An OCR Pipeline for Transforming Parliamentary Debates into Linked Data: Case ParliamentSampo – Parliament of Finland on the Semantic Web

Senka Drobac[1,2], Laura Sinikallio[2,1], and Eero Hyvönen[1,2]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland
https://seco.cs.aalto.fi, firstname.lastname@aalto.fi
[2] Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland

Abstract. This paper presents the OCR pipeline created for *ParliamentSampo – Parliament of Finland on the Semantic Web*, a Linked Open Data (LOD) service, data infrastructure, and semantic portal for studying Finnish political culture, language, and networks of the Members of Parliament (MP). A knowledge graph of linked data has been created based on ca. 967 000 speeches in all plenary sessions of the Parliament of Finland in 1907—2022; the data is also available in XML format, utilizing the new international Parla-CLARIN format. A central part of the historical debates 1907–1999 was available only as PDF documents of fairly low OCR quality and had to be OCRed first; this paper reports lessons learned from this process.

Keywords: OCR, parliamentary studies, linked data, digital humanities

1 Introduction

Parliamentary data are used in many areas of research [3], as they provide a wealth of information about the state and functioning of democratic systems, political life and, more generally, language and culture. The most prominent part of the work of parliaments is the public plenary sessions, in which the Members of Parliament (MP) discuss and vote on issues on the agenda and other topics that arise [11]. Semantic Web (SW) technologies[3] and Linked Data (LD) [8] provide a promising approach for publishing and using parliamentary data in Digital Humanities (DH) [27,4,11]. The LD approach for Cultural Heritage [10] has arguably many advantages: 1) Linked data and ontologies [26] provide a framework for harmonizing heterogeneous distributed datasets and combining them into larger and richer entities. 2) The SW is based on the Predicate Logic [9], which provides an opportunity to enrich data by reasoning new information. 3) When the machine "understands" the content of the data, intelligent

[3] https://www.w3.org/standards/semanticweb/

web services and data analyses can be implemented more easily. 4) Ready-made tools by other actors can be re-used for publishing, processing and analysing the standardized data.

However, using linked data requires that the typically textual, unstructured debates have to be transformed into semantic structured data in several steps: 1) If the minutes are available only in print they have to be first digitized. 2) Text texts have to be OCRed from digitized documents. 3) Metadata about the OCRed texts has to be extracted and represented using RDF[4]. 4) The data can be enriched and interlinked and finally be published and made available in a SPARQL endpoint. 5) Applications on top of the endpoint can be created or the data service can be used for data-analytic research. This paper concerns step 2 in the case of publishing and using Finnish parliamentary speech data. In this case, the digitized data was provided by the open data service of the Parliament of Finland (PoF)[5]. Metadata extraction and enrichment (steps 3–4) are described in [25,16,11]. The speech data outcome described in this paper has already been used as a basis for analyzing concepts in political speeches [6], for network analyses based on MP references in speeches [22], and for data analyses of speeches and for portal application development [11].

2 Related Works

Several parliamentary corpora have been formed from the minutes of the plenary debates, which make it possible to study the content of the speeches and their language; see, e.g., [14] and the CLARIN list of parliamentary corpora[6]. The TEI-based Parla-CLARIN scheme[7] for session minutes has been developed within the CLARIN infrastructure, providing a common way to represent the corpora [21]. The related ParlaMint project[8] brings together Parla-CLARIN-based national corpora. Parliamentary materials have also been transformed into the form of LD when creating the LinkedEP [27] system on the European Parliament's data, the Italian Parliament[9], the LinkedSaeima for the Latvian parliament [4].

The materials of PoF have been digitized in various contexts but are difficult to use, as they have been produced separately from different periods and stored in different formats [25]. The usability of the materials is also hampered by their varying quality and lack of descriptive data [13]. Language corpora have been published on parliamentary debates, such as the Parliamentary Corpus of FIN-CLARIN's Language Bank[10] [15] which covers the years 2008—2016. It contains the speeches in a linguistically annotated form and also synchronized

[4] https://www.w3.org/RDF/
[5] https://avoindata.eduskunta.fi/#/fi/digitoidut/
[6] https://www.clarin.eu/resource-families/parliamentary-corpora
[7] https://github.com/clarin-eric/parla-clarin
[8] https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora
[9] http://data.camera.it
[10] http://korp.csc.fi

links to original plenary session videos [18]. The Voices of Democracy project has produced a research corpus that includes plenary minutes in 1980–2018 annotated grammatically as well as interviews of veteran MPs conducted by the PoF after 1988 [1]. The minutes of the parliamentary debates from 1991 to 2015 can also be found in the International Harvard Parlspeech Corpus [23], but we have identified gaps in the coverage in this corpus.

Some of the most popular open source OCR tools for historical printed texts are Tesseract[11], OCR4All [24] and OCR-D [2]. In Finland, the most OCR efforts have been focused on newspaper material [5] and [12]. A comprehensive post-correction survey is presented in [19].

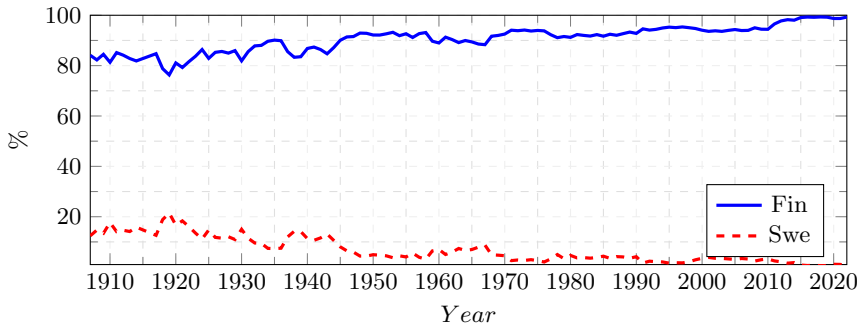

Fig. 1. The graph shows percentages of language representation in speeches through years. The blue line shows the percentage of Finnish text and the red dashed line the percentage of Swedish data. The graph has been calculated on a sentence level.

3 The OCR Pipeline

This section presents the OCR pipeline used in transforming the Finnish debate corpus 1907–1999 into LD.

Data sources The parliamentary speech data is provided in three different file formats. Parliamentary sessions from the period 1907-1999 have been scanned and made available as PDF files. Later data is already in machine-readable formats, with sessions from 1999-2015 in HTML and from 2015 onward in XML format.

The PDF documents are of good quality, which simplified the OCR process. However, the data around 1920 is very noisy with occasional skewed pages. All the data is printed in modern Latin fonts, with a slight variation throughout the time. When it comes to formatting, early minutes are printed in a single column. However, most of the data is in double-column format. In earlier documents,

[11] https://github.com/tesseract-ocr/tesseract

double columns were separated with a black line and in later ones with white space.

The Process We OCRed the data with Tesseract 4, with pre-trained Finnish and Swedish models. The main reasons to choose Tesseract were the availability of high accuracy pre-trained models that can recognize a large variety of modern fonts, as well as the possibility to use multiple language models at the same time.

Skipping the training part enabled us to do the OCR rather quickly and the multi-language support was especially important because the parliamentary speeches in Finland are given mainly in two languages, Finnish and Swedish (both official languages of Finland). Fig. 1 shows the percentages of Finnish and Swedish languages in the speeches through the time.

The first step in the OCR process was to transform PDF files into images. According to Tesseract documentation[12], Tesseract should be working best on images that have a DPI of at least 300 dpi. We performed initial testing on a small amount of data with different image resolutions and found out that we get the best OCR results at the resolution of 350 dpi. Both lower and, surprisingly, higher image resolutions resulted in poorer OCR. Once we had the images ready, we performed the recognition with Tesseract, using `-l fin+swe` option, which prefers Finnish as the recognition language but can also recognize Swedish.

As we were dealing with a large dataset (324,333 page images), we had to perform the recognition using CSC's (CSC -– IT Center for Science[13]) supercomputer Puhti[14]. It not only supports using GPU's for computing, but also allows to submit as many as 100 SLURM batch jobs in parallel (*array jobs*). Once the parallel system was set up (on average around 3,200 pages were recognized on one GPU at the time) the OCR duration was measured in hours (5-8 hours, depending on the size of a job). When taking into account queuing time for the resources, the entire recognition process was done in only a couple of days.

Post-correction and Transformation into Parla-CLARIN

To gather all speeches of the Finnish Parliament in the 20$^{\text{th}}$ century we used pattern recognition and regular expressions on the plain-text version of the OCR results. The OCRed results were satisfactory as they were, but to enhance the reliability of the gathered data we performed a few manual corrections to the OCR results. Each transcript of a plenary session started with a title row that spanned the whole page, whereas the rest of the document was mainly split into two columns. Due to this, the title was sometimes split into two rows or otherwise corrupted in the results. As one file included several transcripts, one after another, these title rows and the information they contained (date, session number) were central in connecting speeches to correct sessions. With a helper script, all distorted title rows were located and manually corrected.

After corrections, we created Python scripts to scrape all relevant data from the OCRed text files. First, we gathered speeches and their metadata in CSV format. Then the data went through several automated correction and enrich-

[12] https://github.com/tesseract-ocr/tessdoc/blob/main/ImproveQuality.md
[13] https://www.csc.fi/en/
[14] https://research.csc.fi/-/puhti

ment steps. A central part of the enrichment was retrieving speaker information from an external *Members of Parliament* (MP) dataset [16] as the transcripts contained only each speaker's title and surname. The correct person was found based on the scraped surname and session date only, so for correct linking, the names needed to be correct. Hence the majority of the correction efforts went into fixing speaker names and titles that had been distorted in the OCR process (e.g. *Procopé* had become *Procop&*).

Typical correction steps were: Handling of missing or extra whitespaces (*MinisteriHuttu* → *Ministeri Huttu, Ministeri Lin n ain maa* → *Ministeri Linnainmaa*), removal of extra trailing characters, such as special characters, and replacing some systematically recurring errors, such as a common surname ending *qvist* having become *gvist*. If the corrections weren't enough to find the right match, we would pick the closest match from the list of all possible surnames from the MP data set.

Finally, we transformed the speeches into two parallel data sets: (1) an RDF (*Resource Description Framework*) [20] format speech knowledge graph, forming linked data and (2) an XML corpus formed according to the Parla-CLARIN v0.2 specification [7]. More on this transformation can be read from [25].

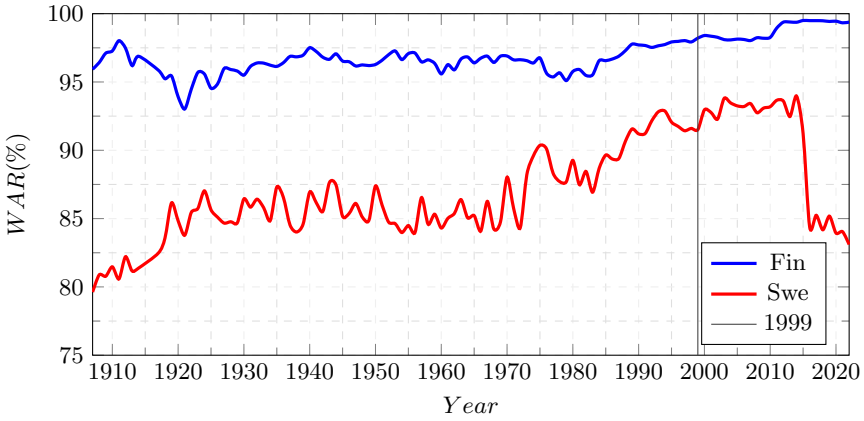

Fig. 2. Word accuracy rates (WAR, %) on Finnish and Swedish speeches. Speeches before 1999 are from OCRed material and after 1999 from already digital documents in HTML and XML formats

4 Results

To evaluate the final results, we calculate Word Accuracy Rate (WAR) – the percentage of correctly recognized words – with the Language Analysis Command-Line Tool (LAS) [17]. The tool has been specifically adapted to work well for

Finnish, however, it also offers support for many other languages. It uses Finite State Machines to check which words (in inflected forms) are present in available morphologies. Since the tool can use only one language for a given string and because many speeches, especially in early years, were bilingual, we couldn't use it on a speech level. Instead, we tokenized speeches with Python's `nltk.tokenize`[15] into sentences and performed the recognition on a sentence level. Conveniently, the tool also performs language recognition.

The results are shown in Fig. 2. The upper (blue) line shows accuracy of the Finnish data, and the lower (red) one accuracy of the Swedish data. The vertical line shows the year 1999, before which the data was OCRed and the data after that was already available in HTML and XMl formats. It can help us determine the quality/broadness of the used morphology.

Finnish WAR is consistently over 95%, except for the period around 1920 where the scanned images ware exceptionally noisy. Swedish WARs are drastically lower, between 79–93%. However they are also low for the control data after 1999, clearly showing that the Swedish morphology likely misses many words used in parliamentary sessions.

5 Discussion

It would be intuitive that doing OCR on images with higher resolution would give better results, but we found that in our case, resolution higher than 350 dpi gives poorer results. Since we used pre-trained models, it is possible that the models were trained on images of that resolution.

The main reason for the lower accuracy in 1920's seems to be image quality. It would be necessary to find a way to remove noise, or possibly train on such noisy data. It would be interesting to see how much improvement would fine-tuning with noisy data bring .

In general, the main reason low accuracy of the Swedish data is that the morphology is quite limited as many words are also not recognized in the native digital data. Furthermore, we tokenized the data into sentences, but occasionally Swedish sentences contain some Finnish words. For example, the speaker may address the parliament in Finnish, and then continue in Swedish. We also don't know how well language recognition works. Because Finnish and Swedish are distant languages, it would be better to combine the two morphological transducers into one and perform the recognition with that instead of tokenizing sentences.

Acknowledgements Our work is funded by the Academy of Finland and is also related to the EU project InTaVia[16] and the EU COST action Nexus Linguarum[17]. We used the computing resources of the CSC – IT Center for Science.

[15] https://www.nltk.org/api/nltk.tokenize.html
[16] https://intavia.eu
[17] https://nexuslinguarum.eu

References

1. Andrushchenko, M., Sandberg, K., Turunen, R., Marjanen, J., Hatavara, M., Kurunmäki, J., Nummenmaa, T., Hyvärinen, M., Teräs, K., Peltonen, J., Nummenmaa, J.: Using parsed and annotated corpora to analyze parliamentarians' talk in Finland. Journal of the Association for Information Science and Technology **185**(1), 1–15 (2021). https://doi.org/10.1002/asi.24500
2. Baierer, K., Büttner, A., Engl, E., Hinrichsen, L., Reul, C.: OCR-D & OCR4all: Two complementary approaches for improved OCR of historical sources. In: 6th International Workshop on Computational History. CEUR Workshop Proceedings (2021)
3. Benoît, C., Rozenberg, O. (eds.): Handbook of Parliamentary Studies: Interdisciplinary Approaches to Legislatures. Edward Elgar Publishing (2020). https://doi.org/10.4337/9781789906516
4. Bojārs, U., Darģis, R., Lavrinovičs, U., Paikens, P.: LinkedSaeima: A linked open dataset of Latvia's parliamentary debates. In: Semantic Systems. The Power of AI and Knowledge Graphs. SEMANTiCS 2019. pp. 50–56. Springer (2019). https://doi.org/10.1007/978-3-030-33220-4_4
5. Drobac, S., Lindén, K.: Optical character recognition with neural networks and post-correction with finite state methods. International Journal on Document Analysis and Recognition (2020). https://doi.org/s10032-020-00359-9
6. Elo, K., Karimäki, J.: Luonnonsuojelusta ilmastopolitiikkaan: Ympäristöpoliittisen käsitteistön muutos parlamenttipuheessa 1960–2020. Politiikka **63**(4) (Nov 2021). https://doi.org/10.37452/politiikka.109690
7. Erjavec, T., Pančur, A.: Parla-CLARIN - a TEI schema for corpora of parliamentary proceedings (May 2022), https://clarin-eric.github.io/parla-clarin/
8. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool (2011), http://linkeddatabook.com/editions/1.0/
9. Hitzler, P., Krötzsch, M., Rudolph, S.: Foundations of Semantic Web technologies. Springer (2010)
10. Hyvönen, E.: Publishing and Using Cultural Heritage Linked Data on the Semantic Web. Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, Palo Alto, CA, USA (2012)
11. Hyvönen, E., Sinikallio, L., Leskinen, P., Mela, M.L., Tuominen, J., Elo, K., Drobac, S., Koho, M., Ikkala, E., Tamper, M., Leal, R., Kesäniemi, J.: Finnish parliament on the semantic web: Using ParliamentSampo data service and semantic portal for studying political culture and language. In: Digital Parliamentary data in Action (DiPaDa 2022), Workshop at the 6th Digital Humanities in Nordic and Baltic Countries Conference, long paper. CEUR Workshop Proceedings, Vol. 3133 (May 2022), http://ceur-ws.org/Vol-3133/paper05.pdf
12. Kettunen, K.T., Koistinen, J.M.O., et al.: Open source Tesseract in Re-OCR of finnish fraktur from 19th and early 20th century newspapers and journals – collected notes on quality improvement. In: Digital Humanities in the Nordic Countries Proceedings of the Digital Humanities in the Nordic Countries 4th Conference. CEUR-WS.org (2019)
13. La Mela, M.: Tracing the emergence of Nordic allemansrätten through digitised parliamentary sources. In: Fridlund, M., M., Oiva, Paju, P. (eds.) Digital histories: Emergent approaches within the new digital history, pp. 181–197. Helsinki University Press (2020). https://doi.org/10.33134/HUP-5-11

14. Lapponi, E., Søyland, M.G., Velldal, E., Oepen, S.: The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016. Lang Resources & Evaluation **52**, 873–893 (2018). https://doi.org/10.1007/s10579-018-9411-5
15. Lennes, M.: FIN-CLARIN and language bank parliamentary data. Workshop 'Digital Parliamentary Data and Research' (2019), https://www2.helsinki.fi/en/helsinki-centre-for-digital-humanities/workshop-digital-parliamentary-data-and-research
16. Leskinen, P., Hyvönen, E., Tuominen, J.: Members of Parliament in Finland knowledge graph and its linked open data service. In: Graphs. Proceedings of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands. pp. 255–269 (2021). https://doi.org/10.3233/SSW210049, https://ebooks.iospress.nl/volumearticle/57420
17. Mäkelä, E.: LAS: an integrated language analysis tool for multiple languages. J. Open Source Software **1**(6), 35 (2016). https://doi.org/10.21105/joss.00035
18. Mansikkaniemi, A., Smit, P., Kurimo, M.: Automatic construction of the Finnish parliament speech corpus. In: Proc. Interspeech 2017. pp. 3762–3766 (2017). https://doi.org/10.21437/Interspeech.2017-1115
19. Nguyen, T.T.H., Jatowt, A., Coustaty, M., Doucet, A.: Survey of post-ocr processing approaches. ACM Computing Surveys (CSUR) **54**(6), 1–37 (2021)
20. Pan, J.Z.: Resource Description Framework. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, pp. 71–90. International Handbooks on Information Systems, Springer, Berlin, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92673-3_3
21. Pancur, A., Erjavec, T.: The siParl corpus of Slovene parliamentary proceedings. In: Proceedings of the Second ParlaCLARIN Workshop. pp. 28–34. European Language Resources Association (2020), https://www.aclweb.org/anthology/2020.509parlaclarin-1.6
22. Pokkimäki, H., Leskinen, P., Tamper, M., Hyvönen, E.: Analyses of networks of politicians based on linked data: Case ParliamentSampo – Parliament of Finland on the Semantic Web (2022), http://seco.cs.aalto.fi/publications/2022/poikkimaki-et-al-2022.pdf, paper under peer review
23. Rauh, C., De Wilde, P., Schwalbach, J.: The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states (V1) (2017). https://doi.org/10.7910/DVN/E4RSP9
24. Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., Puppe, F.: OCR4all – an open-source tool providing a (semi-) automatic OCR workflow for historical printings. arXiv preprint arXiv:1909.04032 (2019)
25. Sinikallio, L., Drobac, S., Tamper, M., Leal, R., Koho, M., Tuominen, J., Mela, M.L., Hyvönen, E.: Plenary debates of the Parliament of Finland as linked open data and in Parla-CLARIN markup. In: 3rd Conference on Language, Data and Knowledge, LDK 2021. pp. 1–17. Schloss Dagstuhl- Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing (August 2021), https://drops.dagstuhl.de/opus/volltexte/2021/14544/pdf/OASIcs-LDK-2021-8.pdf
26. Staab, S., Studer, R. (eds.): Handbook on Ontologies (2nd Ed.). Springer (2009)
27. Van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., Beunders, H.: The debates of the European Parliament as Linked Open Data. Semantic Web – Interoperability, Usability, Applicability **8**(2), 271–281 (2017). https://doi.org/10.1007/s42001-019-00060-w