Turning the Carniolan regional assembly proceedings into an enriched historical corpus

Marolt, Matija

matija.marolt@fri.uni-lj.si University of Ljubljana, Slovenia

Gašparič, Jure

Jure.Gasparic@inz.si Institute of Contemporary History, Slovenia

Mundjar, Aleksander

am9973@student.uni-lj.si University of Ljubljana, Slovenia

Kavčič, Alenka

alenka.kavcic@fri.uni-lj.si University of Ljubljana, Slovenia

Fišer, Darja

Darja.Fiser@inz.si Institute of Contemporary History, Slovenia

Pančur, Andrej

andrej.pancur@inz.si Institute of Contemporary History, Slovenia

Introduction

Parliamentary proceedings reflect the political, societal, and cultural Zeitgeist of a certain period and are as such an invaluable source for a broad range of DH research questions (Blaxhill and Beelen 2016, Blätte et al. 2020, Müller-Hansen et al. 2021). While contemporary parliamentary corpora have been compiled, annotated and made available for a number of languages through projects such as ParlaMint (Erjavec et al., 2022), similar resources with historical data remain rare (Puren et al., 2022). They are essential, however, as in the past two decades the focus of research on parliamentary discussions has shifted from the history of politics to the history of the political. Conceptual history (Begriffsgeschichte) has become increasingly relevant for the understanding of the society and its processes: how certain political concepts have been introduced, used, and understood (Koselleck, 2006). Such in-depth historical analysis, connected with the speech act theory, is of major importance for political history. In this framework, language is not just a tool used for describing the world, but also a significant part of the world that it co-creates (Hampsher-Monk, 1998).

In this abstract, we present the development of a richly annotated corpus of historical proceedings of the Krainer Landtag (The Carniolan Regional Assembly), which was the highest legislative body of the autonomy of the Carniola Herzogtum of the Habsburg Empire. It was introduced with the February patent (February 26, 1861) when the Austrian part of the Empire entered the second constitutional period and ended with the onset of the first world war, comprising altogether 12 parliamentary terms. The Assembly, which was initially composed of 37 delegates and was in 1908 extended to 50 delegates, adopted laws that were within its jurisdiction, which included agriculture, public buildings, charitable institutions, municipal, church and school matters, matters related to the provision of harness, provision and accommodation of the military, the regional budget, and until 1873 appointed delegates to the central parliament in Vienna (Reichsrat).

Procedure

We collected the scanned and OCR processed pdf documents from the The Digital Library of Slovenia - dLib.si. The documents span the period from 1861 to 1913 (903 pdfs, 42746 pages) and include the assembly meeting proceedings (694 pdfs, 15353 pages), as well as supplementary materials (laws, budgets etc.). The documents are bilingual, in Slovenian and German, depending on the speaker. German was first typeset in the Gothic script and later on in Latin.

We first separated the meeting notes from the supplementary materials, as the latter often contain complex layouts (tables etc.), which will be addressed in our future work. We assessed the quality of OCR and found it to be acceptable on most documents (around 2% character error rate, a bit worse for German written in Gothic). Although there were some outlier pages where OCR performed poorly, especially when characters from neighbouring pages were visible in the scans, these pages were also included for further processing, but we plan to automatically detect such cases and perform OCR on them again with appropriate preprocessing.

On a subset of debates spanning 15 years, we devised a set of rules coded in a Python script to detect layout, find headings and text segments, deal with hyphenated words, assign text to speakers, detect events and extract meeting metadata including the date and hour of the beginning/end of each meeting, its agenda and a list of attendants with their titles. We used the Lingua Python library for language detection at sentence level. For lemmatization and part-of-speech tagging, we used Trankit¹ for both languages. We first tried the CLASSLA Slovenian fork ² for lemmatization of Slovenian words and assessed PoS tagging accuracy of both lemmatizers. On a small sample of 3400 words (8 pages of text), CLASSLA performed only 0.5% better than Trankit. The lemmatization F1 scores were at 94.4% and 93.9%, respectively. The results are around 5% worse than for modern-day Slovenian, as the used Slovenian language is quite different (more archaic) to modern-day Slovenian on which the NLP tools were trained.

After fine-tuning the set of rules, we applied them to the entire corpus and made adjustments to handle the changes in document formatting that occurred during the years. Altogether, the processed documents contain over 44.000 segments, over 540.000 sentences (roughly 58% in Slovenian and 42% in German) and approximately 10M words. Earlier debates were predominantly in German, while the language shifted towards Slovenian in the latest years.

We output the documents in the modified TEI XML format according to the schema compliant with Parla-CLARIN (Erjavec and Pančur, 2022). The developed corpus will enable us to identify types of changes that occur in political language in a certain political, national and partisan context (Ihalainen et al. 2016) and so provide new and innovative source-based fundamentals to re-think the political life in the multinational Habsburg empire. When completed, the entire corpus will be made available in the CLARIN.SI repository and importantly contribute to the availability of comparable, interoperable collections of parliamentary debates for diachronic, multilingual and transnational research.

	untimation an Antalas bintiditid has biestatte an art.	Last to bills and some build the state of the state of	Falla
	entsprechenden Antrage binfichtlich bes biesfalls zu erlaf- fenden Laudesgefeses und bes eventuell zu gewährenden Landesbeitrages zu ben Koften bes fraglichen Straffen-	mi je bilo od gosp. ministra pravoslovja obljubljeno ustmeno, da bodo te prošnje v deželni zbor prišle o pravim času, da bode deželni zbor odločevati	& E1
	baues in der naften Seffion einzubringen".	mogel, ko so tudi prišle iz Kostanjevice in Senožeč enake prošnje v deželni zbor. Ker smo pa o teh dveh	Er
	Landeshauptmann : Bunicht Jemand ber herren bas Bort zu biefem	že določno sklenili, mislim samo omeniti, da zarad teh dveh sklepov se nimajo tudi te prošnje še enkrat za-	Sche
	Antrage ? (Rach einer Paufe - po prestanku.) Benn nicht, fo schreiten wir zur Abstimmung und	vračati na deželni odbor, tem več, da jih predloži mi- nisterstvu, tako, da je ta reč rešena in nima već v	rin-e
	ich bitte jene gerren, welche mit Diefem Antrage einver- ftanden find, figen ju bleiben. (niemand erhebt fich -	odbor priti.	Ha
	nobeden ne vstane.) Der Ausschußantrag ift fomit vom hohen gandtage	Landeshauptmann : Bunicht noch Jemand von ben herren das Wort?	
	angenommen.	(Nach einer Baufe, po prestanku) Benn nicht, fo schreiten wir zur Abstimmung, und	cepti
	Run fommt ber Bericht bes Landesausichuffes uber eine Bufchrift bes Juftigminifteriums die Gerichts-	ich bitte jene herren, welche mit dem Ausichugantrage einverftanden find, fiben zu bleiben. (niemand erhebt fich,	tives
	fprengel betreffend. Poročevalec dr. Costa :	Der Ausschutzantrag ift vom h. Haufe genehmiget.	Ih
	Slavni zbor! Še le 29. septembra t. l. je došlo	Die Tagesordnung ift erichopft und zugleich die	and
	deželnemu odboru sledeče pismo naddeželne sodnije v Gradcu: (bere – licšt)	lette Tagesordnung erledigt. In Folge ber Erledigung aller unferer Beschäfte,	cept.
	"In Gemäßheit des herabgelangten hohen Juftig- minifterial-Erlaffes vom 15. September 1868, 3. 14533, hat das f. f. Oberlandesgericht die Ehre bem löblichen	und in Folge A. h. Befehles ift diese Landtags-Seffion zu Ende. Meine Herren! nur ein Baar Worte ! (Die Ber-	Ko
	f. f. Canbesausschuffe bie aus Anlag ber letten Juftig- organistrung vorgefommenen Eingaben, welche eine Aen-	fammlung erhöft fich, zbor se vzdigne) Das Berf ift vollbracht! Groß und fcwierig waren	tik u
	berung ber bestehenden Begirtseintheilung anftreben, in ./. und bie beim hohen Juftigminifterium vorgefommenen	bie Aufgaben, die Gie ju fofen hatten, mit Rudficht auf bie furge Beit, bie uns ju Gebote ftand.	am I
	Betitionen in 1/2 gur gefällig weitern Beranlaffung im Ginne bes §. 2 bes Gefeges vom 11. Juni 1868 R. G. B.	Gie fehren mit bem Bewußtfein vollfommen erfullter Bflicht in 3hre Heimat gurud.	Μ
	3. 59 einzuschließen. Grag am 22. September 1868".	3ch, für meine Perfon, dante Ihnen für die mir bewiesene Rachficht in ber Leitung der Berhandlungen bes	rand
	Kakor je očividno, deželnemu odboru ni bilo mogoče, izdelati natančno sporočilo, kako in ktere pre-	Landtages, ich banke Ihnen insbesonbers und aus vollem gergen für bas mir oft und insbesondere jungt bethätigte	coal
ļ	membe o sodnijskih okrajih se bi koj zdaj nasvetovale:	Wohlwollen.	coal
			ence
			Pu
			tract
	is type"spine"/bedvelie de (cotar (doto) guilden (Listen (Listen)) on all Liste (Listen) (Listen) (Listen) (Listen) on all Liste (Listen) (List		enric
	ses wall.lege"11" או 11:46"-100, 00, 00, 00, 00, 00, 00, 00, 00, 00,		Cont
	<pre>cw xmliid="02X_1868-10-03_09_23_sept2.s2.v4" lemma="september" msd="UPosTag=Case=Gim[Gender=Hass[Humber=Sing" ana="mterHonsg">septembrac/wo cw xmliid="02X_1868-10-03_09_23_sept2.s2.v4" lemma="t," msd="UPosTag=Case=Gim[Gender=Hass")+t.c/wo</pre>		

Figure 1: A short excerpt from the proceedings dated 3.10.1868 in PDF and TEI format.

Acknowledgements

The work was supported by the Slovenian Research Agency research programmes P6-0436:

Digital Humanities: resources, tools and methods (2022–2027) and P6-0281: Political History (2022–2027) and the DARIAH-SI research infrastructure (2022-2027).

Notes

1. https://github.com/nlp-uoregon/trankit

2. https://github.com/clarinsi/classla

Bibliography

Blaxill, L., & Beelen, K. (2016). A Feminized Language of Democracy? The Representation of Women at Westminster since 1945. Twentieth Century British History, 27(3), pp. 412–449.

Blätte, A., Gehlhar, S., & Leonhardt, C. (2020). The Europeanization of Parliamentary Debates on Migration in Austria, France, Germany, and the Netherlands. 66–74.

Erjavec, T., Ogrodniczuk, M., Osenova, P. et al. (2022). The ParlaMint corpora of parliamentary proceedings. Lang Resources & Evaluation, https://doi.org/10.1007/s10579-021-09574-0

Erjavec, T. and Pančur A. (2022). Parla-CLARIN: A TEI Schema for Corpora of Parliamentary Proceedings https://clarin-eric.github.io/parla-clarin/ (accessed 12 October 2022).

Hampsher-Monk, J. (1998). "Speech Act, Languages or Conceptual History?" In History of Concepts, Comparative Perspectives, 37–50. Amsterdam: Amsterdam University Press.

Ihalainen, P., Ilie, C. and Palonen, K., eds. (2016). Parliament and Parliamentarism: A Comparative History of a European Concept. New York: Berghahn Books.

Koselleck, R. (2006). Begriffsgeschichten: Studien zur Semantik und Pragmatik der politischen und sozialen Sprache. Frankfurt am Main: Suhrkamp Verlag.

Müller-Hansen, F., Callaghan, M. W., Lee, Y. T., Leipprand, A., Flachsland, C., & Minx, J. C. (2021). Who cares about coal? Analyzing 70 years of German parliamentary debates on coal with dynamic topic modeling. Energy Research & Social Science, 72, 101869.

Puren, M.A., Vernus, P., Pellet, A., Bourgeois, N. (2022). Extracting and providing online access to annotated and semantically enriched historical data. The AGODA project. Digital Humanities Conference 2022 Book of abstracts.