

Text network analysis and visualization of Hungarian, communist-era political reports

Attila Gulyás
RECENS
Centre for Social Sciences, HAS
gulyas.attila@tk.mta.hu

Martina K. Szabó
RECENS
Centre for Social Sciences, HAS
szabo.martina@tk.mta.hu

István Boros Jr.
RECENS
Centre for Social Sciences, HAS
boros.istvan@tk.mta.hu

Gergő Havadi
RECENS
Centre for Social Sciences, HAS
havadi.gergo@tk.mta.hu

Abstract

This paper is presenting a part of our research project which aims to filter and visualize authority networks embedded in Hungarian communist-era political reports. The structure and development of authority networks are reconstructable owing to the well documented archive materials, reports and recorded interviews. The research focuses on the informal relations latent in authority networks. The corpus of the analysis is built by great amount of textual data mainly originated from reports recorded on party committee meetings. The quality of digitalization of these documents takes place on the wide scale of perfect readability and perfect unusability; therefore, it is a huge challenge to process these documents. Among others, the very basis of the process of text network analysis, its tools and methodology are presented; moreover, a step-by-step argument of visualization techniques is provided. Last but not least, on the basis of pilot analysis, the excellent opportunities of text analysis are demonstrated. Furthermore, the research aims the future application of sentiment and topic analysis in order to support or deny the previous findings.

1. Introduction

1.1 Historical Background

Following the Second World war the Hungarian Workers' Party (in Hungarian: Magyar Dolgozók Pártja, abbreviated MDP) became the governing party of Hungary for almost a decade. In these years between 1948 and 1956 the HWP was basically ran by a small group of power elite that secured its rule through the party hierarchy and its members' networks.¹ These networks include their connections in the political life as well as their informal relationships outside the world of politics as well. Later on a number of examples can be witnessed how the party officials strengthened their political connections due to their informal relationships or how they made political capital right out of these relationships [BoHK12, Majt10, Sík01].

In our research we compare the relationships created by political cooperation with the structures dictated by party hierarchy via the tools of network analysis. The research project is a historical elite research in a way which is being carried out with the tools of social sciences (the method of text, and social network analysis) through the

Copyright © 2018 for the individual papers by the paper's authors. Copying permitted for private and academic purposes. This volume is published and copyrighted by its editors.

In: A. Jorge, R. Campos, A. Jatowt, S. Nunes (eds.): Proceedings of the Text2StoryIR'18 Workshop, Grenoble, France, 26-March- 2018, published at <http://ceur-ws.org>

¹ The Hungarian Workers' Party (in Hungarian: Magyar Dolgozók Pártja, abbreviated MDP) was formed in June 1948 when the Hungarian Communist Party (MKP) and the Social Democratic Party (SZDP) 'merged'. Though officially it was referred to as a merger of the two parties representing the working class in reality the MKP absorbed the by that time weakened and decimated remains of the SZDP. Initially the chairman of the party was a social democrat, Árpád Szakasits, however that position was merely a formality and real power was in the hand of the Mátyás Rákosi, the General Secretary of MDP.

prosopographical examination of historical sources. Some previous promising Hungarian attempts in this field are for instance [Ková11, Rácz14].

Here we work with the definition of ‘elite’ as used by Hungarian sociologist Rudolf Andorka² [Ando06] and not by the classical definition by Mills [Mill56], as we focus elite formation mainly on positions held. The political elite is a type of this and in Stalinist communist dictatorships it is the exclusively dominating group over resources (economy, culture, society and the capital related to all these).³

The political/power elite of the Rákosi era (1948-1956) is relatively easy to identify and define from an organizational aspect: it consists of the members and the alternate members of the leading political bodies of the MDP, namely the Secretariat (‘Titkárság’), the Politburo (‘Politikai Bizottság’) and finally the Organizing Committee (‘Szervező Bizottság’), with the later one existing only until 1953. Even within these bodies primarily (and informally) those can be considered unquestionably and in a longer term perspective as part of the power elite who occupied high priority positions with personal influence and a network of relationships (therefore possessing information as well) in the state administration, in the leadership of mass organizations (e.g. trade unions), in the control of cultural life (like the chief editor of the daily newspaper of the MDP called *Szabad Nép*) or those supervising the law enforcement agencies, political police (ÁVH) and the military. These people can be referred to as the ‘inner circle of the party leadership’.

It is typical about the political elite of the Rákosi era that falling from power was as easy as getting in. In this regard the hysteria about constant vigilance and the show trials based on the Soviet model maintaining concepts of the enemy are excellent examples.

Later on, after the crushed revolution of 1956 the elite group of the Kádár era was largely based on the second and third lines of the elite of the Rákosi era (for instance Apró, Gáspár, Hegedüs, Komócsin, Münnich, Piros, Szalai, Vég or Czinege). Consequently it is especially interesting what kind of networks the members of the elite in the Rákosi era had established which may be interpreted within the party hierarchy, though they did not derive from that.

1.2 Text Network Analysis in a Nutshell

Network structures could be recognized in many aspect of life. The world-famous Hungarian origin author Barabasi argues, that the brain is a network of nerve cells connected by axons, the cells are a network of molecules connected by biochemical reactions. Societies are networks as well, networks variously embedded in technological systems such as internet and electronic networks [Bara06].

At the end of 1990s studies, from diverse fields, were arguing that uncorrelated networks (such as road networks and social networks, etc.) share mutual attributes; moreover, their attributes mathematically describable and analysable [BaBV08, Ková13, Watt04, Watt99].

The concept of text network analysis is defined as a paradigm, whereas texts are interpreted as networks and analysed with the tools of social network analysis (SNA) [Para11]. The method of SNA, alongside with the well-known roots of graph-theory, had been developing on the basis of physics’ more practical approaches [Bara02, ErRé59].

In social sciences SNA is frequently used as a tool for the measurement of social capital [AnTa12, Gran73]. Among others, the popularity of the application of network analysis arises from the method’s ‘social embeddedness’ [Néme14].

However, not only social networks, but also ‘everything could be visualized as networks’ [Para11]. Even the language we use for communication is a network of words connected with syntactic relations [Bara06]. According to this argument, in text representing networks nodes are not people, but pieces of texts, mostly words. Although, bimodal networks are popular, these networks associate keywords or hashtags with authors [Sedi15]. For instance, mapping the publications of a specific scientific field could help to understand the mainstream topics and the role of the authors in the development. The relations between the words have only one “uniplex” [Taká11] condition, the co-occurrence within the specified range of text. Thus, text network analysis is not only a new way of visualization, but does support the understanding of latent attributes and the contents of texts.

2. Research Questions and Sources

The aim of our research project was to create the latent network of the political elite in the Rákosi era (1948-1956) through processing and analysing different types of historical sources. The historical network of the dynamics of relations and of the latent and manifest hierarchy are analyzed.

Several previous papers have covered the developments of the latent relationships in the party elite, excellent examples are available about connections being established through informal channels such as hunting together [BoHK12], or about the considerably successful political activities of György Aczél which were strongly based on

² Referring to a small group on the top of social hierarchy that is smaller than the ruling class.

³ Experts doing research on recent history almost unanimously agree that it is worth analysing the power elite of one-party states not based on value or prestige but on the examination of positions. (Rácz 2014).

his network and personal connections [Sik01]. According to this argument, the main hypothesis of our research project is that latent connections and relationships could have been established between party members engaging together in political or additional activities which relationships shaped their actions in the political sphere parallelly to the party hierarchy.

2.1 Sources and Data Processing

The sources of the paper are the edited notes from the meetings of the leading political bodies (the Secretariat, the Politburo and the Organizing Committee) of the Hungarian Workers' Party, the MDP from between 1948 and 1956. To precisely identify the actors in the network we used other historical sources (such as biographies, records of the party membership and biographical databases) from which the individuals' political positions, their role in the party and numerous other pieces of information (like education, place of residence, participation at politically important events, etc.) are identifiable and furthermore, deductions can be made regarding their informal relationships. The sources and documents at our disposal are typed texts often with handwritten notes. Due to their varying state of conditions processing and analysing them presents a great challenge. Data processing was carried out according to the main working steps presented below.

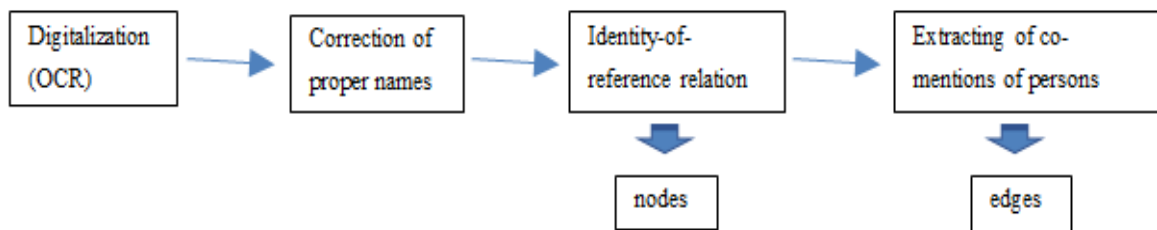


Figure 1: Steps of data processing

The data processing starts with the correction of the scanned and digitized text in a semi-automated way, followed by identifying the names found in the documents. Then with this information a network is created from this text using these names and their collocation as nodes and edges respectively. The following two sections explain these methods. The following two sections explain these methods in deeper detail.

2.2 Text Preprocessing

In order to be able to process and analyze the data the digitalization of the collected text was indispensable. This work was performed by optical character recognition (OCR) software. After digitalization a substantial correction of errors appeared in the texts was carried out. It was particularly essential to correct errors concerning proper names since these language elements are crucial from the point of view of the main goal of the research.

Errors of the proper names were caused by different features of the source texts. On one hand, in some of the cases condition of the paper of the documents or the quality of the ink used was not sufficient. In addition, it is possible that the scanning tools used during the work produced low quality results. On the other hand, in these types of historical documents it is common to use word processing features that are particular compared to the features admitted nowadays. For instance, it is not rare to put spaces in between all the characters of a proper name of a salariat in order to emphasize it, e.g. *P e t r ó c z i*. It is also common that a special form of a punctuation mark is used instead of a conventional one. For instance, in the given data common slash (/) is basically used to denote the bracket function., e.g. /Bencsik/. Evidently, algorithms that are trained on databases representing standard texts are not able to handle with these special text features sufficiently. The following examples demonstrate some of the typical problems, occurring during the digitalization process of the data. On the left side the source texts, on the right side the results of the digitalization are presented.

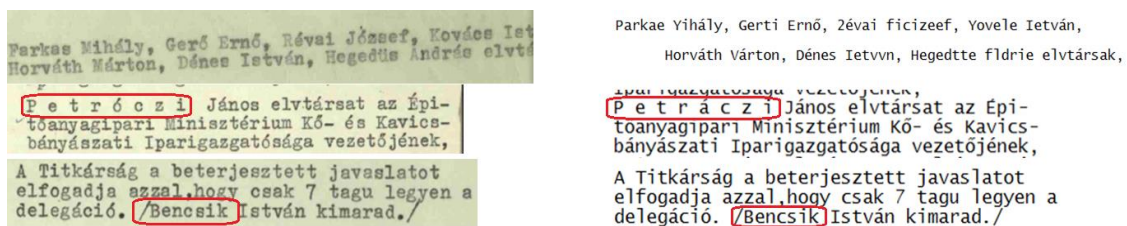


Figure 2: (from top to bottom) 1. Errors caused by bad condition of the documents; 2. A special pattern in order to emphasize a proper name; 3. A special usage of punctuation marks

As a result of all the mentioned conditions, most of proper names had to be corrected manually after digitalization, and the rest of the texts, as well. So as to reduce costs of manual processing of the proper names, an

automatic correction was carried out on the data as follows: on the basis of an available database consisting of proper names, an algorithm compared each character chain of the texts to the elements of this list. If the Levenshtein-distance [Leve66] of two character chains compared was less than 30%, the algorithm changed the wordform to the given element of the list, for instance: *Ger6* → *Gerő*, *K6d6r* → *Kádár*. With this method it was able to correct a notable amount of the typical errors, concerning the proper names. Errors still remaining in the corpus were corrected manually with the help of the original documents. In order to make this process easier and more cost effective, a software was created specifically for this purpose.

The next step of the work was the identity-of-reference relation that is a part-task of coreference resolution [Simo13, ViFa12]. In this phase every proper name that occurred in the database was assigned to the corresponding entity that the given proper name denotes. Coreference resolution of proper names is far not a trivial task in computational linguistics, namely proper names referring to the same entity can occur in different forms in texts. When two or more elements denote the same discourse entity (e.g. person, location, organization etc.) is called *coreference* [ZCCS11]. Due to identification of reference relations of proper names in our database the coreference resolution was accomplished, as well [Simo13, ViFa12]. It is worth noting that in coreference resolution not just proper names are assigned to the corresponding entities, but other referential language elements (e.g. personal pronouns, special verb forms etc.) in the given text, as well. However, it is important to emphasize that in kind of texts represented in our database these elements do not appear, proper names are used to refer to entities.

The identity of reference relation was performed on the basis of a list of proper names with semi-automatic processing method. This involves that human annotators examined all the proper names respectively and the software proposed possible references to all of them. Besides this, to make the decision process of the annotators easier and to increase the efficiency of the work, additional biographical information about suggested references was presented by the software, as well. Based on the results of coreference resolution the software converted each proper name into a unique tag of the given person with the help of the database of proper names (see above). For instance, different forms referring to the same entity like *Rákosi Mátyás* and *Rákosi* were converted to the code `<rakosi_matyas_8538>`.

In order to analyse the relations the first step was to connect the words; therefore, the first task is to create the networks, where the nodes are proper names and the connections between them are based on collocation or the lack of collocation.

Collocation is stated when two identification tag co-occurs in the same paragraph and the distance between them is not more than five words. The process of the analysis now holds at the identification on names. As the quality of the OCR is quite acceptable at numerous records (above 80%), an opportunity rises to apply deeper techniques such as text network analysis on the whole dataset.

3. Methodology

The utmost important result of text network analysis is revelation of the text's structure. This type of analysis focuses on the description of the topics [Para11] (distinguished from topic-analysis [LiNG09]) and their relations in the text. In the following paragraphs, the method is presented throughout the analysis of a text corpus.

The test corpus was built with six, randomly chosen Secretarial records. Previous to the analysis, the corpus was formatted to a tidy-text format in order to ease the input for processing software.

The model presented here is a structural model, since the whole word count (without stop words) of the corpus is included and the main goal is to process huge amount of textual data throughout the analysis and visualization of the structure of text. The most frequent forms of textual data visualized and analysed as networks are semantic networks [Even09].

In semantic networks usually stemming, lemmatization and N-grams are applied. However, text network analysis aims to analyse the original forms of words; therefore, in this scenario the paper does not apply stemming nor lemmatization. In this paper only lexemes are studied, n-grams are disregarded.

The non-informative words, function words as well as the typical words of reports (such as report) were filtered out with the application of a stop list. The stop list was manually built and optimized for the task, on the basis of experiences with the corpus.

The edge list, the list containing the relations between words, were provided by co-occurrence results. Although co-occurrences could be measured in different units of text such as document, paragraph or sentence or Δx distance of words. The Δx distance of words is a distance computed towards both directions from the occurring point of the source word. For the calculation of the matrix the software WORDij [Dano13] was chosen. Word pairs co-occurring in the same sentence less than five words far from each other are included into the analysis. The coloring of the graph figure significantly supports the understanding of the data [FrDu93]. To visualize and analyse the text networks the software Gephi 0.9.1 [BaHJ09] was applied, this software possesses advanced graphical and algorithmical features.

The edges, based on the very nature of co-occurrences in the text, do not have directions; therefore, the results of the text network analysis are interpreted as undirected ones. On the network figures words indicated as nodes, their size equal to the amount of their betweenness centrality [Bran01, FrDu93, Para11]. Modularity algorithm identifies

the communities within the network. Nodes ordered in the same community have more connections than it would be expected on the basis of chance in a random network [ErR659] with the same amount of nodes and density. Therefore, the clusters computed with this formula present the topic-communities of the corpus.

4. Results

In this section the results of the analysis on the test corpus will be presented. The purpose of this description is to introduce text network analysis rather than testing actual hypotheses about our research subject. As it was already mentioned this corpus covers only a very small portion of the actual text to be processed.

The network constructed from the database consists of 806 nodes (words) and 783 edges (co-occurrences). Clearly, this means that we have a sparse network, yet most edges are found in a few clusters that are presented in Figure 3.

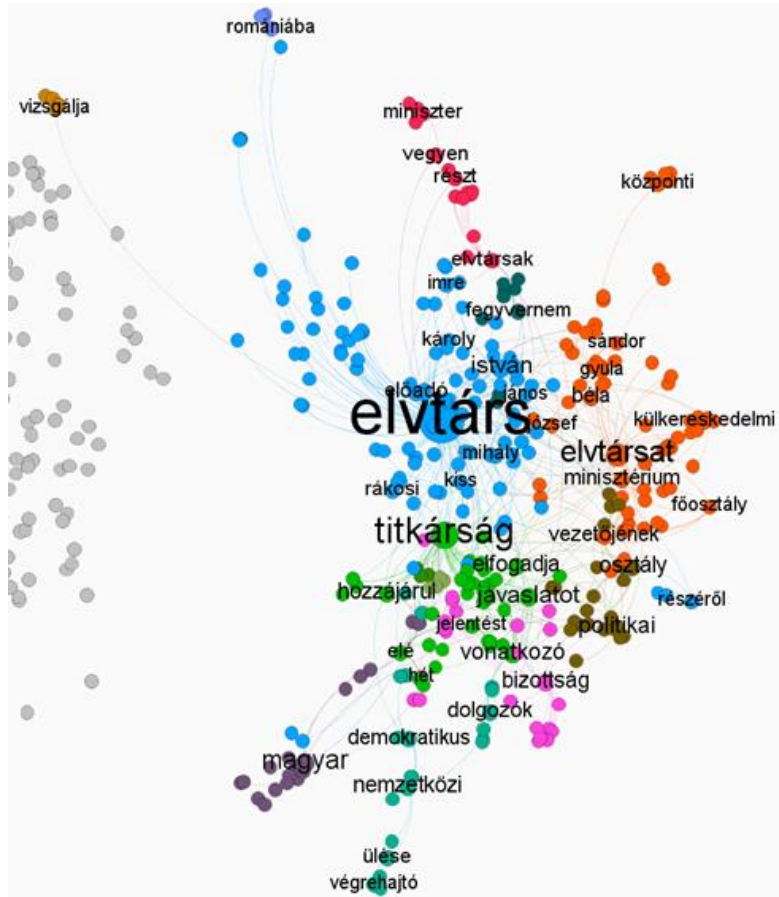


Figure 3: The network representing the corpus⁴

Most nodes in this network have less than 5 edges (the network is undirected) and there are only a few nodes having more than 40 edges. These are the followings: *elvtárs* ‘comrade-NOM’, *elvtársat* ‘comrade-ACC’, *titkárság* ‘secretariat-NOM’ and *magyar* ‘Hungarian-NOM’.

The average degree of the nodes is 1.943, the highest degree present is 104. The results show that the variance of the words in the corpus is relatively low and there are some word-pairs mentioned together with a notable high frequency.

The modularity of this network is 0.65, suggesting that the clusters identified are not random clusters, the nodes therein are more interconnected than those in a random graph. There are 420 cluster in the network in total, with 9 of them making up more than 2% of the whole network. Those nodes that are not connected to any other nodes are considered as individual clusters. Only a few of these nodes are not presented on Figure 3. The clusters associated to specific topics cover almost a third of all words in the corpus.

Figure 3 highlights that analyzing the text as a network enables us to uncover the most important topics in the

⁴ English equivalents of the words of the network are not presented here because translation of language elements without contexts would not be adequate from theoretical point of view. Instead of this, a survey and explanation of the main results of the network analysis is given here, together with samples.

text. For instance, the node *elvtársat* ‘comrade-ACC’ visibly occurs with those words that semantically connected to the topics of directive, nomination and assignment. At the same time, the node *titkárság* ‘secretariat-NOM’ is connected to words, expressing different types of official activity. These elements are verbs in majority like *elfogad* ‘accept-3SING’ and *hozzájárul* ‘consent-3SING’ or nouns in accusative case like *javaslatot* ‘proposal-ACC’ and *jelentést* ‘report-ACC’.

Technique applied in our project proves very useful for the processing of such a small corpus as the current test corpus and in this specific case it has pointed out a very interesting phenomenon.

As mentioned before stemming was not applied during the analysis. As it turned out this had a very important impact: all the wordforms were preserved in the database. For instance, the accusative case form of the word *elvtárs* ‘comrade-NOM’ is *elvtársat* ‘comrade-ACC’ in Hungarian. In the highly inflective, agglutinative Hungarian language the affixes are directly connected to the words, stemming would remove conjugation of the words, therefore, ‘assignments’ did not occur as independent hubs. The central word of this cluster is *elvtárs* in accusative case. This word and therefore this cluster would have been completely lost if the stemming was applied.

5. Conclusions and Future Work

This paper gives a brief overview of applying text network analysis on a small corpus constructed from the material of a current research. Also some linguistic processing aspects of this research were presented to further emphasize the need and use of text network analysis in this work.

The research focuses on the analysis of committee meeting minutes of the Hungarian Workers’ Party (between 1949 and 1956) to uncover the latent political network behind the party hierarchy based on the cooperation of individuals in formal or informal matters. The connections between individuals are modeled with the common mentions in the committee meeting minutes.

The work is done in multiple steps. First and foremost the quality of the text –optical character recognition was performed on documents with different levels of material preservation – required a semi-automated correction of the text. The current phase of the research focuses on correcting the obtained text and creating text networks from the corrected text. Correction requires considerable labor and historical knowledge as well. Text network analysis will provide the final results of this research – this method was presented in this paper with a restricted corpus constructed from a few documents only. The results presented here highlight that analyzing text as a network of words can point out the most important topics in the text.

As an additional outcome of this study it was found that stemming would undermine this method, but this is due to the specificities of the Hungarian language, where affixes are directly attached to the stem of the words. Stemming would mask the the actual functions of the entities denoted by the wordforms of the database, hence important topics could be lost – the same phenomenon was seen in the processing of the test corpus.

One of the following research steps planned is sentiment analysis of the texts of the corpus. The goal of this work is to reveal those semantic contents that express positive or negative evaluation of some target (object, person, or event) of the texts. The analysis is going to be carried out with the help of the bag of words model [BoMo09] that is a cost-efficient method of information retrieval tasks (compared to learning algorithms) [DrSz17]. For this work a sentiment dictionary is required, containing linguistic elements with positive and negative evaluative meaning [Szab15, Szab16]. At the same time, sentiment analysis of these kind of historical texts is far not trivial by reason of the special semantic features connected to historical circumstances. In a consequence of this, a sustancial research should be carried out before executing the work.

6. References

- [Ando06] ANDORKA, RUDOLF: *Bevezetés a szociológiába* (‘Introduction to Sociology’). Budapest : Osiris, 2006
- [AnTa12] ANGELUSZ, RÓBERT ; TARDOS, RÓBERT: A gyenge kötések ereje és gyengesége (‘Potency and weakness of weak ties’). In: *Hálózatok, Stílusok, Kultúrák*. Budapest : ELTE Angelusz Róbert Társadalomtudományi Szakkollégium, 2012, S. 101–127
- [BaBV08] BARRAT, ALAIN ; BARTHÉLEMY, MARC ; VESPIGNANI, ALESSANDRO: *Dynamical Processes on Complex Networks*. Reprint edition. Aufl. Cambridge : Cambridge University Press, 2008 — ISBN 978-1-107-62625-6
- [BaHJ09] BASTIAN, M. ; HEYMANN, S. ; JACOMY, M.: Gephi: an open source software for exploring and manipulating networks. In: , 2009
- [Bara02] BARABÁSI ALBERT-LÁSZLÓ: *Behálózza - A hálózatok új tudománya* (‘In the net – The new science of networks’) : Helikon, 2002

- [Bara06] BARABÁSI, ALBERT-LÁSZLÓ: A hálózatok tudománya: a társadalomtól a webig ('Science of networks: from society to web'). In: *Magyar Tudomány* (2006), Nr. 11, S. 1298–1308
- [BoHK12] BOZSONYI, KÁROLY ; HORVÁTH, ZSOLT ; KMETTY, ZOLTÁN: A hatalom hálójája - A Kádár-kori hatalmi elit hálózati struktúrája az együttvadásási szokások alapján ('The Power Grid. The Social Network of the Hungarian Elite in the Kádár era Based on Hunting Habits'). In: *Korall* (2012), Nr. 47, S. 157–184
- [BoMo09] BOIY, ERIK ; MOENS, MARIE-FRANCINE: A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts. In: *Inf. Retr.* Bd. 12 (2009), Nr. 5, S. 526–558
- [Bran01] BRANDES, ULRİK: A faster algorithm for betweenness centrality. In: *The Journal of Mathematical Sociology* Bd. 25 (2001), Nr. 2, S. 163–177
- [Dano13] DANOWSKI, J. A.: WORDij version 3.0: Semantic network analysis software, University of Illinois at Chicago (2013)
- [DrSz17] DRÁVUCZ, FANNI ; SZABÓ, MARTINA KATALIN: A beszélői szubjektivitás vizsgálata szentiment- és emóciókorpuszokon ('Analysis of subjectivity on the basis of sentiment and emotion corpora'). In: LUDÁNYI, Z. (Hrsg.): *Doktoranduszok tanulmányai az alkalmazott nyelvészet köréből*, 2017, S. 39–49
- [ErRé59] ERDŐS, PAUL ; RÉNYI, ALFRÉD: On Random Graphs I. In: *Publicationes Mathematicae (Debrecen)* Bd. 6 (1959), S. 290–297
- [Even09] EVENS, MARTHA WALTON: *Relational Models of the Lexicon: Representing Knowledge in Semantic Networks*. 1st. Aufl. New York, NY, USA : Cambridge University Press, 2009 — ISBN 978-0-521-10476-0
- [FrDu93] FREEMAN, LINTON C. ; DUQUENNE, VINCENT: A note on regular colorings of two mode data. In: *Social Networks* Bd. 15 (1993), Nr. 4, S. 437–441
- [Gran73] GRANOVETTER, MARK: The Strength of Weak Ties. In: *The American Journal of Sociology* Bd. 78 (1973), Nr. 6, S. 1360–1380. — ArticleType: primary_article / Full publication date: May, 1973 / Copyright © 1973 The University of Chicago Press
- [Ková11] KOVÁCS, I. G.: *Elitek és iskolák, felekezetek és etnikumok - Társadalom- és kultúratörténeti tanulmányok ('Elites and Schools, Denominations and Ethnic Groups. Papers in Social and Culture History')*. Budapest : L'Harmattan, 2011 — ISBN 978-963-236-452-0
- [Ková13] KOVÁCS, LÁSZLÓ: *Fogalmi rendszerek és lexikai hálózatok a mentális lexikonban ('Systems of concepts and lexical networks in mental lexicon')* : Tinta Könyvkiadó, 2013 — ISBN 978-615-5219-35-1
- [Leve66] LEVENSHTAIN, V. I.: Binary Codes Capable of Correcting Deletions, Insertions and Reversals. In: *Soviet Physics Doklady* Bd. 10 (1966), S. 707
- [LiNG09] LIU, YAN ; NICULESCU-MIZIL, ALEXANDRU ; GRYC, WOJCIECH: Topic-link LDA: Joint Models of Topic and Author Community. In: *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*. New York, NY, USA : ACM, 2009 — ISBN 978-1-60558-516-1, S. 665–672
- [Majt10] MAJTÉNYI, GYÖRGY: *K-vonal - Uralmi elit és luxus a szocializmusban - Uralmi elit és luxus a szocializmusban ('Cadre-line. Dominant Elite and Luxury during Socialism')* : Nyitott Könyvműhely, 2010 — ISBN 978-963-9725-74-4
- [Mill56] MILLS, C. WRIGHT: *The power elite*. New York : Oxford University Press, 1956 — ISBN 978-0-19-500680-3
- [Néme14] NÉMETH, RENÁTA: Módszerek a kvantitatív társadalomkutatási paradigmákban ('Methods in quantitative social science paradigms'). In: *SOCIO.HU* Bd. 3 (2014), Nr. 10.18030/SOCIO.HU.2014.3.27, S. 1–42

- [Para11] PARANYUSHKIN, DMITRY: *Identifying the Pathways for Meaning Circulation using Text Network Analysis*. URL <http://noduslabs.com/research/pathways-meaning-circulation-text-network-analysis/>. - abgerufen am 2017-11-23
- [Rácz14] RÁCZ, ATTILA: A budapesti hatalmi elit prozopográfiai vizsgálata 1956-1989 (Prosopographical Study of the Ruling Elite in Budapest, 1956-1989) (ELTE BTK Doktori disszertáció). Budapest, 2014
- [Sedi15] SEDIGHI, MEHRI: Using of co-word analysis method in mapping of the structure of scientific fields(case study: The field of Informetrics). In: *Journal of Information processing and Management* Bd. 30 (2015), Nr. 2, S. 373–396
- [Sík01] SÍK, ENDRE: Aczélhálóban ('In the Net of Aczél. Contribution to Understanding of the Operation of Social Capital'). In: *Szociológiai Szemle* Bd. 3 (2001), S. 64–77
- [Simo13] SIMON, ESZTER: A magyar nyelvű tulajdonnév-felismerés módszerei ('Methods of Named Entity Recognition') (Tézisfüzet). Budapest, 2013
- [Szab15] SZABÓ, MARTINA KATALIN: Egy magyar nyelvű szentimentlexikon létrehozásának tapasztalatai és dilemmái ('Experiences and dilemmas of the creation of a Hungarian sentiment dictionary'). In: GECSŐ, T. ; SÁRDI, C. (Hrsg.): *Nyelv, kultúra, társadalom. Segédkönyvek a nyelvészet tanulmányozásához*. Bd. 177, 2015, S. 278–285
- [Szab16] SZABÓ, MARTINA KATALIN: A nyelvi értékelés mibenlétének kérdése a számítógépes értékelélemzés (szentimentelemzés) szempontjából ('Concept of evaluation in the language usage from computational linguistics point of view'). In: GÉCSEG, Z. (Hrsg.): *LingDok 15. Nyelvészdoktoranduszok dolgozatai*. Szeged : Szegedi Tudományegyetem, Nyelvtudományi Doktori Iskola, 2016, S. 153–172
- [Taká11] TAKÁCS, KÁROLY: *Kapcsolatháló elemzés; Társadalmi kapcsolathálózatok elemzése ('Network analysis; Analysis of social networks')*. Digitális tankönyvtár. Aufl. Budapest : Budapesti Corvinus Egyetem, 2011
- [ViFa12] VINCZE, VERONIKA ; FARKAS, RICHÁRD: Tulajdonnevek a számítógépes nyelvészetben ('Named Entities in Computational Linguistics'). In: *Általános nyelvészeti tanulmányok XXIV.* : Akadémiai Kiadó, 2012, S. 97–119
- [Watt04] WATTS, DUNCAN J.: The "New" Science of Networks. In: *Annual Review of Sociology* Bd. 30 (2004), Nr. 1, S. 243–270
- [Watt99] WATTS, D.J.: *Small Worlds* : Princeton University Press, 1999
- [ZCCS11] ZHENG, JIAPING ; CHAPMAN, WENDY W. ; CROWLEY, REBECCA S. ; SAVOVA, GUERGANA K.: Coreference resolution: A review of general methodologies and applications in the clinical domain. In: *Journal of Biomedical Informatics* Bd. 44 (2011), Nr. 6, S. 1113–1122