Drobac, Senka; Sinikallio, Laura; Hyvönen, Eero

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

Published in:
Digital Humanities in the Nordic and Baltic Countries Publications

DOI:
10.5617/dhnbpub.10670

Published: 01/01/2023

Document Version
Publisher's PDF, also known as Version of record

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

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

[1]*Aalto University (Semantic Computing Research Group (SeCo)), Finland*
[2]*University of Helsinki (HSSH, HELDIG, Cultural heritage studies), 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.

1. Introduction

Parliamentary data are used in many areas of research [1], 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 [2]. Semantic Web (SW) technologies[1] and Linked Data (LD) [3] provide a promising approach for publishing and using parliamentary data in Digital Humanities (DH) [4, 5, 2]. The LD approach for Cultural Heritage [6] has arguably many advantages, including:

- Linked data and ontologies [7] provide a framework for harmonizing heterogeneous distributed datasets and combining them into larger and richer entities.
- The SW is based on the Predicate Logic [8], which provides an opportunity to enrich data by linking new information.
- When machines can "understand" data content, intelligent web services and data analyses can be implemented more easily.

DHNB2023 | Sustainability: Environment - Community - Data. The 7th Digital Humanities in the Nordic and Baltic Countries Conference. Oslo – Stavanger – Bergen, Norway. March 8–10, 2023.
senka.drobac@aalto.fi (S. Drobac)
0000-0002-7645-3079 (S. Drobac); 0000-0003-1695-5840 (L. Sinikallio); 0000-0003-1695-5840 (E. Hyvönen)

DHNB Publications, DHNB2023 Conference Proceedings, https://journals.uio.no/dhnbpub/issue/view/875

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

- 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. Texts have to be OCRed from digitized documents.
3. Metadata about the OCRed texts has to be extracted and represented using RDF.[2]
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).[3] Metadata extraction and enrichment (steps 3-4) are described in [9, 10, 2]. The speech data outcome described in this paper has already been used as a basis for analyzing concepts in political speeches [11], for network analyses based on MP references in speeches [12], and for data analyses of speeches and for portal application development [2].

2. Related Work

The minutes of parliamentary plenary debates have been compiled into several corpora, allowing for analysis of their content and language (e.g., the corpus of the Norwegian parliament [13]; CLARIN list of parliamentary corpora[4]). These corpora are represented using the TEI-based Parla-CLARIN scheme[5], which has been developed within the CLARIN infrastructure to provide a unified standard [14]. The related ParlaMint project[6] is dedicated to creating comparable national parliamentary corpora based on the Parla-CLARIN scheme. Additionally, parliamentary materials have been transformed into Linked Data format for the creation of systems such as LinkedEP [4], which deals with European Parliament data, as well as the Italian Parliament[7] and LinkedSaeima for the Latvian parliament [5].

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 [9]. The usability of the materials is also hampered by their varying quality and lack of descriptive data [15]. Language corpora have been published on parliamentary debates, such as the Parliamentary Corpus of FIN-CLARIN's Language Bank[8] [16] which covers the years 2008-2016. It

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

contains the speeches in a linguistically annotated form and also synchronized links to original plenary session videos [17]. 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 [18]. The minutes of the parliamentary debates from 1991 to 2015 can also be found in the International Harvard Parlspeech Corpus [19], 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[9], OCR4All [20] and OCR-D [21]. In Finland, the most OCR efforts have been focused on newspaper material [22] and [23]. A comprehensive post-correction survey is presented in [24].

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 performance of OCR systems is heavily influenced by the quality of the input data. In our situation, the PDF documents are generally of high quality, which simplifies the OCR process. However, data from the early 1920s is considerably noisy due to the use of low-quality paper during that era, as illustrated in Figure 1. Moreover, there are occasional skewed pages that present difficulties for OCR. All the data is printed in contemporary Latin fonts, with minor variations over time. As for formatting, early minutes are presented in a single-column layout, while the majority of the data is presented in a double-column format. In earlier documents, the double columns were separated by a black line, whereas in later ones, they were separated by white space.

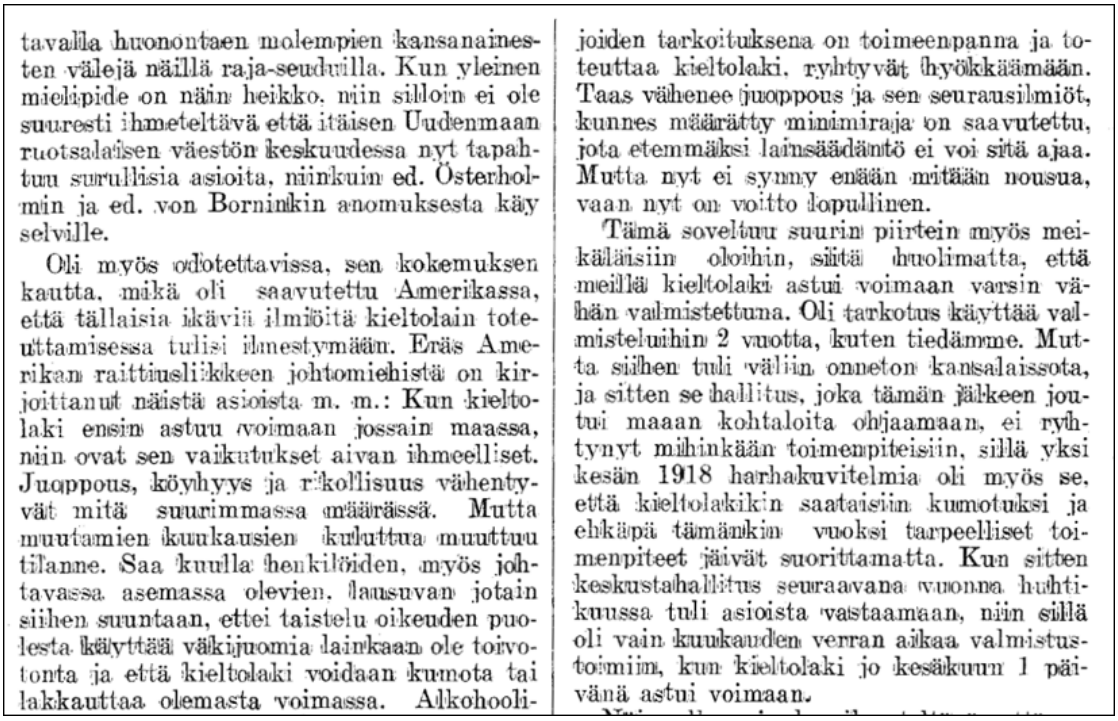
The OCR Process We OCRed the data with Tesseract 4, with pre-trained models for both Finnish and Swedish. We opted for Tesseract due to its high accuracy pre-trained models, capable of recognizing various contemporary fonts, and its ability to use multiple language models concurrently.

We were able to perform OCR quickly by omitting the training stage, and the ability to support multiple languages was particularly crucial since parliamentary speeches in Finland are primarily given in two official languages: Finnish and Swedish. Fig. 2 illustrates the proportion of Finnish and Swedish languages used in these speeches over time. Finnish is the predominant language used, and the percentage of Swedish has gradually decreased over time, from 18-20% to minimal levels today.

To begin the OCR process, we first converted the PDF files into images. Tesseract's documentation[10] suggests that the software performs optimally on images with a minimum DPI of 300. After conducting preliminary tests on a small dataset with various image resolutions, we discovered that the best OCR outcomes were obtained at a resolution of 350 DPI. Curiously, both

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

[10] https://github.com/tesseract-ocr/tessdoc/blob/main/ImproveQuality.md



tavalla huonontaen molempien kansanainesten välejä näillä raja-seuduilla. Kun yleinen mielipide on näin heikko, niin silloin ei ole suuresti ihmeteltävä että itäisen Uudenmaan ruotsalaisen väestön keskuudessa nyt tapahtuu surullisia asioita, niinkuin ed. Österholmin ja ed. von Borninkin anomuksesta käy selville.

Oli myös odotettavissa, sen kokemuksen kautta, mikä oli saavutettu Amerikassa, että tällaisia ikäviä ilmiöitä kieltolain toteuttamisessa tulisi ilmestymään. Eräs Amerikan raittiusliikkeen johtomiehistä on kirjoittanut näistä asioista m. m.: Kun kieltolaki ensin astuu voimaan jossain maassa, niin ovat sen vaikutukset aivan ihmeelliset. Juoppous, köyhyys ja rikollisuus vähentyvät mitä suurimmassa määrässä. Mutta muutamien kuukausien kuluttua muuttuu tilanne. Saa kuulla henkilöiden, myös johtavassa asemassa olevien, lausuvan jotain siihen suuntaan, ettei taistelu oikeuden puolesta käyttää väkijuomia lainkaan ole toivotonta ja että kieltolaki voidaan kumota tai lakkauttaa olemasta voimassa. Alkoholijoiden tarkoituksena on toimeenpanna ja toteuttaa kieltolaki, ryhtyvät hyökkäämään. Taas vähenee juoppous ja sen seurausilmiöt, kunnes määrätty minimiraja on saavutettu, jota etemmäksi lainsäädäntö ei voi sitä ajaa. Mutta nyt ei synny enään mitään nousua, vaan nyt on voitto lopullinen.

Tämä soveltuu suurin piirtein myös meikäläisiin oloihin, siitä huolimatta, että meillä kieltolaki astui voimaan varsin vähän valmistettuna. Oli tarkotus käyttää valmisteluihin 2 vuotta, kuten tiedämme. Mutta siihen tuli väliin onneton kansalaissota, ja sitten se hallitus, joka tämän jälkeen joutui maaan kohtaloita ohjaamaan, ei ryhtynyt mihinkään toimenpiteisiin, sillä yksi kesän 1918 harhakuvitelmia oli myös se, että kieltolakikin saataisiin kumotuksi ja ehkäpä tämänkin vuoksi tarpeelliset toimenpiteet jäivät suorittamatta. Kun sitten keskustahallitus seuraavana vuonna huhtikuussa tuli asioista vastaamaan, niin sillä oli vain kuukauden verran aikaa valmistustoimiin, kun kieltolaki jo kesäkuun 1 päivänä astui voimaan.

Figure 1: A snippet from a 1921 document shows smudged text due to the poor quality of the paper.

lower and higher image resolutions yielded inferior OCR results. After preparing the images, we performed the recognition with Tesseract, using `-l fin+swe` option, which prioritizes Finnish as the language for recognition while also having the capability to recognize Swedish.

Due to the extensive size of our dataset, consisting of 324,333 page images, we used CSC's supercomputer Puhti[11] provided by CSC - IT Center for Science[12] for OCR. This supercomputer enabled us to leverage GPU computing and perform up to 100 SLURM batch jobs in parallel using its *array jobs* feature. Once we set up the parallel system (with an average of around 3,200 pages recognized per GPU at a time), the OCR process took only a few hours (typically between 5 to 8 hours, depending on the job size). Accounting for the queuing time for the resources, the entire recognition process was completed in a matter of days.

Post-correction and Transformation into Linked Data and Parla-CLARIN

To gather all speeches of the Finnish Parliament in the 20th 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

[11] https://research.csc.fi/-/puhti
[12] https://www.csc.fi/en/

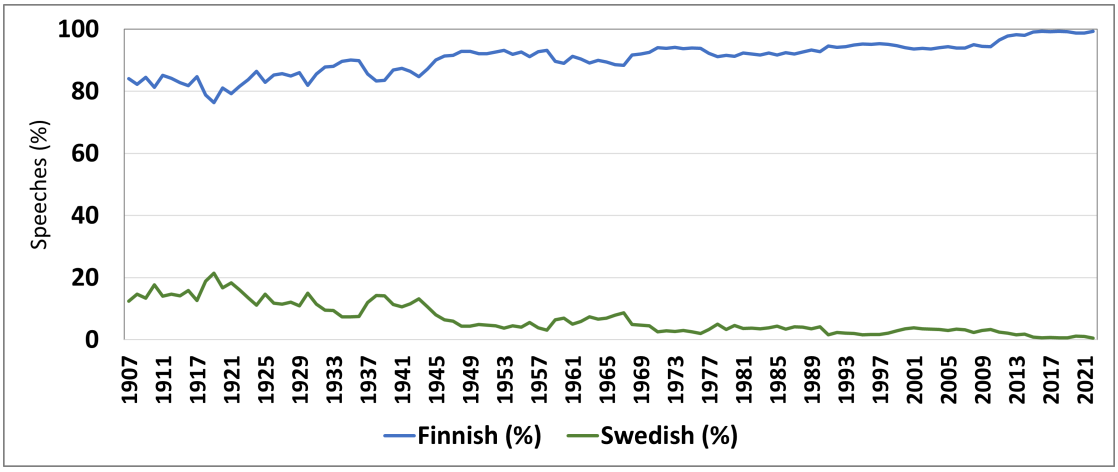


Figure 2: The graph shows percentages of language representation in speeches through years. The blue line shows the percentage of Finnish text and the green line the percentage of Swedish data. The graph has been calculated on a sentence level.

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 enrichment steps. A central part of the enrichment was retrieving speaker information from an external *Members of Parliament* (MP) dataset [10] 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*) [25] format speech knowledge graph, forming linked data and (2) an XML corpus formed according to the Parla-CLARIN v0.2 specification [26]. More on this transformation can be read from [9].

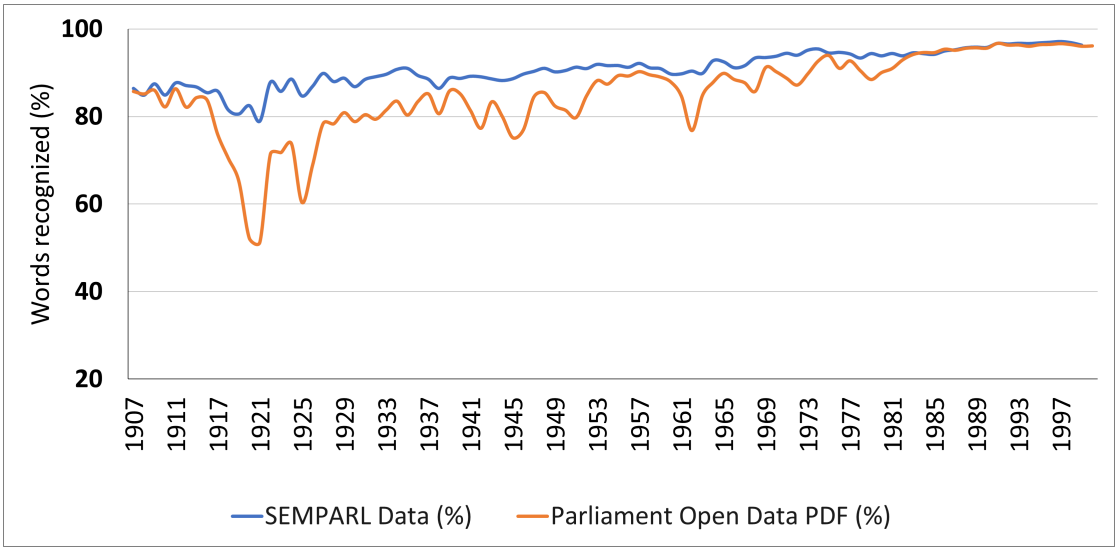


Figure 3: The percentage of recognized words with LAS tool on PoF OCR results (orange) and our new OCR (blue) results.

4. Evaluation

In order to evaluate the final results, we utilized the Language Analysis Command-Line Tool (LAS) [27] to calculate the percentage of correctly recognized words. LAS has been specifically adapted to work well for Finnish and uses Finite State Machines to verify the presence of words in Finnish morphological lexical database Omorfi [28].

This tool is particularly useful for a task like this because it covers all possible grammatical word forms in a language. However, a limitation of such a tool is the dictionary's scope, particularly with regards to specialized and historical language. Although the LAS tool has expanded the original Finnish morphology to cover historical spelling variations, it has not been adapted to legal vocabulary.

In the first experiment, we conducted a comparison between the accuracy of our OCR texts and those provided in the PoF original documents. Figure 3 shows the percenteges of recognized words with LAS tool on the received material (orange line) and our results (blue line). This evaluation was carried out on the entire dataset, not just speeches, and using only Finnish morphology. Recognizing text in complete documents is generally more challenging than in speeches, because they may contain structures other than running text (eg., tables and lists). Additionally, the dataset is bilingual, resulting in lower overall accuracy. Nevertheless, this experiment provides us with an indicator of the improvement we are pleased to report in OCR accuracy, particularly during the early 1920s when the dataset was most challenging

In addition to the evaluation on the entire dataset, we also conducted a separate evaluation solely for speeches. However, since the tool can only use one language for a given string and many speeches, particularly in the early years, were bilingual, we tokenized speeches

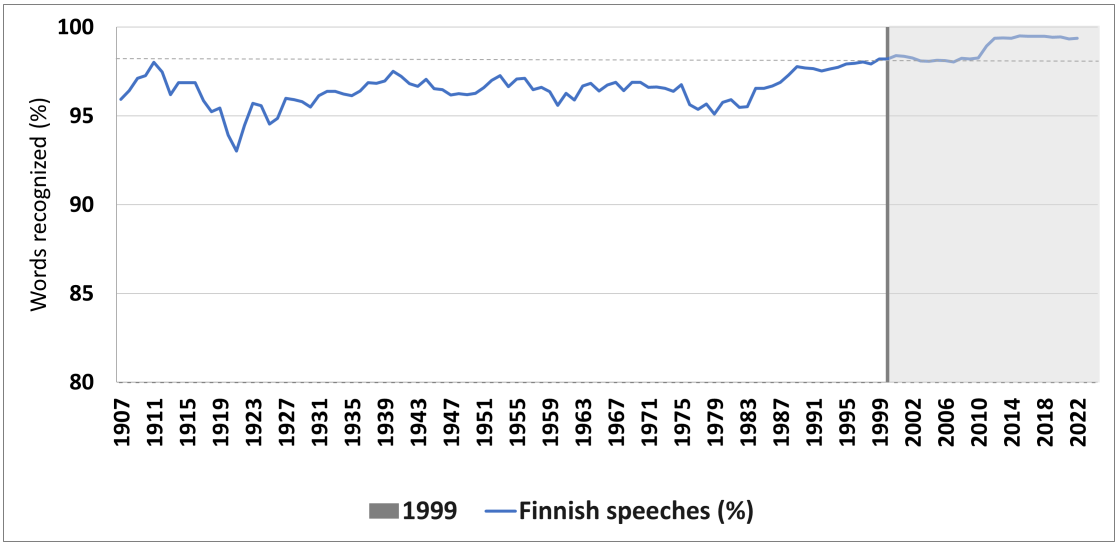


Figure 4: The percentage of recognized words with LAS tool on Finnish speeches (blue line). Results prior to 1999 indicate the performance on OCR text, where those after were based from native digital data.

into sentences using Python's `nltk.tokenize`[13] and performed recognition on a sentence level. Conveniently, the tool also performs language recognition, making this approach both convenient and effective.

The results of the evaluation on the Finnish speeches are presented in Figure 4. The blue line represents the accuracy of the Finnish data. A vertical line denotes the year 1999, which signifies the point before the data was OCRed and after which it was available in HTML and XML formats. This information helps in evaluating the quality and scope of the morphology employed.

The graph demonstrates that the percentage of correctly recognized words remains consistently above 95%, except for the period around 1920 when the scanned images were particularly noisy. On the right side of the graph, the results for natively digital data illustrate that the benchmark in the early 2000s is approximately 98%. This indicates that the accuracy of the OCR data is only slightly lower, ranging from 0-3% below the benchmark.

Similarly, we conducted an evaluation on Swedish speeches; however, the benchmark was low, with recognition rates mostly ranging from 86% to 93%. The OCR accuracy rates were mostly between 85% and 92%, with early data up to 1916 showing recognition rates of 80-82%.

5. Discussion and Conclusion

Although higher resolution images are often thought to yield better OCR results, we discovered that in our case, using resolutions greater than 350 dpi led to poorer accuracy. This could be

[13] https://www.nltk.org/api/nltk.tokenize.html

because the pre-trained models we used were developed using images with that resolution.

The most practical way we found to evaluate the data was to use the LAS tool since we did not have any Ground Truth (GT) data. Creating GT data would have been a time-consuming process, especially considering the variations throughout the years. We would have to create GT data for each year, and to obtain a good evaluation, we would need to sample a significant amount of data from every year. It would not be feasible to capture entire pages, so we would need to select lines from different pages, create GT and evaluate against the produced OCR. This would require annotating at least 150-200 lines from each year, totaling around 13,800-18,000 lines. It is uncertain whether this effort would yield more information on data quality, especially for Finnish text. Our approach provided us with a reasonable understanding of OCR quality, despite not having good conclusive quality estimation for the Swedish proportion of the corpus.

The results with LAS tool on the Swedish speeches are difficult to interpret since the low benchmark makes it unclear whether the unrecognized words come from the the OCR errors or the small scope of the morphological acceptor. The majority of the data falls within the benchmark range, but since the range is so large, it is impossible to draw any definitive conclusions based solely on these results.

The Finnish OCR quality is outstanding, with only a 0-3% difference from the benchmark in terms of the number of recognized words, except for the early 1920s period where lower accuracy was due to poor image quality. One potential solution for future work could be to train recognition models on noisy data from that period, or explore the extent of improvement that could be achieved through fine-tuning existing models. Furthermore, to avoid the need for manual post-correction of titles, we could experiment with using an XML version of the OCR results. It includes text coordinates in images and it could enable us to automatically identify split titles. This would make the entire process fully automatic and easily reproducible.

Acknowledgments

Our work is funded by the Academy of Finland and is also related to the EU project InTaVia[14] and the EU COST action Nexus Linguarum[15]. We used the computing resources of the CSC - IT Center for Science.

References

[1] C. Benoît, O. Rozenberg (Eds.), Handbook of Parliamentary Studies: Interdisciplinary Approaches to Legislatures, Edward Elgar Publishing, 2020. doi:10.4337/9781789906516.

[2] E. Hyvönen, L. Sinikallio, P. Leskinen, M. L. Mela, J. Tuominen, K. Elo, S. Drobac, M. Koho, E. Ikkala, M. Tamper, R. Leal, J. Kesäniemi, 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, 2022. URL: http://ceur-ws.org/Vol-3133/paper05.pdf.

[14] https://intavia.eu

[15] https://nexuslinguarum.eu

[3] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space (1st edition), Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, 2011. URL: http://linkeddatabook.com/editions/1.0/.

[4] A. Van Aggelen, L. Hollink, M. Kemman, M. Kleppe, H. Beunders, The debates of the European Parliament as Linked Open Data, Semantic Web – Interoperability, Usability, Applicability 8 (2017) 271–281. doi:10.1007/s42001-019-00060-w.

[5] U. Bojārs, R. Darģis, U. Lavrinovičs, P. Paikens, LinkedSaeima: A linked open dataset of Latvia's parliamentary debates, in: Semantic Systems. The Power of AI and Knowledge Graphs. SEMANTiCS 2019, Springer, 2019, pp. 50–56. doi:10.1007/978-3-030-33220-4_4.

[6] E. Hyvönen, 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.

[7] S. Staab, R. Studer (Eds.), Handbook on Ontologies (2nd Ed.), Springer, 2009.

[8] P. Hitzler, M. Krötzsch, S. Rudolph, Foundations of Semantic Web technologies, Springer, 2010.

[9] L. Sinikallio, S. Drobac, M. Tamper, R. Leal, M. Koho, J. Tuominen, M. L. Mela, E. Hyvönen, 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, Schloss Dagstuhl- Leibniz-Zentrum fur Informatik GmbH, Dagstuhl Publishing, 2021, pp. 1–17. URL: https://drops.dagstuhl.de/opus/volltexte/2021/14544/pdf/OASIcs-LDK-2021-8.pdf.

[10] P. Leskinen, E. Hyvönen, J. Tuominen, Members of Parliament in Finland knowledge graph and its linked open data service, in: of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands, 2021, pp. 255–269. URL: https://ebooks.iospress.nl/volumearticle/57420. doi:10.3233/SSW210049.

[11] K. Elo, J. Karimäki, Luonnonsuojelusta ilmastopolitiikkaan: Ympäristöpoliittisen käsitteistön muutos parlamenttipuheessa 1960–2020, Politiikka 63 (2021). doi:10.37452/politiikka.109690.

[12] H. Pokkimäki, P. Leskinen, M. Tamper, E. Hyvönen, Analyses of networks of politicians based on linked data: Case ParliamentSampo – Parliament of Finland on the Semantic Web, 2022. URL: http://seco.cs.aalto.fi/publications/2022/poikkimaki-et-al-2022.pdf, paper under peer review.

[13] E. Lapponi, M. G. Søyland, E. Velldal, S. Oepen, The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016, Lang Resources & Evaluation 52 (2018) 873–893. doi:10.1007/s10579-018-9411-5.

[14] A. Pancur, T. Erjavec, The siParl corpus of Slovene parliamentary proceedings, in: Proceedings of the Second ParlaCLARIN Workshop, European Language Resources Association, 2020, pp. 28–34. URL: https://www.aclweb.org/anthology/2020.509parlaclarin-1.6.

[15] M. La Mela, Tracing the emergence of Nordic allemansrätten through digitised parliamentary sources, in: M. Fridlund, M., Oiva, P. Paju (Eds.), Digital histories: Emergent approaches within the new digital history, Helsinki University Press, 2020, pp. 181–197. doi:10.33134/HUP-5-11.

[16] M. Lennes, FIN-CLARIN and language bank parliamentary data. Workshop 'Digital Parliamentary Data and Research', 2019. URL:

https://www2.helsinki.fi/en/helsinki-centre-for-digital-humanities/workshop-digital-parliamentary-data-and-research.

[17] A. Mansikkaniemi, P. Smit, M. Kurimo, Automatic construction of the Finnish parliament speech corpus, in: Proc. Interspeech 2017, 2017, pp. 3762–3766. doi:10.21437/Interspeech.2017-1115.

[18] M. Andrushchenko, K. Sandberg, R. Turunen, J. Marjanen, M. Hatavara, J. Kurunmäki, T. Nummenmaa, M. Hyvärinen, K. Teräs, J. Peltonen, J. Nummenmaa, Using parsed and annotated corpora to analyze parliamentarians' talk in Finland, Journal of the Association for Information Science and Technology 185 (2021) 1–15. doi:10.1002/asi.24500.

[19] C. Rauh, P. De Wilde, J. Schwalbach, 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. doi:10.7910/DVN/E4RSP9.

[20] C. Reul, D. Christ, A. Hartelt, N. Balbach, M. Wehner, U. Springmann, C. Wick, C. Grundig, A. Büttner, F. Puppe, OCR4all – an open-source tool providing a (semi-) automatic OCR workflow for historical printings, arXiv preprint arXiv:1909.04032 (2019).

[21] K. Baierer, A. Büttner, E. Engl, L. Hinrichsen, C. Reul, OCR-D & OCR4all: Two complementary approaches for improved OCR of historical sources, in: 6th International Workshop on Computational History, CEUR Workshop Proceedings, 2021.

[22] S. Drobac, K. Lindén, Optical character recognition with neural networks and post-correction with finite state methods, International Journal on Document Analysis and Recognition (2020). doi:s10032-020-00359-9.

[23] K. T. Kettunen, J. M. O. Koistinen, 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.

[24] T. T. H. Nguyen, A. Jatowt, M. Coustaty, A. Doucet, Survey of post-ocr processing approaches, ACM Computing Surveys (CSUR) 54 (2021) 1–37.

[25] J. Z. Pan, Resource Description Framework, in: S. Staab, R. Studer (Eds.), Handbook on Ontologies, International Handbooks on Information Systems, Springer, Berlin, Heidelberg, 2009, pp. 71–90. doi:10.1007/978-3-540-92673-3_3.

[26] T. Erjavec, A. Pančur, Parla-CLARIN - a TEI schema for corpora of parliamentary proceedings, 2022. URL: https://clarin-eric.github.io/parla-clarin/.

[27] E. Mäkelä, LAS: an integrated language analysis tool for multiple languages., J. Open Source Software 1 (2016) 35. doi:10.21105/joss.00035.

[28] T. A. Pirinen, Omorfi—free and open source morphological lexical database for Finnish, in: Proceedings of the 20th Nordic conference of computational linguistics (NODALIDA 2015), 2015, pp. 313–315.