Extracting Knowledge from Parliamentary Debates for Studying Political Culture and Language

Minna Tamper[1,2], Rafael Leal[2], Laura Sinikallio[1,2], Petri Leskinen[2,1], Jouni Tuominen[1,2], and Eero Hyvönen[1,2]

[1] University of Helsinki (HELDIG and HSSH), Finland
https://heldig.fi, firstname.lastname@helsinki.fi
[2] Aalto University, Department of Computer Science, Finland
firstname.lastname@aalto.fi
Semantic Computing Research Group (SeCo)
https://seco.cs.aalto.fi

Abstract. This paper presents knowledge extraction and natural language processing methods used to enrich the knowledge graph of the plenary debates (textual transcripts of speeches) of the Parliament of Finland. This knowledge graph includes some 960 000 speeches (1907–2021) interlinked with a prosopographical knowledge graph about the politicians. A recent subset of the speeches was used to extract named entities and topical keywords for semantic searching and browsing the data and for data analysis. The process is based on linguistic analysis, named entity linking (NEL), and automatic subject indexing. The results were included into the ParliamentSampo knowledge graph in a SPARQL endpoint. This data can be used for studying parliamentary language and culture in Digital Humanities research and for developing applications, such as the ParliamentSampo portal.

1 Introduction

Parliaments enact new laws, oversee the work of the government, and decide on the state budget. Parliamentary data are used in many areas of research [3], as they provide a wealth of information on the state and functioning of democratic systems, political life and, more generally, language and culture. For these reasons, a lot of parliamentary materials have been digitized in recent decades [1]. Digitized parliamentary materials offer a wide range of perspectives on different research topics and have been used in a variety of fields, such as linguistics, political science, media studies, economics, and history. A most important research material for parliament studies are the debates in the parliaments, i.e., sequences of transliterated speeches (minutes) of Members of Parliament (MP) and other politicians, through which one can study the language and its changes itself as well as the underlying societal phenomena at large [8].

This paper argues and shows that by enriching textual parliamentary speeches with linked data using knowledge extraction methods [30], it is possible to support Digital Humanities (DH) research and enhance the usability of the data

in applications, such as semantic search, browsing, and data analysis. As a case study, a part the ca. 960 000 speeches of the system *ParliamentSampo – Finnish Parliament on the Semantic Web* [16,17] are used. In an earlier work, the speeches covering the whole history 1907–2021 of the Parliament of Finland (PoF) were extracted from original heterogeneous data sources and transformed into a speech knowledge graph (S-KG) [36] (and also into the Parla-CLARIN format[3]). At the same time, the S-KG was interlinked with a prosopographical KG (P-KG) representing detailed biographical data and networks of the ca. 2800 MPs and politicians involved in the PoF activities [25]. Both graphs were published as a LOD service on the Linked Data Finland platform LDF.fi [11], including a SPARQL endpoint[4]. In this paper, the textual speeches of this ParliamentSampo dataset are enriched further using knowledge extraction techniques in order to support Digital Humanities (DH) [10] analysis and for further development of the semantic portal ParliamentSampo on top of the endpoint.

In this paper, we first shortly overview the speech data in focus and then focus on the new data enrichments using Natural Language Processing (NLP) methods (Section 2). Section 3 discusses how the new data can be utilized in the ParliamentSampo portal. Lastly, the contributions of this work are summarized and related works are discussed (Section 4).

2 Datasets and Knowledge Extraction

2.1 Core Datasets

The ParliamentSampo system includes data about the MPs, parliamentary speeches, and political organizations within the PoF. The data covers not only the speeches that form the official debates but also the comments of the Speaker (President) of the PoF and all other small comments recorded in the minutes, for example, in connection with voting proceedings. The ParliamentSampo data contains two major parts:

1. The Prosopographic Knowledge Graph The Prosopographic Knowledge Graph [25] covers all MPs of Finland since the year 1907. At its core lies a RDF conversion of data about MPs from the originally XML-formatted Open Data service[5] of PoF. In addition to basic information, such as times and places of birth and death, the data includes detailed information about politicians' life events, such as studies, working life, political career, and their written publications. In addition to people, the graph contains information about organizations, professions, and positions, as well as places. Organizations include, e.g., parties, ministries, parliamentary groups, committees, and constituencies, as well as schools, organizations, and companies outside the political community.

[3] https://github.com/clarin-eric/parla-clarin

[4] The endpoint will be published openly using the CC BY 4.0 license by the end of 2022.

[5] https://avoindata.eduskunta.fi/#/fi/dbsearch

2. The Parliamentary Speeches Knowledge Graph The knowledge graph of parliamentary speeches contains speeches collected from all the minutes of the plenary sessions of the PoF since 1907 [36]. This knowledge graph was compiled from the documents available on the Open Data services[6] and web sites[7] of the PoF. Depending on the time period they covered, the documents were available in different formats: PDF, HTML, or XML. PDF documents were transformed into text with OCR.

In addition to the actual speeches, the speech graph contains all the relevant metadata attached to the minutes, such as interjections, information about the session where the speech was given (time, date, serial number, etc.), speaker information (name, role, party) and possible topic of discussion, and supporting documents (e.g. committee report). Based on the metadata, the speeches were linked to the MPs P-KG. For example, speakers and the parties they represent are resources with URI identifiers described in the P-KG.

2.2 Knowledge Extraction

In our work, the speech knowledge graph was enriched using various NLP methods. In this work, the toolset that was used in enriching the BiographySampo dataset [15,41] was re-used together with new methods for NER, lemmatization, and automatic subject indexing. The parliamentary debates were enriched with named entity recognition (NER) and linking (NEL), subject indexing, and by creating a linguistic knowledge graph containing linguistic details for the speeches. Here, NLP methods were used on a subset of the speeches dataset, consisting of speeches from parliamentary session 2015 to the end of parliamentary session 2021, totaling in a little over 114 000 speeches. This covers about 12% of the whole speeches dataset.

Lemmatization and Subject Indexing The under-development library Secompling[8], which aims at integrating different Finnish NLP tools, was used for the tasks of lemmatization and subject indexing.

Lemmatization can be seen of as a kind of text normalization, especially for a language as morphologically rich as Finnish, which has 15 inflectional cases and a rich system for derivative words. Lemmatization enables more exact term-based search instead of wildcard-based stemming. Lemmatization allows word count-based algorithms, such as TF-IDF, to work with more precision. Secompling employs the Turku Neural parser pipeline [22,23] for lemmatization, and Voikko[9] and uralicNLP [19] to check and possibly fix errors regarding these base forms.

Subject indexing allows texts to be described succinctly by focusing on keywords that best characterize their contents. In our work, the new subject indexing tool Annif [37], developed by the National Library of Finland, is re-used as the backend for this task. As Annif is based on machine learning, it may sometimes

[6] https://avoindata.eduskunta.fi/#/fi/home
[7] https://www.eduskunta.fi/fi/Sivut/default.aspx
[8] https://version.aalto.fi/gitlab/seco/secompling
[9] https://voikko.puimula.org/

suggest named entities not mentioned in the texts. In our case the focus is on entities that are actually mentioned in the speeches, and the named entities in the annotation results of Annif were therefore ignored. The other subject keywords are filtered out according to their weight as provided by Annif. The keywords in the Annif model used were taken from the General Finnish Ontology YSO[10], part on the national Finnish LOD infrastucture [13], with ready to use URIs for data linking.

Named Entity Recognition and Linking NEL was performed on the speeches dataset in order to improve data browsing and searching in the ParliamentSampo portal. Similarly to the analytics done for the textual biographies in the BiographySampo [41] system, NEL enables more detailed data analytics in the ParliamentSampo dataset, too. Named entities were extracted using the Nelli [42,39] tool and its results linked using the ARPA tool [28]. Unlike in the BiographySampo dataset, here Nelli was configured to use FinBERT's combined NER model [27], Reksi [42], and the Turku Neural parser pipeline. FinBERT's NER tool that is coupled with Reksi to be able to also pick up links to legislation and references to various dates and identifiers (e.g., references to laws). The Turku Neural parser was selected for morphological analysis based on its evaluated good performance [22]. These tools extracted entities that were later linked using ARPA to the ParliamentSampo datasets and to few other external datasets, such as the Kanto ontology of Finnish actors[11], the Place Name Register PNR ontology of contemporary Finnish places[12], and the YSO places[13] ontology that contains also historical places mentioned in the speeches.

Morpholinguistic Knowledge Graph Lastly, the speeches were transformed into a separate morpholinguistic knowledge graph containing detailed linguistic and morphological information about the speeches using a pipeline previously used for BiographySampo [40]. This graph can be used for linguistic analysis of the parliamentary speeches similarly to the work done in BiographySampo [41]. For example, in the BiographySampo dataset, it was noticed that biographies of women contained more family-related terminology while biographies about men used more words related to war and religion. In order to apply same methods to parliamentary speeches, a similar pipeline was used, updated to use the Turku Neural parser pipeline, and adjusted to handle larger datasets in smaller chunks of text. In this case, it was configured to process data by year. The results are also linked to the ParliamentSampo speeches dataset to enable analysis of speeches using the speech metadata.

3 Using the Enriched Data in ParliamentSampo

The enriched ParliamentSampo data is used in the development of the ParliamentSampo portal [18] which is based on the Sampo model [12] and the

[10] https://finto.fi/yso/fi/new?clang=en
[11] https://finto.fi/finaf/fi/
[12] https://www.ldf.fi/dataset/pnr/
[13] https://finto.fi/yso-paikat/en/

Sampo-UI framework [20]. The portal demonstrates how the data service can be used for developing applications for DH research. In this application the data can be browsed using ontology-based faceted search, and the results can then be analyzed with the integrated visualization and data analysis tools.

The enriched data is initially used to enhance the browsing and searching capabilities of the portal. For instance, named entities and keywords can be used via facets to find speeches that mention a specific topic or a named entity, such as a place or an organization. Coupled with the speeches, their metadata, and the prosopographical data, this enables studying, for example, how MPs talk about matters related to their constituency. At the moment mentioned organizations and places have already been added into portal as facets to test the data.

Currently, the morpholinguistic KG about the speeches is small, limited to a few years. It is not yet included in the ParliamentSampo portal, but we plan to add it and develop similar linguistic analysis views as in BiographySampo. This enables, for example, to study the vocabulary used by the MPs and parties in their speeches. It is also possible to compare speeches of men and women and their differences in vocabulary.

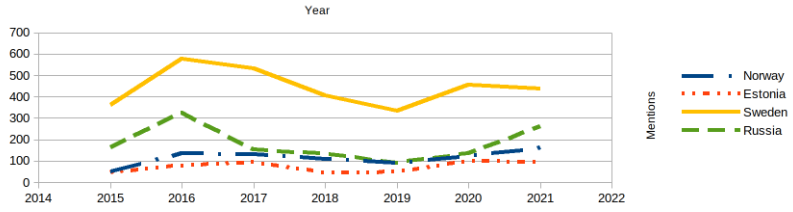

Fig. 1. Frequency of place mentions in parliamentary debates from the 2015 parliamentary session until end of 2021.

In addition, the SPARQL endpoint underlying the ParliamentSampo portal can be used for querying, analyzing, and visualizing the enriched data. In Fig. 1, for example, the speeches mentioning Finland's neighbouring countries Norway, Sweden, Estonia, and Russia are counted on a yearly basis and plotted from 2015 to 2021. The figure shows that Sweden appears most frequently in the speeches than the other neighbours. Russia is also mentioned increasingly in 2020 and 2021. Both Russia and Sweden have shared history with Finland at different periods of time but it remains as future work to study the context where these mentions are made. Similarly, by linking to various place ontologies it is possible to leverage the benefit of organized information and create visualizations that cluster all, e.g., Russia-related place names as mentions about Russia or any other country. It is also possible to use map-based visualizations.

4 Discussion

Contributions In this paper, we presented ongoing work done for enriching the ParliamentSampo dataset in order to suport searching and browsing of the data and using it for DH research. Currently, the ParliamentSampo portal is still under development and data enrichment processes continue for the data. This is ongoing work that still requires adjustments and evaluation. The tools used in the enrichment have been previously evaluated with different corpora, but not for the parliamentary data yet. The FinBERT NER tool has achieved an accuracy of 93.11 % using the combined model in cross-corpus evaluation [27]. Similarly, the Turku Neural Parser pipeline is evaluated based on CoNLL 2018 UD Shared Task [45] with accuracy of LAS[14] 86.60%, UPOS[15] 96.66%, and XPOS[16] 97.63% [22]. Subject indexing is difficult to evaluate, however based on evaluation done for the Annif tool, its accuracy is 30–50% depending on the test corpus [38]. The Secompling lemmatization module has not been formally evaluated yet. These results have been produced on formal Finnish language texts similar to the Finnish parliamentary debates corpus.

The data has been partially added to the ParliamentSampo knowledge graph and utilized already in the facets of the semantic portal. With the current data it is possible already to improve search and browsing of the data. Similarly, the data with the enrichment enables DH research through topics and named entities. Also, the data can be already used to study topics and mentions in the parliamentary speeches to some extent. The enrichments help to find interesting phenomenon in the ParliamentSampo dataset. It remains future work to create around it applications for the DH community to enable to study them in more detail.

Related Work Significant amount of parliamentary data have been digitized in recent decades [1]. A most important part has been the minutes of the plenary debates, see, e.g., [24] and the CLARIN list of parliamentary corpora[17] in different countries. Parliamentary materials have also been transformed into the form of linked data, too. A prominent example of this is the LinkedEP [43] system on the European Parliament's data. Linked data has also been used in the Italian Parliament[18], and the LinkedSaeima for the Latvian parliament [4] in addition to the Finnish ParliamentSampo system [16,17] whose data [36,25] was re-used in the paper.

Knowledge extraction has been applied to enrich datasets to enable distant reading approaches to studying parliamentary debates. For example, the Latvian LinkedSaeima dataset has utilized named entity linking to enrich their metadata. Similarly, the Dutch parliamentary debates dataset [21] has been enriched with

[14] Labeled attachment score (LAS) is the proportion of words that have connected correctly the head word with the right dependency relation.
[15] Universal part- of-speech tagging
[16] Language-specific part-of-speech tagging
[17] https://www.clarin.eu/resource-families/parliamentary-corpora
[18] http://data.camera.it

named entities. The Slovenian siParl corpus [31] has been enriched with linguistic information about the parliamentary debates.

The NLP methods used in this work have been developed mainly for handling Finnish texts. With respect to NER, some of the most relevant tools are StanfordNER [9], FiNER [33] and the FinBERT based NER tool, of which the last one is currently estimated to be the most accurate [27,44,34]. Similarly, there are morphological analyzers for Finnish besides the Turku Neural parser, such as the two used in this paper, Voikko and uralicNLP, the latter of which employs Omorfi [32]. Regarding entity linking, there are few tools available for Finnish, such as ARPA and Annif. Various tools have been created also for other languages to link named entities to different datasets, such as [2,6,7].

In Finland, parliamentary materials have been digitized and utilized to some extent in digital humanities and social science research. For example, [35] examines the differences in political speech between parties throughout the parliamentary period 1907–2018. In [29], the content of the plenary speeches given in Parliament in 1999–2014 were studied by using topic modeling. Also, in [26] the debates were examined. However, data have so far been used only in a few studies that deploy methods from corpus linguistics, language technology, or computer science [1].

Previous search applications for the Finnish parliamentary speech data are based mostly on traditional text search. However, search applications have been developed for other digitized and enriched Cultural Heritage datasets [14,5]. The data analysis tools to examine the results are few, such as the concordance analysis of the Language Bank of Finland[19], where the words found are visualized in their textual contexts and show some statistics of words occurrences in the search results. The Language Bank's tool has many corpora and one small corpus covering a small part of the entire time series of the Finnish parliamentary speeches.

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/https://doi.org/10.1002/asi.2450
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: Dbpedia: A nucleus for a web of open data. In: The Semantic Web: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007, pp. 722–735. Springer, Berlin, Heidelberg (2007). https://doi.org/https://doi.org/10.1007/978-3-540-76298-0{_}52
3. Benoît, C., Rozenberg, O. (eds.): Handbook of Parliamentary Studies: Interdisciplinary Approaches to Legislatures. Edward Elgar Publishing (2020). https://doi.org/10.4337/9781789906516

[19] https://www.kielipankki.fi/support/access/

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. Brandsen, A., Verberne, S., Lambers, K., Wansleeben, M.: Can bert dig it?–named entity recognition for information retrieval in the archaeology domain. arXiv preprint arXiv:2106.07742 (2021)
6. Damljanovic, D., Bontcheva, K.: Named Entity Disambiguation using Linked Data. In: The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings. pp. 231–240. Springer-Verlag Berlin Heidelberg (2012), http://2012.eswc-conferences.org/sites/default/files/eswc2012_submission_334.pdf
7. Derczynski, L., Maynard, D., Rizzo, G., Van Erp, M., Gorrell, G., Troncy, R., Petrak, J., Bontcheva, K.: Analysis of named entity recognition and linking for tweets. Information Processing and Management **51**(2), 32–49 (3 2015). https://doi.org/10.1016/j.ipm.2014.10.006, https://doi.org/10.1016%2Fj.ipm.2014.10.006
8. Elo, K., Karimäki, J.: Luonnonsuojelusta ilmastopolitiikkaan: Ympäristöpoliittisen käsitteistön muutos parlamenttipuheessa 1960–2020. Politiikka **63**(4) (Nov 2021). https://doi.org/10.37452/politiikka.109690
9. Finkel, J.R., Grenager, T., Manning, C.D.: Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In: Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), June, 25-30, 2005, University of Michigan, Ann Arbor, Michigan, USA. pp. 363–370. Association for Computational Linguistics (2005). https://doi.org/10.3115/1219840.1219885
10. Gardiner, E., Musto, R.G.: The Digital Humanities: A Primer for Students and Scholars. Cambridge University Press, New York, NY, USA (2015), https://doi.org/10.1017/CBO9781139003865
11. Hyvönen, E., Tuominen, J., Alonen, M., Mäkelä, E.: Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In: The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers. pp. 226–230. Springer-Verlag (2014), https://doi.org/10.1007/978-3-319-11955-7_24
12. Hyvönen, E.: Digital humanities on the Semantic Web: Sampo model and portal series (2021), submitted
13. Hyvönen, E.: How to create a national cross-domain ontology and linked data infrastructure and use it on the semantic web (2021), https://seco.cs.aalto.fi/publications/2021/hyvonen-dcmi-2021.pdf, keynote presentation for the DCMI 2021 conference
14. Hyvönen, E., Ikkala, E., Koho, M., , Leal, R., Rantala, H., Tamper, M.: How to search and contextualize scenes inside videos for enriched watching experience: Case stories of the second world war veterans (February 2022), under peer review
15. Hyvönen, E., Leskinen, P., Tamper, M., Rantala, H., Ikkala, E., Tuominen, J., Keravuori, K.: BiographySampo - publishing and enriching biographies on the semantic web for digital humanities research. In: Hitzler, P., Fernández, M., Janowicz, K., Zaveri, A., Gray, A.J., Lopez, V., Haller, A., Hammar, K. (eds.) The Semantic Web. ESWC 2019. pp. 574–589. Springer-Verlag (June 2019). https://doi.org/10.1007/978-3-030-21348-0_37
16. Hyvönen, E., Sinikallio, L., Leskinen, P., Drobac, S., Tuominen, J., Elo, K., Mela, M.L., Koho, M., Ikkala, E., Tamper, M., Leal, R., Kesäniemi, J.: Parlamenttisampo: eduskunnan aineistojen linkitetyn avoimen datan palvelu ja sen

käyttömahdollisuudet. Informaatiotutkimus **40**(2) (November 2021). https://doi.org/10.23978/inf.107899
17. 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 (2022), accepted, forth-coming
18. 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 WS Proceedinsg (2022)
19. Hämäläinen, M.: UralicNLP: An NLP library for Uralic languages. Journal of Open Source Software **4**(37), 1345 (2019). https://doi.org/10.21105/joss.01345
20. Ikkala, E., Hyvönen, E., Rantala, H., Koho, M.: Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces. Semantic Web – Interoperability, Usability, Applicability **13**(1), 69–84 (2022). https://doi.org/10.3233/SW-210428
21. Juric, D., Hollink, L., Houben, G.J.: Bringing parliamentary debates to the semantic web. In: DeRiVE@ ISWC. pp. 51–60 (2012)
22. Kanerva, J., Ginter, F., Miekka, N., Leino, A., Salakoski, T.: Turku neural parser pipeline: An end-to-end system for the conll 2018 shared task. In: Proceedings of the CoNLL 2018 Shared Task: Multilingual parsing from raw text to universal dependencies. pp. 133–142 (2018)
23. Kanerva, J., Ginter, F., Salakoski, T.: Universal lemmatizer: A sequence to sequence model for lemmatizing universal dependencies treebanks. Natural Language Engineering pp. 1–30 (2020). https://doi.org/10.1017/S1351324920000224
24. 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
25. Leskinen, P., Hyvönen, E., Tuominen, J.: Members of parliament in finland knowledge graph and its linked open data service. In: Further with Knowledge 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
26. Lillqvist, E., Kavonius, I.K., Pantzar, M.: "velkakello tikittää": Julkisyhteisöjen velka suomalaisessa mielikuvastossa ja tilastoissa 2000—2020. Kansantaloudellinen Aikakauskirja **116**(4), 581—607 (2020)
27. Luoma, J., Oinonen, M., Pyykönen, M., Laippala, V., Pyysalo, S.: A broad-coverage corpus for finnish named entity recognition. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 4615–4624 (2020)
28. Mäkelä, E.: Combining a REST Lexical Analysis Web Service with SPARQL for Mashup Semantic Annotation from Text. In: The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers. pp. 424–428. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-11955-7{_}60
29. Makkonen, K., Loukasmäki, P.: Eduskunnan täysistunnon puheenaiheet 1999–2014: Miten käsitellä LDA-aihemalleja? Politiikka **61**(2), 127—159 (2019)

30. Martinez-Rodriguez, J.L., Hogan, A., Lopez-Arevalo, I.: Information extraction meets the semantic web: A survey. Semantic Web – Interoperability, Usability, Applicability **11**(2), 255–335 (2020)
31. Pancur, A., Erjavec, T.: The siparl corpus of slovene parliamentary proceedings. In: Proceedings of the Second ParlaCLARIN Workshop. pp. 28–34 (2020)
32. Pirinen, T.A.: Development and use of computational morphology of finnish in the open source and open science era: Notes on experiences with omorfi development. SKY Journal of Linguistics **28**, 381–393 (2015)
33. Ruokolainen, T., Kauppinen, P., Silfverberg, M., Lindén, K.: A finnish news corpus for named entity recognition. Language Resources and Evaluation **54**(1), 247–272 (2020)
34. Ruokolainen, T., Kettunen, K.: À la recherche du nom perdu–searching for named entities with stanford ner in a finnish historical newspaper and journal collection. In: 13th IAPR International Workshop on Document Analysis Systems (2018)
35. Simola, S.: A century of partisanship in Finnish political speech (2020), part of PhD thesis: Essays in Labor and Political Economics, Aalto University
36. Sinikallio, L., Drobac, S., Tamper, M., Leal, R., Koho, M., Tuominen, J., La Mela, M., 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). vol. 93, pp. 8:1–8:17 (2021). https://doi.org/10.4230/OASIcs.LDK.2021.8
37. Suominen, O.: Annif: DIY automated subject indexing using multiple algorithms. LIBER Quarterly **29**(1), 1–25 (2019). https://doi.org/10.18352/lq.10285
38. Suominen, O., Lehtinen, M., Inkinen, J.: Annif and finto ai: developing and implementing automated subject indexing. Annif and Finto AI: Developing and Implementing Automated Subject Indexing pp. 265–282 (2022)
39. Tamper, M., Hyvönen, E., Leskinen, P.: Visualizing and analyzing networks of named entities in biographical dictionaries for digital humanities research. In: Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICling 2019). Springer (April 2019), accepted
40. Tamper, M., Leskinen, P., Apajalahti, K., Hyvönen, E.: Using Biographical Texts as Linked Data for Prosopographical Research and Applications. In: Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection. 7th International Conference, EuroMed 2018, Nicosia, Cyprus. pp. 125–137. Springer-Verlag (November 2018). https://doi.org/10.1007/978-3-030-01762-0_11
41. Tamper, M., Leskinen, P., Hyvönen, E., Valjus, R., Keravuori, K.: Analyzing biography collection historiographically as linked data: Case national biography of finland. Semantic Web – Interoperability, Usability, Applicability (2021), accepted
42. Tamper, M., Oksanen, A., Tuominen, J., Hietanen, A., Hyvönen, E.: Automatic Annotation Service APPI: Named Entity Linking in Legal Domain. In: The Semantic Web: ESWC 2020 Satellite Events. pp. 110–114. Springer-Verlag (2020)
43. 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
44. Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., Pyysalo, S.: Multilingual is not enough: Bert for finnish (2019)
45. Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., Nivre, J., Petrov, S.: CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In: Proceedings of the CoNLL 2018 Shared Task:

Multilingual Parsing from Raw Text to Universal Dependencies. pp. 1–21. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/K18-2001