implies a better insight into users' needs, the social relevance of research and the democratization of science in general.

The presentation will report the workflow of the project based on the cooperation of academia (university teachers and students), citizens (retired sports journalists), and heritage experts in documentation, research, and communication of knowledge related to the collection of more than 16.500 digitised negatives from the Croatian Sports Museum (the photographs are documenting local and international sports events in Croatia and Yugoslavia from 1970 to 2000).

The first part of the project covered digitising photographs (negatives) and structuring photo series according to the principle of the original order with their multilevel descriptions. All that a retired sports photographer carried out. Then, an information architect and developer created the initial functional prototype of the annotation application. The following steps were prototype testing, initial assessment and participatory formative evaluation. Considering the results, researchers, students and developers improved the application and design methodology for future project activities. Sports journalists were assigned tasks such as tagging, annotating and evaluating photos, and adding any contextual information they were familiar with. The collected information was verified and appraised by a collection custodian. Information experts were further working on semantic approaches to automatic recognition of subject tags in their semi-structured descriptions, linking and enriching data. As a final result, the museum provides open access to the photos through its digital collections catalogue and the national digitisation portal eKultura.

Creating a Corpus of the National Representation of the First Yugoslavia

Alenka Kavčič, Matija Marolt, Andrej Pančur, Darja Fišer

Parliamentary debates have been an important source for political and comparative historians for several decades. They are also relevant for scholars in the fields of social, economic and religious history, history of law, political science, sociology and linguistics. While corpora of contemporary parliamentary proceedings are readily available for several national and regional parliaments (e.g., ParlaMint (Erjavec et al. 2022)), historical parliamentary records are still mostly used in the form of PDF documents or unstructured plain textual documents that are difficult to query and exploit at scale. The availability of linguistically processed, metadata-rich and structurally encoded historical parliamentary corpora (e.g., Hansard[1] (Coole et al. 2020), AGODA (Puren et al. 2022)) opens up new research opportunities and methodological advances across a wide range of disciplines, as well as engagement of the public interested in cultural heritage.

This paper describes the OCR processing, annotation and encoding of the meeting proceedings of the National Representation of the First Yugoslavia from 1919 to 1939. The corpus includes 714 sessions of the Temporary National Representation of the Kingdom of Serbs, Croats and Slovenes (in 1919 and 1920), the Legislative Committee of the National Assembly of the Kingdom of Serbs, Croats and Slovenes (in 1921 and 1922), and the National Representation (National Assembly and Senate) of the Kingdom of Yugoslavia (from 1931 to 1939). The documents comprise 15403 pages and over 13 million words.

Our source data were scanned images (TIFF images @300 dpi) of printed stenographic minutes from the History of Slovenia – SIstory[2] portal. The documents were multilingual, in Serbo-Croatian and Slovenian, depending on the speaker. Serbo-Croatian was typeset in Cyrillic (Serbian) and Latin (Croatian) script. The images were OCR-processed using ABBYY FineReader,[3] and the resulting docx and txt files were automatically processed using rule-based Python scripts to extract metadata such as

1 https://hansard.parliament.uk/.

2 https://www.sistory.si.

3 https://pdf.abbyy.com/.

title, participants, agenda, beginning and end of the meeting, speakers, and events. The OCR quality was generally acceptable, although some errors and inconsistencies were found, especially in the texts written in Cyrillic script.

Once the text of each meeting and its metadata were extracted, Lingua[4] was used for language detection on the sentence level. About 59% of the sentences were in Serbian (Cyrillic script), 38% in Croatian (Latin script) and 3% in Slovenian, while some sentences in German and French were also detected. The sentences in Serbian, Croatian and Slovenian were linguistically annotated using CLASSLA[5] (Ljubešić and Dobrovoljc, 2019) for tokenisation, MSD tagging and lemmatisation.

In the resulting documents, each session was stored in a file and encoded in Parla-CLARIN[6] (Erjavec and Pančur, 2021) compliant TEI XML format. The corpus is accessible in the CLARIN.SI repository and can be freely downloaded together with the corresponding pdf facsimile (Kavčič et al. 2023a). The availability of this corpus contributes significantly to a collection of comparable parliamentary debates ranging from the Carniolan Provincial Assembly (Kavčič et al. 2023b) in 1861 to the present-day Slovenian Parliament (Pančur et al. 2022).

References

Erjavec, T., et al. (2022). "The ParlaMint corpora of parliamentary proceedings." *Language Resources and Evaluation*, 57(1), 415-448. DOI: 10.1007/s10579-021-09574-0.

Coole, M., Rayson, P. and Mariani, J. (2020). "Unfinished Business: Construction and Maintenance of a Semantically Tagged Historical Parliamentary Corpus, UK Hansard from 1803 to the present day." *Proceedings of the Second ParlaCLARIN Workshop*, May 2020, Marseille, France, European Language Resources Association, 23-27. https://aclanthology.org/2020.parlaclarin-1.5.

4 https://github.com/pemistahl/lingua-py.

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

6 https://github.com/clarin-eric/parla-clarin.

Puren, M., Vernus, P., Pellet, A., Bourgeois, N. and Lebreton, F. (2022). "Extracting and providing online access to annotated and semantically enriched historical data. The AGODA project." Zenodo. DOI: 10.5281/zenodo.6594666.

Ljubešić, N. and Dobrovoljc, K. (2019). "What does Neural Bring? Analysing Improvements in Morphosyntactic Annotation and Lemmatisation of Slovenian, Croatian and Serbian." *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, August 2019, Florence, Italy, Association for Computational Linguistics, 29-34. https://www.aclweb.org/anthology/W19-3704, DOI: 10.18653/v1/W19-3704.

Erjavec, T. and Pančur, A. (2021). "The Parla-CLARIN Recommendations for Encoding Corpora of Parliamentary Proceedings." *Journal of the Text Encoding Initiative* [Online], 14, April 2021 – March 2023. https://journals.openedition.org/jtei/4133, DOI: 10.4000/jtei.4133.

Kavčič, A., Mundjar, A. and Marolt, M. (2023). "Parliamentary corpus of first Yugoslavia (1919-1939) yu1Parl 1.0." Slovenian language resource repository CLARIN.SI, ISSN 28204042, http://hdl.handle.net/11356/1845.

Kavčič, A., Mundjar, A. and Marolt, M. (2023). "Carniolan Provincial Assembly corpus Kranjska 1.0." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1824.

Pančur, A., Erjavec, T., Meden, K., Ojsteršek, M., Šorn, M. and Blaj Hribar, N. (2022). "Slovenian parliamentary corpus (1990-2022) siParl 3.0." Slovenian language resource repository CLARIN.SI, ISSN 2820-4042, http://hdl.handle.net/11356/1748.

Acknowledgements

The work was supported by the Slovenian Research Agency research programmes P6-0436: Digital Humanities: Resources, Tools and Methods (2022–2027) and the DARIAH-SI research infrastructure (2022-2027).