract-ocr>). The output of the OCR procedure, i.e., the raw textual data for the “Pártélet” pages, was further processed by removing the page numbers, whitespaces as well as hyphenations. We used self-written Linux shell scripts and Python routines for this step.

The resulting text was then processed by “magyarlanc”² (Zsibrita, Vincze, and Farkas 2013), a toolkit written in JAVA for the linguistic processing of Hungarian texts. With this tool, the text was first split into sentences, then tokenized (further split the sentences into words), and finally the tokens were lemmatized (converted the word to its dictionary form). A token in NLP represents a semantic unit, a sequence of characters usually separated by spaces from each other. A token can be a word, a number, or punctuation as well. Our final corpus was obtained after removing punctuation and stopwords from the tokenized and lemmatized document. We removed stopwords (cf. Ullman 2011) and check all the remaining words with a hungarian spell checker, called hunspell. Hunspell was unable to recognize 14 percent of the unique words. We manually checked all the unknown words that occurred at least 10 times and consisted of at least 3 characters. Then we replaced those words with the corrected ones, where it was possible to find the original word. With this

method we were able to decrease the ratio of unknown words to 10.7 percent. The manual correction clearly showed, that the corpus contains era-specific words, which the spell checker was unable to identify (like a special abbreviation or words unique to this era). Based on the manual correction phase, more than 10 percent of the inspected words belonged to this category. Thus, we assume that the overall ratio of misspelled words is under 10 percent in the corpus.

As we have already mentioned before (see 2.1), the complete “Pártélet” journal consists of 33 volumes in total published between 1956 and 1989 with 12 issues every year. As we proceeded our analysis on a yearly basis, we had to ensure having enough unique words in every volume, because word embedding methods we applied are sensitive for small corpus size. After the initial evaluation of the corpus, we decided to omit the first 3 volumes and the last one as they do not fit the criteria containing enough unique words. So our analysis covers the period of 1959 and 1988. The distribution of tokens in the corpus is well balanced among the years; that is, the number of tokens is roughly the same for each year. Our final and pre-processed corpus contains a total of 9,432,200 tokens and 609,905 unique words.

Method of the analysis: the word embedding model

Our aim in this work is to map the semantic change of selected concepts in the time period, when “Pártélet” was one of the major sources of information for the leading officers of the party and the government. To this end, we have applied word embedding linguistic modeling method from the field of NLP.

Word embedding is basically the dense vector representation of words in a vocabulary, where the dimension of a word vector should be lower than the size of the vocabulary itself. The vocabulary represents the document or corpus, and is usually the list of unique words extracted from the corpus. The resulting vectors—in contrast to the results of a simple co-occurrence analysis—capture semantic relations between the language elements analyzed: according to the distributional hypothesis, semantically similar words tend to have similar contextual distributions (Harris 1954). In connection with this, applying a dynamic approach (Bamler and Mandt 2017) we can extract, for instance, the changes of position of social groups and identify social changes in a historical

period (Hamilton, Leskovec, and Jurafsky 2016a, 2016b; Garg et al. 2018).

The vector representation is useful in many aspects. For example, we can perform mathematical operations on these vectors, or more importantly, we can define similarities or, from another perspective, distances between two vectors, or, correspondingly, two words. This means that in a given embedding model, similar words have similar vector representations, or, in other words, vectors for words in similar context are mapped to nearby regions in the vector space.

A number of word embedding algorithms has been introduced so far, such as word2vec by Google (Mikolov et al. 2013a; Mikolov et al. 2013b; Mikolov, Yih, and Zweig 2013c), FastText (Joulin et al. 2016) or GloVe (Pennington, Socher, and Manning 2014). They all produce distributed representations for each word type in the form of a dense, n -dimensional vector (a string of real numbers) by incorporating local (and sometimes global) corpus context (Liu, Huang, and Gao 2018). These methods provide fairly similar results in many areas. We chose the one of GloVe for two reasons. On the one part, the literature suggested that GloVe provides more stable and robust results in the case of smaller corpus (Spirling and Rodriguez 2019) compared with other embeddings; on the other part, practical reasons have also played a role in the selection, namely that this method has a well-written implementation in R.

Since working with large vocabularies is computationally expensive, we first reduced our vocabulary by removing words, which appeared less than five times in the initial dictionary (all the published articles of “Pártélet” journal). GloVe is an unsupervised machine learning algorithm that is trained by using a global word-word co-occurrence statistics of the corpus.

We applied the following procedure. We ran the embedding models for each year with 6 parameter combinations, 10 times for each combination. Each of the six unique combinations had one parameter for the embedding size (200, 250, and 300 dimensions) and another for the window size (for 4, 7, and 10). We calculated the stability of the key concepts (agriculture and industry) between the 10 embeddings of the same year and same parameter combination. For the stability test, we aligned the 10 vector spaces with procusteres matrix rotation and calculated the cosine similarity of the same concepts in the aligned vector spaces. The stability tests showed that the most stable results arise from models with 10 window size and 200 dimensions.³ Therefore these were the parameters we used in the final models.

We used the GloVe algorithm implemented in the text2vec package (Selivanov and Wang 2016) in R, and the maximum number of training iterations was 10, i.e., the machine learning algorithm went through the training dataset 10 times.

The similarity of words was measured by the cosine similarity of their representing vectors. This is the most commonly used distance metric in embedding analyses. The similarity of words was measured by the cosine similarity of their representing vectors. This is the most commonly used distance metric in embedding analyses. Other metrics like Euclidean distance could have been misleading here, because the length of each word vector is strongly correlates with the frequency of the given word in the corpora (and also with the context variability) (Schakel and Wilson 2015). Stated briefly, the maximum cosine similarity is 1, if the orientation of two word vectors is the same, i.e., they point into the same direction; 0 if the word vectors are perpendicular; and -1 if the two word vectors point into the opposite direction. Note that by definition, this metric is independent of the magnitude of the vectors, only their orientation is important.

In order to capture the contextual shifts of words, we decided to create different embeddings for all the volumes (years). To ensure that the vector spaces are comparable (and not just randomly changing from one year to the other), we applied an alignment method. We aligned the vector space of a given volume to the vector space of the previous volume with Procrustes rotation (Kulkarni et al. 2014; Hamilton, Leskovec, and Jurafsky 2016a; Hamilton, Leskovec, and Jurafsky 2016b). Using this technique, we were able to calculate the stability of key concepts in the given historical period (we will deal with the stability of concepts in the following part of the text). In order to get robust and a statistically reliable outcome, we trained 100 embedding models for each time period (year) and present the average similarity values of the same concepts (Antoniak and Mimno 2018) with 95 percent confidence intervals. We evaluated the context of the key concepts (agriculture and industry) both from a data driven (Hengchen, Ros, and Marjanen 2019) and from a theory driven perspective.

Discussion

Two key concepts of the era: agriculture and industry

The historical period examined in this study is characterized by a specific ideological language, including central and related concepts. During this era, the

main concepts were necessarily static from a semantic point of view, hence “Pártélet” journal represents an ideology-driven discourse of the era in question. However, the semantic content of the related concepts may shifted over time to some extent as a result of temporal political and economic changes.

Our analysis examines the changing context of the concepts *industry* and *agriculture*. Our aim is to explore how strongly the two concepts were related to other concepts indicating the development of state economic policy during certain periods of the Kádár era.

The average occurrence frequency was 0.61 for *agriculture* and 0.5 for *industry* per 1000 words. Both words were more frequently used in the first period of the era (see Figure A1 in the Appendix) and except the last years of the regime, *agriculture* was used more frequently than *industry*. We can see a high increase in the usage of *agriculture* around the end of the seventies, which is followed by a steep decrease. The cause of this peak is the two oil crises: as a result of them, the country's export capacity has declined and living standards have deteriorated. As a consequence, it became necessary to reintroduce certain elements of the 1968 reform affecting the agricultural sector in order to increase the level of consumption through the rise of the supply.

Using the aligned vector spaces, we calculated the temporal stability of the key concepts, by measuring the cosine similarity of the same words between consecutive years. The average stability was 0.44 for *agriculture* and 0.42 for *industry*. The raw stability values decreased over time, but it is partly caused by the decreasing frequency of the words. When we controlled for this effect, the residual trend did not imply any change, which means that the average stability of these words was the same in the whole period. However, it does not mean there were no changes in the collection of closest words around the key concepts.

We collected the 20 closest words of the two key concepts for each year.⁴ The 5 closest concepts to *agriculture* were *industry*, *production*, *development*, *progress* and *socialist*. These words are close to *agriculture* in almost every year. We also found words, which appeared in one period and then disappeared in another. Words like *plan* or *objective* were only close in the 60s, *result* was a close word in the 70s and *product* only appeared in the 80s.

As for *industry*, the closest word was *agriculture*, which was followed by *production*, *progress*, *development* and a special word: *economy of people* (in

Hungarian: *népgazdaság*). *Plan* and *socialist* were close to *industry* in the first years of the regime. In the 70s, the word *employed* appeared but then disappeared for the 80s, where new words emerged, such as *domestic* (in Hungarian: *hazai*).

As the above results shows, *agriculture* and *industry* were semantically close concepts in the entire period: each of them appeared in the collection of the other's closest words. The dynamic analysis of these two concepts (see Figure A2 in the Appendix) revealed a quite stable pattern of closeness throughout the whole era, with a sharp decrease at the end of the period. This significant change of the discourse analyzed may be closely related to the upcoming free-market capitalism (Kmetty et al. 2020).

From progress to direction—shifting concepts around agriculture and industry

In this section we examine the temporal similarity of some selected concepts with *agriculture* and *industry*. The concepts analyzed were selected based on their frequency distribution and on the consideration of relevant literature on Hungarian history (cf. e.g., Kornai 1992; Romsics 2007; Varga 2002; Földes 2019). We start with the words *development* and *progress* which were among the 5 closest words for both *agriculture* and *industry*.

The connection between the term „development” and the two economic branches is a good indication of the direction of the state economic policy, as well as the bargaining power of the companies of the given economic branches. Mária Csanádi (2004) emphasized the importance of relations between the various levels of state and party leadership. The major economic politicians of the era, the leaders of state-owned industrial enterprises and large agricultural holdings were necessarily part of the state- and party hierarchy. Their access to budgetary resources depended primarily on their positions within the two hierarchies. The amount and the distribution of development resources allocated to the agriculture and to the industry were determined by the influence of their leading actors within the state and party hierarchy. Based on the above reasoning, the official discourse of the technocracy of the era primarily follows rather than voluntarily shapes economic processes.

Due to the soft budget constraint (described by Kornai 2014), access to public development funds depended on the position of the given sector in the political space. The sector leaders' political positions and their relations to influential politicians were

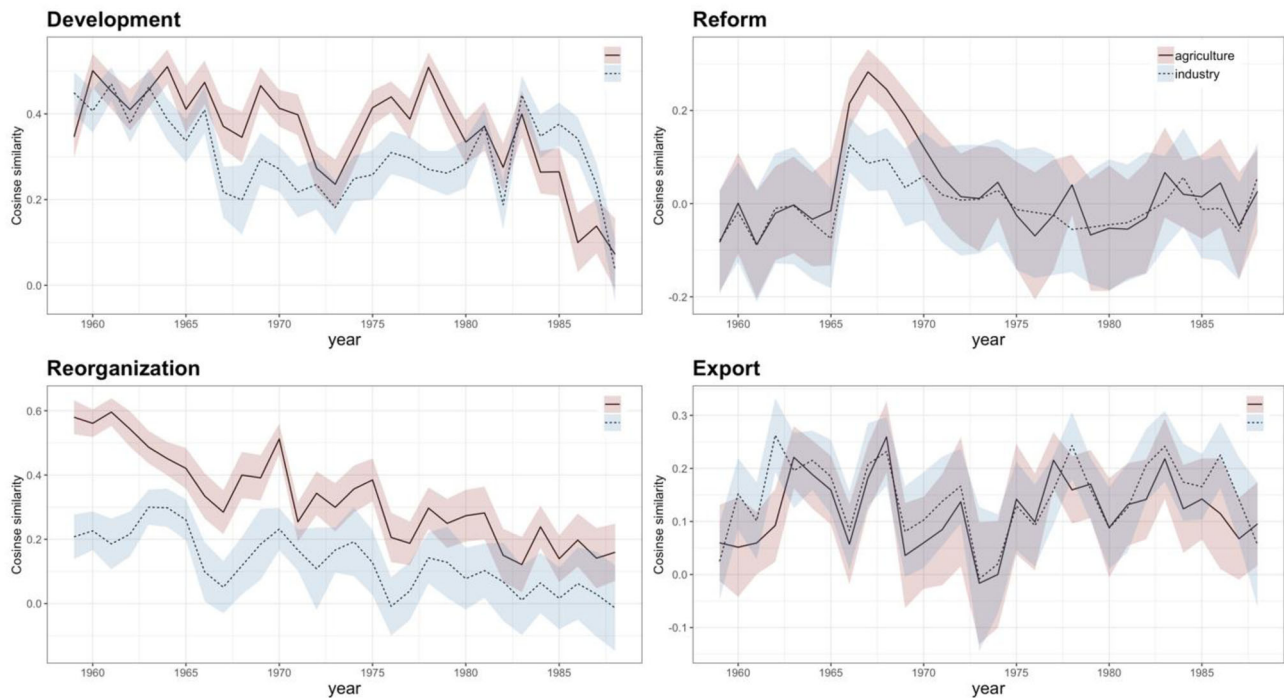


Figure 1. The cosine similarities of *development* and *reform* and *reorganization* and *export* with *agriculture* and *industry* between 1959 and 1988 ((we calculated the 95% confidence interval based on the 100 embedding trained for each year with the following formula: $MEAN \pm (SD/\sqrt{100}) \cdot 1.96$)).

extremely important. The state has not allowed its companies to go bankrupt. However, success and profitability were not necessarily available to all entities. The soft budget constraint of the production units collided with the state budget's "hard" constraint, the ceiling of distributable development resources at the national level. These processes led to serious struggles for development resources, not only between different productive units and sectors, but also between regional units, counties and cities (Vági 1982).

Through the development resources, the centrally managed economic policy ultimately forced industrial and agricultural companies and cooperatives into the same "development field". Thus, the question arises: which economic sector shows stronger relationship with the word *development*, in the articles of a journal for technocracy. Answering this question can bring us closer to the understanding, that which sector had priority in the competition for resources.

In Figure 1, we see that the relationships between the two economic sectors and the concept of *development* show almost the same dynamics. However, in some periods, the strength of the relationship is very different for the two sectors.

We can see three periods when the development of the agricultural sector had much more emphasis—based on the texts of the party outlet. The first period

is the beginning of the 1960s, which was the period of collectivization of agriculture, forcing it to socialist production conditions. The second period is related to the years of 1968–71, where the experiment of the "new economic mechanism" was conducted.⁵ The third period is observable in the late 1970s, when the failed economic reform was forcibly revived.

Development in the discourse of the given era was closely linked to *reform*. Analyzing the closeness of the word *reform* with the two economic sectors, we can observe relationships between the communicated economic policy processes and the position of the sectors (see Figure 1). Until the early 1960s, the term *reform* showed almost no connection with the two sectors (whose position relied on the development resources). With other words, the notion of *reform* as a political will was not part of the political communication. However, we see a notable change after 1964. Although we know that the economic change did not "begin" on the date, when the related legislation was promulgated, it is noteworthy, that the topic of economic reform (which then was introduced in 1968) has already appeared prominently in the political communication in 1964.

Initially, the relationship between the words *reform* and *industry*, such as *reform* and *agriculture* were strong, indicating that both sectors were strongly involved in the preparation of the economic reform

process. However, after 1966, only agricultural reform remained in the focus and the strength of the relationship between *industry* and *reform* decreased.⁶ After 1968, communication on the agricultural reform has also declined rapidly. Changes in economic governance—introduced in 1976—appeared not as a *reform*, but as a necessary correction of existing policy in communication. The term reform will return to political communication only at the end of the era.

Since collectivization was a fundamental feature of the system in question, we analyzed the concept *reorganization* as well.⁷ The usage of the word *reorganization* follows different temporal trend compared to the one we observed at the word *reform*. Overall, we can see a decreasing pattern in the relationship of *reorganization* with *industry*; and *reorganization* and *agriculture* (Figure 1). *Reorganization* and *agriculture* shared the very same contexts in the first years of the Kádár era. This strong connection in the early 1960s refers to an important transformation, namely the collectivization of agriculture.

Both the distribution of development resources and the central economic policy guidelines (such as *reforms*) affected the export capacity of individual sectors (Figure 1).

For socialist economic entities operating within soft budgetary constraints, exports were one of the most important areas of competition. The main questions for these economic entities were the following: who can produce for export; whose ability of exporting is supported by central government actors; and what happens to the profit of the export, who distributes and shares it. Strategic actors of the Hungarian foreign trade sector were large, state-owned companies, operating in a monopoly within their own field. Although the direction of export of industrial products was only significant to the socialist and developing countries, maintaining the export capacity of the Hungarian industry was an important goal for large companies, whereas the bargaining position of the exporting companies was much stronger than that of the others. This bargaining position determined—even in the period between 1968 and 1972—the amount of profits remaining by the companies.

We can also see on Figure 1, that *agriculture* had some extent weaker relationship with export in the text of the official party outlet. It is especially interesting as beside industrial companies, large agricultural holdings were also concerned with exports. According to earlier analyzes (Juhász 2001, 36–7), the viability of Hungarian industry largely depended on the success of agricultural exports, and the export revenues from agricultural products—as industrial

imports were largely covered by exported agricultural products.⁸

If we examine the strength of the relationship between the words *central*, *governance*, *efficiency*, and *modernization* with *industry* and *agriculture* (Figure 2), we can observe intensive dynamics. The figure shows the strong centralization aspirations of the early Kádár era (the period between 1960 and 1965). This was a period of the creation of “trusts” from economic entities. Political leadership simultaneously had to deal with the new investment needs of inherited priority corporations (“trusts”); the socialist transformation of agriculture; and also with the shaping of the financial foundations for the consolidation of the living standards to make people satisfied. These processes were implemented under increased central control, therefore the independence of the companies and the number of individual production units were reduced, while the size of central administration staff was expanded.

The economic decentralization of the mid-sixties also has its mark in the party’s official communication. At the time, the independence of production units increased in both sectors, and the role of central management has radically decreased. At the same time, *efficiency*, another concept of economic policy communication began to rise in the discourse. In the late 1950s, this concept has not appeared in the official communication yet, but since 1964, more and more emphasis has been on production efficiency, especially in the context of the economic reform in 1968 (Varga 2002; Romsics 2007). While we see fluctuating relationships between other examined concepts and the two sectors, the relationships with “efficiency” reflect to the economic- and economic policy crises and changes of the 60s and 70s very clearly. Only the importance of its relationship with *agriculture* declined by the mid-1980s, signaling the relegation of agricultural production to the background of economic policy, the onset of the post-1990 crisis in the agricultural sector. The term *modern* was a slogan for the launch of the consolidation policy in the early 1960s and then for the economic reform in 1968. In a peculiar way, the overthrow of the reform is also marked by that concept. The peaks in the relationship of this concept with the two sectors mirrors these communication patterns.

Conclusions and future work

In this paper, we examined the dynamic changes in the political discourse between 1959 and 1989 in Hungary on the basis of the official journal of the Hungarian governing party of the era. Despite the fact

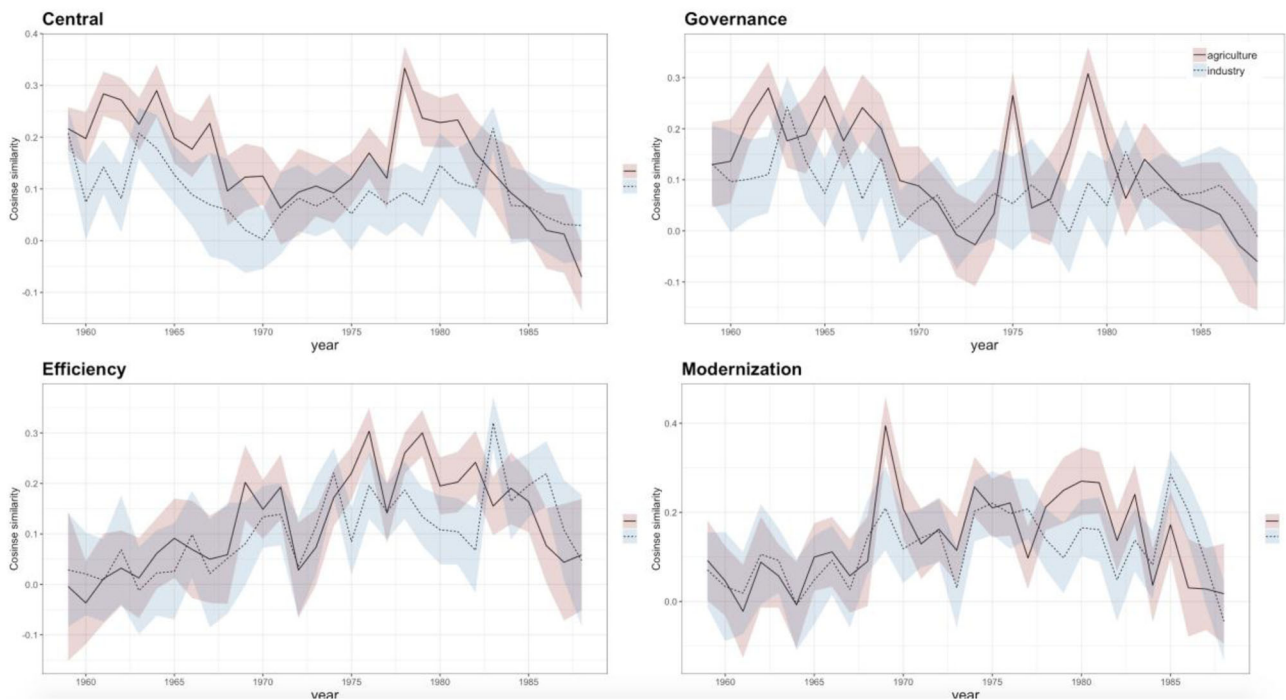


Figure 2. The cosine similarities of *central*, *governance*, *efficiency* and *modernization* with *agriculture* and *industry* between 1959 and 1988 ((we calculated the 95% confidence interval based on the 100 embedding trained for each year with the following formula: $MEAN \pm (SD/\sqrt{100}) \cdot 1.96$)).

that the given period of Hungarian history is a widely researched topic in the historical, political and socio-logical sciences, a longitudinal, computer-based, automated text analysis of the discourse of this era has not been made so far. We believe that NLP driven analysis of this era brings new evidence and insights to the organization of the regime and the most relevant policy shifts within the era.

As a first step of the research work we compiled a large, digitized corpus from the Hungarian journal “Pártélet,” the official issue of the Central Committee of the MSZMP, the governing party in Hungary in the investigated era. After the basic pre-processing steps of the raw texts we processed the data with a word embedding method. This method allowed us to analyze the dynamic changes of some of the key concepts of the examined discourse.

Two key concepts, which were analyzed thoroughly were *mezőgazdaság* (“agriculture”) and *ipar* (“industry”), which were in the focal point of the official discourse of the era. Our main research goal was to map the temporal dynamic of selected concepts in the political discourse during the years of the Kádár era, between 1956 and 1989 in Hungary. Our findings show that there was close relationship between the policy decisions and their communications. The change in the discourse always precedes the evaluation and modification of processes. Compared to analysis based on

the history of events, this method is more suitable for describing fine granular processes. Using quantitative text mining methods, it makes it possible to study large volumes of texts in a way that we would not be able to do by reading the articles one by one.

With the computer-assisted analysis of all volumes in this journal, we have the opportunity to analyze the leading political discourse of the examined period. This type of corpora and methods make it possible to longitudinally analyze the discourses of the era on large scale textual data and thus, to understand the different social historical processes even on a latent level. What is more, our corpus provides a unique opportunity for a systematic analysis of a huge amount of propaganda texts which is one of the most challenging tasks in the field of Natural Language Processing (Rashkin et al. 2017; Barrón-Cedeno et al. 2019; Kmetty et al. 2020).

As a next step of the research project, we would like to investigate the differences and similarities between the dynamic features of the discourse of “Pártélet” and of some other journals like “Népszabadság”/“Szabad Nép” that were intended for a different audience and purpose and thus compare the different narratives of the same phenomena in different timepoints.

In our paper, we argued and showed that the creation of such corpora, like the one we built from “Pártélet” journal can lead to more valid and reliable

analysis of historical texts and thus contribute to the deeper understanding of different historical eras.

Notes

1. From Russian language: *kol(lektivnoe) hoz(yaĭstvo)* “collective household economy.”
2. <http://www.inf.u-szeged.hu/rgai/magyarlanc>
3. This was both true for the words agriculture and industry. With the parameters of 200 dimensions and 10 size window the mean similarity of agriculture between same year embeddings was 0.67, and the same value for industry was 0.64. Based on our results, in our corpus, the similarity value depends more on the window size than on the number of dimensions.
4. Here we used a robust approach. As described earlier, we ran 100 embeddings for each year. Based on the cosine similarities, we extracted the 50 closest words from every run and then checked how many times a word was extracted. Thus, the method is eventually a robust rank order of the closest words.
5. “In the analysis of the situation of the national economy, the Congress has devoted much space to boosting agriculture. The report of the Central Committee stated that the development of agriculture should be made a matter for the whole party, the people. One of the most important tasks of the new 5-year plan is to boost agriculture.” *Pártélet*, 1966/5.
6. “One of the basic goals of the economic governance reform is to make the most efficient use of production opportunities in agriculture.” *Pártélet*, 1968/6. “The reform of the new economic mechanism has removed the barriers that previously hampered the desirable development of agriculture and the food economy.” *Pártélet*, 1969/4.
7. “The struggle for the socialist reorganization of agriculture in the Hungarian villages is successfully completed.” *Pártélet*, 1961/4.
8. “Although some industrial companies have made significant efforts to expand capitalist exports non-ruble exports were lower than in 1981.” *Pártélet*, 1984/1. “In 1984, it will be a basic requirement for industry to play a key role in increasing exports.” *Pártélet*, 1984/1. “The industry has to pursue a fundamentally export-oriented, restructuring-oriented development policy.” *Pártélet*, 1984/1

Source

Arcanum DigiTheca (<https://adtplus.arcanum.hu/en/collectio/Partelet/>)

Funding

This work was supported by the Hungarian Research Fund (NKFIH / OTKA, grant number FK 131826).

References

- Antoniak, M., and D. Mimno. 2018. Evaluating the stability of embedding-based word similarities. *Transactions of the Association for Computational Linguistics* 6:107–19. doi: [10.1162/tacl_a_00008](https://doi.org/10.1162/tacl_a_00008).
- Bamler, R., and S. Mandt. 2017. Dynamic word embeddings. In *Proceedings of the International Conference on Machine Learning (ICML 2017)*, 380–389, Sydney, NSW, Australia.
- Barrón-Cedeno, A., I. Jaradat, G. Da San Martino, and P. Nakov. 2019. Propy: Organizing the news based on their propagandistic content. *Information Processing & Management* 56 (5):1849–64.
- Cristianini, N., T. Lansdall-Welfare, and G. Dato. 2018. Large-scale content analysis of historical newspapers in the town of Gorizia 1873–1914. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 51 (3):139–64. doi: [10.1080/01615440.2018.1443862](https://doi.org/10.1080/01615440.2018.1443862).
- Csanádi, M. 2004. A comparative model of party-states: The structural reasons behind similarities and differences in self-reproduction, reforms and transformation. IEHAS Discussion Papers, No. MT-DP - 2004/7, Hungarian Academy of Sciences, Institute of Economics, Budapest.
- Csurgó, B. 2007. Építészek és falvak [Architects and villages]. In *Vidék- és falukép a változó időben [Country and rural images in changing times]*, ed. I. Kovách. 99–114. Budapest: Argumentum.
- Csurgó, B., and L. Kiss. 2007. A Kádár-kor [The Kádár-era]. In *Vidék- és falukép a változó időben [Country and rural images in changing times]*, ed. I. Kovách. 133–156. Budapest: Argumentum.
- Desagulier, G. 2019. Can word vectors help corpus linguists?. *Studia Neophilologica*, Taylor & Francis (Routledge): SSH Titles, 2019, [ff10.1080/00393274.2019.1616220](https://doi.org/10.1080/00393274.2019.1616220). [ff10.1080/00393274.2019.1616220](https://doi.org/10.1080/00393274.2019.1616220).
- Földes, G. 2019. Economic reform, ideology, and opening, 1965–1985. *Múltunk*, Special issue: Openness and Closedness – Culture and Science in Hungary and the Soviet Bloc after Helsinki 4–27.
- Gábor, R. I. 1989. Second economy and socialism: The Hungarian experience. In *The underground economies: Tax evasion and information distortion*, ed. E. Feige, 339–60. Cambridge: Cambridge University Press.
- Garg, N., L. Schiebinger, D. Jurafsky, and J. Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115 (16):E3635–44. doi: [10.1073/pnas.1720347115](https://doi.org/10.1073/pnas.1720347115).
- Guldi, J., and D. Armitage. 2014. *The history manifesto*. Cambridge: Cambridge University Press.
- Gyáni, G. 2016. *A történelem mint emlékmű [History as a memory/monument]*. Budapest: Kalligram.
- Győri, Z. 2013. Diskurzus, hatalom és ellenállás a késő Kádár-kor filmszociográfiáiban [Discourse, power, and resistance in late Kádár-era film sociographies]. *Apertúra*. Tavasz. https://uj.apertura.hu/2013/tavasz/gyori_diskurzus-hatalom-es-ellenallas-a-keso-kadar-kor-filmszociografiaiban (last accessed August 17, 2020).
- Hamilton, W. L., J. Leskovec, and D. Jurafsky. 2016a. Cultural shift or linguistic drift? comparing two computational measures of semantic change. In *Proceedings of*

- the 2016 Conference on Empirical Methods in Natural Language Processing, 2116–2121. Austin, TX: Association for Computational Linguistics.
- Hamilton, W. L., J. Leskovec, and D. Jurafsky. 2016b. 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)*, ed. K. Erk and N. A. Smith, Vol. 1, 1489–1501. Berlin: Association for Computational Linguistics.
- Harris, Z. S. 1954. Distributional structure. *Word* 10 (2–3): 146–62. doi: [10.1080/00437956.1954.11659520](https://doi.org/10.1080/00437956.1954.11659520).
- Hengchen, S., R. Ros, and J. Marjanen. 2019. A data-driven approach to the changing vocabulary of the nation in English, Dutch, Swedish and Finnish newspapers, 1750–1950. In *Proceedings of the Digital Humanities (DH) Conference, Utrecht*.
- Herczog, N. 2017. *Feljelentő színikritika a Kádár-korban. Kísérlet egy sztálinista kultúráirányítási modell átalakulásának megragadására* [“Reporting theater criticism” in the Kádár-era. An attempt to demonstrate the transformation of a Stalinist cultural governance model]. PhD diss., Színház-és Filmművészeti Egyetem.
- Hoekstra, R., and M. Koolen. 2019. Data scopes for digital history research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 52 (2):79–94. doi: [10.1080/01615440.2018.1484676](https://doi.org/10.1080/01615440.2018.1484676).
- Huistra, H., and B. Mellink. 2016. Phrasing history: Selecting sources in digital repositories. *Historical Methods: A Journal of Quantitative and Interdisciplinary History* 49 (4):220–9. doi: [10.1080/01615440.2016.1205964](https://doi.org/10.1080/01615440.2016.1205964).
- Jakab, A. Z. 2012. *Emlékkállítás és emlékezési gyakorlat. A kulturális emlékezet reprezentációi Kolozsváron* 1–316. [Remembrance and memory practice. Representations of cultural memory in Kolozsvár]. Cluj-Napoca. Kriza János Néprajzi Társaság–Nemzeti Kisebbségkutató Intézet.
- Jatowt, A., and K. Duh. 2014. A framework for analyzing semantic change of words across time. In *Proceedings of ACM/IEEE-CS Conference on Digital Libraries*, 229–238. London: IEEE Press.
- Joulin, A., E. Grave, P. Bojanowski, and T. Mikolov. 2016. *Bag of tricks for efficient text classification*. arXiv preprint arXiv:1607.01759. <https://arxiv.org/pdf/1607.01759.pdf> (last accessed September 2, 2020).
- Juhász, P. 2001. Az agrárscoda vége és a tulajdonreform. [The end of the ‘agricultural miracle’ and the property reform]. *Beszélő* 2001 (2):36–41.
- Kim, Y., Y. I. Chiu, K. Hanaki, D. Hegde, and S. Petrov. 2014. Temporal analysis of language through neural language models. <https://arxiv.org/pdf/1405.3515.pdf> (last accessed September 4, 2020).
- Kiss, L. 2007. Vidékmeghatározási viták az ezredfordulón [Countryside definition disputes at the turn of the millennium]. In *Vidék- és falukép a változó időben [Country and rural images in changing times]*, ed. I. Kovách. 233–254. Budapest: Argumentum.
- Kmetty, Z., V. Vincze, D. Demszky, O. Ring, B. Nagy, and M. K. Szabó. 2020. Pártélet: A Hungarian corpus of propaganda texts from the Hungarian socialist era. In *Proceedings of the 12th Language Resources and Evaluation Conference, Marseilles*, 2381–2388.
- Kornai, J. 1992. *The socialist system: The political economy of communism*. New Jersey: Princeton University Press.
- Kornai, J. 2014. The soft budget constraint. *Acta Oeconomica* 64 (Suppl. 1):25–79. doi: [10.1556/aoecon.64.2014.s1.2](https://doi.org/10.1556/aoecon.64.2014.s1.2).
- Kovai, M. 2016. *Lélektan és politika. Pszichotudományok a magyarországi államszocializmusban 1945–1970*. 1–514. [Psychology and politics. Psychosocial sciences in Hungarian state socialism 1945–1970]. Károli Gáspár Református Egyetem - L'Harmattan.
- Kozłowski, A. C., M. Taddy, and J. A. Evans. 2018. The geometry of culture: Analyzing meaning through word embeddings. <http://arxiv.org/abs/1803.09288> (last accessed August 7, 2020).
- Kulkarni, V., R. Al-Rfou, B. Perozzi, and S. Skiena. 2014. Statistically significant detection of linguistic change. In *Proceedings of WWW*, 625–635. <https://arxiv.org/pdf/1411.3315.pdf> (last accessed September 4, 2020).
- Kutuzov, A., L. Øvrelid, T. Szymanski, and E. Velldal. 2018. Diachronic word embeddings and semantic shifts: A survey. arXiv preprint arXiv:1806.03537. <https://arxiv.org/pdf/1806.03537.pdf> (last accessed September 4, 2020).
- Liu, Q., H. Huang, and Y. Gao. 2018. Task-oriented word embedding for text classification. In *27th International Conference on Computational Linguistics, 2023–2032*, New Mexico.
- Medgyesi, K. 2017. *Demokrácia-diskurzus(ok) Magyarországon az 1945 és 1949 közötti időszakban és időszakról* [Democracy discourse(s) in Hungary during and from 1945 to 1949]. PhD diss., Szegedi Tudományegyetem.
- Mikolov, T., K. Chen, G. Corrado, and J. Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781. <https://arxiv.org/pdf/1301.3781.pdf> (last accessed July 15, 2020).
- Mikolov, T., I. Sutskever, K. Chen, G. Corrado, and J. Dean. 2013b. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546. <http://arxiv.org/abs/1310.4546> (last accessed July 15, 2020).
- Mikolov, T., W.-T. Yih, and G. Zweig. 2013c. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL-HLT*, 746–751. <https://www.aclweb.org/anthology/N13-1090> (last accessed September 4, 2020).
- Miller, I. M. 2013. Rebellion, crime and violence in Qing China: A topic modeling approach. *Poetics* 41 (6):626–49. doi: [10.1016/j.poetic.2013.06.005](https://doi.org/10.1016/j.poetic.2013.06.005).
- Nagy, K. I. 2007. A dolgozó parasztság [The working peasantry]. In *Vidék- és falukép a változó időben [Country and rural images in changing times]*, ed. I. Kovách. 115–132. Budapest: Argumentum.
- Pap, M. 2015. *Kádár demokráciája. Politikai ideológia és társadalmi utópia a Kádár-korszakban [Democracy of Kádár. Political ideology and societal utopia in the Kadar era]*. Budapest: Nemzeti Közszolgálati Egyetem.
- Pap, M. 2017. A népitől a szocialista demokráciáig A korai Kádár-korszak demokráciafogalma a pártfolyóiratok tükrében [From the people’s democracy to the socialist democracy. The concept of democracy in the early Kádár-era in the light of party journals]. *Múltunk* 2017/1: 202–226.

- Pennington, J., R. Socher, and C. D. Manning. 2014. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), 1532–1543. <https://www.aclweb.org/anthology/D14-1162>, DOI: 10.3115/v1/D14-1162 (last accessed September 5, 2020).
- Rashkin, H., E. Choi, J. Y. Jang, S. Volkova, and Y. Choi. 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, 2931–2937, Copenhagen, Denmark: Association for Computational Linguistics, September.
- Romsics, I. 2007. Economic reforms in the Kádár Era. *The Hungarian Quarterly* 187:69–79.
- Schakel, A. M., and B. J. Wilson. 2015. Measuring word significance using distributed representations of words. <https://arxiv.org/pdf/1508.02297.pdf> (last accessed April 21, 2020).
- Sebők, M. 2016. *Kvantitatív szövegelemzés és szövegbányászat a politikatudományban [Quantitative text analysis and text mining in political science]*. Budapest: L'Harmattan.
- Selivanov, D., and Q. Wang. 2016. text2vec: Modern text mining framework for R. Computer software manual (R package version 0.4. 0). <https://CRAN.R-project.org/package=text2vec> (last accessed April 21, 2020).
- Spirling, A., and P. L. Rodriguez. 2019. Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *Journal of Politics*, Working paper. <https://www.nyu.edu/projects/spirling/documents/embed.pdf>, accepted for publication.
- Szabó, M. 2007. A dolgozó, mint állampolgár. Fogalomtörténeti tanulmány a magyar szocializmus három korszakáról [The worker as a citizen. A Concept History Study on the Three Era of Hungarian Socialism]. *Korall*, 2007/27: 151–171.
- Vági, G. 1982. *Versengés a fejlesztési forrásokért. Területi elosztás – társadalmi egyenlőtlenségek. [Competition for development resources. Territorial distribution - Social inequalities]*. Budapest: Közgazdasági és Jogi.
- Ullman, J. 2011. Data mining. In *Mining of massive datasets*, ed. J. Leskovec, A. Rajaraman and J. Ullman. 1–20. New York: Cambridge University Press.
- Varga, Z. 2002. Agriculture and the new economic mechanism. *Hungarologische Beiträge* 14:201–17.
- Xu, Y., and C. Kemp. 2015. A computational evaluation of two laws of semantic change. In Proceedings of 37th Annual Conference on Cognitive Science Society. <https://cogsci.mindmodeling.org/2015/papers/0463/paper0463.pdf> (last accessed September 5, 2020).
- Zsibrita, J., V. Vincze, and R. Farkas. 2013. magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In Proceedings of RANLP 2013, 763–771. <http://publicatio.bibl.u-szeged.hu/3981/1/Zsibrita-Vincze-Farkas.pdf> (last accessed September 9, 2020).

Appendix



Figure A1. The normalized frequency (per 1000 words) of agriculture and industry between 1959 and 1988 (we calculated the 95% confidence interval based on the 100 embedding trained for each year with the following formula: $MEAN \pm (SD/\sqrt{100}) \cdot 1.96$).

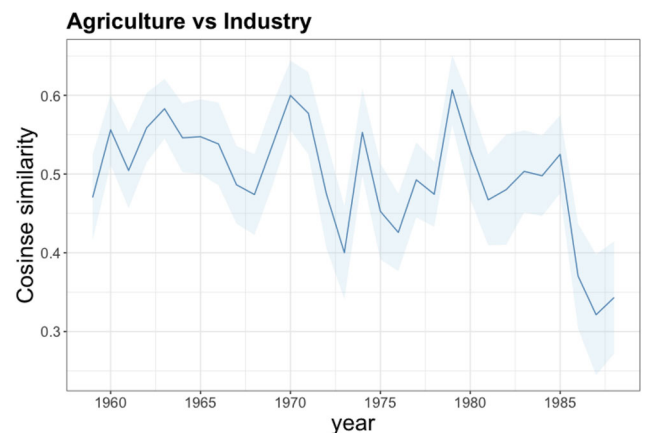


Figure A2. The cosine similarities of agriculture and industry between 1959 and 1988 (we calculated the 95% confidence interval based on the 100 embedding trained for each year with the following formula: $MEAN \pm (SD/\sqrt{100}) \cdot 1.96$).