1.01 UDC: 328: 930.25:004.9

Andrej Pančur,* Mojca Šorn**

Smart Big Data: Use of Slovenian Parliamentary Papers in Digital History

IZVLEČEK

PAMETNI MASIVNI PODATKI: UPORABA SLOVENSKIH PARLAMENTARNIH DOKUMENTOV V DIGITALNI ZGODOVINI

Avtorja v prispevku opozorita na problem velikih količin digitalnih zgodovinskih virov, s katerim se bodo srečavali raziskovalci sodobne zgodovine. Bolj nadrobno predstavita slovenske parlamentarne dokumente kot primer pametnih masivnih podatkov. Avtorja menita, da velikih količin tega digitalnega gradiva zgodovinarji ne bodo mogli obdelovati samo z uporabo klasičnih zgodovinskih metod, temveč bodo morali začeti uporabljati še metode in orodja, ki jih razvija digitalna zgodovina, digitalna humanistika in tudi jezikoslovne tehnologije.

Ključne besede: digitalna humanistika, digitalna zgodovina, Slovenija, parlament

ABSTRACT

The paper calls attention to the problem of massive amounts of digital historical sources that will eventually be faced by researchers of contemporary history. Slovenian parliamentary papers are then presented in detail as an example of smart big data. The authors believe that historians will be unable to process massive amounts of such digital materials using only standard historiographical methods and will be forced to start using methods and tools developed by digital history, digital humanities and also language technologies.

Keywords: digital humanities, digital history, Slovenia, parliament

Big Data

Big data is the buzzword of the decade. However, the very ubiquity of the term both in the industry, the media as well as in the academic community has led to big data being defined in various ways. Furthermore, big data in the humanities is not

* **Research Associate, PhD, Institute of Contemporary History, Kongresni trg 1, SI-1000 Ljubljana, Slovenia, andrej.pancur@inz.si**

** **Research Associate, PhD, Institute of Contemporary History, Kongresni trg 1, SI-1000 Ljubljana, Slovenia, mojca.sorn@inz.si**

the same as big data in natural sciences.[1] This paper uses the definition according to which historical big data is described as follows;

- large volumes of data, particularly texts that cannot be read within a reasonable time frame, and
- information that only allows us to reach new conclusions through the use of digital methods.[2]

Large amounts of computer-readable data are gradually becoming the reality of modern historiography. The late Roy Rosenzweig, one of the pioneers of digital history, warned as early as 2003 that historiography, instead of working within the scarcity paradigm of historical records, will have to start facing the problem of an excess of resources.[3] Up to now, historiography has mostly had to deal with the lack of resources, their incompleteness and often also with high costs of acquiring additional sources. Today, on the other hand, historians can access new digitized and digital sources quickly and effectively. Although a lion's share of analogue materials has not been and is unlikely to be digitized within reasonable time,[4] the materials emerging today are increasingly created in the digital form. This is the reason why, for example, the Slovenian archives are currently working hard on establishing a Slovenian electronic archive – e-ARH.si.[5] Consequently, the problem of (over)abundance of historical resources is also emerging in the studies of contemporary history,[6] including the history of the Republic of Slovenia after 1991.[7]

The massive amount of digital materials has spurred the creation and spreading of digital humanities, which use digital methods and tools to address new research questions.[8] In the following sections, the paper uses the example of Slovenian parlia-

---

1 Christof Schöch, "Big? Smart? Clean? Messy? Data in the Humanities," *Journal of Digital Humanities* 2, no. 3 (2013), accessed on 25 September 2016, http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/.

2 Shawn Graham, Ian Milligan and Scott Weingart, *Exploring Big Historical Data. The Historian's Macroscope* (London: Imperial College Press, 2015), accessed on 28 September 2016. http://www.themacroscope.org.

3 Roy Rosenzwig, "Scarcity or Abundance? Preserving the Past in a Digital Era," *American Historical Review* 108, no. 3 (2003): 735–62.

4 Gerben Zaagsma, "On Digital History," *BMGN – Low Countries Historical Review* 128, no. 4 (2013): 19–23.

5 Tatjana Hajtnik, "Strategija razvoja slovenskega javnega elektronskega arhiva 'e-ARH.si'," *Knjižnica* 55, no. 1 (2011): 40, 41, 44. Bojan Cvelfar et al., *Strategija in izvedbeni načrt razvoja slovenskega elektronskega arhiva 2016–2020* (Ljubljana: Archives of the Republic of Slovenia, 2016), 10.

6 Peter Haber, "Zeitgeschichte und Digital Humanities," in: *Zeitgeschichte – Konzepte und Methoden*, eds. Frank Bösch and Danyel Jürgen (Göttingen: Vandenhoeck & Ruprecht, 2012), 47–66.

7 Jure Gašparič, "Pisati politično zgodovino Republike Slovenije," in: *Četrt stoletja Republike Slovenije – izzivi, dileme, pričakovanja*, eds. Jure Gašparič and Mojca Šorn (Ljubljana: Institute of Contemporary History, 2016), 30.

8 Sandra Collins et al., *ALLEA E-Humanities Working Group Report. Going Digital: Creating Change in the Humanities* (Berlin: All European Academies, 2015), 9, accessed 19 July 2016, http://www.allea.org/Content/ALLEA/WG%20E%20Humanities/Going%20Digital_digital%20version.pdf. Devon Elliot, Robert MacDougall and William J. Turkel, "New Old Things. Fabrication, Physical Computing, and Experiment in Historical Practice," *Canadian Journal of Communication* 37 (2012):

mentary papers to present the advantages and drawbacks of the digital historiography methods in the analysis of big historical data. In this context, the digital history is understood as part of the digital humanities, which is primarily concerned with online distribution and presentation of historical sources using various computer tools, especially for mapping, network software and, last but not least, text analysis.[9] The paper focuses mainly on the possibilities for the use of text analysis methods and tools.

Slovenian Parliamentary Papers

Parliamentary papers are a rich source of data used by different academic disciplines including historiography. In some European countries, a large part of these papers is already accessible in digital form, mostly in PDF format.[10]

Both researchers and the public can avail themselves of the materials from parliamentary institutions located within today's Slovenia and from parliamentary institutions whose members once included representatives from Slovenia. While it is true that most of the materials are currently only available in analogue form, an increasing amount has already been digitized and made available to the public:

- Austrian National Assembly (1861–1918);[11]
- Styrian Provincial Assembly (1848–1914);[12]
- Carniolan Provincial Assembly (1861–1869);[13]
- Yugoslav legislative bodies 1919–1939, 1942–1953;[14]
- People's Assembly of the People's Republic of Slovenia (1947–1963);[15]
- Assembly of the Socialist Republic of Slovenia (1963–1990);[16]
- National Assembly of the Republic of Slovenia, from 1990 until today.[17]

122, accessed 19 July 2016, http://www.cjc-online.ca/index.php/journal/article/view/2506.

9 Stephen Robertson, "The Differences between Digital Humanities and Digital History," in: *Debates in Digital Humanities 2016*, eds. Matthew K. Gold and Lauren F. Klein (Minneapolis and London: University of Minnesota Press, 2016).

10 Agiatis Benardou, Alastair Dunning, Martin Schaller and Nephelie Chatzi Chatzidiakou. *Research Themes for Aggregating Digital Content. Parliamentary Papers in Europe* (Europeana Cloud, 2015), 6.

11 "Parlamentaria," *ALEX – Historische Rechts- und Gesetzestexte*, accessed on 30 September 2016, http://alex.onb.ac.at/sachlichegliederung.htm.

12 "Landtag Steiermark – stenographische Sitzungsberichte," *Das Land Steiermark*, accessed on 30 September 2016, http://www.landesarchiv.steiermark.at/cms/ziel/111284715.

13 "Provincial Assembly of Carniola 1861–1918," *SIstory – History of Slovenia*, accessed on 30 September 2016, http://hdl.handle.net/11686/menu719.

14 "Stenographical minutes of the executive and legislative bodies, Yugoslavia," *SIstory – History of Slovenia*, accessed on 30 September 2016, http://hdl.handle.net/11686/menu396.

15 "Shorthand minutes of the People's Assembly of the People's Republic of Slovenia (1947–1963)," *SIstory – History of Slovenia*, accessed on 30 September 2016, http://hdl.handle.net/11686/menu407.

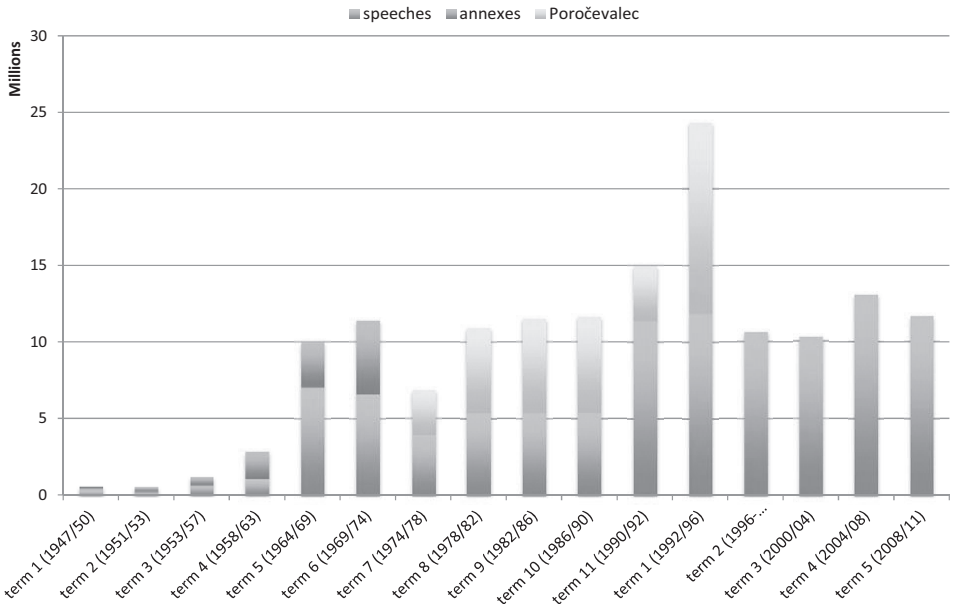
16 "Assembly of the Socialist Republic of Slovenia (1963–1990)," *SIstory – History of Slovenia*, accessed on 30 September 2016, http://hdl.handle.net/11686/menu408.

17 "Seje Državnega zbora – Po datumu," *Republic of Slovenia: National Assembly*, accessed on 30 September 2016, https://www.dz-rs.si/wps/portal/Home/deloDZ/seje/sejeDrzavnegaZbora/PoDatumuSeje/.

With the exception of the documents from the National Assembly, which are being published in the HTML format, all documents have been published in PDF.

Chart 1 shows the number of words contained in parliamentary speeches for each parliamentary term of the Slovenian parliament which are archived in PDF at the History of Slovenia – SIstory portal (36 million words) and in HTML at the National Assembly website (69 million words). Up until 1974, minutes of the sessions also included extensive attachments (10.5 million words), which were later published in a special serial publication titled *Poročevalec*. From 2010 onward, *Poročevalec* has been regularly accessible at the National Assembly website,[18] where visitors can also look through archived issues from 2006 to 2010.[19] Other issues of *Poročevalec* (1974–2006) are going to be digitized within three years by the Institute of Contemporary History that manages the History of Slovenia – SIstory portal. As of this moment, the Institute has already managed to digitize issues of *Poročevalec* from 1974 to 1996, which together contain almost 37 million words (see Chart 1).

Chart 1: Number of words in parliamentary speeches (1947–1990), attachments (1947–1974) and Poročevalec (1974–1996) at the History of Slovenia – SIstory portal and the number of words in parliamentary speeches at the National Assembly website (1990–2011); in millions of words per parliamentary term



18 "Gradivo DZ," *Republic of Slovenia: National Assembly*, accessed on 30 September 2016, https://www.dz-rs.si/wps/portal/Home/deloDZ/Porocevalec/GradivaDZ.

19 "Arhiv Poročevalec od 14.4.2006 do 15.7.2010," *Republic of Slovenia: National Assembly*, accessed on 30 September 2016, https://www.dz-rs.si/wps/portal/Home/deloDZ/Porocevalec/arhivPorocevalec.

Searching

It is clear that no researcher is able to read that much text in its entirety. Researchers thus only read those parts that they consider relevant for their research. In doing so, they read the selected parts of the text carefully, from word to word, or quickly skim over the pages looking for relevant parts of the text. However, such research is generally based on the assumption that the researchers will find what they are looking for in the text. Researchers thus determine the text they are looking for, as well as the context of their research, in advance. In doing so, they necessarily lean on their previous comprehensive or inadequate knowledge of the area of study.[20]

In history as well as in other humanities, such methods are of course completely acceptable and often yield useful results. A number of good studies has been created in such a manner using materials from the National Assembly of the Republic of Slovenia.[21] Most researchers in the humanities thus primarily understand digital materials in terms of easier and quicker access to desired information.[22] In the case of materials from the National Assembly, researchers can make use of a search engine that allows them to filter their results according to search modules. In a similar manner, researchers can search for older parliamentary materials at the History of Slovenia – SIstory portal.

Most researchers in the humanities first use search engines to identify sources and then do a full text keyword search. This means that studies are no longer conducted only vertically, from top to bottom, where a researcher only reads canonical texts and browses previously organized data collections. Rather, studies are being conducted in a bottom-up manner, with researchers looking for parts of text pertinent to their research that they would otherwise not have read. However, this research method has its limitations. The researcher must know the search queries in advance, and these can differ from modern thought patterns. Inasmuch as the researcher does not carefully examine every search result, the results are always lacking proper context.[23]

Despite these limitations, a full text keyword search can yield very useful results. A good example of such research is the article *War in Parliament: What a Digital Approach Can Add to the Study of Parliamentary History* whose authors used carefully

---

20 Federico Nanni, Hiram Kumper, and Simone Paolo Ponzetto, "Semi-Supervised Textual Analysis and Historical Research Helping Each Other. Some Thoughts and Observations," *International Journal of Humanities and Arts Computing* 10, no. 1 (2016): 73, 74.

21 Jure Gašparič, *Državni zbor 1992–2012. O slovenskem parlamentarizmu* (Ljubljana: Institute of Contemporary History, 2012). Jure Gašparič, *Slovenski parlament. Politično-zgodovinski pregled od začetka prvega do konca šestega mandata (1992–2014)* (Ljubljana: Institute of Contemporary History, 2014). Rosvita Pesek, *Osamosvojitev Slovenije* (Ljubljana: Nova revija, 2007).

22 Lisa Spiro, "Access, Explore, Converse: The Impact (and Potential Impact) of the Digital Humanities on Scholarship," in: *Keys for architectural history research in the digital era*, eds. Juliette Hueber and Antonio Mendes da Silva 3 (2014).

23 Robertson, "The Differences between Digital Humanities and Digital History." Bob Nicholson, "The Digital Turn. Exploring the methodological possibilities of digital newspaper archives," *Media History* 19, no. 1 (2013): 66, 67.

selected search queries and a search engine to systematically check to which extent the Boerenpartij (Farmers' Party) was described as "wrong" in all Dutch parliamentary debates between 1958 and 1982.[24] The article was written as part of the *War in Parliament* project. The results of this project clearly showed that satisfactory research results can only be obtained if we are familiar not only with the advantages but also with the shortcomings of digital research methods.[25]

Smart Big Data

One of the prerequisites for the *War in Parliament* project to be successful was the use of partly structured data in the XML format, which allowed for the search results to be filtered by speaker's name, by party, by time period, structure of the text, etc.[26] What was used was thus not big data in the form of plain text, but rather smart data. Smart data may be structured or partly structured, and compared to implicit big data, smart data is explicit, marked, enriched and described by metadata. The creation of smart data is often a labour-intensive process that requires human intervention.[27]

As we have seen, parliamentary papers represent extremely extensive data collections. We thus cannot expect to be able to reveal their explicit content merely through precise manual annotation. The dilemma necessarily encountered by researchers in the use of digital parliamentary papers was succinctly stated by Christof Schöch:

> *"I believe the most interesting challenge for the next years when it comes to dealing with data in the humanities will be to actually transgress this opposition of smart and big data. What we need is bigger smart data or smarter big data, and to create and use it, we need to make use of new methods. So, how can we enrich big data sufficiently to make more intelligent queries possible? How can we speed up the process of creating smart data so that we can produce larger volumes of it?"*[28]

At the same time, Schöch calls our attention to two possible ways of making big data smarter: through automatic annotation and through crowdsourcing. In practice, parliamentary papers proved very suitable for automatic annotation. In particular, parliamentary debates were written down in a format that has changed very little

24 Hinke Piersma et al., "War in Parliament. What a Digital Approach Can Add to the Study of Parliamentary History," *DHQ: Digital Humanities Quarterly* 8, no. 1 (2014).

25 Hinke Piersma and Kees Ribbens, "Digital Historical Research. Context, Concept and the Need for Reflection," *BMGN – Low Countries Historical Review* 128, no. 4 (2013): 87–90, 100, 101.

26 "War in Parliament," *NIOD*, accessed on 30 September 2016, http://www.niod.nl/en/projects/war-parliament.

27 Schöch, "Big? Smart? Clean? Messy? Data in the Humanities," 4.

28 Ibid., 10.

with time.[29] This is one of the reasons why various research projects often annotate parliamentary debates using the XML markup language. Among others, sessions of the British parliament (Hansard) from 1803,[30] the Dutch parliament from 1803,[31] the Spanish parliament from 1977,[32] the Czech parliament from 1993[33] and the Polish parliament from 1993[34] are all available in the XML format.

The following sections of this article will present the use of Slovenian parliamentary papers, particularly the minutes of parliamentary debates, in digital history. A number of cases will be presented to illustrate the huge potential of smart big data in contemporary history studies. As an example, 2.7 million words of the minutes of parliamentary debates in the Chamber of Associated Labour of the Assembly of the Republic of Slovenia from 1990 to 1992 have been annotated using the XML format.[35] In doing so, it was decided that the Text Encoding Initiative (TEI) Guidelines should be used,[36] as these are the *de facto* standard for the encoding of texts in digital humanities.[37]

Automatic conversions were carried out using XSL stylesheets created specifically for the project. However, annotation was also carried out by hand, not just automatically. The reason for this was that automatic conversions can also contain annotation errors. Attempts were made to find these errors and remove them through an upgrade of XSL stylesheets. There were also some parts of text that could only be annotated manually. Using such semi-automatic annotation, brief sessions could be marked up in 30 minutes, while those of medium length usually took up to two hours and the longest (over 200,000 words) up to four hours. Speeches were marked in accordance with the TEI module for performance texts (speech, speaker, stage direction). Other annotations included the structure of the assemblies and type of

---

29 Maarten Marx, "Advanced Information Access to Parliamentary Debates," *Texas Digital Library* 10, no. 6 (2009): 2, 3.

30 "Hansard archive (digitised debates from 1803)," *www.parliament.uk*, accessed on 30 September 2016, http://www.hansard-archive.parliament.uk/.

31 Maarten Marx and Anne Schuth, "DutchParl. The Parliamentary Documents in Dutch," in: *Proceedings of the International Conference on Language Resources and Evaluation*, eds. Nicoletta Calzolari et al. (Valetta: LREC, 2010), 3670–77.

32 Carlos Martin-Dancausa and Maarten Marx. "Parliamentary documents from Spain," in: *Proceedings of the International Conference on Language Resources and Evaluation*, eds. Nicoletta Calzolari et al. (Valetta: LREC, 2010).

33 Miloš Jakubíček and Vojtěch Kovář. CzechParl, "Corpus of Stenographic Protocols from Czech Parliament," in: *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010*, eds. Petr Sojka and Aleš Horák (Tribun EU, 2010), 41–46.

34 Maciej Ogrodniczuk, "The Polish Sejm Corpus," in: *LREC 2010, Eight International Conference on Language Resources and Evaluation*, eds. Nicoletta Calzolari et al. (Istanbul, 2012), 2219–23.

35 "SlovParl," *GitHub*, accessed on 30 September 2016, https://github.com/SIstory/SlovParl.

36 TEI Consortium, *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (Text Encoding Initiative Consortium, 2016).

37 For example, the German Research Community (Deutsche Forschungsgemeinschaft) demands that any texts being digitized be encoded using the TEI guidelines, if at all possible. Deutsche Forschungsgemeinschaft, *DFG Practical Guidelines on Digitisation* (Bonn: Deutsche Forschungsgemeinschaft, 2013), 31.

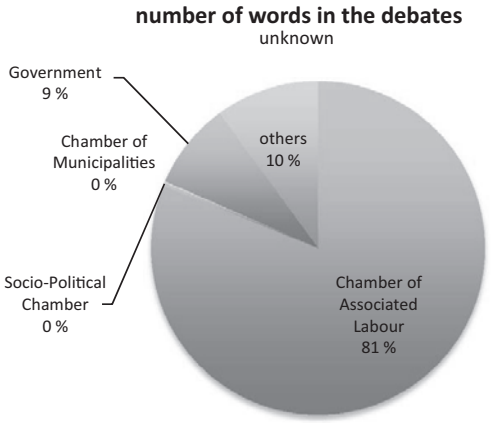
sessions, individual sessions, topics of individual sessions and dates and duration of sessions. Links were created to tables of contents and lists of speakers.[38]

Based on such annotated minutes of sessions, researchers can carry out various types of fundamental analyses.[39] Over 2.7 million words were thus spoken in 13,894 speeches at 54 sessions. At the longest, 36th session, which lasted eight days, the total duration of speeches was 29 hours and 3 minutes, and the session was adjourned no less than 21 times. On the other hand, the total duration of speeches at the briefest, 9th session was only one hour. The longest uninterrupted span was 460 minutes, while the average (median) duration of speeches between two interruptions was 90 minutes.

Connection to External Data

However, TEI documents annotated in such a manner also have some shortcomings that preclude precise analysis of the speeches. Initial analyses of the speeches of various speakers were carried out based on their first and last names. A single person, whose name might be written differently in other cases, is thus treated as two or more different people. On the other hand, people with identical first and last names are automatically considered the same person. Various historical records were thus used to manually verify and sanitize the lists of MPs, ministers and other invited speakers. These data are contained in a separate TEI document.

Chart 2: Number of words spoken in the Chamber of Associated Labour of the Assembly of the Republic of Slovenia (1990/92) by organization membership; in %



38 Andrej Pančur, "Označevanje zbirke zapisnikov sej slovenskega parlamenta s smernicami TEI," in: *Zbornik konference Jezikovne tehnologije in digitalna humanistika*, eds. Tomaž Erjavec and Darja Fišer (Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani, 2016), 142–48.

39 Jure Gašparič, "Slovenian Socialist Parliament on the Eve of the Dissolution of the Yugoslav Federation. A feeble "ratification body" or important political decision-maker?," *Prispevki za novejšo zgodovino* 55, no. 3 (2015): 54.

Data connected in such a manner can, for example, be used to determine the provenance of speakers in the Chamber of Associated Labour (see Chart 2). In doing so, we find that almost 20 % of the words were spoken by various representatives of the Government and other rapporteurs associated with the legislation that was being passed by the Assembly. As President and Vice-President of the Chamber of Associated Labour, Jože Zupančič (President, 735,166 words) and Bogo Rogina (Vice-President, 114,418 words) together spoke as much as 38.3 % of all words spoken by the MPs of the Chamber of Associated Labour. Among other MPs of the Assembly, Jože Arzenšek (106,000 words), Roman Jakič (69,111 words) and Andrej Šter (67,005) were the most verbose. Then there was the silent Jože Košak who only managed to say 14 words during his term.

Speeches of the MPs can obviously also be analysed according to the parties they belonged to (see Table 1). MPs who were, at the start of their term, members of the DEMOS coalition thus spoke 21.5 % of all words, opposition MPs spoke 23.3 %, while the numerous independent MPs (including the President of the Chamber) spoke as much as half of all words.

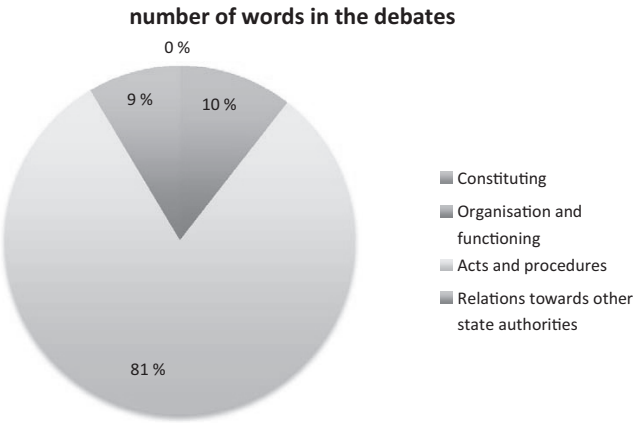
Table 1: Number of words spoken by members of political party; Chamber of Associated Labour of the Assembly of the Republic of Slovenia (1990/92)

Political parties	No. speekers	No. words	Percent	Percent
SDZ → DS	3	30762	1,4	
SDZ → NDS	2	85246	3,8	
SDSS	9	118899	5,4	
SKD	6	86135	3,9	21,5
SKZ → SLS	8	88439	4	
ZS	1	11901	0,5	
DEMOS	1	56858	2,6	
ZKS-SDP → SDP	17	231148	10,4	
ZSMS → LDS	9	236388	10,6	23,3
SZS → SSS	3	49708	2,2	
Independent	19	1109636	50	50
SOPS → Independent	1	114418	5,2	5,2
Unknown	1	736	0	0

Lists of MPs, members of the Government and other speakers can also be used to formulate research questions connected to additional variables: first and last name, gender, date and place of birth, date and place of death, education, profession, residence, organization membership.

Answers to complicated research questions are possible in part thanks to the newly created TEI file that includes a thematic index of the topics dealt with by the Chamber. The creation of this table of contents used data from existing topics and tables of contents, which were then annotated according to the new scheme. The

Chart 3: Thematic index of the speeches in the Chamber of Associated Labour of the Assembly of the Republic of Slovenia (1990/92); No. of words by topic as per the categorization in the Rules of Procedure of the National Assembly



data was first categorized in accordance with the Rules of Procedure of the National Assembly (see Chart 3).[40]

The largest category, *Acts and Procedures,* was classified in accordance with the thematic index of the Legal Information System of the Republic of Slovenia.[41] Based

Table 2: Main topic categories within Acts and Procedures as per the thematic index of the Legal Information System of the Republic of Slovenia

	Number of words	Percent
Constitutional regime of the Republic of Slovenia	497575	21,8
Foreign affairs and international relations	27417	1,2
Interior and administrative law	67156	2,9
Civil law	24017	1,1
Criminal law	5925	0,3
Economic order	403880	17,7
Public finance	482764	21,1
Economic activities	165752	7,3
Non-economic activities	550466	24,1
Environment an spatial planning	43803	1,9
Protection against natural and other disasters	14960	0,7

40 "Rules of Procedure of the National Assembly (PoDZ-1)," *Republic of Slovenia: National Assembly*, accessed on 30 September 2016, https://www.dz-rs.si/wps/portal/en/Home/ODrzavnemZboru/PristojnostiInFunkcije/RulesoftheProcedureText.

41 "Tematsko kazalo" [Thematic Index], *PIS: Pravno-informacijski sistem* [PIS: Legal Information System], accessed on 30 September 2016, http://www.pisrs.si/Pis.web/pravniRedRSDrzavniNivoKazalaTematskoKazalo.

Table 3: Longest discussions on individual topics (in the category of Acts and Procedures)

	No. of words	No. of speeches
Ownership Transformation of Companies Act	103870	520
The Law on Budget of the Republic of Slovenia for the year 1992	97175	511
Cooperatives Act	61616	379
The Law on Budget of the Republic of Slovenia for the year 1991	52121	314
Military Service Act	51112	365
Pension and Disability Insurance Act	47801	261
The Constitution of the Republic of Slovenia	43004	267
Sales Tax Act	37795	192
Health Services Act	37238	201

on data annotated in such a manner, it is easily determined that over a fifth of all speeches were associated with the legislation pertaining to the constitutional arrangements in the Republic of Slovenia (see Table 2).

The greatest amount of discussion was stirred up by the Ownership Transformation of Companies Act, which resulted in 520 speeches containing over 100,000 words (see Table 3). On the other hand, debate regarding the Act Ratifying the Agreement between the Government of the Republic of Slovenia and the Federal Council of the Swiss Confederation on the Abolishment of Visas was extremely brief, consisting of only 46 words.

Data Enrichment: Natural Language Processing (NLP)

Another extremely extensive category was the category named *Initiatives, Suggestions and Questions from the MPs* (102,858 words). However, the category is too broad to allow any conclusions about its actual content based solely on the title. Natural language processing technologies can be of some help in this regard. For example, the topic modelling method can be used to search for word patterns in the text, which can in turn assist with determining the semantic meaning of various parts of the text. One of the most popular (among historians as well as other researchers)[42] tools used in such analyses is MALLET.[43] Although the results were incomplete, MALLET was nonetheless successfully used to discern a number of topics within the *Initiatives, suggestions and questions* category: customs duties, healthcare, the environment, strikes, banks, etc.

A tool that searched the text for named entities (people, places and organizations)

42 Shawn Graham, Scott Weingart and Ian Milligan, "Getting Started with Topic Modeling and MALLET," *Programming Historian* (2 September 2012).

43 Andrew McCallum, *MALLET*: *A Machine Learning for Language Toolkit* (2002).

Table 4: List of the 30 most common place names in speeches in the Chamber of Associated Labour of the Assembly of the Republic of Slovenia (1990/92) as identified by the Stanford NER for the Slovenian language

place name	No.	Percent
Slovenija	2978	26,2
Republika Slovenija	2153	18,9
Jugoslavija	579	5,1
Evropa	534	4,7
Ljubljana	270	2,4
Maribor	244	2,1
Hrvaška	241	2,1
Nemčija	182	1,6
Italija	170	1,5
Avstrija	167	1,5
Poročevalca	132	1,2
Celje	103	0,9
Skupščine	94	0,8
Beograd	91	0,8
Irak	89	0,8
Komisija	82	0,7
Srbija	82	0,7
Koper	55	0,5
Lendava	53	0,5
Ptuj	51	0,4
Republika Hrvaška	51	0,4
Demos	49	0,4
Kranj	42	0,4
Logatca	38	0,3
Republiškem	38	0,3
Francija	37	0,3
Švica	36	0,3
Madžarska	34	0,3
Združene države Amerike	33	0,3
Piran	31	0,3
other named entities	2639	23,2
	11378	100,0

yielded much better results. Named entity recognition was carried out using the Stanford NER for the Slovenian language.[44] It should come as no surprise that those

[44] Nikola Ljubešić et al., "Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene," *Slovenščina 2.0* 1, no. 2 (2013): 35–57.

who were most often named in the speeches of MPs were the other MPs. It is also unsurprising that the place name used most often by speakers in the Slovenian parliament was Slovenia. Table 4 thus shows that the Stanford NER for the Slovenian language recognized 11,378 place names in parliamentary speeches, 45 % of which were identifiable as Slovenia or the Republic of Slovenia. However, a detailed look at the table quickly reveals that the place names identified by Stanford NER for the Slovenian language also included names of organizations (Assembly, Commission, DEMOS) and other names (Poročevalec).

These results clearly indicate that history researchers should never simply copy the results of natural language processing (NLP) technologies. At its current level of development, the technology is far from infallible. For example, the Stanford NER for the Slovenian language has 85 % precision when annotating persons, while its precision is below 80 % when it comes to places.[45] The following warning must thus truly be taken to heart: "Historians need to be aware that, in addition to verifying reliability of sources as is common in their field, they also need to take the reliability of NLP methods into account when working with automatically extracted information."[46]

The authors of this paper are certain that the most effective way to achieve this would be through close collaboration with computational linguists. The existing TEI documents that had been encoded in accordance with the TEI module used for performance texts were thus subsequently converted into TEI documents wherein the text was annotated in accordance with the speech transcription TEI module. Such documents can then be furnished with linguistic annotations at a later time. Computational linguists have thus already provided part-of-speech tagging for the text of the speeches. The corpus has been imported into the No Sketch Engine concordance base[47] and all TEI documents are accessible at the CLARIN.SI repository.[48]

---

45 Ljubešić, "Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene," 48.

46 Antske Fokkens et al., "BiographyNet: Methodological issues when NLP supports historical research," in: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, eds. Nicoletta Calzolari et al. (Reykjavik: European Language Resources Association (ELRA), 2014), 3734.

47 "SlovParl (parlament RS 1990–1992)," *NoSketch Engine*, accessed on 30 September 2016, http://nl.ijs.si/noske/all.cgi/corp_info?corpname=slovparl. Tomaž Erjavec, "Korpusi in konkordančniki na strežniku nl.ijs.si," *Slovenščina 2.0* 1, no. 1 (2013): 24–49.

48 Tomaž Erjavec, Jan Jona Javoršek and Simon Krek, "Raziskovalna infrastruktura CLARIN.SI," in: *Proceedings of the 17th International Multiconference Information Society – IS 2014: Language Technologies*, eds. Tomaž Erjavec and Jerneja Žganec Gros (Ljubljana: IJS, 2014), 19–24. Andrej Pančur, Mojca Šorn and Tomaž Erjavec, *Slovenian parliamentary corpus SlovParl 1.0* (2016), distributed by Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1075.

Conclusion

This article has attempted to show, based on a number of fundamental analyses of the minutes of the sessions of the Chamber of Associated Labour of the Assembly of the Republic of Slovenia (1990/92), that future historians will be unable to process the increasing amounts of digital materials using standard historiographical methods and will be forced to start supplementing these with methods and tools developed by digital humanities. In doing so, most historians will still rely on various tools developed by digital historians in order to simplify the work of their colleagues who are unfamiliar with digital humanities. At the same time, digital historians will also have to be aware of the limitations of the use of tools developed by other fields, including language technologies. On the other hand, historians who have started learning the basics of programming languages are increasingly establishing themselves in the field.[49] The reason for this is that this is the only way to extract additional useful research results from existing digital sources. This is also the manner in which most of the analyses presented by this paper were carried out. At the same time, the authors of this article are well aware that in addition to new knowledge supplied by digital history, it remains indispensable for researchers to be very familiar with research domain. It can thus be anticipated that future research of contemporary history will take place through fruitful collaboration of experts from a multitude of different fields.

Sources and Literature

Dataset:

- Pančur, Andrej, Mojca Šorn and Tomaž Erjavec. *Slovenian parliamentary corpus SlovParl 1.0* (2016). Distributed by Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1075.

Literature:

- Benardou, Agiatis, Alastair Dunning, Martin Schaller and Nephelie Chatzidiakou. *Research Themes for Aggregating Digital Content. Parliamentary Papers in Europe*. Europeana Cloud, 2015. Accessed 28 September 2016. http://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/Europeana_Cloud/WP1%20Research%20Needs/research-themes-for-aggregating-digital-content-parliamentary-papers.pdf.
- Collins, Sandra, Natalie Harrower, Dag Trygve Truslew Haug, Beat Immenhauser, Gerhard Lauer, Tito Orlandi, Laurent Romary and Eveline Wandl-Vogt. *ALLEA E-Humanities Working Group Report. Going Digital: Creating Change in the Humanities*. Berlin: All European Academies, 2015. Accessed 19 July 2016. http://www.allea.org/Content/ALLEA/WG%20E%20Humanities/Going%20Digital_digital%20version.pdf.
- Cvelfar, Bojan, Tatjana Hajtnik, Miroslav Novak, Nada Čibej and Drago Trpin. *Strategija in izvedbeni načrt razvoja slovenskega elektronskega arhiva 2016 – 2020.* Ljubljana: Arhiv Republike Slovenije, 2016.
- Deutsche Forschungsgemeinschaft. *DFG Practical Guidelines on Digitisation.* Bonn:

49 Nanni et al., "Semi-Supervised Textual Analysis and Historical Research Helping Each Other," 63–68. Graham et al., *Exploring Big Historical Data: The Historian's Macroscope*.

Deutsche Forschungsgemeinschaft, 2013. Accessed 19 July 2016. http://www.dfg.de/formulare/12_151/12_151_en.pdf.

- Elliot, Devon, Robert MacDougall and William J. Turkel. "New Old Things: Fabrication, Physical Computing, and Experiment in Historical Practice." *Canadian Journal of Communication* 37 (2012): 122. Accessed 19 July 2016. http://www.cjc-online.ca/index.php/journal/article/view/2506.
- Erjavec, Tomaž, Jan Jona Javoršek and Simon Krek. "Raziskovalna infrastruktura CLARIN.SI." In: *Proceedings of the 17th International Multiconference Information Society – IS 2014: Language Technologies*, edited by Tomaž Erjavec and Jerneja Žganec Gros, 19–24. Ljubljana: IJS, 2014. Accessed 30 September 2016. http://nl.ijs.si/isjt14/proceedings/isjt2014_03.pdf.
- Erjavec, Tomaž. "Korpusi in konkordančniki na strežniku nl.ijs.si." *Slovenščina 2.0* 1, no. 1 (2013): 24-49. Accessed 19 July 2016. http://slovenscina2.0.trojina.si/arhiv/2013/1/Slo2.0_2013_1_03.pdf.
- Fokkens, Antske, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne and Guus Schreiber. "BiographyNet: Methodological issues when NLP supports historical research." In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis, 3728–35. Reykjavik: European Language Resources Association (ELRA), 2014. Accessed 30 September 2016. http://www.lrec-conf.org/proceedings/lrec2014/pdf/1103_Paper.pdf.
- Gašparič, Jure. "Slovenian Socialist Parliament on the Eve of the Dissolution of the Yugoslav Federation. A feeble "ratification body" or important political decision-maker?." *Prispevki za novejšo zgodovino* 55, no. 3 (2015): 41–59. Accessed 19 July 2016. http://ojs.inz.si/pnz/article/view/123.
- Gašparič, Jure. "Pisati politično zgodovino Republike Slovenije." In: Četrt *stoletja Republike Slovenije – izzivi, dileme, pričakovanja*, edited by Jure Gašparič and Mojca Šorn, 27–37. Ljubljana: Institute of Contemporary History Institute of Contemporary History, 2016.
- Gašparič, Jure. *Državni zbor 1992 – 2012. O slovenskem parlamentarizmu*. Ljubljana: Institute of Contemporary History, 2012.
- Gašparič, Jure. *Slovenski parlament: politično-zgodovinski pregled od začetka prvega do konca šestega mandata (1992-2014)*. Ljubljana: Institute of Contemporary History, 2014. Accessed 19 July 2016. http://hdl.handle.net/11686/26950.
- Graham, Shawn, Ian Milligan and Scott Weingart. *Exploring Big Historical Data: The Historian's Macroscope*. London: Imperial College Press, 2015. Accessed 28 September 2016. http://www.themacroscope.org.
- Graham, Shawn, Scott Weingart and Ian Milligan. "Getting Started with Topic Modeling and MALLET." *Programming Historian* (2 September 2012). Accessed 19 July 2016. http://programminghistorian.org/lessons/topic-modeling-and-mallet.
- Haber, Peter. "Zeitgeschichte und Digital Humanities." In: *Zeitgeschichte – Konzepte und Methoden*, edited by Frank Bösch and Danyel Jürgen, 47–66. Göttingen: Vandenhoeck & Ruprecht, 2012. Accessed 19 July 2016. http://dx.doi.org/10.14765/zzf.dok.2.269.v1.
- Hajtnik, Tatjana. "Strategija razvoja slovenskega javnega elektronskega arhiva 'e-ARH.si'." *Knjižnica* 55, no. 1 (2011): 39–56. http://revija-knjiznica.zbds-zveza.si/Izvodi/K1101/Hajtnik.pdf
- Jakubíček, Miloš and Vojtěch Kovář. CzechParl. "Corpus of Stenographic Protocols from Czech Parliament." In: *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010*, edited by Petr Sojka and Aleš Horák, 41–46. Tribun EU, 2010. Accessed 19 July 2016. http://www.muni.cz/research/publications/914313.
- Ljubešić, Nikola, Marija Stupar, Tereza Jurić and Željko Agić. "Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene." *Slovenščina 2.0* 1, no. 2 (2013): 35–57.
- Martin-Dancausa, Carlos and Maarten Marx. "Parliamentary documents from Spain." In: *Proceedings of the International Conference on Language Resources and Evaluation*, edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias. Valetta: LREC, 2010. Accessed 19 July 2016. https://www.researchgate.net/publication/239585911.

- Marx, Maarten and Anne Schuth. "DutchParl. The Parliamentary Documents in Dutch." In: *Proceedings of the International Conference on Language Resources and Evaluation*, edited by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias, 3670–77. Valetta: LREC, 2010. Accessed 19 July 2016. http://www.lrec-conf.org/proceedings/lrec2010/pdf/263_Paper.pdf.
- Marx, Maarten. "Advanced Information Access to Parliamentary Debates." *Texas Digital Library* 10, no. 6 (2009): 1–11. Accessed 19 July 2016. https://journals.tdl.org/jodi/index.php/jodi/article/view/668.
- McCallum, Andrew Kachites. *MALLET*: *A Machine Learning for Language Toolkit*. 2002. Accessed 19 July 2016. http://mallet.cs.umass.edu/.
- Nanni, Federico, Hiram Kumper and Simone Paolo Ponzetto. "Semi-Supervised Textual Analysis and Historical Research Helping Each Other: Some Thoughts and Observations." *International Journal of Humanities and Arts Computing* 10, no. 1 (2016): 63–77. Accessed 19 July 2016. http://dx.doi.org/10.3366/ijhac.2016.0160.
- Nicholson, Bob. "The Digital Turn: Exploring the methodological possibilities of digital newspaper archives." *Media History* 19, no. 1 (2013): 59–73. Accessed 30 September 2016. http://dx.doi.org/10.1080/13688804.2012.752963.
- Ogrodniczuk, Maciej. "The Polish Sejm Corpus." In: *LREC 2010, Eight International Conference on Language Resources and Evaluation*, edited by Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk and Stelios Piperidis, 2219–23. Istanbul, 2012. Accessed 19 July 2016. http://www.lrec-conf.org/proceedings/lrec2012/pdf/653_Paper.pdf.
- Pančur, Andrej. "Označevanje zbirke zapisnikov sej slovenskega parlamenta s smernicami TEI." [Encoding the Slovenian Parliament Session Minutes in Line with the TEI Guidelines] In: *Zbornik konference Jezikovne tehnologije in digitalna humanistika* [Proceedings of the Conference on Language Technologies & Digital Humanities], edited by Tomaž Erjavec and Darja Fišer, 142–48. Ljubljana: Znanstvena založba Filozofske fakultete v Ljubljani, 2016. Accessed 5 October 2016. http://nl.ijs.si/isjt16/proceedings-en.html.
- Pesek, Rosvita. *Osamosvojitev Slovenije*. Ljubljana: Nova revija, 2007.
- Piersma, Hinke and Kees Ribbens. "Digital Historical Research: Context, Concept and the Need for Reflection." *BMGN – Low Countries Historical Review* 128, no. 4 (2013): 78–102.
- Piersma, Hinke, Ismee Tames, Lars Buitinck and Maarten Marx. "War in Parliament. What a Digital Approach Can Add to the Study of Parliamentary History." *DHQ: Digital Humanities Quarterly* 8, no. 1 (2014). Accessed 30 September 2016. http://www.digitalhumanities.org/dhq/vol/8/1/000176/000176.html.
- Robertson, Stephen. "The Differences between Digital Humanities and Digital History." In: *Debates in Digital Humanities 2016*, edited by Matthew K. Gold and Lauren F. Klein. Minneapolis and London: University of Minnesota Press, 2016. Accessed 25 September 2016. http://dhdebates.gc.cuny.edu/debates/text/76.
- Rosenzwig, Roy. "Scarcity or Abundance? Preserving the Past in a Digital Era." *American Historical Review* 108, no. 3 (2003): 735–62.
- Schöch, Christof. "Big? Smart? Clean? Messy? Data in the Humanities." *Journal of Digital Humanities* 2, no. 3 (2013). Accessed 25 September 2016. http://journalofdigitalhumanities.org/2-3/big-smart-clean-messy-data-in-the-humanities/.
- Spiro, Lisa. "Access, Explore, Converse. The Impact (and Potential Impact) of the Digital Humanities on Scholarship." In: *Keys for architectural history research in the digital era*, edited by Juliette Hueber and Antonio Mendes da Silva. 2014. Accessed 25 September 2016. https://inha.revues.org/4925.
- TEI Consortium. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, Text Encoding Initiative Consortium, 2016. Accessed 19 July 2016. http://www.tei-c.org/Guidelines/P5/.
- Zaagsma, Gerben. "On Digital History." *BMGN – Low Countries Historical Review* 128, no. 4 (2013): 3–29. http://www.bmgn-lchr.nl/articles/10.18352/bmgn-lchr.9344/.

Andrej Pančur, Mojca Šorn

PAMETNI MASIVNI PODATKI: UPORABA SLOVENSKIH PARLAMENTARNIH DOKUMENTOV V DIGITALNI ZGODOVINI

P O V Z E T E K

Avtorja v uvodu opozoriva na dejstvo, da velika količina računalniško berljivih podatkov postaja stvarnost in neizogibno dejstvo tudi v zgodovinopisju, pri čemer poudariva, da je prav ta fenomen spodbudil nastanek in uveljavitev digitalne humanistike, ki s pomočjo digitalnih metod in orodij odgovarja na nova raziskovalna vprašanja.

Ker so parlamentarni dokumenti bogat vir podatkov, ki ga uporabljajo različne discipline v humanistiki, med drugim tudi v zgodovinopisju, avtorja v nadaljevanju predstaviva primer uporabe parlamentarnih debat v digitalni zgodovini. Predstavljeni primeri so ilustrativni vzorci, s pomočjo katerih hočeva prikazati ogromen potencial, ki ga lahko imajo pametni množični podatki v raziskavah sodobne zgodovine. Vzorčno sva v XML formatu označila 2,7 milijona besed parlamentarnih debat v Zboru združenega dela Skupščine Republike Slovenije v letih 1990–1992. Pri tem sva se odločila uporabiti Smernice Text Encoding Initiative (TEI), ki so v digitalni humanistiki *de facto* standard za kodiranje tekstovnih besedil. Avtomatske pretvorbe sva izvajala s pomočjo XSLT stilov, napisanih posebej za ta projekt. Ker avtomatske pretvorbe lahko vsebujejo tudi napačne označbe, je označevanje potekalo delno tudi ročno. S tem delno avtomatskim označevanjem sva krajše seje lahko označila v pol ure, za daljše seje sva po navadi porabila do dve uri, za najdaljše (več kot 200000 besed) pa do štiri ure. Govore sva označila v skladu s TEI modulom za dramska besedila (govor, govorec, didaskalija). Označila sva še strukturo zborov in vrste sej, posameznih sej, vsebinskih sklopov posameznih sej, datumov in časovnega poteka sej. Naredila sva povezave na kazala vsebine in sezname govorcev.

Na podlagi tako označenih zapisnikov sej lahko raziskovalci naredijo različne vrste osnovnih analiz. V 13894 govorih je bilo na 54 sejah tako skupaj izgovorjenih več kot 2,7 milijonov besed. Na najdaljši – 36. seji – so v osmih dnevih skupaj govorili 29 ur in 3 minute, pri čemer so sejo kar enaindvajsetkrat prekinili. Na najkrajši – 9. seji – pa so nasprotno skupaj govorili samo eno uro. Največ so neprekinjeno govorili 460 minut, v povprečju (mediana) pa so neprekinjeno govorili uro in pol.

Osnovne analize zapisnikov torej pokažejo, da v prihodnosti zgodovinarji vedno večjih količin digitalnega gradiva ne bodo več mogli obdelovati samo z uporabo klasičnih zgodovinskih metod, temveč bodo morali začeti uporabljati metode in orodja, ki jih razvija digitalna humanistika. Večina zgodovinarjev se bo pri tem (še naprej) zanašala na različna orodja, ki jih razvijajo digitalni zgodovinarji z namenom, da olajšajo delo svojim kolegom, ki se ne ukvarjajo z digitalno humanistiko. Pri tem pa bodo digitalni zgodovinarji morali poznati omejitve, ki jih prinaša uporaba orodij, razvitih v okrilju drugih disciplin, med drugim tudi jezikovne tehnologije. Avtorja prispevka se dobro zavedava, da je za kakovostne analize poleg novih znanj, ki jih prinaša digitalna zgodovina, neobhodno potrebno tudi temeljito poznavanje raziskovalne domene. Zato lahko predvidimo, da bo raziskovanje sodobne zgodovine potekalo v znamenju plodnega sodelovanja strokovnjakov iz različnih področij.