WarMemoirSampo : A Semantic Portal for War Veteran Interview Videos

Leal, Rafael

CEUR-WS.org
2022

Leal , R , Rantala , H , Koho , M , Ikkala , E , Tamper , M , Merenmies , M & Hyvönen , E 2022 , WarMemoirSampo : A Semantic Portal for War Veteran Interview Videos . in K Berglund , M La Mela & I Zwart (eds) , Proceedings of the 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022) . CEUR Workshop Proceedings , vol. 3232 , CEUR-WS.org , Aachen , pp. 317-325 , Digital Humanities in the Nordic and Baltic Countries Conference , Uppsala , Sweden , 15/03/2022 .

http://hdl.handle.net/10138/350315

WarMemoirSampo: A Semantic Portal for War Veteran Interview Videos

Rafael Leal[1], Heikki Rantala[1,2], Mikko Koho[1,2], Esko Ikkala[1], Minna Tamper[1,2], Markus Merenmies[3] and Eero Hyvönen[1,2]

[1]Semantic Computing Research Group (SeCo), Aalto University, Finland
[2]Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland
[3]National Archives of Finland

Abstract

This paper presents WarMemoirSampo, a portal that provides semantic search and navigation of video interviews with Finnish World War II veterans. The portal associates video fragments with contextual data extracted from the video transcriptions, enabling users to find suitable video segments via faceted search and highlighting relevant content in the video being watched. This is carried out by processing natural language texts in order to extract named entities, keywords and lemmas. The result is a Linked Data Knowledge Graph that underpins the portal. We describe the collaboration between Natural Language Processing and Semantic Web technologies used in order to produce these results.

Keywords

Linked Open Data, Named entity recognition, Named entity linking, Military history, War veterans

1. Introduction

WarMemoirSampo is a Linked Open Data (LOD) resource of Finnish Second World War (WW2) veteran interview videos, as well as a semantic portal for easy access to them. It hosts a collection of 159 videos realized by the veteran organization Tammenlehvän Perinneliitto, with approximately 400 hours in total, in which veterans recollect their lives during and after wartime. The system is being realized using the Sampo model [1], by enriching the videos with related information from the WarSampo knowledge graph [2] and Wikidata. Rough transcriptions of the interviews, which form the basis of the textual information presented in the portal and the metadata extracted from them, were provided by Tammenlehvän Perinneliitto and the National Archives of Finland.[1] The videos and their segments are indexed [3, 4] with the extracted metadata.

This work addresses the key technical challenge of extracting semantic linked data from the textual descriptions of videos: it describes a system in which natural language text is processed

The 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), Uppsala, Sweden, March 15-18, 2022.

rafael.leal@aalto.fi (R. Leal)
0000-0001-7266-2036 (R. Leal); 0000-0002-4716-6564 (H. Rantala); 0000-0002-7373-9338 (M. Koho); 0000-0002-9571-7260 (E. Ikkala); 0000-0002-3301-1705 (M. Tamper); 0000-0001-6144-1473 (M. Merenmies); 0000-0003-1695-5840 (E. Hyvönen)

[1]More information about the project can be found at: https://seco.cs.aalto.fi/projects/war-memoirs/en.

in order to produce a harmonized knowledge graph (KG) – including named entities, keywords and lemmas – with timestamp information. This enables the videos to be accessed at different points according to their semantic content, and additional contextual information to be provided while the video is being watched. In order to achieve this, we created a KG containing enriched data for the interviews as well as information about the interviewees. The data enrichment and lemmatization is carried out using Natural Language Processing (NLP) techniques and knowledge extraction, resulting in new metadata about keywords and mentioned named entities – e.g., people, places, and events. These are integrated into the KG as entities and properties. The data is then stored in a SPARQL endpoint for open access.

The interview videos were edited by adding titles and copyright notes in the beginning and by merging multi-part interviews. They were then published on the YouTube platform for streaming, as well as archived in a repository at the National Archives for later use. The WarMemoirSampo Portal[2] was published on December 3, 2021, at the National Archives of Finland.

This system is related to previously published works of knowledge extraction from text to linked data [5]. The idea of providing contextual information while watching videos has been suggested already in the 80's in systems such as Hypersoap[3] that demonstrated the possibility of interactive product placement in a broadcast setting. There are also existing linked data-based metadata models for videos and video segmentation [3, 4, 6]. However, there does not seem to exist linked data-based systems for publishing and viewing videos while showing time-segmented contextual metadata.

The paper is organized as follows: first, the knowledge extraction and transformation pipeline for the underlying data service is explained. After this, functionalities of the portal application on top of the data service are briefly delineated. In conclusion, the main results of the work are discussed and directions for future research outlined.

2. Text Processing and Data Enrichment

The interviews were carried out in Finnish, which presents additional difficulties for natural language processing in comparison to better-researched European languages: its rich morphology[4] complicates lemmatization and Named Entity Recognition (NER). Moreover, Finnish NLP models are not as varied as for example those for English or French, and since the interviews were realized in colloquial, dialectal Finnish, not even state-of-the-art speech-to-text algorithms proved robust enough to handle them: at this point in time, only the standard variant of the language (known as *yleiskieli*) leads to acceptable results in speech-to-text tasks.

As an alternative, rough transcriptions made post-hoc by the interviewers were used for each of the 159 interviews. These vary widely in length and level of detail, with an average character count of 5599 ± 2458, which corresponds to a correlation of 0.718 with the interview lengths. Due to their nature, these notes are ultimately not meant to be used as a source of information

[2]The portal is in use at: https://sotamuistot.arkisto.fi.

[3]www.media.mit.edu/hypersoap/

[4]Finnish contains around 15 cases and various suffixes, which result in a number of surface forms that may reach well over 2000 for a single word. For example, the webpage http://www.ling.helsinki.fi/~fkarlsso/genkau2.html lists 2253 automatically generated forms for the word *kauppa* 'shop'

for others than the interviewers themselves: they do not provide full-fledged texts – or even sentences necessarily – and contain abbreviations as well as orthographic and grammatical errors, which compounds the challenge posed by the Finnish language. Nevertheless, we found the notes mostly adequate for the purposes of this project, although we have not thoroughly assessed their accuracy.

The interviewers' notes were provided as a spreadsheet document containing also other types of metadata, such as interviewees' name, date and place of birth, length and place of the interviews, as well as links for the interview videos uploaded to the YouTube platform. These notes are divided into rows, containing each one roughly a sentence. Each row also has a timestamp related to their location in the YouTube video. However, these timestamps are not unique: usually several consecutive rows have the same timestamp. These rows were grouped together, resulting in video segments of variable length but typically several minutes long, which became our data unit for this project.

An overview of the subsequent data processing and transformation process is shown in Figure 1. The original spreadsheet – containing one tab per interview, as well as metadata for all of them – is initially split into multiple CSV files. Each interview is then divided into shorter segments, as explained above. Three different procedures are carried out: the notes related to these segments are lemmatized in order to facilitate searching and other tasks; their keywords are generated; and their named entities are extracted and linked. The result of the data transformation process[5] are RDF files, which contain the KG that portrays all the different entities found in the data, their relations and properties.

The lemmatization of the texts is carried out by our Secompling library[6], while keywords are obtained via Annif [7], a subject indexing tool developed by the National Library of Finland. An Annif pre-trained model is used, and from the results obtained all named entities and keywords below a certain threshold are discarded. Three different NER/NEL (named entity linking) systems are used in this project: Secompling-NER, which uses the TurkuNLP Finnish NER tool [8] to extract a large quantity of named entities but does not provide entity linking; Nelli, which in this project is performing linking to Wikidata; and Warsa-linkers, which links to entities in WarSampo. These tools compliment each other by extracting entities and linking them to two different knowledge bases. The process and configurations for each tool will be explained in the next sections, as well as the procedure adopted to harmonize them and the overarching data model.

Lemmatization and NER with Secompling. Secompling is a library that aims at combining various Finnish NLP tools, including third-party ones, in order to provide an easy and integrated interface for methods such as lemmatization, NER, relevance feedback and unsupervised classification [9], and keyword extraction. The module responsible for lemmatization and NER relies on two tools developed by the TurkuNLP research group[7]: the Finnish NER tool for named entity recognition and the Neural Parser pipeline [10] for tokenization and lemmatization [11]. Both are based on FinBERT [12], a deep learning Finnish language model the TurkuNLP group has trained from scratch following the BERT guidelines [13]. Secompling

[5] WarMemoirSampo data transformation process: https://version.aalto.fi/gitlab/seco/veterans_manager
[6] https://version.aalto.fi/gitlab/seco/secompling
[7] https://turkunlp.org/

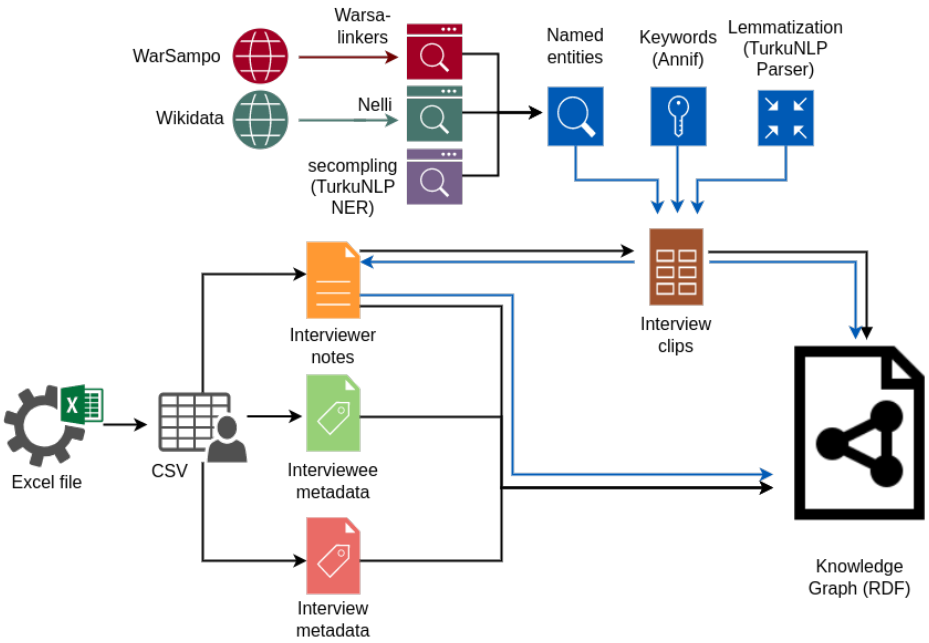

Figure 1: Overview of the data extraction and enrichment pipeline

complements the Neural parser with the third-party tools Voikko[8] and uralicNLP [14], which are used to check and correct the lemmas based on the part-of-speech (POS) tags provided by the parser itself. Some heuristics and automatic document editing are also applied in order to fix errors commonly made by the parser, such as wrong tokenization of strings containing punctuation marks, which may result in erroneous POS tags and lemmas.

The results of the NER tool and the parser are then aligned. Since the former does not provide lemmatization, this step provides the means for obtaining basic word forms for the named entities. Moreover, this procedure helps to fix errors in both tools, such as combining tokens that should be split, or assigning different named entity classes to parts of the same entity. The results of a brief examination with 100 random entities indicate that the Neural parser achieved a lemmatization accuracy of 0.84, while Secompling raised it to 0.96 by also applying named entity-specific heuristics [15].

Only a part of the eighteen entity categories recognized by the NER tool were judged to be of interest in WarMemoirSampo: Person, Product, Organization, Event, and a combination of Location (LOC), Facilities (FAC) and Geopolitical entities (GPE) as Place. Warsa-linkers adds Military Units to this list.

Since at this point Secompling-NER does not perform entity linking, it does not have the means to tell an accurate named entity from an error in either recognition or lemmatization. As a temporary solution, manual correction is applied to the candidates, so that it is possible to remove an entity, or edit its category or lemma. As a result, a total of 336 entities were ignored

[8] https://voikko.puimula.org/

and around 270 had their category and/or lemma changed.

NEL Using Nelli. The named entity linking tool Nelli [16, 17] is used to identify and link entities to different knowledge bases. In WarMemoirSampo this tool is configured to use the Turku neural parser pipeline to lemmatize the text prior to linking, and ARPA [18] to identify and then link named entities to the given vocabularies. ARPA is a configurable entity linking tool that consists of a list of linking configurations for different services; in the case of WarMemoirSampo the configurations are used to link entities to Wikidata.

In the interviews, the veterans mentioned some of their comrades in arms and superiors. Correctly identifying the most notable officers is often easy; disambiguating regular soldiers is not. The latter were often mentioned only by first name and/or surname, so that more information would have been necessary to differentiate them. Hence, the linking strategy centered around notable figures present in Wikidata.

Regarding place linkage, several Wikidata ARPA configurations are used to link places such as continents, countries, cities, as well as smaller places such as towns and villages. It can be challenging to match former Finnish place names to Wikidata due to varying practices in classifying place entities (e.g., towns or urban settlements) or missing labels, e.g., *Enso* or *Viipuri*. Moreover, places that were annexed to Russia often changed their names from Finnish to Russian, and some place names in Wikidata were lacking their original Finnish names mentioned in the interviews.

Warsa-linkers. WarSampo contains a data and ontology infrastructure for Finland in World War II. The WarSampo KG consists of around 100 000 persons, 51 000 places, 16 000 military units, 166 000 photographs, 26 000 war diaries, and 3000 war veteran memoir articles, among others, with large amounts of links between entities.

The data processing infrastructure of WarSampo contains a NEL process for linking descriptions of events and photographs to persons, places, and military units mentioned in texts [19]. The same process is reused in WarMemoirSampo to link the interviewers' notes to persons, places, and military units in WarSampo. After this NEL phase, the linked entities are pulled with SPARQL queries from the WarSampo KG, and new entities are created in WarMemoirSampo based on their metadata, with links to the original entities.

Entity Reconciliation. The generated named entities from different steps are disambiguated between the different data sources based on 1) matching the sets of URIs received for them from LOD data sources, and 2) matching entity types and names. WarSampo and Wikidata often contain alternative labels which are helpful in entity matching. In Wikidata, it is also possible to manually add missing labels in order to improve entity matching. After matching, the entities are reconciled by merging the duplicates found. The reconciled named entities consist of 310 persons, 1840 places, 66 events, 51 products, 117 military units, and 610 other organizations. These include linked entities as well as those unlinked which were deemed important, which also receive their own reference page (as explained in section 4).

We evaluated the outcome of the entity recognition pipeline by inspecting twenty random interview segments containing a total of 99 recognized entities [15]: 91 entities were correctly identified, 2 entities were mistakenly linked, 2 mistakenly recognized, 4 wrongly categorized, and 5 were not recognized. The final Precision is 0.919 and the Recall is 0.875, producing an F1 score of 0.897.

3. WarMemoirSampo: Knowledge Graph

The WarMemoirSampo KG, which is the result of the data transformation pipeline, is contained in RDF files and served via a SPARQL endpoint[9]. It is the only point of contact between the system's NLP backend and its user interface. It contains 323 371 triples, each one describing an entity or linking it to another. The main classes are (with the *:* namespace referring to *http://ldf.fi/schema/warmemoirsampo/*).

- *:Interview* (159 instances), corresponding to one video interview;
- *:PersonRecord* (159 instances), which collect the interviewee's personal information, such as given name, family name, gender and date of birth;
- *:TimeSlice* (2417): a video segment from start timestamp to end timestamp. An *:Interview* is divided into *:TimeSlices*, as explained in section 2;
- *:NamedEntity* (2994): Named Entity instances. Each named entity type also receives its own subclass;
- *skos:Concept* (3127): keywords as provided by Annif. They are instances of another KG, the General Finnish Ontology (YSO);

Both Interviews and TimeSlices are indexed with named entities and keywords. The named entities and keywords contain links to the same resources in other information sources.

4. WarMemoirSampo Portal

In order to provide an intuitive and user-friendly access to the WarMemoirSampo KG, a new web-based user interface had to be created. To avoid starting from scratch, the Sampo-UI JavaScript framework [20], which has been used for the creation of user interfaces for several domain-specific semantic portals in recent years[10], was chosen as the basis.

The structure of the WarMemoirSampo Portal[11] is based on the three main classes in the KG: *:Interview*, *:TimeSlice*, and *:NamedEntity*. Guided by principle 4. of the Sampo model [1], a separate faceted search perspective was created for each of these classes. Thus, instead of providing the end-user with only a single text search field to start with, these perspectives offer three complementary ways to search and browse the KG:

1. The **Interviews** perspective is meant for faceted searching of full interview videos using a combination of the following facets: text search, name and gender of the interviewee, and a named entity (place, person, military unit, organization, event, or product).

2. The **Parts of interviews** perspective offers similar search functionalities as the Interviews perspective, but in this case the result set consists of interview segments (*TimeSlices*). The search result links can be used to access the interview at the specific point that meets the search criteria. It is possible to for example list all interview segments where the best-known Finnish military leader, Carl Gustaf Emil Mannerheim, is mentioned.

[9] The WarMemoirSampo SPARQL endpoint is published at https://ldf.fi/warmemoirsampo/sparql

[10] See the full list of semantic portals powered by Sampo-UI at https://seco.cs.aalto.fi/applications/sampo

[11] The source code of the user interface is available on GitHub: https://github.com/SemanticComputing/veterans-web-app

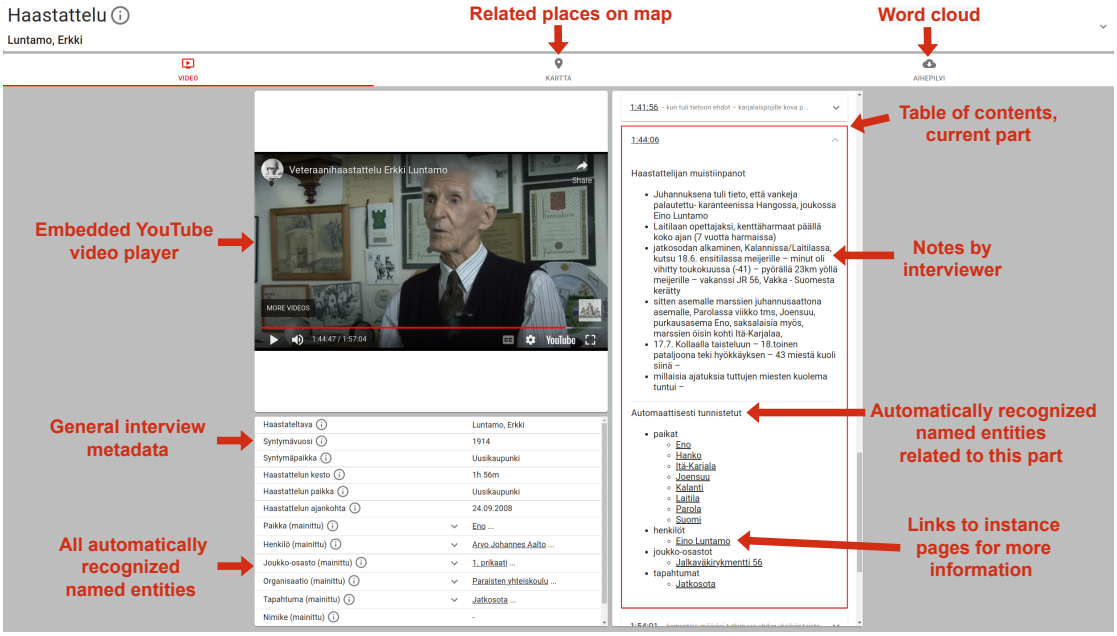

Figure 2: A video page with dynamic table of contents functionalities, provided by the portal.

3. The **Index** perspective includes faceted search and browsing functionalities for the nearly 3000 named entities mentioned in the summary notes of the interviews. For each named entity, the search results include links to specific part(s) of the interview where this entity is mentioned.

Each search result in the aforementioned faceted search perspectives acts as link to an instance page, which aggregates information about the particular instance (e.g. an interview or a place) and offers links to related instances in WarSampo and Wikidata. These pages offer the end-user the possibility to learn more about e.g. the military units, persons or locations mentioned.

For the purposes of interactive viewing of interview videos, the Sampo-UI framework's instance page component was extended for embedded videos with dynamic table of contents functionalities using the YouTube IFrame player API[12]. Figure 2 shows an instance page of an interview. On the top left there is an embedded YouTube video player, while on the bottom left there is metadata related to the interview. On the right there is a dynamic table of contents (TOC) of the interview, which can be used to navigate to a specific segment of the interview video. The time position of the YouTube video is synchronized with the TOC, so that the active segment is expanded automatically. The expanded TOC part contains the interviewer notes as well as the recognized named entities related to the current part of the interview, which are also links to instance pages.

[12]https://developers.google.com/youtube/iframe_api_reference

5. Discussion

WarMemoirSampo is the outcome of a collaboration between Linked Open Data and NLP techniques. Its backend capitalizes on various language-processing tools in order to handle texts and extract information from them, while the result of this process, a linked data knowledge graph, is able to store the information and swiftly deliver relevant parts to a faceted search-based user interface. The resulting WarMemoirSampo Portal offers the general public as well as specialists a way to grasp the contents of the interviews through keywords and named entities and the means to easily search and access relevant parts of the videos.

The portal is also provided with a feedback button so that the developers can gain insights based on user experience, since the public did not have access to it during development. Nevertheless, more features are planned for the future: the links to the entities will be integrated into the text instead of being listed separately. An event detection tool is to be developed which extracts event entities based on times and places mentioned in the interviews. Moreover, when Finnish speech-to-text technology advances to the point that everyday dialectal speech can be reliably transcribed, the same tools could be used on these new transcriptions.

Acknowledgements Ilpo Murtovaara provided the original transcripts of the videos and was influential in filming the videos. Kare Salonvaara edited the videos. Tammenlehvän Perinneliitto ry funded our work. CSC – IT Center for Science provided computational resources.

References

[1] E. Hyvönen, Digital humanities on the semantic web: Sampo model and portal series, Semantic Web — Interoperability, Usability, Applicability (2022). Submitted.

[2] M. Koho, E. Ikkala, P. Leskinen, M. Tamper, J. Tuominen, E. Hyvönen, Warsampo knowledge graph: Finland in the second world war as linked open data, Semantic Web – Interoperability, Usability, Applicability 12 (2021) 265–278. doi:10.3233/SW-200392.

[3] J. Hunter, R. Iannella, The application of metadata standards to video indexing, in: C. Nikolaou, C. Stephanidis (Eds.), Research and Advanced Technology for Digital Libraries, Springer Berlin Heidelberg, Berlin, Heidelberg, 1998, pp. 135–156.

[4] C. Ribeiro, M. L. Mucheroni, Dynamic indexation in video metadata, Procedia-Social and Behavioral Sciences 73 (2013) 551–555.

[5] J. L. Martinez-Rodriguez, A. Hogan, I. Lopez-Arevalo, Information extraction meets the semantic web: A survey, Semantic Web Journal 11 (2020) 255–335.

[6] H. Q. Yu, C. Pedrinaci, S. Dietze, J. Domingue, Using Linked Data to annotate and search educational video resources for supporting distance learning, IEEE Transactions on Learning Technologies 5 (2012) 130–142.

[7] O. Suominen, Annif: DIY automated subject indexing using multiple algorithms, LIBER Quarterly 29 (2019) 1–25. doi:10.18352/lq.10285.

[8] J. Luoma, M. Oinonen, M. Pyykönen, V. Laippala, S. Pyysalo, A broad-coverage corpus for Finnish named entity recognition, in: Proceedings of the 12th Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 4615–4624.

[9] R. Leal, J. Kesäniemi, M. Koho, E. Hyvönen, Relevance Feedback Search Based on Automatic Annotation and Classification of Texts, in: D. Gromann, G. Sérasset, T. Declerck, J. P. McCrae, J. Gracia, J. Bosque-Gil, F. Bobillo, B. Heinisch (Eds.), 3rd Conference on Language, Data and Knowledge (LDK 2021), Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2021, pp. 18:1–18:15. doi:10.4230/OASIcs.LDK.2021.18.

[10] J. Kanerva, F. Ginter, N. Miekka, A. Leino, T. Salakoski, 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, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 133–142. URL: https://aclanthology.org/K18-2013. doi:10.18653/v1/K18-2013.

[11] J. Kanerva, F. Ginter, T. Salakoski, Universal lemmatizer: A sequence to sequence model for lemmatizing universal dependencies treebanks, Natural Language Engineering (2020) 1–30. doi:10.1017/S1351324920000224.

[12] A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter, S. Pyysalo, Multilingual is not enough: Bert for Finnish, 2019. arXiv:1912.07076.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1, Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.

[14] M. Hämäläinen, UralicNLP: An NLP library for Uralic languages, Journal of Open Source Software 4 (2019) 1345. doi:10.21105/joss.01345.

[15] M. Koho, R. Leal, E. Ikkala, M. Tamper, H. Rantala, E. Hyvönen, Building lightweight ontologies for faceted search with named entity recognition: Case WarMemoirSampo, in: International Workshop on Knowledge Graph Generation from Text (TEXT2KG 2022), Proceedings, Springer, 2022. Forthcoming.

[16] M. Tamper, A. Oksanen, J. Tuominen, A. Hietanen, E. Hyvönen, Automatic Annotation Service APPI: Named Entity Linking in Legal Domain, in: The Semantic Web: ESWC 2020 Satellite Events, Springer-Verlag, 2020, pp. 110–114.

[17] M. Tamper, E. Hyvönen, P. Leskinen, 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-Verlag, 2019. Forthcoming.

[18] E. Mäkelä, Combining a REST Lexical Analysis Web Service with SPARQL for Mashup Semantic Annotation from Text, in: The Semantic Web: ESWC 2014 Satellite Events, Springer International Publishing, 2014, pp. 424–428. doi:10.1007/978-3-319-11955-7_60.

[19] E. Heino, M. Tamper, E. Mäkelä, P. Leskinen, E. Ikkala, J. Tuominen, M. Koho, E. Hyvönen, Named entity linking in a complex domain: Case second world war history, in: J. Gracia, F. Bond, J. P. McCrae, P. Buitelaar, C. Chiarcos, S. Hellmann (Eds.), Language, Data, and Knowledge (LDK 2017), Springer International Publishing, Cham, 2017, pp. 120–133. doi:10.1007/978-3-319-59888-8_10.

[20] E. Ikkala, E. Hyvönen, H. Rantala, M. Koho, Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces, Semantic Web – Interoperability, Usability, Applicability 13 (2022) 69–84. doi:10.3233/SW-210428.