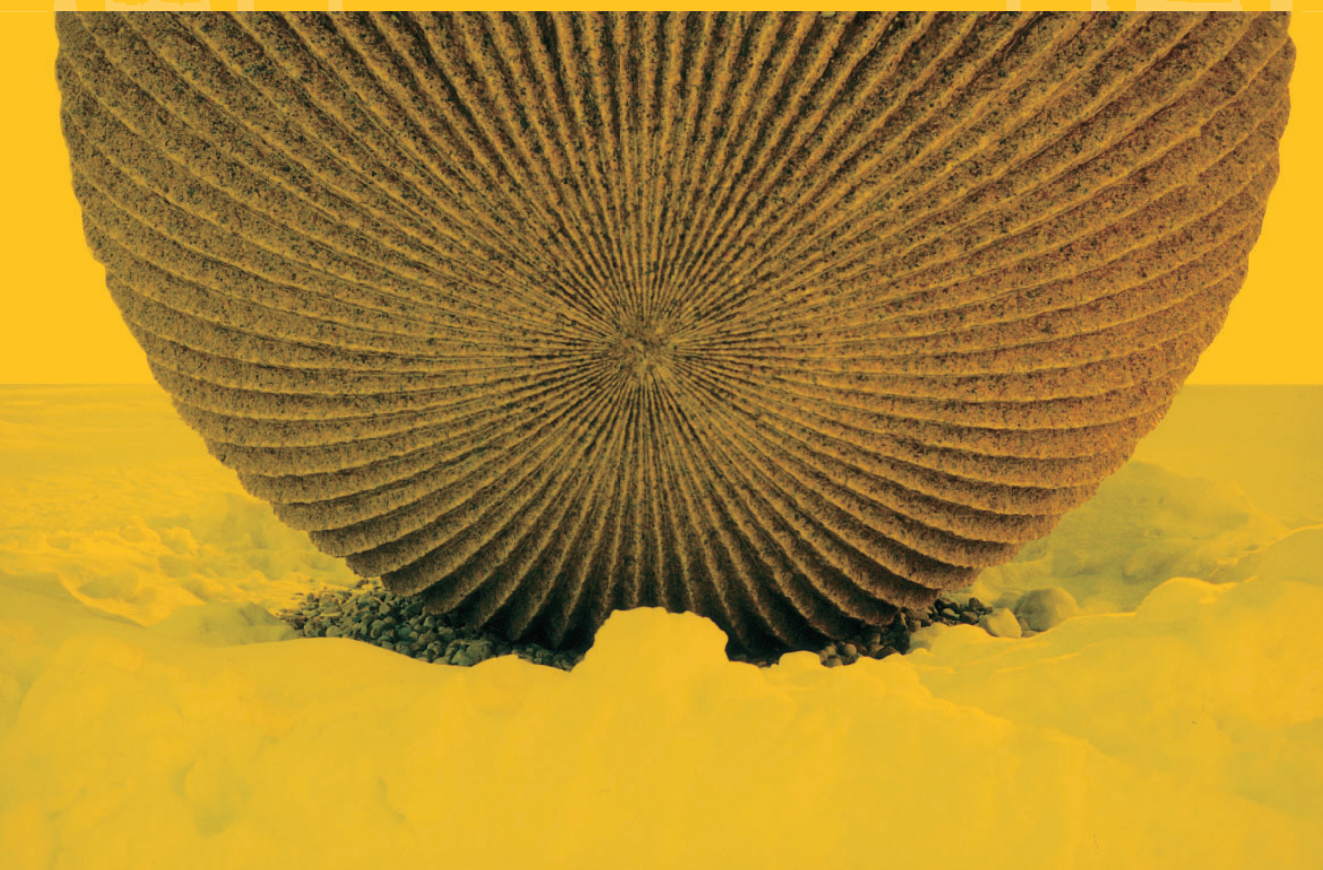
Studia humaniora ouluensia

Jarmo Harri Jantunen, Sisko Brunni, Niina Kunnas, Santeri Palviainen and Katja Västi (eds)

PROCEEDINGS OF THE RESEARCH DATA AND HUMANITIES (RDHUM) 2019 CONFERENCE: DATA, METHODS AND TOOLS



Jarmo Harri Jantunen, Sisko Brunni, Niina Kunnas, Santeri Palviainen and Katja Västi (eds)

PROCEEDINGS OF THE RESEARCH DATA AND HUMANITIES (RDHUM) 2019 CONFERENCE: DATA, METHODS AND TOOLS

Studia humaniora ouluensia 17

Editor-in-chief: Santeri Palviainen

Publishing office and distribution:
Faculty of Humanities
Linnanmaa
P.O. Box 1000
90014 University of Oulu
Finland

ISBN: 978-952-62-2320-9
ISSN: 1796-4725

Electronic version:
ISBN: 978-952-62-2321-6

Cover Design: Raimo Ahonen

OULU 2019

Preface

RDHum 2019, the Research Data and Humanities Conference, takes place August 14–16, 2019 at the University of Oulu, Finland. RDHum 2019 is jointly organised by the University of Oulu and the University of Jyväskylä, in collaboration with FIN-CLARIN and The Language Bank of Finland. The event is the first in the series of conferences taking place biennially in one of the universities within the FIN-CLARIN Consortium. The first RDHum Conference is hosted by the University of Oulu, where the Oulu Corpus, a comprehensive and widely used digital research resource at the time, was collected and compiled in a project led by professor Pauli Saukkonen 50 years ago.

Digital resources and technology are used more and more within the humanities and the social sciences. Researchers in digital humanities gather, administer, share and study rapidly accumulating digital resources. They also need various research methods and tools in analysing these resources. The conference Research Data and Humanities gathers researchers around these themes, and the scientific program of the Conference includes numerous topics related to digital data, digital methods and analysis in the Humanities. In this first Conference, the subjects of the presentations, posters and workshops come from several disciplines, such as linguistics, literary studies, computer science and information science. Thus the languages and societal phenomena under study, data and methods vary widely in the conference.

The peer reviewed articles published in these proceedings are grouped into three categories according to their main focus: data, methods and tools. New data and corpora are presented in the following papers: Kurki et al. present Digilang, a joint venture to combine six different digital corpora. The corpora represent different kinds of data in various modalities. Ijaz seeks to determine editions analytically from bibliographic metadata. Lahti et al. describe the use of bibliograpghic data science in the study of bibliographic metadata collections. Pääkkönen presents challenges the end-user face with digital presentation systems and discusses the issues relating to metadata. Salonen et al. describe the collection and process of establishing the Corpus of Finlands Sign Language. They a lso discuss the storage, metadata and publication of the corpus. Jauhiainen presents Wanca in Korp, a sentence corpus for under-resourced Uralic languages and the process how the corpus was collected.

New methods for digital humanities are presented in the following papers: Laippala in her paper discusses how to classify texts collected from the internet by

means of automatic identification. Ryynänen and Hyyryläinen analyze the concept of Digital Humanities and propose a concept of "practical digital humanities" for describing research utilising a humanist approach to practical problem solving with digital technology development in the digital humanities context. Mikhailov compares texts by their frequency lists. He uses two different types of frequency word lists, unlemmatized and lemmatized, to conduct an experiment with. He observes the different outcomes of the two lists in the experiment. Ivaska presents an analysis of machine learning to identifying translated and non-translated Finnish texts and how to identify the source language of the translated text. Drobac and Linden discuss the issues relating to optical character recognition (OCR) in historical newspaper and journal text and assert that font families need to be recognized. They present an experiment relating to recognizing text in two different fonts. Cohrs and Petersen propose experimental methods of guessing a persons political party based on his tweets. Ijaz presents possibilities of analytical determination of editions from bibliographic metadata. Pääkkönen, Kettunen and Kervinen discuss findings made from user observations in searching digitized serial publications

The following papers introduce new tools in digital humanities: Kettunen presents an analysis of semantic annotation of texts in the context of other automated tools for analyzing languages. He introduces a new tool, FiST, that has been developed to annotate semantically texts in Finnish. Huttunen describes digital games in reinforcing linguistic and socioemotional skills of children with communicative disabilities. She describes the properties of two versions of the game Tunne-etsivät and collection of research data from the users of the game.

We gratefully acknowledge the financial and technical support from The Federation of Finnish Learned Societies, FIN-CLARIN consortium, The Language Bank of Finland, and the Universities of Oulu and Jyväskylä, and the city of Oulu, which made this event possible. We would also like to thank all the members of the FIN-CLARIN steering group, the members scientific and organising committees and the local students in the University of Oulu who encouraged to organize this event and worked hard to make this conference a reality. Finally we wish to thank all reviewers for their work and the Faculty of Humanities for agreeing to publish proceedings in Studia Humaniora Ouluensia.

Jarmo Harri Jantunen (chair), Sisko Brunni, Niina Kunnas, Santeri Palviainen and Katja Västi

Table of contents

I Data

Analytical determination of editions from bibliographic metadata

Ali Zeeshan Ijaz (University of Turku), Mikko Tolonen (University of Turku), Leo Lahti (University of Helsinki) and Iiro Tiihonen (University of Turku)

Abstract
Analytical bibliography aims to understand the production of books. Systematic methods can be used to determine an overall view of the publication history. In this paper, we present the state of the art analytical approach towards the determination of editions using the ESTC meta data. The preliminary results illustrate that metadata cleanup and analysis can provide opportunities for edition determination. This would significantly help projects aiming to do large scale text mining.

1 Introduction

Analytical bibliography studies books as material objects and aims to understand how they were produced (Tanselle, 1977). Large scale library catalogues, where millions of documents have been cataloged, can be utilized to develop an automated framework that can highlight book publication and production (Tolonen et al., 2015; Lahti et al., 2019). Systematic methods can provide a more thorough view of the publishing history (Eliot and Rose, 2009;Tolonen et al. 2018). At the early times of book printing, publishers could reprint the same book several times based on the public reception to the literary work. Additionally, the expiration of the Licensing Act of 1695 and the resultant increase in book production meant that some publishers resorted to duplicating their own issues. The term “edition” was then used for successive issues as means to keep on selling the same work (Todd, 1951). Furthermore, by the 18th century, the octavo format became the most popular for books, paving the way for cheaper production (Lahti et al., 2019). For these reasons, the number of editions and their ordering are integral to understanding book production (Howsam, 2014).

Contemporary text mining approaches generally ignore edition level information, or provide generic solutions that may omit important details. Projects such as “Commonplace Cultures” (Morrissey, 2016) have performed large-scale text mining of the Eighteenth Century Collections Online (ECCO), a collection of books printed in the United Kingdom during the 18th century. “BookSampo”, a semantic portal which uses the FRBRoo ontology (Riva et al., 2008), covers

metadata on Finnish fiction literature, though only at the work level (Mäkelä et al., 2011). Furthermore, standard algorithms used in this field, such as latent Dirichlet allocation are time agnostic (Blei et al., 2003), and while later algorithms have become time aware, they only focus on topics. Hence, such methods may fall short in contextualizing historical developments in book printing and publishing.

2 Background

According to Joseph Dane, printed books are not individual objects but are in fact, instances of an historical process that produces similar members of a defined collection (Dane, 2016). Hence contextualizing their production chronologically and in relation to their close relatives is more important than just viewing each book in isolation.

Considering how fundamentally important editions are for this research work, it is essential to determine what is meant by an edition of a book and the concepts related to it. Fredson Bowers defined editions as the combined number of books printed from the same type-pages and therefore includes both the issues and variants. In this manner, issues were a form of editions put on sale by the publisher at planned times, while variants represented alterations in a book but with no change in the title page (Bowers, 1994).

Libraries have begun moving from the traditional print oriented model to having digitization projects being conducted under them. For many digital humanities projects, the library's source material serves as the groundwork on which research is conducted (Cunningham, 2010). This is especially true for catalogs and bibliographic metadata that various libraries hold, which become essential tools in information science and provide avenues for significant research work (Lahti et al., 2019).

Historically, the hand press era, circa 1450 to 1825, had a starkly different means of production of books as compared to the more recent publishing eras. Technologically, this time period saw little change in the method of book production. However, the more specific requirements of cataloging books from this era necessitated a modification of cataloging rules developed mainly for modern publication of books. This required much more sophisticated record keeping, where identification of an edition is no small task. Nonetheless, several new editions were discovered when more comprehensive bibliographic metadata were utilized (Snyder and Hutchinson, 2014).

Analysis of large textual datasets such as ECCO conducted under projects such as Commonplace cultures project tends to be incomplete as only the earliest edition information is utilized. This significantly hampers the ability to properly contextualize and analyze book production, as the patterns in the historical production of printed books is as important as the written text itself. Furthermore, this resulted in a significant reduction in usable data, as later editions were discarded. Coverage wise, ECCO consists of around 200,000 volumes of texts from 18th century, while bibliographic metadata such as ESTC provides information for around 480,000 titles, from 14th to 18th century.

Hence bibliographic metadata provides a great opportunity for pattern discovery, the ability to determine editions and works, and exploratory analysis to complement earlier studies in the history of knowledge production. Unfortunately, metadata cannot be readily used for research due to remarkable shortcomings in the raw data quality, and has to be first properly harmonized. The quality and overall value of metadata can be significantly improved by transforming it into a more standardized format. Bibliographic data science is a field that aims to produce high quality, consistent and improved metadata for further analysis (Lahti et al., 2019). Given that edition information is crucial for determining patterns in historical book publication and production, we have developed an automated pipeline that can process and clean up the raw metadata into a more usable format in order to determine edition level information for various works of authors. This provides the foundation for exploratory analysis and elucidating significant patterns in book production.

3 Material and Methods

The English Short Title Catalogue (ESTC) provides a wealth of knowledge concerning the books published in the early modern period. Nonetheless, it uses the Machine Readable Cataloging standard (MARC, 1999), where the raw data is unsuitable for research, due to the presence of spurious and erroneous information, as well as differences in standards and languages (Nilsson, 2010; Lahti et al., 2019). The ESTC contains metadata for more than 460,000 documents, covering the hand press era (1470-1800). It contains various fields that can be used to determine each book uniquely, while providing rich details on the editions.

For this work in progress, our aim has been to develop a pipeline that can be used to determine the chronological ordering of editions of various books in the ESTC metadata. Before such ordering is determined, it is imperative to describe

the “work” under which these editions of books are found. Considering the difficulty in determining the definition of work that is widely applicable, we opted for a collection oriented approach. Here, books of similar titles are collected in a collection, henceforth known as the work field. Discriminating information such as publication date, edition number (if available), publisher details, and more can be used to determine more precise edition information, as well as to further subset the collection as per requirement of the end users. The harmonization process begins with selecting the edition field, and other supporting metadata fields that are needed during the harmonization. The “edition statement”, for instance, provides the edition number as well as fields related to book title, publisher information, date of publication and more. Additionally, this work is a part of a larger project, which includes several collaborators who use the harmonized ESTC metadata for various research purposes (Lahti et al., 2019; Tolonen et al., 2018).

The raw data is harmonized in an iterative and progressive manner, with various processing steps performed to handle certain aspects of the harmonization process. This is done on a per author basis to ensure consistency and minimize any errors, by linking titles to their respective authors. Harmonization is then performed for each author individually. Figure 1 illustrates the flow diagram for the harmonization process, while the step by step details are described below.

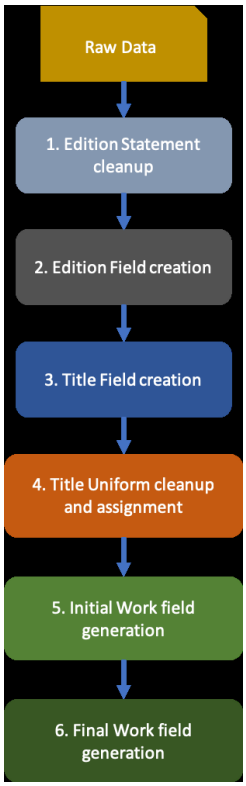


Fig. 1. Flow diagram for the harmonization process.

The process begins with the "edition statement" field[1], which is processed to determine the edition number of the book. Currently, only a small subset of the whole of ESTC metadata contains information in this field. Hence the publication date is combined with the edition number, if available, to create a new edition field. This provided the ability to distinguish various editions on chronological basis.

The book titles[2] are cleaned up, removing any unwanted characters and non-important words. Furthermore, the "title uniform" field[3], which provides a representative title for the work undergoes a similar cleanup as well. An initial work

[1] MARC Field 250a
[2] MARC Fields 245a and 245b
[3] MARC Field 240a

field is then generated by combining the author identifier with the harmonized title. This provided a unique work field identifier for each book. A final work field is then created by combining similar titles into a collection. We have created custom algorithms in order to combine similar titles, by assigning a represetative title of the collection, alongside the author identifier. Source code will be accessible via the COMHIS homepage.

At the moment, the algorithms employed in the harmonization process as well as the creation of collections are being worked upon to improve their applicability across different genres in ESTC metadata. Considering the scale of ESTC, during the initial development and tuning of the pipeline, a small subset of works of 7 popular authors was selected. The list of authors included William Shakespeare, David Hume, Jonathan Swift, John Locke, Isaac Watts, Alexander Pope and Daniel Defoe. The harmonized entries were manually checked to determine if the titles were being properly assigned the correct work field. Corrective steps were then taken to improve the performance of the harmonization process, by tuning the various methods and parameters of the underlying algorithms.

However, the performance and reliability of the newly designed harmonization and analysis algorithms need to be validated against a known ground truth. Hence, such gold standard was created as follows:

Constructing the Gold Standard

- 250 authors were randomly selected from ESTC, with varying title counts. Additionally, records from seven prolific authors were also selected.
- All records with the same content were carefully inspected and assigned to works.
- The main fields we used for this assignment were Title Uniform (MARC field 240a), Title statement and Remainder of title (245a and 245b). We also consulted the MARC fields for page count, physical extent, edition statement, general note and genre.
- We encountered several issues including spelling mistakes or word replacements in the titles, which we combined into a single work.
- Another issue was the handling of different or multiple volumes of the same work. Although a work could exist in the ESTC as a multi-volume book, each of the volumes could also be listed separately. We resolved this

issue by creating an additional layer, where all the different volumes were combined into a single work.

- A layer was added for handling calendars and music performance handouts.
- A collection oriented layer was added, which described if the book was a collection of other works. In most cases these contain reprints of earlier works; separating them from other material reduces redundancy.
- We also made a crude genre classification for each record. Altogether, 25.3% of the records have an annotated genre.
- Using formally structured documents, such as meeting minutes, dictionaries or court case reports in creating word embeddings could potentially skew the outcomes.

4 Results

The gold standard would serve as the main evaluation method for the harmonization process. Currently, the system performs well on collecting various titles into a single generic collection for some authors. We evaluated the harmonization process for three different authors in the gold standard in terms of precision and recall against the manually constructed ground truth. The results are illustrated in Figure 2. Additionally, the number of titles for each author are listed in Table 1. We define true positives as those titles that were assigned to the correct work field, false positives as those titles that were incorrectly assigned to a work field and lastly, false negatives as those titles who were not assigned to the correct work field and instead had their own work field generated.

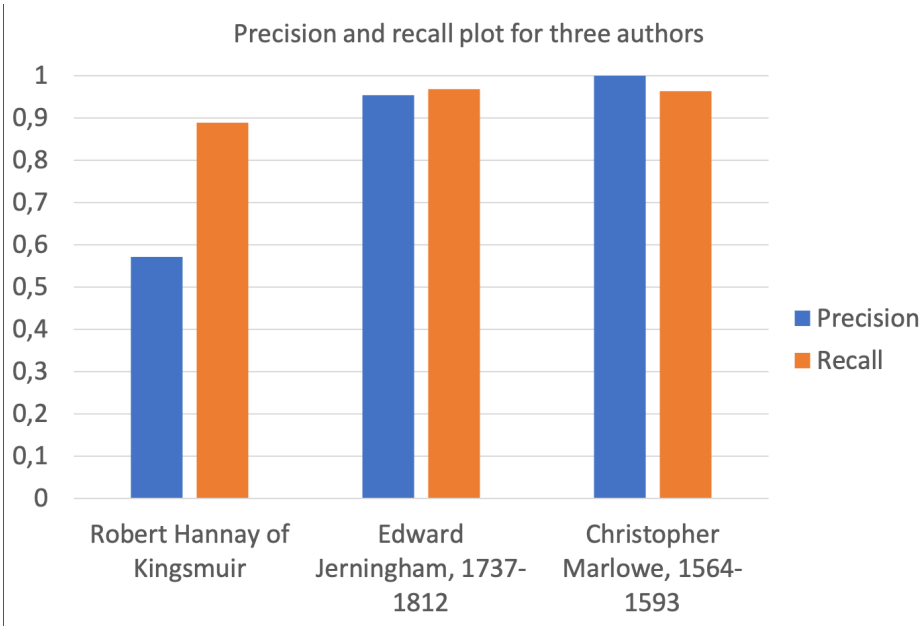


Fig. 2. Precision and recall for three authors.

Table 1. Number of titles for three authors.

Author	Number of Titles
Robert Hannay of Kingsmuir	15
Edward Jerningham, 1737-1812	66
Christopher Marlowe, 1564-1593	55

The results illustrate that the system is capable of assigning titles to the correct work fields in most of the cases, with an overall high recall. For the authors Edward Jerningham and Christopher Marlowe, the precision was at 0.95 and 1 respectively. However, it suffered in the case of Robert Hannay of Kingsmuir, where it decreased to 0.57. Recall was 0.89 for Robert Hannay of Kingsmuir, 0.97 for Edward Jerningham and 0.96 for Christopher Marlowe. In general, titles that differ by a few words increase the chances of false positives. While on the other hand, longer titles tend to reduce this possibility.

In less polished data, both precision and recall would greatly suffer as there are no direct means of assigning each title to a correct work field, unless the titles are exactly the same. This is confounded by the fact that the differences in title length, variations in the title themselves, spelling mistakes and more, complicates the issue further. Hence, developing a work field without such harmonization techniques entails manually assigning each title to a work, a significantly labor-intensive task.

The next steps would include expanding this proof-of-concept study from the few selected authors to the complete ESTC collection.

5 Conclusions

Considering the scope of the ESTC, spelling variations and different styles of writing, we opted for a collection-oriented approach towards the work field. As the work is currently in development, the first version of the pipeline was developed using the dataset of popular authors. Nonetheless, the results illustrate the applicability of this work towards the research goal of determining a correct ordering for editions.

There are various issues that still need to be addressed. The collection work field is only meant to ensure that similar titles are collected into the same collection. This ignores more detailed information such as volumes, or titles that may be belonging to different works, depending on how work is defined. This will necessitate finer grained work field to be generated from the current collection-oriented work field for downstream analysis. Considering that discriminating information is available for each book title, the collection-oriented work field provides a suitable starting point for such analysis.

Future steps will include improving the harmonization process for other types of textual artefacts, such as pamphlets and other document types. Furthermore, generation of sub-work fields, so as to develop more specific datasets required for downstream statistical analysis would also be done. Finally, statistical analysis on the harmonized dataset can be used to determine edition ordering. This would enable determination of editions in a more precise manner for each work in a chronological fashion.

As the development on the harmonization process moves forward, the overarching aim has been to provide an automatic and reproducible system for harmonization and analysis of the ESTC metadata. This research work enhances the overall analysis via investigating the harmonization and cleanup of the edition field.

Large scale text mining projects would significantly benefit from this, as combining the harmonized metadata with text mining of large data sets such as ECCO would lead to a finer description of what was the first edition. Furthermore, changes between different editions can also be explored. Overall, a more descriptive analysis can then be performed.

Researchers interested in charting the book publication over time of certain authors in the early modern era would benefit from this research work, as well as those interested in exploratory analysis concerning historical book production in general. Libraries would benefit too, as this would grant them the ability to perform better information retrieval and provide contextual information on the books present in their catalog.

References

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3, pp.993-1022.

Bowers, F. (1994). Principles of bibliographical description. St. Paul's Bibliographies. Oak Knoll Press.

Cunningham, L. (2010). The librarian as digital humanist: the collaborative role of the research library in digital humanities projects. Faculty of Information Quarterly, 2(1).

Dane, J. A. (2016). Abstractions of evidence in the study of manuscripts and early printed books. Routledge.

Eliot, S. and Rose, J. (eds.) (2009). A Companion to the History of the Book (Vol. 98). John Wiley & Sons.

ESTC. English Short Title Catalogue. http://estc.bl.uk/ (Accessed 27 November 2018).

Howsam, L. (ed.) (2014). The Cambridge companion to the history of the book. Cambridge University Press.

Lahti, L., Marjanen, J., Roivainen, H., & Tolonen, M. (2019). Bibliographic Data Science and the History of the Book (c. 1500–1800). Cataloging & Classification Quarterly, 1-19.

Mäkelä, E., Hypén, K. and Hyvönen, E., (2011), October. BookSampo—lessons learned in creating a semantic portal for fiction literature. In International Semantic Web Conference (pp. 173–188).Springer, Berlin, Heidelberg.

MARC. (1999). MARC 21 Format for Bibliographic Metadata. https://www.loc.gov/marc/bibliographic/ (Accessed 27 November 2018).

Morrissey, R. (2016). Commonplace Cultures: Mining Shared Passages in the 18th Century using Sequence Alignment and Visual Analytics. Retrieved from https://hcommons.org/deposits/item/hc:12365/

Nilsson, M. (2010). 'From Interoperability to Harmonization in Metadata Standardization: Designing an Evolvable Framework for Metadata Harmonization', Doctoral Thesis, KTH School of Computer Science and Communication. https://www.divaportal.org/smash/get/diva2:369527/FULLTEXT02.pdf (Accessed 27 November 2018).

Riva, P., Doerr, M. and Zumer, M., (2008), August. FRBRoo: enabling a common view of information from memory institutions. In World Library and Information Congress: 74th IFLA General Confrence and Council.

Snyder, H. L., & Hutchinson, H. L. (2014). Cataloging of the hand press: a comparative and analytical study of cataloging rules and formats employed in Europe (Vol. 1). Walter de Gruyter GmbH & Co KG.

Tanselle, G. T. (1977). 'Descriptive Bibliography and Library Cataloguing', Studies in Bibliography, 30: 1-56.

Todd, W. (1951). Bibliography and the Editorial Problem in the Eighteenth Century. Studies in Bibliography, 4, 41-55. Retrieved from http://www.jstor.org/stable/40371090

Tolonen, M., Lahti, L. and Ilomäki, N. (2015). A quantitative study of history in the English shorttitle catalogue (ESTC), 1470-1800. Liber quarterly.

Tolonen, M., Lahti, L., Roivainen, H., & Marjanen, J. (2018). A Quantitative Approach to Book-Printing in Sweden and Finland, 1640–1828. Historical Methods: A Journal of Quantitative and Interdisciplinary History, 1–22.

Wanca in Korp: Text corpora for underresourced Uralic languages

Heidi Jauhiainen, Tommi Jauhiainen and Krister Lindén
University of Helsinki

Abstract

This paper introduces "Wanca in Korp", a set of sentence corpora for under-resourced Uralic languages, the pipeline used in creation of the corpora, as well as the various tools used as part of the pipeline.

The *Ethnologue* recognizes 38 different Uralic languages. The Uralic language group includes mostly linguistically under-resourced languages and only three of the Uralic languages are used as national majority languages: Hungarian, Finnish, and Estonian. We have chosen all the minority languages of the Uralic branch as languages of interest and created new sentence corpora for most of them using texts openly available on the internet.

For gathering the texts from the internet, we used an open-source web-crawling software, Heritrix, developed and used by the Internet Archive in cooperation with several National Libraries. In addition to conducting our own crawling, we also used the pre-crawled corpus distributed by the Common Crawl Foundation.

In order to determine which pages were written in one or more of the relevant languages, we used a state-of-the-art language identification software developed within the project. For post-processing the identified pages, we developed a web-service, Wanca, where experts and native speakers of each language can participate in manual curation of the crawled links.

The process of sentence corpora creation begins from the pages tagged with relevant languages in Wanca. All the texts available behind existing Wanca links are downloaded. A language set identifier is used for whole texts in order to verify that the downloaded web pages still include texts in relevant languages. Then complete sentences are extracted from the texts. The language of each sentence is again identified and the sentence added to the sentence collection of the respective language. For the most rare languages, a manual curation and additional manual processing is conducted for those pages where possible sentences were found.

The end product is a sentence corpus collection for 28 less-resourced Uralic languages ranging in size from 20 sentences of Vod to 214,226 sentences of North Saami. The corpus is available in the Language Bank of Finland. The work has

been conducted within the Kone Foundation funded project "The Finno-Ugric Languages and The Internet" at the University of Helsinki as part of FIN-CLARIN.

1 Introduction

In this paper, we present the work done in creation of a set of sentence corpora—"Wanca, Korp version"[1]—for under-resourced Uralic languages. The end product will be new sentence corpora for 28 less-resourced Uralic languages using texts openly available on the internet. For gathering the texts from the internet, we used an open-source web-crawling software, Heritrix (Mohr *et al.* 2004), developed and used by the Internet Archive.[2] In addition to conducting our own crawling, we used the pre-crawled corpus distributed by the Common Crawl Foundation.[3]

In order to find pages written in the relevant languages, we used a state-of-the-art language identification software developed within the project. For post-processing the identified pages, we developed a web-service, Wanca[4], where experts and native speakers of each language can participate in manual curation of the crawled links.

The process of sentence corpora creation begins from the links tagged with relevant languages in Wanca. A language identifier is used to make certain that the pages still include texts in relevant languages. The texts pass through an intricate workflow where sentences are extracted from them. The language of each extracted sentence is automatically identified and the sentence added to the sentence collection of the respective language. For the most rare languages, a manual curation and additional manual processing is performed.

The work has been conducted within the project "Finno-Ugric Languages and The Internet" at the University of Helsinki from 2013 to 2019. The project was funded from the Kone Foundation Language Programme (Kone Foundation 2012) and was part of FIN-CLARIN.

This paper is divided into four parts. The first part describes some related previous work. The second part is a description of what we have done in the project leading up to and including the Wanca website. In the third part, we introduce the

[1] http://urn.fi/urn:nbn:fi:lb-2019052401
[2] https://archive.org
[3] http://commoncrawl.org
[4] http://suki.ling.helsinki.fi/wanca/

workflow for creating sentence corpora using the Wanca link collections as the foundation. The fourth part is a short description of the corpora, conclusions, and suggestions for future work.

2 Previous work

As we are not experts in Uralic languages, we have used the *Ethnologue* (Simons & Fennig 2018) as our source for the division of Uralic languages and the ISO-639-3 standard (SIL 2013) as our guide to which are considered languages of their own. At the moment, the *Ethnologue* recognizes 38 different Uralic languages. We have chosen all the minority languages of the Uralic branch as languages of interest.[5]

2.1 Text corpora for under-resourced Uralic languages

Text corpus creation for under-resourced Uralic languages has been considered earlier and also concurrently with our efforts. Suihkonen (1998) documents the corpora for Uralic languages available at the University of Helsinki at the time. Novák (2008) presents resources created during several projects for Udmurt, Komi-Zyrian, Mari, Mansi, Khanty, Nenets, and Nganasan. Endrédi *et al.* (2010) describe a corpus they used for creating resources for Nganasan. Prószéky (2011) presents a workshop where corpora for some Uralic languages were discussed. Jokinen (2014) describes efforts to improve North Saami Wikipedia, an important text source for the language. Vincze *et al.* (2015) describe the FinUgRevita project aiming to develop language technology tools for Udmurt and Mansi. Arkhangelskiy & Medvedeva (2016) describe methods to create morphologically annotated corpora for minority languages used in Russia, among them the Udmurt language. Simon & Mus (2017) give preliminary results for building a linguistic corpus for some Uralic languages. Arkhangelskiy (2019) describes corpora of social media texts for minority Uralic languages, as well as the pipeline used to develop the corpora.

[5] See Table 1 where all language names are followed by their corresponding ISO 639-3 codes.

2.2 Creating language specific corpora from web-pages

Many scholars have searched the web in order to build text corpora in various languages. Early builders of web corpora queried for search engine results; however, nowadays using a web crawler is more common (Kristoffersen 2017, 11). The way the language of a page considered for the corpus is verified varies. Schäfer *et al.* (2014) used language identification as a pre-processing step when trying to find pages in specific languages to be used as seeds for a crawl. Baykan *et al.* (2008) extracted words from URLs and used various machine learning algorithms to distinguish pages in different languages from each other already before downloading them.

The metadata of the page to be downloaded has also been used for determining the language of that page. Priyatam *et al.* (2012) used the metadata in the HTML tags and in the URLs while Somboonviwat *et al.* (2005) used only the encoding information of the page. A language identifier can be used to complement the information on the encoding (Tamura *et al.* 2007, Mon & Mikami 2010, Mon *et al.* 2011).

Language of the page has also been identified during crawling by Medelyan *et al.* (2006), Suchomel & Pomikálek (2012), and Barbaresi (2013b,a), and as a post-processing step by Ghani *et al.* (2001), Boleda *et al.* (2006), Kornai *et al.* (2006), Baroni & Kilgariff (2006), Ferraresi *et al.* (2008), Baroni *et al.* (2009), Emerson & O'Neil (2006), and Pomikálek *et al.* (2009).

2.3 Sentence boundary disambiguation

Sentence boundary disambiguation is not a trivial undertaking as the most common sentence boundary marker period '.' is also used for various other tasks (Palmer & Hearst 1994, Kiss & Strunk 2006). The use of period with numbers is not problematic to detect, but abbreviations are more language specific and often the list of them is long (Grefenstette & Tapanainen 1994). Many NLP systems use approaches that are build to a specific text and use hand-made lists of, for example, abbreviations (Palmer & Hearst 1994). In order to find sentence boundaries in an unannotated corpus, an unsupervised method is needed for detecting the abbreviations.

Grefenstette & Tapanainen (1994) introduced a set of abbreviation guessing rules for English. Mikheev (2000, 2002) used abbreviation guessing heuristics

similar to the ones proposed by Grefenstette & Tapanainen (1994). We used his heuristics as part of our pipeline and they are introduced in Section 1.4.5. He collected a list of known abbreviations and used the list to tag tokens as abbreviations in other corpora. In his heuristics, each word containing at most four letters is considered a possible abbreviation if it is followed by a period but is not yet on the list. To make a decision, frequencies of the token in ambiguous and unambiguous situations are used with and without the previous token. Kiss & Strunk (2002a,b, 2006) presented an unsupervised method for sentence boundary disambiguation, which relies on word collocation statistics. Every token ending in a period is considered a possible abbreviation and the number of occurrences of the token with and without the period is calculated from the corpus/document to be processed.

2.4 Language identification

Language identification is the task of determining the language in which a text, of any length, is written. Automatic language identification of digital text has been researched for over 50 years. Language identification can be considered a subspecies of general text categorization and most of the methods that can be used in categorizing text according to their topic can also be used for language identification. The size of the text where the language should be identified can vary from parts of words to complete books. In language identification of monolingual texts, a text is tagged with the language best fitting for the whole text. In language set identification, the set of languages used in a given piece of text is identified. In multilingual segmentation by language, the exact parts of text for each language are identified. As part of our project, we were involved in writing a comprehensive survey article on language identification (Jauhiainen *et al.* 2018d).

3 The Finno-Ugric Languages and the Internet project

The Kone Foundation funded "Finno-Ugric Languages and The Internet" project[6] was active from 2013 to 2019. One of the main goals of the project was to collect

[6] http://suki.ling.helsinki.fi/eng/project.html

texts written in under-resourced Uralic languages from the internet (Jauhiainen *et al.* 2015a). During the project, we had several periods of intensive web harvesting and downloaded hundreds of millions of web pages. While crawling, we identified the language of the page and afterwards the languages used in the downloaded web pages were re-identified using a language set identification method developed within the project. The links to the Uralic pages found during harvesting were published in a web service called Wanca, where experts of the languages in question could curate the links.

3.1 Language identification of Uralic web pages

In order to find out which of the downloaded pages are written in one of the under-resourced Uralic languages, we use a language identifier developed within the project.[7] Currently the language identifier in production can distinguish between c. 400 languages and dialects. The language identifier implements a state-of-the-art language identification method, HeLI, originally developed by Jauhiainen (2010). We have continued to develop the method within the project and the HeLI method has proven to be very robust (Jauhiainen *et al.* 2017b) and has fared very well in several shared tasks for distinguishing between close languages (Jauhiainen *et al.* 2015b, 2016, 2017a, 2018a,b,c, 2019).

At the beginning of the project we were able to find suitable training material for 34 of the 38 Uralic languages presented in Table 1.[8] Hungarian, Finnish, and Estonian are not considered to be minority languages and thus we have 31 languages of interest, which we refer to as the relevant languages in this paper. Some of these languages do not have any official written form and the writers of these languages use several more or less different orthographies. In the beginning of the project, we chose only one orthography per language for training the language identifier. Later, we added some additional orthographies, e.g. the different ways of writing the non-ASCII vowels in the South Saami language used in Sweden and Norway.

In order to cope with multilingual documents, we developed a new language set identification method (Jauhiainen *et al.* 2015c). The basic idea of the method is

[7] https://github.com/tosaja/HeLI

[8] No digitally encoded texts were found for Akkala, Ter, and Pite Saami languages nor for the Kamas language.

to slide a window of a certain size through the document in steps of one or more bytes. The text in each window is sent to the language identifier, which gives the most likely language for the window. There is a variable storing the current language and if enough consecutive identifications have given a differing language for the window, the current language is changed. The document is given a label for each language that was set as the current language at some point.

For the purposes of the project, we have set up two language identification servers. One server, HeLI, implements the language identification for monolingual documents and is used when fast language identification is needed during the web crawling. The other server, MultiLI, implements the language set identification method. When a piece of text is sent to one of the servers, they return either the language or the set of languages for the text.

3.2 Harvesting the web

In order to crawl for pages written in small Uralic languages, we use Heritrix (Mohr *et al.* 2004), an open source web archiving system developed by the Internet Archive.[9] Heritrix is the outcome of many years of development by the Internet Archive and it is still being maintained. It was a product of cooperation between Internet Archive and the Nordic national libraries and it is still used by several national libraries around the world to collect national web archives. It has also been successfully used for collecting similar corpora to ours by Baroni & Kilgariff (2006), Baroni *et al.* (2009) and Pomikálek *et al.* (2009).

The goal of the Internet Archive is to archive the web sites as usable collections for future generations. In this project, we are only interested in collecting the text content of the pages written using one of the relevant Uralic languages. The version of Heritrix we are currently using downloads all text files as well as pdf files it finds. We have made some custom changes to the code of the crawler. From each page downloaded, the HTML code is removed and up to three excerpts of 100 characters are sent to HeLI. If the identified language of at least one of the excerpts is one of the Uralic languages we are interested in, the whole text is, furthermore, sent to be identified. If this identification still indicates a relevant Uralic language,

[9] http://www.archive.org

the whole text of the page is archived. The links found on such pages are given precedence over links from other pages in the frontier queue of the crawler.

We chose to start collecting the material by crawling the national domains most likely to contain material written in the relevant Uralic languages, i.e. .ee, .fi, .hu, .lv, .no, .ru, and .se. As seeds, we used the main pages of the universities in the countries in question. Furthermore, we conducted a two month crawl which also included the .com domain in addition to the national domains. As seeds in the .com crawl, we used the pages containing relevant languages found in the previous crawls.

Afterwards, all the texts found during crawling were re-identified with MultiLI. Using the language set identifier, we could better find the pages containing any of the target languages. For each page, we received from MultiLI the set of languages present and their approximate percentages of the text of the whole page. Using that information, we tagged each page with a language using the following heuristics:

1. if more than 9 languages were returned by MultiLi, the page was marked with xxx, indicating an unknown language or junk
2. else
   - if any of the relevant Uralic languages were present
     - any relevant Uralic language which formed at least 2% of the page's text was considered
     - the page was tagged with the relevant Uralic language with most text on the page
   - else the page was tagged with the language with most text on the page.

Later, we downloaded the Common Crawl archive from December 2014.[10] The size of the archive was over 160TB and we initially used HeLI to identify the languages of almost two billion pages and then MultiLi for more precise analysis of the 155.000 texts indicated to be of interest by HeLI. This way we found many new relevant links from outside the national domains that we had crawled ourselves.

[10] http://commoncrawl.org/2015/01/december-2014-crawl-archive-available/

3.3 Wanca

In order to be able to get feedback from those who are more familiar with the relevant Uralic languages, we created a crowd-sourcing platform called Wanca.[11] Wanca contains links to the relevant pages identified during the project. The links lead to the actual pages currently active on the internet and the pages can be viewed together with their respective information in Wanca. Since 2015, thousands of links have either been verified or discarded by us or the experts in the languages in question using the Wanca platform.

4 Pipeline for creating corpora

When creating sentence corpora from the web, one of the greatest challenges we have is that many of the downloaded pages are multilingual. This is especially true for documents containing texts written using the relevant Uralic languages as they are minority languages in the countries where they are used. We needed, hence, to divide the texts into sentences and identify the language of each sentence before adding it to the corpus of one of the languages in question. For this end, we created a pipeline which can be downloaded from our GitHub page[12]. We wanted each step of the workflow to be independent and do only one thing for better quality control and modularity.

4.1 Downloading the relevant pages

Web pages tend to move or disappear over time. In 2016, we wanted to see which of the pages found in 2014 and 2015 were still available. For this end, we used Heritrix to download all the links in Wanca which at that moment were marked as containing text in one of our target languages. The language set identification was then performed on the texts found and Wanca updated accordingly. The unavailable links were, furthermore, hidden from the users of the database. Another re-crawl was done in 2017. When compiling the sentence corpora, we decided to use the downloaded texts from the 2016 re-crawl as it contained more texts than the one from 2017.

[11] http://suki.ling.helsinki.fi/wanca/
[12] https://github.com/uhdigihum/SUKISentencePipeline

4.2 Language set identification for pages

Since 2016, we had further developed the language set identification method in use. Therefore, we decided to start by re-identifying the language of the pages downloaded in the 2016 re-crawl. We started by removing most of the unwanted non-unicode characters as such existed in the downloaded pages. Then we sent the text of each page to MultiLI as one continuous string. For MultiLI, we used the parameters resulting in the most accurate identifications from our earlier research (Jauhiainen et al. 2015c). For a long web page, this means a huge number of identifications which ends up taking a long time. MultiLI utilizes only one computing core when identifying a single text, but can serve several requests simultaneously making better use of the available computing resources. Hence, we decided to divide the documents in the Heritrix produced WARC files into several approximately equally large files and sent them to be identified at the same time.

After receiving the identification results, we redivided the texts into files corresponding to each individual relevant Uralic language and discarded the pages that were not written in any of the target languages.[13] The sentence detector used in this pipeline is currently not able to handle sentences over line borders. In order to minimize further processing, we used this redivision step to remove those lines that did not contain sentences detectable by our sentence detection algorithm: we discarded all those lines which did not contain at least one instance from our end-of-sentence punctuation list (. ? ! . . . : ;) and at least one uppercase letter.

4.3 From collection of texts to collection of lines

We divided the texts into individual lines and removed extra spaces from each line. When encountering any punctuation usually marking the end of a sentence (] . ? ! . . . : ;) between a lowercase letter and an uppercase letter followed by a lowercase letter, as in ’lauantaina.Joka’, we added a space after the punctuation. Such instances were surprisingly common appearing in around 2% of the over two million lines. Some of these instances occurred within email addresses, but most seemed to be just obvious mistakes. Each line was accompanied by a list of all the

[13] This step is not necessary in the pipeline and was only done for quality control purposes.

languages present on the original page according to MultiLI, as well as the URL of that page.

The next phase was the removal of duplicate or near duplicate lines. We created a thumbprint for each line by removing spaces and all non-alphabetic characters except apostrophes from a copy of the line. We generated a list of web addresses that were not available in our 2017 re-crawl and when a duplicate line was found, the URL of the currently stored line was checked against this list of unavailable links. If that page was not available in 2017, the URL and the corresponding line is replaced with the ones of the duplicate. In the end, one thumbprint was represented by only one original line - with numbers, spaces, punctuation, etc. - and its URL. The set of languages corresponding to each thumbprint was the union of all the languages on the lists of its duplicates.

4.4 Language set identification for lines

The lines were then divided into several equally sized files and the lines in each file were sent to MultiLI separately. The list of languages of the original page, or the union of them in case duplicates were found, was also passed to MultiLI. Of the Uralic languages, MultiLI was only allowed to consider the ones on that list. Discriminating between very close languages, such as Tornedalen and Kven Finnish, becomes challenging with very short snippets of text. When classifying such short snippets, we wanted to take the earlier analysis of the larger context into account. The heuristics of choosing the language tags for the line were the same as when identifying the language of a page described above. The lines not identified as containing texts in one of the relevant languages were discarded at this point.

4.5 Making sentences

The next phase is extracting the sentences from the lines. We considered each word of a line at a time. If the word ended in full stop, the word was considered against the abbreviation guessing heuristics presented by Mikheev (2000, 2002). He used heuristics similar to the ones proposed by Grefenstette & Tapanainen (1994) to collect a list of abbreviations from a corpus. Token (space delimited) ending with a period can be guessed to be an abbreviation, if it is:

- a token without vowels, but not all capital letters
- a token consisting of single letters separated by periods

• a token consisting of a single letter.

Furthermore, words with four characters or less that are followed by a period and then by (1) a comma, (2) a lower-cased word, or (3) a number are considered by Mikheev (2000, 2002) to be abbreviations. We were not looking for abbreviations per se but rather the end of a sentence. Hence, we did not consider any tokens ending with a comma but moved on to the next word. However, if the following word did not start with a lowercase letter (possibly preceded by a character (" ” “ «)) and the token in consideration

• ended with full stop but did not adhere to Mikheev’s heuristics (i.e. not an abbreviation)

• ended with other punctuation (? ! ...) followed by zero or more of these characters:) " ” » "

• ended with full stop but did not have a letter before the punctuation character (e.g. word".)

• ended with colon (:)

we added a new line after the token.

After this, we still had lines that did not start with an uppercase letter or did not end in punctuation. We then selected, first of all, all lines that contained at least 2 letters but did not start with any of the symbols

ˆ © , | { C

Then we removed all spaces and the following symbols

. ↑ /) → ←ˆ ; _] &

from the beginning of a line as well as the symbols

ˆ → _ | ´ <

and spaces from the end of a line. After these replacements, we proceeded only if the line ended with punctuation (. ? ! ... : ;) and started with an uppercase letter, number or the symbol ¿, when ignoring various kinds of parentheses in the beginning and the end of the line. Since we had been dividing text into sentences without knowing whether two or more sentences were, for example, part of a line spanning quote, we removed all hyphens from the beginning and end of the sentence when the sentence did not include a matching pair.

During the previous steps, we always kept track of the URL of the page the sentence was originally found in as well as the list of languages provided by MultiLI for the original line tested. At this point, we removed duplicate sentences in a similar manner as duplicate lines were removed earlier. The remaining unique

sentences had the union of the lists of languages of the possible duplicates and the URL of the retained sentence as described above.

4.6 Language set identification for sentences

Finally, the language of each sentence was identified using MultiLI. The language identifier was given the list of languages attached to the sentence and was only allowed to consider the relevant Uralic languages present on that list as was done when identifying the language of the lines. However, unlike when identifying the language of a web page or a line, the language that was most present in the sentence was considered the "winning" language. Each sentence identified as written with a relevant Uralic language was saved and added to the corresponding sentence collection.

4.7 Manual post-processing

The automatic process described in the earlier sections is good enough for those languages where language material can be found in larger amounts. As the number of false positive language identifications is not relative to the number of true positives, it is worthwhile to manually investigate and clean the smallest automatically created sentence collections. Therefore, we manually inspected those 8 collections which consisted of less than 600 sentences. We removed those sentences that we were able to identify as erroneously identified from the collections. In addition, we manually processed the original pages from which sentences in the correct languages were found and added new sentences that had not been detected by our current heuristics to their respective collections.

5 Wanca in Korp - conclusions and future work

Table 1 shows the sizes of the sentence collection for each language. In Korp, each sentence is linked to the original web page from which the sentence was found. As web pages tend to disappear over time, we suggest the users check the Internet Archive Wayback Machine[14] for the unavailable pages.

[14] https://archive.org/web/

In the future, we plan to implement a more intelligent sentence detection algorithm for the relevant Uralic languages, that would be able to detect sentences over line borders using the hfst-pmatch formalism.

We intend to process the 2017 re-crawl also using the same pipeline, as well as perform another re-crawl in 2019. The corpora resulting from these crawls will also be available in the Korp service.

Table 1. The number of sentences, words, and characters in the corpora for each Uralic language.

	#unique sentences	#words	#characters
Finnic			
Estonian, Standard (**ekk**)			
Finnish (**fin**)			
Finnish, Kven (fkv)	2,156	20,659	167,347
Finnish, Tornedalen (fit)	5,203	52,753	432,229
Ingrian (izh)	81	582	4,318
Karelian (krl)	2,593	28,214	258,731
Liv (liv)	705	6,927	59,589
Livvi-Karelian (olo)	9,920	108,154	891,246
Ludian (lud)	771	7,043	54,921
Veps (vep)	13,461	127,929	987,541
Vod (vot)	20	139	1,129
Võro (vro)	66,878	708,654	4,845,649
Hungarian (**hun**)			
Khanty (kca)	1,006	13,720	132,291
Mansi (mns)	904	9,684	120,289
Mari			
Mari, Hill (mrj)	30,793	191,342	2,260,225
Mari, Meadow (mhr)	110,216	1,073,115	14,109,134
Mordvin			
Erzya (myv)	28,986	294,065	4,231,531
Moksha (mdf)	21,571	214,052	3,101,193
Permian			
Komi-Permyak (koi)	8,162	77,666	905,400

Komi-Zyrian (kpv)	21,786	191,255	2,300,145
Udmurt (udm)	56,552	499,239	6,467,702
Sami			
Sami, Akkala (**sia**)			
Sami, Inari (smn)	15,469	171,178	1,689,962
Sami, Kildin (sjd)	132	1,239	16,821
Sami, Lule (smj)	10,605	113,699	1,002,804
Sami, North (sme)	214,226	2,300,119	20,456,823
Sami, Pite (**sje**)			
Sami, Skolt (sms)	7,819	77,758	867,104
Sami, South (sma)	15,380	170,826	1,602,025
Sami, Ter (**sjt**)			
Sami, Ume (sju)	124	1,536	14,558
Samoyed			
Enets, Forest (**enf**)			
Enets, Tundra (**enh**)			
Kamas (xas)			
Nenets (yrk)	443	3331	51,486
Nganasan (nio)	62	1,379	22,326
Selkup (**sel**)			

Acknowledgments

This work has been funded by the Kone Foundation as part of its Language Programme (Kone Foundation 2012). We thank Jussi-Pekka Hakkarainen, Sulev Iva, Ilja Moshnikov, Julia Normanskaja, Tommi Pirinen, Michael Riessler, Jack Rueter, and Trond Trosterund for their invaluable assistance during the work.

References

Arkhangelskiy T (2019) Corpora of social media in minority Uralic languages. Proc. of the Fifth International Workshop on Computational Linguistics for Uralic Languages (IWCLUL 2019), Tartu, Estonia, 125–140.

Arkhangelskiy T & Medvedeva M (2016) Developing morphologically annotated corpora for minority languages of Russia. Proc. of Corpus Linguistics Fest (CLiF 2016), Bloomington, Indiana, 1–6.
Barbaresi A (2013a) Challenges in web corpus construction for low-resource languages in a post-BootCaT world. Proc. of the 6th Language & Technology Conference, Less Resourced Languages special track (LT-LRL 2013), Poznan, Poland, 69–73.
Barbaresi A (2013b) Crawling microblogging services to gather language-classified URLs: Workflow and case study. Proc. of the Student Research Workshop (ACL 2013), Sofia, Bulgaria, 9–15.
Baroni M, Bernardini S, Ferraresi A & Zanchetta E (2009) The WaCky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation 43(3): 209–226.
Baroni M & Kilgariff A (2006) Large linguistically-processed Web corpora for multiple languages. Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006), Trento, Italy, 87–90.
Baykan E, Henzinger M & Weber I (2008) Web Page Language Identification Based on URLs. Proc. of the 34th International Conference on Very Large Data Bases (VLDB '08), Auckland, New Zealand, 176–187.
Boleda G, Bott S, Meza R, Castillo C, Badia T & Lopez V (2006) CUCWeb: a Catalan corpus built from the Web. Proc. of the 2nd International Workshop on Web as Corpus (WAC '06), Trento, Italy, 19–26.
Emerson T & O'Neil J (2006) Experience Building a Large Corpus for Chinese Lexicon Construction. In: Baroni M & Bernardini S (eds) WaCky! Working papers on the Web as Corpus, 41–62.
Endrédi I, Fejes L, Novák A, Oszkó B, Prószéky G, Szeverényi S, Várnai Z & Wagner-Nagy B (2010) Nganasan - Computational Resources of a Language on the Verge of Extinction. Proc. of the 7th SaLTMIL Workshop at LREC-2010, Valetta, Malta, 41–44.
Ferraresi A, Zanchetta E, Baroni M & Bernardini S (2008) Introducing and evaluating ukWaC, a very large web-derived corpus of English. Proc. of the 4th Web as Corpus Workshop (WAC-4), Marrakech, Morocco, 47–54.
Ghani R, Jones R & Mladenic` D (2001) Mining the Web to Create Minority Language Corpora. Proc. of the 10th International Conference on Information and Knowledge Management (ACM CIKM 2001), Atlanta, Georgia, 279–286.
Grefenstette G & Tapanainen P (1994) What is a word, What is a sentence? Problems of Tokenization. Proc. of the 3rd Conference on Computational Lexicography and Text Research (COMPLEX'94), Budapest, Hungary.

Jauhiainen H, Jauhiainen T & Lindén K (2015a) The Finno-Ugric Languages and The Internet Project. Proc. of the 1st International Workshop on Computational Linguistics for Uralic Languages (IWCLUL 2015), (2): 87–98.
Jauhiainen T (2010) Tekstin kielen automaattinen tunnistaminen. Master's thesis, University of Helsinki, Helsinki.
Jauhiainen T, Jauhiainen H & Lindén K (2015b) Discriminating Similar Languages with Token-Based Backoff. Proc. of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial), Hissar, Bulgaria, 44–51.
Jauhiainen T, Jauhiainen H & Lindén K (2018a) HeLI-based Experiments in Discriminating Between Dutch and Flemish Subtitles. Proc. of the 5th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), Santa Fe, New Mexico, 137–144.
Jauhiainen T, Jauhiainen H & Lindén K (2018b) HeLI-based experiments in Swiss German dialect identification. Proc. of the 5th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), Santa Fe, New Mexico, 254–262.
Jauhiainen T, Jauhiainen H & Lindén K (2018c) Iterative Language Model Adaptation for Indo-Aryan Language Identification. Proc. of the 5th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018), Santa Fe, New Mexico,66–75.
Jauhiainen T, Jauhiainen H & Lindén K (2019) Discriminating between Mandarin Chinese and Swiss-German varieties using adaptive language models. Proc. of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2019), Minneapolis, Minnesota, 178–187.
Jauhiainen T, Lindén K & Jauhiainen H (2015c) Language Set Identification in Noisy Synthetic Multilingual Documents. Proc. of the Computational Linguistics and Intelligent Text Processing 16th International Conference, (CICLing 2015), Cairo, Egypt, 633–643.
Jauhiainen T, Lindén K & Jauhiainen H (2016) HeLI, a Word-Based Backoff Method for Language Identification. Proc. of the 3rd Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2016), Osaka, Japan, 153–162.
Jauhiainen T, Lindén K & Jauhiainen H (2017a) Evaluating HeLI with Non-Linear Mappings. Proc. of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2017), Valencia, Spain, 102–108.
Jauhiainen T, Lindén K & Jauhiainen H (2017b) Evaluation of Language Identification Methods Using 285 Languages. Proc. of the 21st Nordic Conference on Computational

Linguistics (NoDaLiDa 2017), Linköping University Electronic Press, Gothenburg, Sweden, 183–191.
Jauhiainen T, Lui M, Zampieri M, Baldwin T & Lindén K (2018d) Automatic Language Identification in Texts: A Survey. arXiv preprint arXiv:1804.08186v2.
Jokinen K (2014) Open-domain Interaction and Online Content in the Sami Language. Proc. of the Language Resources and Evaluation Conference (LREC-2014), Reykjavik, Iceland, 517–522.
Kiss T & Strunk J (2002a) Scaled log likelihood ratios for the detection of abbreviations in text corpora. Proc. of the 19th International Conference on Computational Linguistics (COLING 2002), Taipei, Taiwan, 1228–1232.
Kiss T & Strunk J (2002b) Viewing sentence boundary detection as collocation identification. Proc. of 6th Konferenz zur Verarbeitung natürlicher Sprache (KONVENS 2002), Saarbrücken, Germany, 75–82.
Kiss T & Strunk J (2006) Unsupervised Multilingual Sentence Boundary Detection. Computational Linguistics 32(4): 485–525.
Kone Foundation (2012) The Language Programme 2012-2016. Http://www.koneensaatio.fi/en.
Kornai A, Halácsy P, Nagy V, Oravecz C, Trón V & Varga D (2006) Web-based frequency dictionaries for medium density languages. Proc. of the 2nd International Workshop on Web as Corpus (WAC '06), Trento, Italy.
Kristoffersen KB (2017) Common Crawled web corpora. Constructing corpora from large amounts of web data. Master's thesis, University of Oslo, Oslo.
Medelyan O, Schulz S, Paetzold J, Poprat M & Marcó K (2006) Language Specific and Topic Focused Web Crawling. Proc. of the 5th International Conference on Language Resources and Evaluation (LREC'06), Genoa, Italy, 865–868
Mikheev A (2000) Tagging Sentence Boundaries. Proc. of the 1st North American chapter of the Association for Computational Linguistics conference (ANLP-NAACL 2000), Seattle, Washington, U.S.A., 264–271.
Mikheev A (2002) Periods, Capitalized Words, etc. Computational Linguistics 28(13): 289–318.
Mohr G, Stack M, Rnitovic I, Avery D & Kimpton M (2004) Introduction to Heritrix. 4th International Web Archiving Workshop (at ECDL2004).
Mon PY, Choong CY & Mikami Y (2011) Language Specific Crawler for Myanmar Web Pages. International Journal of Computer Science Issues 8(2): 127–135.

Mon PY & Mikami Y (2010) Myanmar Language Search Engine. Proc. of the 11th International Conference on Advances in ICT for Emerging Regions (ICTer 2010), Colombo, Sri Lanka, 69–74.
Novák A (2008) Language resources for Uralic minority languages. Proc. of the SaLTMIL Workshop at LREC-2008: Collaboration: Interoperability between People in the Creation of Language Resources for Less-resourced Languages, Marrakech, Morocco, 27–32.
Palmer DD & Hearst MA (1994) Adaptive Sentence Boundary Disambiguation. Proc. of the 4th Conference on Applied Natural Language Processing (ANLP 1993), Stuttgart, Germany, 78–83.
Pomikálek J, Rychlý P & Kilgarriff A (2009) Scaling to Billion-plus Word Corpora. Advances in Computational Linguistics 41(3): 3–13.
Priyatam PN, Vaddepally S & Varma V (2012) Domain Specific Search in Indian Languages. Proc. of the first workshop on Information and knowledge management for developing regions (IKM4DR'12), Maui, Hawaii, 23–30.
Prószéky G (2011) Endangered Uralic Languages and Language Technologies. Proc. of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage (DigHum 2011), Hissar, Bulgaria, 1–2.
Schäfer R, Barbaresi A & Bildhauer F (2014) Focused Web Corpus Crawling. Proc. of the 9th Web as Corpus Workshop (WaC-9), Gothenburg, Sweden, 9–15.
SIL (2013) ISO 639-3 Codes for the representation of names of languages. SIL International.
Simon E & Mus N (2017) Languages under the influence: Building a database of Uralic languages. Proc. of the 3rd International Workshop for Computational Linguistics of Uralic Languages (IWCLUL 2017), St. Petersburg, Russia, 10–24
Simons GF & Fennig CD (eds) (2018) Ethnologue: Languages of the World, Twenty-first edition. SIL International, Dallas, Texas.
Somboonviwat K, Tamura T & Kitsuregawa M (2005) Simulation Study of Language Specific Web Crawling. Proc. of the 21st International Conference on Data Engineering Workshops (ICDEW'05), Tokyo, Japan, 1254.
Suchomel V & Pomikálek J (2012) Efficient Web Crawling for Large Text Corpora. Proc. of the 7th Web as Corpus Workshop (WAC7), Lyon, France, 39–43.
Suihkonen P (1998) Documentation of the Computer Corpora of the Uralic Languages at the University of Helsinki. Technical report, University of Helsinki, Helsinki, Finland.

Tamura T, Somboonviwat K & Kitsuregawa M (2007) A Method for Language-Specific Web Crawling and Its Evaluation. Systems and Computers in Japan 38(2): 10–20.
Vincze V, Nagy Á, Horváth C, Szilágyi N, Kozmács I, Bogár E & Fenyvesi A (2015) FinUgRevita: Developing Language Technology Tools for Udmurt and Mansi. Proc. of the 1st International Workshop on Computational Linguistics for Uralic Languages (IWCLUL 2015), Tromsø, Norway, 108–118.

Digilang – Turun yliopiston digitaalisia kieliaineistoja kehittämässä

Tommi Kurki, Nobufumi Inaba, Annekatrin Kaivapalu, Maarit Koponen, Veronika Laippala, Christophe Leblay, Jorma Luutonen, Maarit Mutta, Markku Nikulin ja Elisa Reunanen
Turun yliopisto, kieli- ja käännöstieteiden laitos

Tiivistelmä
Turun yliopiston kieli- ja käännöstieteiden laitoksen Digilang-hankkeessa täydennetään ja kehitetään laitoksen digitaalisia kieliaineistoja. Samalla kieliaineistojen näkyvyyttä lisätään keräämällä ne yhteen ja luomalla yhteinen käyttäjäportaali, jonka avulla tutkijat ja opiskelijat löytävät entistäkin paremmin tarvitsemiaan aineistoja. Yliopiston rehtori on myöntänyt hankkeelle 580 000 euroa aineistojen kehitystyöhön, ja hanke toimii vuosina 2018–2021.

Turun yliopiston kieliaineissa on koostettu, kehitetty ja ylläpidetty digitaalisia aineistoa tutkimuksen tarpeisiin vuodesta 1967, jolloin suomen kielen oppiaineen yhteyteen perustettiin Lauseopin arkisto (LA). Lauseopin arkiston murrekorpus on Suomen ensimmäinen digitaalinen annotoitu kieliaineisto. Varsinkin viime vuosikymmenten aikana alkuperäisen murrekorpuksen rinnalle on kieliaineissa luotu useita muita korpuksia.

Digilang-hankkeessa parannetaan nykyisten aineistojen käytettävyyttä kehittämällä niiden ns. metatietoja, kun esimerkiksi kunkin sanan, lauseen, virkkeen, intonaatiojakson ja diskurssin rakenteesta ja visualisoinnista lisätään tietoja. Näin aineiston käyttäjät pystyvät löytämään helpommin yhä useammasta laajasta puhe- tai tekstimassasta tarvitsemansa tapaukset. Kieliaineistojen saavutettavuutta ja näkyvyyttä lisätään keräämällä ne yhteen ja luomalla yhteinen käyttäjäportaali, jonka avulla tutkijat ja opiskelijat löytävät entistäkin paremmin tarvitsemiaan aineistoja ja saattavat samalla löytää heille entuudestaan tuntemattomia mutta hyödyllisiä aineistoja.

Ensi vaiheessa Digilang-hankkeessa yhdistyy kuusi TY:ssa eri tahoilla kehitettyä kieliaineistoa: Lauseopin arkiston (LA) aineistoja, muita suomen kielen ja suomalais-ugrilaisen kielentutkimuksen oppiaineen kieliaineistoja, kielentutkimusta ja kieliteknologiaa yhdistävän TurkuNLP-ryhmän kehittämiä Universal Parsebanks -aineistoja sekä eri kielten ja kääntämisen tutkijoiden LOG-aineisto.

a) Satakuntalaisuus puheessa -korpus (Sapu; 246 tuntia nykysatakuntalaisen puhekielen äänitteitä, joista 210 tunnista kielitieteelliset transkriptiot, annotoimaton aineisto, osa Lauseopin arkistoa)

b) Suomen kielen prosodian alueellisen ja sosiaalisen variaation korpus (Prosovar; n. 430 puhujalta n. 5700 prosodista äänitekatkelmaa; annotoimaton aineisto; osa Lauseopin arkistoa)

c) Fennougristiset korpukset: Volgan alueen kielten tutkimusyksikön morfologisesti annotoitu mordvalaiskielten korpus Mormula (n. 35 000 virkettä); annotoimattomat tekstikorpukset 7 kielestä (ersä, moksha, mari, udmurtti, komipermjakki, tshuvassi, tataari; yht. n. 17 milj. sanaa); kirjakielen historian korpukset (mari, mordvalaiskielet, yht. n. 1000 tekstiä); paralleelitekstikorpukset (2 tekstiä, 14 kieltä).

d) Akateemisen suomen korpus, joka on rakennettu TY:n strategisen rahoituksen turvin (Edistyneiden suomen oppijoiden korpus ja Ensikielisten suomalaisten akateemisten tekstien korpus sekä nyt koostettava tutkimusartikkelikorpus; osa Lauseopin arkistoa).

e) Universal Parsebanks (UP), joka on 45 erikielistä Internetistä koneellisesti kerättyä aineistoa sisältävä datakokoelma automaattisesti syntaksijäsennettynä. Kielikohtaisten aineistojen koot ovat miljardeja sanoja, ja ne ovat vapaasti käytettävissä. Kehittynein niistä on suomenkielinen Finnish Internet Parsebank.

f) LOG-aineisto, jonka avulla kirjoittamista ja kääntämistä voi tarkastella prosessilähtöisesti: millaisessa prosessissa teksti syntyy, ja esimerkiksi missä järjestyksessä sen osat tuotetaan ja miten sitä muokataan. Aineisto koostuu nyt yhdistettävistä eri tutkijoiden eri kielillä keräämistä tuotoksista.

1 Digilang-hanke

Turun yliopiston kieli- ja käännöstieteiden laitoksessa on erityisaloillaan kansallisesti ja kansainvälisesti ainutlaatuisia kieliaineistoja, joilla on jo entuudestaan oma kotimainen ja kansainvälinen käyttäjäkuntansa. Laitoksessa on koostettu, kehitetty ja ylläpidetty digitaalisia aineistoja tutkimuksen tarpeisiin vuodesta 1967, jolloin suomen kielen oppiaineen yhteyteen perustettiin Lauseopin arkisto (LA).

Laitoksessa on muodostettu annotoituja digitaalisia korpuksia yli 50 vuoden ajan, ja erityisesti viime vuosikymmenten aikana on koostettu yhä useampia uusia kieliaineistoja tutkimuksen tarpeisiin. Jo Lauseopin arkiston perustaminen ja hanke

muodostaa atk-pohjainen kielentutkimuksen korpus ovat itsessään kansallisesti merkittäviä: Lauseopin arkisto on ensimmäinen [1] suomalainen digitaalinen annotoitu kielentutkimuksen korpus (Ikola 2001: 164). Se kattaa kaikki suomen murrealueet ja muodostuu morfologisesti ja syntaktisesti koodatuista 134 pitäjänmurteen osa-aineistoista. Korpuksessa on yli 70 000 virkettä ja yli miljoona sanaa. Lauseopin arkiston pohjalta on tehty vuosikymmenten aikana useita tutkimuksia ja tutkielmia, ja varsinkin sen 2000-luvun alussa tehtyjen uudistusten jälkeen se on edelleen yksi merkittävimmistä suomen kielen annotoiduista kieliaineistoista. (LAO 1985.)

Lauseopin arkiston malli ja korpusten muodostamisesta karttunut kokemus on vaikuttanut vahvasti siihen, että vastaavanlaisia aineistoja on koostettu Turun yliopistossa tämän jälkeen useita. Jo 1970-luvulla yliopiston suomalais-ugrilaisessa kielentutkimuksen oppiaineessa luotiin edellä mainitun mallin pohjalta mordvalaiskielistä vastaavanlainen digitaalinen kielikorpus Mormula. Suomen kielen oppiaineessa alkuperäisen murrekorpuksen (LA) rinnalle on kehitetty puolestaan Mikael Agricolan morfosyntaktinen tietokanta, Akateeminen suomi -korpus (LAS1), Edistyneiden suomenoppijoiden korpus (LAS2) sekä arkikeskustelujen morfosyntaktinen Arkisyn-korpus. Tulevaisuuden teknologioiden laitoksen ja kieli- ja käännöstieteiden laitoksen yhteistyönä on kehitetty Turkulainen suomen kielen puupankki. Nämä kaikki edellä mainitut ovat annotoituja digitaalisia aineistoja, ja ne ovat Kielipankin tai Turun yliopiston välityksellä laajemmin tutkijakunnan käytettävissä.

Edellä mainittujen lisäksi laitoksessa on koostettu muuten kansallisesti ja kansainvälisesti merkittäviä laajoja kieliaineistoja ja -kokoelmia. Tällainen on esimerkiksi Turun yliopiston suomen kielen äänitearkisto, joka koostuu yli 5 000 tunnista kielitieteellisiä alkuperäisäänitteitä. Osa vanhoista kokoelmista on digitoimatta, osa vaatii uudelleen järjestämistä tai yhdenmukaistamista ja esimerkiksi osan käytettävyyttä olisi mahdollista huomattavasti parantaa.

Turun yliopistossa alkuvuodesta 2018 perustettu Digilang-hanke kokosi yhteen kuusi sellaista Turun yliopiston hanketta ja digitaalista kieliaineistoa, joita päätettiin ensi vaiheessa ryhtyä kehittämään edelleen. Turun yliopiston rehtori myönsi keväällä 2018 laitoksessa toimivalle Digilangille yli 580 000 euroa digitaalisten kieliaineistojen kehitystyöhön vuosille 2018–2021. Tavoitteena

[1] Samoihin aikoihin Oulun yliopistossa alettiin muodostaa Oulun korpusta.

hankkeessa on, että tulevaisuudessa nämä aineistot tavoittavat yhä laajemman käyttäjäkunnan, kun aineistoja kehitetään, niiden näkyvyyttä parannetaan ja ne ovat saavutettavissa saman portaalin kautta. Näin vahvistetaan ja terävöitetään samalla Turun yliopiston ja laitoksen brändiä kieliaineistojen tuottajana.

Digilang-hankkeessa parannetaan nykyisten aineistojen käytettävyyttä kehittämällä niiden ns. metatietoja, kun esimerkiksi kunkin sanan, lauseen, virkkeen, intonaatiojakson ja diskurssin rakenteesta ja visualisoinnista lisätään tietoja. Näin aineiston käyttäjät pystyvät löytämään helpommin yhä useammasta laajasta puhe- tai tekstimassasta tarvitsemansa tapaukset. Lisäksi kieliaineistojen näkyvyyttä lisätään keräämällä ne yhteen ja luomalla yhteinen käyttäjäportaali, jonka avulla tutkijat ja opiskelijat löytävät entistäkin paremmin tarvitsemiaan aineistoja ja saattavat samalla löytää heille entuudestaan tuntemattomia mutta hyödyllisiä aineistoja. Aineistojen metatiedot on koottu yhteen portaaliin, mutta aineistot on fyysisesti tallennettu yliopistossa - ainakin toistaiseksi - hankkeiden omiin digitaalisiin arkistoihin. Osa laitoksen kieliaineistoista on liitetty osaksi kansallista Kielipankkia. Myös näistä korpuksista liitetään tiedot portaaliin, ja portaalin käyttäjät löytävät nämäkin aineistot helposti. Laitoksen omasta portaalista ei tule siis kilpailijaa Kielipankille, vaan se täydentää sitä.

Esittelemme seuraavassa Digilang-hankkeessa mukana olevia osahankkeita ja korpuksia. Tarkoituksemme on toisaalta syventää potentiaalisten käyttäjien tietoja näistä korpuksista ja toisaalta havainnollistaa esimerkinomaisesti, kuinka Digilang-portaali tarjoaa potentiaalisille käyttäjilleen hyvin erilaisia kielentutkimuksen digitaalisia aineistoja. Pohdimme samalla lyhyesti näiden aineistojen erilaisia käyttömahdollisuuksia. Tarkoitus on, että tulevaisuudessa Turun yliopistossa muodostettuja digitaalisia kieli- ja käännösaineistoja (esim. teksti, kuva, video, multimodaalinen aineisto) lisätään samaan portaaliin.

2 Osahankkeet ja korpukset

2.1 Yleistä

Digilang-hankkeessa on mukana kuusi laitoksessa eri tahoilla kehitettyä kieliaineistoa: Satakuntalaisuus puheessa -korpus, Suomen kielen prosodian alueellisen ja sosiaalisen variaation korpus, erilaisia fennougristisia korpuksia (mm. Mormula ja marin ja mordvalaiskielten kirjakielen historian korpukset), Akateemisen suomen korpus, Universal Parsebanks -korpus (joka sisältää mm.

suomenkielisen Finnish Internet Parsebank -korpuksen) sekä ranska–suomi–ranska- ja englanti–suomi-kirjoitus- ja kääntämisprosessien LOG-korpus.

2.2 Satakuntalaisuus puheessa

Satakuntalaisuus puheessa -korpus on sosiolingvistisessä samannimisessä tutkimushankkeessa kerätty puhekielen aineisto, joka on kerätty vuosina 2007–2013 ja 2016–2017. Hanketta ovat rahoittaneet sekä Suomen Kulttuurirahaston Satakunnan rahasto että Turun Yliopistosäätiö.

Satakunnalla on maakuntana vuosisatoja pitkä yhteinen historia, mutta kielellisesti – murteellisesti – se on jakautunut vahvasti toisaalta lounaismurteiden piiriin kuuluviin ja toisaalta hämäläisvaikutteisten välimurteiden ja hämäläismurteiden piiriin kuuluviin paikkakuntiin. Dialektometrisesti tarkastellen yksi suomen vahvimmista murrerajoista on halkonut juuri tätä maakuntaa, vaikka muut yhtä vahvat rajat ovat lähes poikkeuksetta myötäilleet suomen länsi- ja itämurteiden välistä rajalinjaa. Hankkeessa on tallennettu vapaamuotoista arkista puhekieltä 16 paikkakunnalla, jotka sijaitsevat Satakunnassa vanhan lounaismurteiden ja lounaisten murteiden välisellä rajavyöhykkeellä. (Kurki, Siitonen, Väänänen, Ivaska & Ekberg 2011; Kurki & Siitonen 2014.)

Hankkeessa on haastateltu eri-ikäisiä ja sosiaaliselta taustaltaan erilaisia informantteja, jotka ovat joko syntyneet keruualueella tai muuttaneet sinne. Informantteja on 302, ja 303 äänitiedoston kokonaiskesto on 246 tuntia. Tästä aineistosta on litteroitu puolikarkealla SU-transkriptiolla 210 tunnin osuus. Aineisto on kerätty dialektologista ja variationistista sosiolingvististä tutkimusta varten, mutta sitä on käytetty myös keskusteluanalyyttisissä ja vuorovaikutuslingvistisissä tutkielmissa ja tutkimuksissa.

Digilang-hankkeeessa äänitteistä ja litteraateista koostuvaa Satakuntalaisuus puheessa -korpusta kehitetään a) litteroimalla loputkin hankkeen äänitteet ja b) annotoimalla aineisto syntaktisesti ja morfologisesti Lauseopin arkiston murrekorpuksen tapaan. Kustakin litteraatista jokainen sana hakusanoitetaan sekä siitä kirjataan syntaktiset (esim. lauseenjäsen) ja morfologiset (esim. sanaluokka ja sijamuoto) tiedot. Rahoituskauden päättyessä edustavaan osaan korpusta pystytään kohdistamaan syntaktis- ja morfologisehtoisia hakuja. Mahdollisuuksien mukaan korpus liitetään osaksi Kielipankin tarjontaa.

2.3 Suomen kielen prosodian alueellinen ja sosiaalinen variaatio (Prosovar)

Suomen kielen prosodian alueellisen ja sosiaalisen variaation tutkimushanke (Prosovar) on Koneen säätiön rahoittama projekti, jonka yksi päätavoitteista on ollut koostaa ensimmäinen suomen kielen prosodian variaation tarkasteluun tarkoitettu korpus. Taustana tälle on, että vaikka puhuttua suomea ja sen variaatiota on tutkittu runsaasti, prosodian ja sen variaation tutkimus on ollut yksittäisiä poikkeuksia lukuun ottamatta niukkaa. Fennistisessä sosiolingvistiikassa, murteentutkimuksessa ja variaationtutkimuksessa on tutkittu tavallisesti fonologisia, morfofonologisia ja morfologisia ilmiöitä, jonkin verran myös syntaktisia ilmiöitä. Sosiofoneettinen tutkimus on – ainakin vielä toistaiseksi – jäänyt Suomessa marginaaliin. Prosodisia piirteitä on tutkittu ja sivuttu keskusteluntutkimuksessa, mutta siinäkään prosodian ei voi väittää olleen tutkimuksen tai sen vastaanoton keskipisteessä. Lähinnä foneetikkojen harjoittama varsinainen prosodian tutkimus on käsitellyt puolestaan tyypillisesti yleiskieltä tai tarkastellut funktionaalista (diatyyppistä) variaatiota. (Kurki, Nieminen, Kallio & Behravan 2014.) Tähänastiset havainnot prosodian alueellisesta ja sosiaalisesta variaatiosta perustuvatkin yksittäisiin – sinänsä arvokkaisiin – tutkimuksiin (ks. esim. Ylitalo 2004; Yli-Luukko 2010).

Suomen naapurimaista Virossa ja Ruotsissa on jo jonkin aikaa kehitetty prosodiikkaan soveltuvia puhetietokantoja. Ruotsissa tällaisia korpuksia on ollut jo toistakymmentä vuotta, ja Virossakin on virinnyt runsaasti tutkimusta. Suomen kielestä vastaavanlainen korpus on kuitenkin puuttunut. (Kurki, Nieminen, Kallio & Behravan 2014.)

Prosovar-hankkeessa äänitteitä kartutettiin uutta korpusta varten elisitoiduin äänitystehtävin verkkokeruun avulla. Keruun toteuttamiseksi hankkeessa luotiin kokonainen oma aineistonkeruusivusto ja kehitettiin tätä varten äänentallennussovelluksia, jotka mahdollistivat informanttien äänen tallentamisen omilta tietokoneiltaan ja mobiililaitteiltaan. Sivustolle kuka tahansa kiinnostunut pystyi luomaan oman käyttäjätunnuksen, kunhan hän hyväksyi käyttöehdot ja antoi luvan käyttää tieteelliseen tarkoitukseen omalta tietokoneeltaan äänittämiään äänitekatkelmia. Elisitoiduissa tehtävissä informanteille esitettiin visuaalisia, auditiivisia ja tekstuaalisia ärsykkeitä, joihin informanttien piti reagoida verbaalisesti. Nämä reaktiot äänitettiin. Vapaaehtoisia osallistujia kehotettiin käyttämään sellaista kieltä kuin he käyttäisivät tavallisessa arkipäiväisessä

vuorovaikutuksessa. Sivustolla kerättiin aineistoa 8.4.2014–31.12.2016, ja ääntään kävi tallentamassa 440 informanttia. Heiltä karttui äänitekorpukseen eri elisitoiduista tehtävistä yhteensä yli 5 700 näytettä.

Aineisto on koostettu ennen muuta suomen prosodian ja sen variaation tutkimukseen, mutta yhtä lailla se soveltuu esimerkiksi kansanlingvististen tutkimusten aineistoksi (ks. esim. Kurki 2018). Aineistoa on mahdollista käyttää myös dialektologisissa tai sosiolingvistisissä tutkimuksissa, joissa tarkasteltavana ovat segmentaaliset tai suprasegmentaaliset ilmiöt.

Digilang-hankkeessa prosodiaäänitteet järjestetään ja niistä valitaan alueellisesti ja sosiaalisesti edustavat otokset. Samalla aineistosta karsitaan ne näytteet, joiden äänenlaatu ei riitä akustiseen analyysiin. Aluksi äänenlaadultaan riittävän hyvistä ja alueellisesti ja sosiaalisesti edustavasta ääniteotoksesta segmentoidaan äänitekohtaisesti sanat, tavut ja foonit. Tämän jälkeen tavoitteena on segmentoida prominenssit (aksentit ja lausepainot). Odotuksenmukaista on, että Digilang-rahoituksen aikana kaikkea aineistoa ei pystytä segmentoimaan, vaan lopulle aineistolle ja sen käsittelyyn haetaan lisärahoitusta.

Rahoituskauden päättyessä tarkoituksena on, että prosodiahankkeeseen pystytään tekemään koneellisia hakuja paitsi erilaisten elisitoitujen tehtävien perusteella myös ääninäytteiden prosodisten ominaispiirteiden perusteella..

2.4 Fennougristiset korpukset

Myös Turun yliopiston suomalais-ugrilaisessa kielentutkimuksessa on pitkä kokemus sähköisistä kielikorpuksista, sillä jo 1970-luvun lopulla luotiin mordvalaiskielten folkloretekstejä ja kirjakielisiä tekstejä sisältävä Mormula-korpus. Se on morfologisesti koodattu, ja sen teksteissä on joko suomen- tai saksankieliset käännökset. Mordvalaiskielten tutkijoille Mormula on ollut jo lähes 40 vuotta keskeinen tutkimusaineisto. Se mahdollistaa mordvalaiskielten rakenteen tutkimisen sanaston, morfologian ja syntaksin tasoilla. Erityisen paljon sitä on käytetty morfosyntaksin tutkimukseen.

Edellä mainitun lisäksi saman yliopiston Volgan alueen tutkimusyksikössä on koostettu 1980-luvun lopulta yli miljoonan sanan laajuisia sähköisiä tekstikorpuksia marista, mordvalaiskielistä, udmurtista, tšuvassista ja tataarista. Oman aineistotyyppinsä muodostavat paralleelitekstit, joita sisältävä korpus on muodostettu 2000-luvun alussa. Se käsittää kaksi tekstiä: Vitali Gubarevin 1950-luvun nuorisoromaanin Pavlik Morozov (n. 10 000 sanaa) sekä Kaisa Häkkisen ja

Seppo Zetterbergin Suomi eilen ja tänään -teoksen (n. 30 000 sanaa). Ensin mainitusta on korpuksessa tekstit 14:llä ja viimeksi mainitusta 7 kielellä. Oman aineistotyyppinsä muodostavat mordvan ja marin kirjakielten historian korpukset. Ne ovat tekstitiedostoista koostuvia pitkittäisaineistoja. Mikrofilmatuista sanomalehdistä on siirretty tekstitiedostoihin eri vuosikymmeniltä 1920-luvulta lähtien ersän-, mokšan- ja marinkielisiä tekstejä, joiden avulla voi saada eksaktia tietoa kirjakielissä tapahtuneista muutoksista. Toistaiseksi näitä annotoimattomia aineistoja on käytetty tekemällä merkkijonohakuja. Tällaisilla hauilla voidaan helposti tutkia esimerkiksi sanojen käyttötiheyksiä ja sanojen esiintymisympäristöjä tekstissä. Kirjakielen historian korpuksia tarkasteltaessa nähdään, millaista kirjakieli oli syntyvaiheessaan, miten se vakiintui ja miten kielipoliittiset päätökset ovat vaikuttaneet kirjoittamiseen.

Volgan alueen vähemmistökielten sanojen rakenteen tutkimisen avuksi ja kieliteknologisten sovellusten käyttöön koostettiin vuosituhannen vaihteen jälkeen sähköiset sanaluettelot, joihin on koottu yleis- ja erikoissanakirjoista löytyvää sanastoa. Kukin sana on varustettu tiedolla sanan sanaluokasta. Sanalistoja on koostettu marista, mordvalaiskielistä, udmurtista, komista, tšuvassista ja tataarista, ja niiden laajuudet vaihtelevat 31 000 sanasta 75 000 sanaan. Sanalistat ovat CSV-muotoisia tekstitiedostoja, joita voi käsitellä monilla ohjelmilla. Erityisesti niitä varten on myös kehitetty oma työkalunsa, SFOu WordListTool -ohjelma. Sanalistat ovat vapaasti saatavissa Suomalais-Ugrilaisen Seuran sivuilta: https://www.sgr.fi/fi/items/show/404.

Digilang-hankkeessa kaikki fennougristiset aineistot muunnetaan XML-muotoon, jolloin materiaalien struktuuri tulee eksplisiittiseksi ja korpusstandardien mukaiseksi. Kieliopillisesti annotoidut Mormula-korpuksen tekstit pyritään tekemään mahdollisimman pitkälle yhteensopiviksi Lauseopin arkiston korpusten kanssa, mikä mahdollistaa mm. samojen hakutyökalujen käytön. Korpus myös hakusanoitetaan, mikä helpottaa hakujen tekemistä. Osaan Mormulan teksteistä lisätään syntaktinen annotointi. Syntaktisesti annotoitavan materiaalin lopullinen laajuus selviää vasta, kun nähdään, missä määrin operaatio voidaan tehdä automaattisesti ja missä määrin se edellyttää käsityötä.

Yleisemmin tavoitteena on lisätä Turun yliopiston fennougrististen korpusten näkyvyyttä ja saavutettavuutta yhä laajemmalle käyttäjäkunnalle sekä luoda edellytykset materiaalien entistä monipuolisempaan käyttöön.

2.5 Akateemisen suomen korpus

Akateemisen suomen korpus (LAS1) syntyi vuonna 2015, jolloin Turun yliopiston professori Kirsti Siitosen johtamassa Akateeminen suomi kansainvälisen oppijan haasteena -hankkeessa ryhdyttiin keräämään ensikielisten kirjoittajien suomenkielisiä pro gradu -töitä erillistä korpusta varten. Nykyisessä Akateemisen suomen korpus -projektissa jatketaan tätä työtä, ja korpuksesta pyritään luomaan tiedekunnittain ja aihepiireittäin tasapainotettu tekstikokonaisuus tutkimuksen käyttöön. Vaikka korpus toimii edelleen myös vertailuaineistona aiemmalle, S2-kirjoittajien kirjoituksista koostuvalle LAS2-korpukselle, korpus on nykyisellään itsenäistynyt omaksi tutkimukselliseksi kokonaisuudekseen ja koostuu kahdesta alakorpuksesta, joista toinen on em. pro gradu -töistä koostuva korpus, toinen taas tieteellisistä tutkimusartikkeleista koostuva tutkimusartikkelikorpus. Korpus sisältää opinnäytetöitä kaikista Turun yliopiston tiedekunnista ja tarkoituksena on saattaa se – kuten myös tutkimusartikkelikorpus – kattamaan tasapainotetusti kaikki tieteenalat.

Akateemisen suomen korpuksen sanastollisesti, morfologisesti ja syntaktisesti annotoitu korpus soveltuu monenlaisen tutkimuksen kohteeksi aina seminaaritöistä tieteellisiin tutkimuksiin, ja sen avulla voi tutkia kielen ilmiöitä niin määrällisestä kuin laadullisestakin näkökulmasta. Tutkimuskohteena voivat olla yhtä lailla akateemisen kielen sanastolliset piirteet kuin myös morfologian, morfosyntaksin tai syntaksin ilmiöt. Korpuksen aineisto edustaa ennen kaikkea opinnäytetöiden ja tutkimusartikkelien akateemista suomen kieltä ja soveltuu näiltä osin erityisen hyvin tieteellisten – tai tieteenalakohtaisten – tekstien tutkimusmateriaaliksi, mutta on käytettävissä luonnollisesti myös laajemmin asiaproosan ja asiakielen tutkimusaineistona.

Hankkeessa aineisto kerätään jo valmiiksi digitaalisessa muodossa, joten sitä ei tarvitse erikseen digitoida, ainoastaan muuttaa prosessoinnin kannalta sopivaan tiedostomuotoon. Pro gradu -aineistoa kerätään keväästä 2019 alkaen yhteistyössä Turun yliopiston kirjaston kanssa. Digitaaliset tiedostot raakakoodataan aluksi koneellisesti, minkä jälkeen parserin tekemä leksikaalinen, morfologinen ja syntaktinen annotointi vielä tarkistetaan manuaalisesti, virheet korjataan ja mahdollisia erityishuomioita (esim. subjektin puuttuminen, ellipsi jne.) lisätään koodaukseen. Näin varmistetaan, että aineiston koodaus on tasoltaan tutkimuskäyttöön soveltuvaa.

Hankkeen aikana kerätty aineisto siirtyy kokonaisuutena osaksi Turun yliopiston Lauseopin arkiston digitaalisia kokoelmia, josta se on helposti saatavilla tutkimuskäyttöä varten. Kuten LAS2-korpus myös LAS1-korpus on kuitenkin tarkoitus siirtää myös osaksi kansallista Kielipankkia, josta se on vielä näkyvämmin tutkijayhteisön saavutettavissa. Korpus on muodoltaan yhteensopiva merkittävimpien akateemisen kielen korpusten kanssa (Michigan corpus of upper-level student papers (micusp), British Academic Written English (BAWE) sekä Corpus of Academic Learner English (CALE)). Tutkimusluvan saaneilla opiskelijoilla ja tutkijoilla tulee olemaan mahdollisuus käyttää korpuksen aineistoja oman tutkimuksensa apuna.

2.6 Universal Parsebanks

Universal Parsebanks on Tulevaisuuden teknologioiden laitoksen ja kieli-ja käännöstieteiden laitoksen monitieteisen TurkuNLP-tutkimusryhmän yhteistyössä kehitettämä datakokoelma, joka sisältää 45 eri kielistä Parsebankia eli internetistä koneellisesti koottua aineistoa. Kielikohtaisten aineistojen koot ovat useita miljardeja sanoja. Käytetyin aineistoista on suomenkielinen Finnish Internet Parsebank, jonka kokoaminen on aloitettu vuonna 2013. Muunkieliset osiot on kerätty vuonna 2017. Aineiston syntaksimerkinnät on tehty Universal dependencies -mallilla, jossa virkerakenteet on kaikissa kielissä merkitty mahdollisimman samalla tavalla. Tässä hankkeessa keskitytään ensisijaisesti suomen-, englannin-, ruotsin- ja ranskankielisten aineistojen kehittämiseen.

Finnish Internet Parsebank on alusta asti kehitetty sekä kielentutkimuksen että kieliteknologian tarpeisiin. Muunkieliset UP-kokoelmat on koostettu alun perin erityisesti kieliteknologian tarpeisiin, mutta Digilang-hankkeen avulla aineistojen käytettävyyttä myös muille aloille saadaan lisättyä.

Koneellisesti internetistä kootut aineistot tarjoavat ainutlaatuisen näkökulman kieleen kahdesta syystä. Ensinnäkin ne sisältävät erittäin laajan valikoiman erilaisia tekstejä vaihtelevista alkuperistä. Tämä tarjoaa erinomaiset mahdollisuudet esimerkiksi kielen vaihtelun tutkimukselle ja editoimattomassa tekstissä esiintyvien rakenteiden tarkastelulle. Monet näistä teksteistä ei koskaan päätyisi perinteisiin, käsin koottuihin kieliaineistoihin. Toinen internetistä koneellisesti koottujen aineistojen merkittävä etu on niiden koko. Tämän ansiosta niiden avulla on mahdollista tutkia esimerkiksi hyvin harvinaisia kielen rakenteita ja sanoja.

Lisäksi valtava koko mahdollistaa UP-kokoelman käyttämisen kieliteknologian sovellusten kehittämiseen.

UP-aineistot soveltuvat käytettäviksi kaikilla kieliaineistoja soveltavilla aloilla. Kielentutkimuksen ja kieliteknologian lisäksi esimerkiksi kokeellisen psykologian kielen prosessointia käsittelevissä tutkimuksissa laajoja kieliaineistoja käytetään vertailuaineistoina.

Toisin kuin monissa muissa kieliaineistoissa, UP-kokoelmassa on kattavat merkinnät sanaluokista ja virkerakenteista Universal dependencies -mallin mukaan. Näiden ansiosta UP-aineistot soveltuvat erinomaisesti monipuolisten kielen rakenteiden tarkasteluun. Koska Universal dependencies tarjoaa samankaltaiset kuvaukset eri kielille, niiden avulla on mahdollista myös tehdä vertailevaa tutkimusta eri kielten rakenteiden välillä.

Projektin aikana aineiston käytettävyyttä parannetaan: 1) lisäämällä metadataa eli tietoa dokumenttien alkuperästä, Internet-osoitteista ja hakuajasta sekä aineistokokoelmien koista ja 2) parantamalla käyttöliittymää niin, että yllä mainitut tiedot tulevat näkyviin kaikissa alakokoelmissa ja että hauista saadaan myös numeerisia tuloksia esimerkiksi hakusanan kokonaisfrekvenssistä. Lisäksi teksteihin pyritään saamaan metatietoja niiden edustamista tekstilajeista. Tämä tehdään yhteistyössä kieli- ja käännöstieteiden laitoksella alkavan “Uutinen, mielipide vai jotain muuta? Erilaiset tekstit ja niiden automaattinen tunnistus monikielisestä internetistä” -hankkeen kanssa.

UP-aineistot ovat vapaasti käytettävissä TurkuNLP-tutkimusryhmän sivuilla osoitteessa bionlp-www.utu.fi/dep_search/.

2.7 LOG-korpus

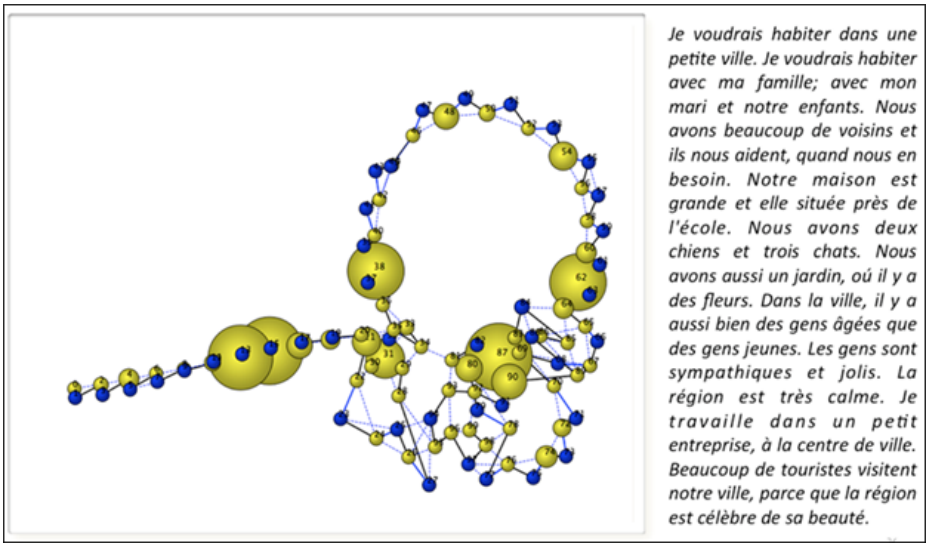
LOG-aineisto eroaa merkittävästi muista hankkeessa olevista aineistoista siinä, että sen avulla tutkitaan prosessilähtöisesti sekä *kirjoittamista* että *kääntämistä*. Prosessilähtöinen lähestymistapa tarkastelee, miten/millaisessa prosessissa kirjoitettu tai käännetty teksti syntyy, esimerkiksi missä järjestyksessä tekstin/käännöksen osat tuotetaan ja miten jo kirjoitettua/käännettyä tekstiä muokataan. LOG-aineistoja lähestytään kahden eri tieteenalan näkökulmasta. Ero on siis kahden lähestymistavan välillä: *kognitio* (kognitiivisen kielentutkimuksen näkökulmasta) vs. *geneesi* (genetiikan tutkimuksen näkökulmasta). *Kognitiivisessa metodologiassa* kartoitetaan mm. seuraavanlaisia asioita: miten teksti ja/tai käännösprosessi rakentuu, millaisia kirjoittaja- ja/tai kääntäjäprofiileja voidaan

tunnistaa ja miten näitä prosesseja voidaan kuvata visuaalisesti (tauot ja niiden pituus, sujuva kirjoittaminen jne.). *Geneettinen metodologia* puolestaan kartoittaa tekstinlaadintaprosessin kirjoitusjälkiä, kirjoitusjälkien välisiä yhteyksiä ja kirjoitusjälkien järjestämistä, esimerkiksi miten kronologinen jatkumo ilmentää tekstin tuottamisen eri vaiheita.

Kerätyt aineistot ovat tapaustutkimusaineistoja (ranska–suomi–ranska-aineistot 43 kpl tekstin tuottaminen L1/L2-kielellä; esim. Mutta, 2017), jotka on kerätty vuodesta 2002 alkaen sekä konekäännöksen jälkieditoinnista vuosina 2016–2017 kerättyjä aineistoja (englanti–suomi, 33 kpl; Koponen, Salmi & Nikulin, 2019). Aineiston keräämisessä on käytetty kahdenlaisia ohjelmia, joista toiset sijoittuvat kognitiivisen (*ScriptLog*, *Translog*) ja toiset geneettisen kielitieteen (*GenoGraphiX*) alalle. Osahankkeessa tehdään kehitystyötä kahdessa vaiheessa. Ensimmäisessä vaiheessa kerätään yhteen kaikki jo tallennetut aineistot ja standardisoidaan ne sellaiseen muotoon, että ne ovat helposti käytettävissä portaalin kautta. Aineiston muokkaaminen tarkoittaa muun muassa sitä, että aiemmin kootut aineistot, joiden kokoamisessa on käytetty nykyisin jo vanhentunutta teknologiaa, täytyy muokata muotoon, jota nykyiset visualisointiohjelmat pystyvät tulkitsemaan. Toisessa vaiheessa tuotetaan visualisointiohjelma, joka on räätälöity sekä kirjoittajien että kääntäjien tarpeisiin, todennäköisesti *GenoGraphiX* (GGX)-ohjelman beta-versio.

Aineiston lopputuotos (l. valmis teksti) voidaan annotoida samoin kuin muissa aineistoissa, mutta aineistoa voida kuvata myös visualisoinnin avulla. Tallennustyökaluilla (GenographiX, Scriptlog, Translog) voidaan ilmaista sekä paikkatietoa (paikkatietojärjestelmät, engl. *Geographic Information System* (GIS); vrt. Geohumanities) että graafeja matemaattisen graafiteorian pohjalta (engl. *graph-oriented protocol*/graph-based knowledge; Caporossi & Leblay, 2011). Paikkatietojärjestelmät edustavat staattisuutta, kun taas graafit dynaamisuutta, ja yhdessä ne tarjoavat dynaamista kuvatietoa tekstin tuottamisesta ja sen uudelleen muokkauksen prosesseista. Visualisointimetodia voidaan soveltaa myös muiden aineistojen kuvaamiseen.

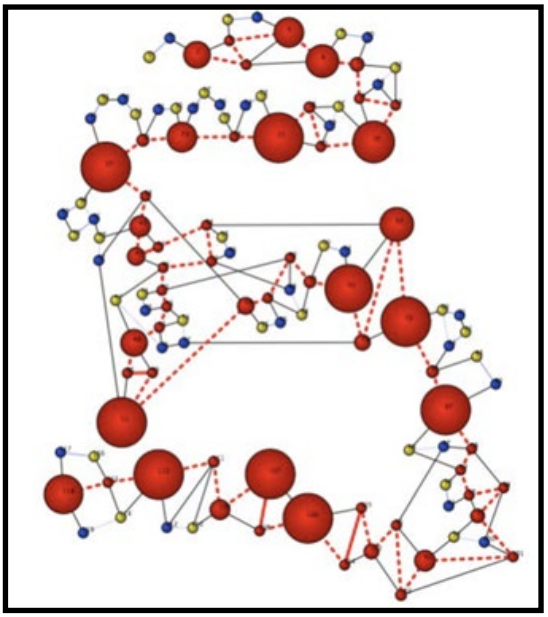
Kuvassa oikealla (ks. kuva 1) näkyy lopullinen teksti. Vasemmalla näkyy saman tekstin rakentuminen, joka on visualisoitu verkkoväritysalgoritmien keinoin (Caporossi & Leblay, 2011; Caporossi & Leblay, 2015).



Kuva 1. LOG-hankkeen kirjoitusprosessin visualisointi. Tekstin kirjoittamisen eri vaiheet (poistot, korjaukset yms.) taukoineen kuvataan suhteessa esim. poistettujen sanojen määrään

Kuvassa 1 näkyvät numerot ilmaisevat tekstin kirjoittamisen jatkumoa siinä järjestyksessä, kun kirjoittaja on solmut kirjoittanut. Mitä suurempi solmu, sitä enemmän elementtejä on lisätty (keltainen väri) tai poistettu (sininen väri). Solmujen muodostama graafin yleismuoto ilmaisee, kuinka kirjoittaja palaa tekemään muutoksia jo kirjoitettuun tekstiinsä.

Graafin muodon (ks. kuva 2) pohjalta voidaan myös havainnoida, että kyseessä on asiantuntijakirjoittajan tuotos. Graafista näkyvät asiantuntijakirjoittajan tekstille ominaiset ja toisiinsa liittyvät *toistuvat silmukat* (engl. *iterative loops*), jotka viestivät asiantuntijakirjoittajan jatkuvasta tekstinsä työstämisestä.



Kuva 2. Kirjoitusprosessin visualisointi, Becotte et al. (2019).

Kuten yllä mainittiin, prosessien kuvauksen visualisoinnin avulla saadaan tarkempaa ja monipuolisempaa tietoa kirjoitus-, käännös- ja editointiprosesseista. Graafin käyttö ilmentää prosessin vaiheiden lisäksi, millaisesta kirjoittajasta on kyse ja millainen profiili hänellä on. Tätä tietoa voidaan käyttää sekä tutkimuksen että opetuksen tukena.

3 Lopuksi

Kuten edellä esitetyistä osahankkeiden kuvauksista käy ilmi, Digilang-projekti tuo yhteen hyvin erilaisia kielitieteellisiä aineistoja sekä niiden luomisesta ja kehittämisestä vastaavia henkilöitä tai työryhmiä. Tämä mahdollistaa materiaalien tallennusmuotoja, annotointia, metatietoja ja yhteensopivuutta koskevan hedelmällisen vuorovaikutuksen tutkijoiden välillä. Projektin puitteissa samaan portaaliin tuotavat materiaalit kuvastavat monitahoisesti kohdekielten ja niiden varieteettien ominaispiirteitä sekä tekstien tuottamisprosesseja. Näin aineistot muodostavat osan sitä digitaalista kielitieteellistä infrastruktuuria, joka on välttämätön tulevaisuuden kielentutkimuksen ja kieliteknologian kehitykselle.

Digilang edistää omalta osaltaan avointa tiedettä ja tutkimusaineistojen saatavuutta. Jokaisella aineistolla on omat erityispiirteensä, ja jo tutkimuseettisistä syistä eri aineistojen käyttömahdollisuudet ovat erilaiset: osa korpuksista sisältää sensitiivistä ainesta, jonka käyttöluvat ovat rajatummat kuin toisissa. Kuten Kielipankissa osa Digilang-portaalin aineistoista tulee olemaan julkisesti saatavilla, osa edellyttää kirjautumista ja osa henkilökohtaista käyttöoikeutta.

Myös aineistojen käyttöön vaadittavat ohjelmistot vaihtelevat: osa aineistoista on valmiita käytettäviksi ilman erityisiä hakutyökaluja, ja toisten hyödyntäminen edellyttää ulkopuolisten työkalujen lataamista (esim. konkordanssiohjelmat). Mahdollisen jatkorahoituksen avulla aineistoihin pyritään kehittämään entistä saavutettavampia käyttöympäristöjä.

Turun yliopisto juhlii 100-vuotistaivaltaan vuonna 2020. Satavuotisen historian aikana aineiston keräämis- ja hallintatavat ovat kehittyneet ja monipuolistuneet: jalan tehtävät raskaat keruumatkat ja hitaat kokoelmien koostamis- ja luokittelutyöt ovat saaneet rinnalleen muun muassa digitaaliset menetelmät, jotka täydentävät ja monipuolistavat aineistotyön kirjoa. Digitaaliset nykyaineistot käsittävät sekä pieniä tapaustutkimuksia että miljardien sanojen big data -kokoelmia ja kaikkea näiden väliltä. Digilang yhdistää aineistotyön perinteet nykytutkimuksen digitaalisiin infrastruktuureihin ja mahdollistaa kielenaineistojen kehittämisen seuraavinakin vuosikymmeninä..

Lähteet

Becotte, H.S., Caporossi, G., Leblay, C. & Hertz, A. 2019. Writing and rewriting: Keystroke logging's colored numerical visualization. K. P. H. Sullivan & E. Lindgren (eds.) *Observing writing: logging handwriting and computer keystrokes* (96–124). Leyde: Brill Academic Publishers.

Caporossi, G. & Leblay, C. (2015). A graph theory approach to online writing data visualization. G. Cislaru (ed.) *Writing(s) at the Crossroads*: *The Process-Product Interface* (171–181). Amsterdam: John Benjamins.

Caporossi, G. & Leblay, C. (2011). Online Writing Data Representation: A Graph Theory Approach. J. Gama, E. Bradley et J. Hollmén (eds.) *Lecture Notes in Computer Sciences* 7014 (80–89). Advances in Intelligent Data Analysis X. Springer: Heidelberg, Dordrecht, London, New York.

Ikola, O. (2001). Suomen kielentutkimuksen historiaa. *Virittäjä*, 162–168.

Kielitaidon tasojen kuvausasteikot (liite 2). Saatavilla www02.oph.fi/ops/taitotasoasteikko.pdf
Kurki, T., Siitonen, K., Väänänen, M., Ivaska, I. & Ekberg, J. (2011). Ensi havaintoja Satakuntalaisuus puheessa -hankkeesta. *Sananjalka* 53, 84–108.
Kurki, T. & Siitonen, K. (2014). 2000-luvun hanke Satakunnan puhutusta kielestä – pronominit ja sananrajat nykysatakuntalaisten puheen ilmentäjinä. S. Heikkilä (toim.) *Virtaa läpi vainioiden – Kokemäenjoki Satakunnan historiassa* (165–191). Harjavalta: Satakunnan Historiallinen Seura.
Kurki, T., Nieminen, T., Kallio, H. & Behravan, H. (2014): Uusi puhesuomen variaatiota tarkasteleva hanke: Katse kohti prosodisia ilmiöitä. *Sananjalka* 56, 186–195.
Kurki, T. (2018). Kielikäsitysten mosaiikki – havaintoja puhekielen näytteistä. *Sananjalka* 60, 71–94.
LAO 1985 = Lauseopin arkiston opas. Toim. O. Ikola. Turku: Turun yliopisto.
Mutta, M. (2017). La conscience métapragmatique et l'attitude métacognitive épistémique des scripteurs universitaires: la révision de texte en temps réel. *Pratiques* 173–174. URL : http://pratiques.revues.org/3313
Koponen, M., Salmi, L. & Nikulin, M. (2019). A product and process analysis of post-editor corrections on neural, statistical and rule-based machine translation output. *Machine Translation*. Volume 33, Issue 1–2, 61–90.
Yli-Luukko, E. (2010): Hämäläisten laulu – puheen melodinen jaksottaminen: Intonaationtutkimusta 50-vuotiaassa Suomen kielen nauhoitearkistossa. *Virittäjä*, 396–409.
Ylitalo, R. (2004): Toisen tavun vokaalin puolipidennyksestä oulunseutulaisten puheessa. *Virittäjä*, 414–422.

Best practices in bibliographic data science

Leo Lahti (University of Turku), Ville Vaara (University of Helsinki), Jani Marjanen (University of Helsinki) and Mikko Tolonen (University of Helsinki)

Abstract
Bibliographic data science aims to quantify historical trends in knowledge production based on the rich information content in library catalogue metadata collections. Compared to the earlier attempts in book history, advances in data science are now making it possible to automate remarkable portions of such analyses, while maintaining or improving data quality and completeness. Such quantitative approaches can support the analysis of classical questions in intellectual history. Here, we discuss best practices in this emerging research field.

1 Introduction

Large-scale integration and analysis of bibliographic data collections and supporting information sources across time, geography, or genre could shed new light on classical hypotheses and uncover overlooked historical trends. This quantitative research potential has been long recognized [1–5]. Systematic large-scale research use of bibliographic collections has proven to be challenging, however. Bibliographic data science (BDS) [6] develops systematic methods for the harmonization, analysis and interpretation of bibliographic metadata collections. Scaling up bibliographic data science can remarkably benefit from the developments in open data science [7–9]. Automated harmonization can enhance the quality and commensurability between metadata collections, thus complement- ing linked open data and other technologies that focus on data management and distribution. Unique to our efforts is the focus on historical interpretation and the necessity of incorporating prior knowledge on the data collection processes which may introduce biases in the analyses. However, ensuring data quality and completeness are critical for research, and opening up the research workflows can support collaboration across institutional and national borders.

In order to address these challenges, we have recently proposed the concept of bibliographic data science (BDS) [6] and provided the first case studies to demonstrate its research potential. Here, we discuss challenges and best practices that can facilitate cumulative research efforts in this field.

2 Best practices

2.1 Reproducible data harmonization and analysis

Bibliographic metadata is often manually entered in the databases, and seldom sufficiently standardized for systematic quantitative analysis. Harmonization of the raw data entries is the first step towards reliable research use. Biases, gaps, and varying standards pose challenges for data integration both within and across catalogues. Heterogeneity of the data fields, from time intervals to persons, physical dimensions, or geographical locations form challenges for analysis [7]. The metadata fields can often have complex dependencies, and harmonization of one field may influence the harmonization and enrichment of the other.

In order to efficiently manage these multi-layered data harmonization processes, we have implemented a data science ecosystem that integrates various individual workflows (Figure 1), incorporating pragmatic approaches from data science [10, 11]. These workflows remove spelling errors, disambiguate and standardize terms, augment missing values, and incorporate manually curated information on pseudonymes, duplicate entries, and other information types. Reproducible workflows support transparent data processing which can be efficiently monitored and improved over time. This can further benefit from the design of dedicated software packages[1]. We have used these tools to extensively harmo- nize selected fields of the Finnish and Swedish National Bibliographies (FNB and SNB, respectively), the English Short-Title Catalogue (ESTC), and Heritage of the Printed Book database (HPBD). Altogether, these collections cover over 6 million entries of print products printed in Europe and elsewhere. Since all bib- liographies are MARC compatible[2], we can apply largely identical processing across all collections, thus facilitating transparent and scalable data processing [6, 12, 7].

[1] https://github.com/COMHIS/bibliographica
[2] Library of Congress web document https://www.loc.gov/marc/bibliographic/

2.2 Semi-automated data curation

The scale of data harmonization greatly exceeds our ability to manually correct and verify individual entries. Automated workflows can be used to standardize data treatment across multiple catalogues, allowing iterative corrections. In addition to inaccurate or erroneous entries, duplicates and missing information can induce systematic bias in the analyses, in particular when they are unevenly distributed. Such biases can be detected by systematic quality monitoring (Figure 2). Missing information can be often augmented based on other information sources. For instance, missing country information can be inferred when the publication place is known, and information on authors can be found from historical databases.

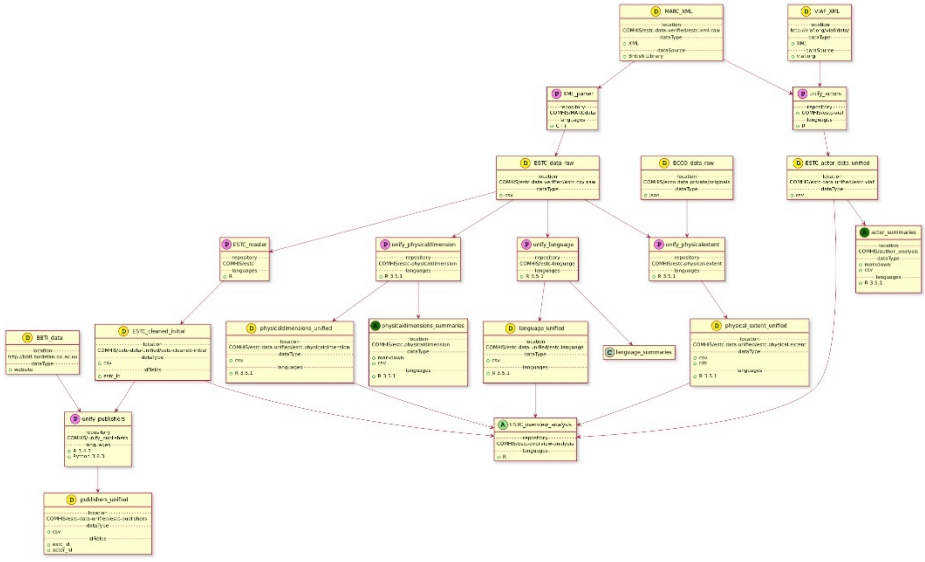


Fig. 1. Dependency chart describing the multi-layered data harmonization process of the English Short Title Catalogue.

Whereas manual verification remains a key component in data curation, this can be greatly facilitated by automation. We have implemented systematic approaches to assess and improve data reliability in a semi-automated fashion based on unit tests, outlier detection, and external verification. Automatically generated summary reports of the harmonized data are a key part of this process, and can be accessed for the different bibliographic catalogues, such as the ESTC and FNB, through the

homepage of Helsinki Computational History group. An example of this is the comparison of page counts between a subset of the ESTC entries from our workflow, and a distinct database, the Eighteenth Century Collections Online (Figure 3). Furthermore, we generate conversion tables that show how common entries have been harmonized. By visualizing complementary aspects of the data such as document dimensions, publication years, author life spans, or gender distributions based on timelines, scatterplots, histograms and heatmaps we can obtain further insights to potential inconsistencies (Figure 4). Such overviews can be invaluable for the detection of inaccuracies or biases in the harmonized data sets.

Moreover, data integration across catalogues enables the detection of robust patterns that are supported by multiple information sources. Joint analysis can be useful in terms of assessing the historical representativity, thus paving the way for future data integration and analysis. Our recent analyses provide examples of the research potential of this approach [6, 12]. For instance, all metadata collections that we have analysed show a systematic rise of the octavo format during the eighteenth century and a parallel, steadily declining trend in Latin publications towards the eighteenth century [6].

2.3 Open science and collaboration

Academic research can benefit enormously from open sharing of data and algorithms [13]. There are differences between research fields, however. Open data sharing in the humanities is still less well developed than in natural sciences but this might be gradually changing.

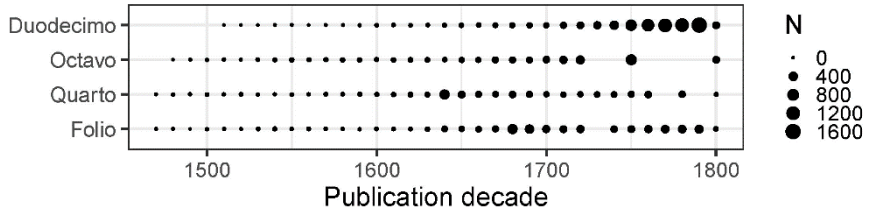


Fig. 2. Visualization of missing page count information by decade (horizontal axis) for different book formats (vertical axis) reveals systematic biases in data availability in the ESTC.

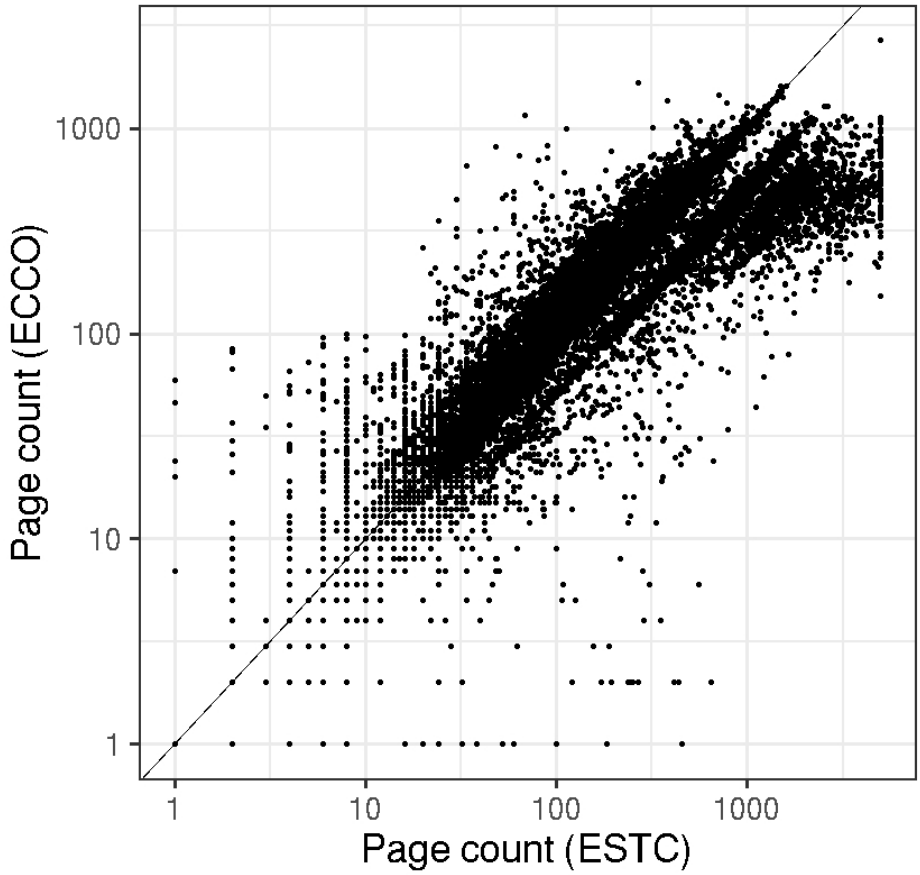


Fig. 3. Comparison of page count estimates to an external validation set. The English Short Title Catalogue (ESTC) includes 175727 documents that have a page count estimate and can be matched with the Eighteenth Century Collections Online (ECCO) database. The ECCO database contains page count information for the same works. The Spearman correlation between the two sources is ρ = 0.99. The comparison quantifies the quality of data harmonization and can potentially highlight issues that would benefit from further improvements in the data harmonization workflow.

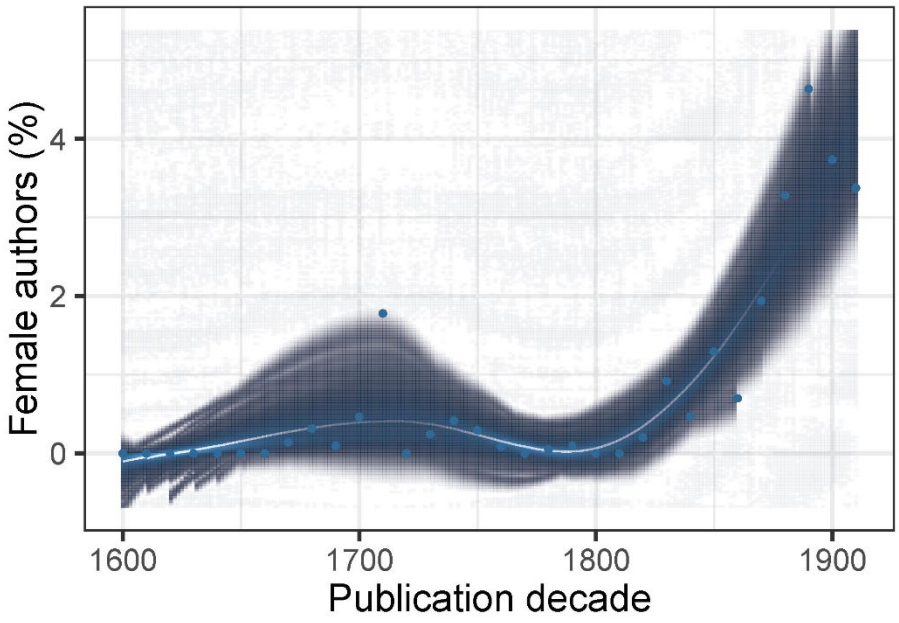


Fig. 4. Proportion of female authors in the Finnish national bibliography (FNB) in the period 1600-1900. The visualization can also highlight unexpected patterns in the data, such as outlier points and the temporary increase in the early 18th century, which may indicate inaccuracies in the author harmonization and gender identification algorithms.

In Finland, for instance, The National Library has recently released the FNB under an explicit open data license[3]. This has al- lowed us to generate and share further, harmonized versions that can be accessed via Helsinki Computational History Group website [4]. Combining such large-scale harmonization with existing infrastructures could open up new opportunities for research.

Open sharing of research data and algorithms can facilitate collaboration across institutional and national borders and increases transparency and efficiency in research. By making our key algorithms, software packages, and work- flows available under open licenses [14] we are allowing others to search, detect and

[3] http://data.nationallibrary.fi/
[4] https://www.helsinki.fi/en/researchgroups/computational-history

report inaccuracies, or make further corrections and utilize these resources in independent research and methods development.

2.4 Contributions and interaction between fields

The research methods discussed in this paper contributes to traditionally distinct fields. The study is motivated by classical questions in book history, intellectual history, and early modern history. In addition to describing broad trends in knowledge production, it is becoming possible to provide exact large-scale quantification of historical trends using library metadata catalogues. This can be contrasted with the existing knowledge in order to confirm earlier hypotheses and as well as to highlight trends that so far received less attention. The new quantitative techniques bring up new historical information and evidence, thus complementing more traditional qualitative methods in history and related fields. At the same time, the work contributes to methodological research from linguistics and natural language processing to machine learning, data science, and interaction design. Quantitative methods can support data harmonization and analysis and integration of the metadata collections with complementary data types such as full texts and geographical information. The new applications can inspire further research in the methodological fields.

3 Conclusion

This short paper has reviewed current and emerging best practices in bibliographic data science. This is an emerging research paradigm in the digital humanities, which focuses on the research use of bibliographic metadata and potentially other related data types such as full text collections and complementary information on persons, organizations and places. Whereas this overall research theme is opening up opportunities to study a variety of traditional research questions in history, sociolinguistics or related fields from a new angle, the focus on the present work has been in demonstrating the challenges for methodological research [6, 12, 15]. There are a number of technical research questions on how to optimally design, choose, implement, validate, or utilize the research algorithms in general and in the specific context our application domain (see e.g. [8]); a full discussion of these aspects is beyond the scope of this short paper, however. Specifically relevant for our application are the questions on how to combine qualitative and quantitative

elements in a reliable and unbiased manner; or at least how to identify and deal with the possible biases, inaccuracies, and gaps in the research data and analysis process. Whereas the concept of bibliographic data science has been proposed only recently [6], it builds on the existing traditions of bibliographic research [1–5] and open data science (see e.g. [14, 16, 8]). Bib- liographic data science is expanding the research potential of large-scale library catalogues. Adaptations of this method can be used in other related fields, such as the study of materiality in newspapers [17]. Automation and quality control are essential when the data collections cover millions of documents. Advances in machine learning and artificial intelligence are now providing further means to scale up the analysis as the already curated data sets provide ample training material for supervised machine learning techniques. Open sharing of research data and analysis methods can support collaborative and cumulative research efforts. We have demonstrated how specifically tailored open data analytical ecosystems can help to address these challenges.

Acknowledgments

We would like to thank members of Helsinki Computational History group for their contributions to data harmonization, Hege Roivainen for his support in the analysis, and Eetu Mäkelä for providing the ECCO page count information. This work has been supported by Academy of Finland (decision 293316).

References

1. Tanselle GT (1974) Bibliography and science. Studies in Bibliography 27: 55–90.
2. Giesecke M (1991) Der Buchdruck in der frühen Neuzeit: eine historische Fallstudie über die Durchsetzung neuer Informations- und Kommunikationstechnologien. Suhrkamp, Frankfurt am Main.
3. Bozzolo C & Ornato E (1980) Pour une histoire du livre manuscrit au Moyen Age: trois essais de codicologie quantitative. Equipe de recherche sur l'humanisme francais des XIVe et XVe siecles. Editions du Centre national de la recherche scientifique, Paris.
4. Bell M & Barnard J (1992) Provisional Count of STC Titles, 1475–1640. Publishing History 31(1): 4764.

5. Horstbøll H (1999) Menigmands medie: det folkelige bogtryk i Danmark 1500–1840: en kulturhistorisk undersøgelse. Danish humanist texts and studies, volume 19. Det Kongelige Bibliotek & Museum Tusculanum, Copenhagen.
6. Lahti L, Marjanen J, Roivainen H & Tolonen M (2019) Bibliographic data science and the history of the book (c. 1500–1800). Cataloging & Classification Quarterly pp. 1–19. Special issue.
7. Tolonen M, Marjanen J, Roivainen H & Lahti L (2019) Scaling up bibliographic data science. In: Proceedings of the Digital Humanities in the Nordics (DHN2019).
8. Lahti L (2018) Open data science. In: Advances in Intelligent Data Analysis XVII. Lecture Notes in Computer Science 11191.
9. Borgman CL (2015) Big data, little data, no data: scholarship in the networked world. The MIT Press, Cambridge, Massachusetts; London, England.
10. Wilson G, Bryan J, Cranston K, Kitzes J, Nederbragt L & Teal TK (2017) Good enough practices in scientific computing. PLoS Computational Biology 13(6): e1005510.
11. Wickham H (2014) Tidy data. Journal of Statistical Software 59(10): 1–23.
12. Tolonen M, Lahti L, Roivainen H & Marjanen J (2018) A quantitative approach to book-printing in Sweden and Finland, 1640–1828. Historical Methods: A Journal of Quantitative and Interdisciplinary History pp. 1–22.
13. Morin A, Urban J, Adams P, Foster I, Sali A, Baker D & Sliz P (2012) Research priorities. shining light into black boxes. Science 336(6078): 159–60.
14. Morin A, Urban J & Sliz P (2012) A Quick Guide to Software Licensing for the Scientist-Programmer. PLoS Computational Biology 8(7): e1002598.
15. Lahti L, Ilomäki N & Tolonen M (2015) A Quantitative Study of History in the English Short-Title Catalogue (ESTC) 1470-1800. LIBER Quarterly 25(2): 87–116.
16. Ioannidis J (2014) How to make more published research true. PLoS Medicine 11(10): e1001747.
17. Marjanen J, Vaara V, Kanner A, Roivainen H, Mäkelä E, Lahti L & Tolonen M (2017) Analysing the language, location and form of newspapers in finland, 1771–1910. Technical report, Digital Humanities in the Nordics, Gothenburg. Conference abstract.

Digital heritage presentation system development + new material types: early findings

Tuula Pääkkönen,
National Library of Finland

Abstract

Recently the National Library of Finland has initiated a development project to include books in the presentation and content search system of its digital archive, Digi (https://digi.nationallibrary.fi). Before year 2019 Digi contained newspapers, journals and ephemera (technical). The new project, nicknamed 'Books to Digi', aimed to streamline the digitisation process so that books can be imported to the presentation system automatically after the post-processing phase of digitisation.

Besides the new production process steps, the project considered a number of other requirements submitted by end-users and stakeholders (different units of the research library). As part of the project, we analysed the requirements, identified user segments and condensed the requirements to the core set. During implementation, use cases, Pugh matrix (Pugh, 1991) and agile software development methods were used to keep system development running and to maintain a continuous throughput of tasks.

Apart from requirement gathering and implementation, the third aspect of the project was to utilise the metadata lifecycle within the library in the best possible way. The national bibliography uses standardised rules for metadata creation, but these rules have changed throughout the years. For an information system, this required some normalisation of the metadata. We normalised publisher names and the publication places to make the content search easier to use. This enrichment of contents was also seen as potential improvement, which in the long run could also be extended to the digitized ephemera, which would then be similarly well-organized as the digitized newspapers and journals.

The implementation of the project has been launched alongside with the new digitized materials (Lehmikoski-Pessa, 2019)(Lehmikoski-Pessa, 2019). One survey was made on 2018 (Pääkkönen & Kettunen, 2018) to get a baseline of user aptitude towards Digi and a new survey will be done on 2019, to evaluate the impact of the new materials and search capabilities. We hope that this project will enable people to do interesting findings via full text search capabilities towards the digitized contents.

1 Introduction

According to one estimate from the National Library of Finland (NLF), around 2% of the material in all its collections is available in digitised format (Kansalliskirjasto, 2018). The digitisation process follows the digitisation policy (National Library of Finland, 2010). There are long-term digitisation plans where the goal is to broaden the large corpus systematically. The annual digitisation plan, is more versatile and includes, for example, material which has been requested by researchers or via customer service of the NLF. Often if a material is used a lot in its physical form, it is added to the digitisation list, in order to preserve it for future users. In several separate projects, researchers' digitisation needs, for example, have been advanced by digitising Finnish Classics and History of the Books collection over the years. In these digitisation projects, the selections for the digitisation have been done by the researchers from the collection of the NLF.

These digitised materials are available to researchers and the public in the digital presentation system called digi.kansalliskirjasto.fi. This system allows users to do full text searches of the contents of the digitised material, to visualize search terms on the page via highlights and to view the metadata of the bindings to mention a few. In addition, it is possible to create clippings, which are a way to store interesting subcomponents of the page.

Recently, a new agreement was made for digitising 2,000 books. There about half were planned to be made available for researchers and around 1,000 to the general public. The agreement was signed in late 2017 by the National Library, and the copyright organisations Kopiosto and Sanasto. After the agreement was signed, selection of the material began, as did creation of metadata, followed by the digitisation. Modifications to the presentation system (digi.kansalliskirjasto.fi) were also started in order to incorporate different kinds of book materials (books, maps, catalogues and manuscripts). The modifications were necessary to fulfil the requirements of the agreement and the stakeholders. The existing platform, which supports newspapers and journals, required several changes to address all of the metadata and usage-specific details required by the users. The information about the agreement, the new features and the new materials were published on the 16th of May 2019 (Lehmikoski-Pessa, 2019).

Part of the digitisation has also been about expanding and exploring various material types, which exist in the collection of the National Library. Books, sheet music and ephemera have been digitised, and all of these require specific

configuration in the post-processing to ensure that the outcome from digitisation is the same regardless of original material. All these new demands are also applicable to the presentation and search of the contents of digitised materials. Our research questions were to a) identify changes needed to the user interface to enable it to support the book material, and b) how to implement those changes, so that the future operations would be as streamlined as possible.

In this paper we focus on the overall development project of 'Books to Digi', and to the special enrichment steps done for the books. We highlight the normalisation of the publishers and publication places. As they are in national bibliography as in the original work, they are versatile in spelling, in the digitised materials of books, maps, theses and sheet music. We learnt that it is useful to do some of these enrichment steps at the time of digitisation when the digital object is viewable. Digital access to the object enables the validation of data between different works and enables having a general understanding across them.

1.1 Copyrights and contents

In Finland, the author's copyright lasts for 70 years after their death. For newspapers and journals, this means that every article writer or illustrator holds the copyright to his or her material. For books, the situation is slightly less complicated, as there are fewer copyright owners, typically just authors and illustrators.

The 'Books to Digi' project ventured into a new area as its goal was to make in-copyright materials available, partly to researchers and partly to the general public (Lehmikoski-Pessa, 2019). This means that it requires modifications to the system to show usage terms alongside the digitised material. The agreement is a pilot project for the NLF and the copyright organisations to analyse how much this kind of content is used and by how many users.

The number of individual users can only be estimated. People can use computers offered by public libraries and utilise multiple devices or proxy services. The copyright organisations also required lists of the selected books so they could review them before their access is opened in the presentation system.

The 2,000 books selected for this project were selected by the researchers from the 'Classics library' and 'Book history' projects within the University of Helsinki (Biström, 2019; Laine, 2019). The books were either fiction or non-fiction. This was a significant divider, as the non-fiction books were made available to the public, whereas the fiction books were restricted to researchers of all universities. The

newsletter of the launch was well received and for example the ABC-books got even a mention in an editorial in the largest newspaper in Finland (HS, 2019).

1.2 Benchmarking existing presentation systems

Prior to any major development efforts, we briefly analysed the core features of digital presentation systems of various national libraries. The analysis was done to keep up-to-date with the newest developments and get an alternative and broader view of the development work done within libraries.

We viewed and observed some key functions of digital presentation systems in the national libraries of the Nordic countries, Estonia, Poland, Great Britain, France, Australia, the La Stampa archive of Italy and the Bavarian State Library to mention a few. We did not do a full requirement analysis across them, but targeted the specific features for the benchmark we were considering. We also looked at whether there are some new features that could be applied. As a general summary, there seems to be a trend that the focus shifts to the presentation system only after the digitised material is available. This is what happened for digi.kansalliskirjasto.fi as well. Digi was launched in 2001, but after that it has been developed in multiple externally funded projects. The evolution of the presentation system has followed the needs of the digitisation process of material and the end-user demands.

The studied systems have several similarities. When digitised material is available, the same page level features exist, for example zooming in and out, swapping pages, downloading PDFs and usually also showing the copyright terms for the binding. These page level features are all available in Digi too. Some systems also featured multi-page view, which showed either a spread of pages or multiple pages. We did not include this feature in the scope of this project, but left it as something to consider in a separate project later on. Bookmarks and author information were only implemented in a couple of systems, so we left these for future consideration as well. All the systems analysed had some unique features or ideas, which we could learn from and utilise when considering our existing and potential user types. For NLF, tight integration with the existing digitisation was important. This meant that we could not use an open source tool, as the features of such a tool would have been limited. Further, using an open source tool would mean that we could not utilise the features that already exist and work for newspapers, journals, and ephemera.

1.3 From requirements to user features

Beside the internal requirements it is important to think for the end-user requirements. Based on end-user feedback in light-weight social media discussions, we hear the need "could be get…, why wouldn't you..". [1] The social media discussions confirmed those feedbacks that we had got also via the Feedback functionality of Digi and user surveys. Social media has very active expert users, who have used the presentation system for a long time, so we do need to listen to their opinion. Constant question were kept pondering was that which features we should be offering? For example, (Golderman & Connolly, 2009) use in their review of innovative "database products" use four criteria: content, searchability, pricing and end-users. Content and searchability have always been targets for the digital collections of National Library of Finland, and e.g. the main portal Finna.fi was recently praised in the end-user survey as having search, which end-user like to recommend to the other users.[2] However the idea of "generous interfaces" of the digital collections, as suggested by (Whitelaw, 2015), is also very appealing as it would be way to bring the content out in the way it deserves. In the strategy of development of the digital collections further, there are various alternatives also for improving the user interface, so the step towards that way depends on what kind of development projects are started in the coming years. However, as we aim to find low-resource ways to improve the quality of optical character recognition of text content (Kettunen et al., 2014), we also need to be frugal in defining new features to the presentation system and fulfil only the most important requirements.

2 Methods

The research methods for this project were chosen based on their suitability for a) analysing the user (and stakeholder) requirements of totally new material types and b) tackling the technical aspects of the needed features. The requirements and expectations of external users were identified from surveys. Internal stakeholder

[1]https://www.facebook.com/groups/360776620709181/permalink/839188389534666/?comment_id=839188712867967&comment_tracking=%7B%22tn%22%3A%22R2%22%7D (Discussion in Facebook OpenGlam-group, September 2015

[2] http://kansalliskirjasto.fi/fi/valtaosa-finnan-k%C3%A4ytt%C3%A4jist%C3%A4-suosittelisi-hakupalvelua-muillekin (Newsletter of survey results, in Finnish, 22.12.2015)

requirements were captured through workshops. Throughout the project, intranet web pages were available to discuss and manage the expectations of internal users.

2.1 Analysing requirements

Before the project started, there were internal discussions regarding whether a project was needed or if the existing solutions would suffice. This discussion phase was a prime source of original requirements for the NLF presentation system. We also analysed existing user surveys to identify the expectations from the presentation system. When analysis and approvals were complete, the 'Books to Digi' project started in early 2017 with a specification phase. Even though the agreement was not yet signed at this point, we started early to ensure a buffer in the schedule to deal with surprises. The requirements were written in one-page use cases. Experts of their area in the library commented on the use cases as did those who belonged to the identified user segments.

After the main comments were analysed, 66 main features and four basic user segments were identified (readers, metadata experts, maintainers and researchers). The aim of identifying these user segments was to have slightly different viewpoints on the materials and also to incorporate the expectations of the agreement, which was being drafted at that time. We had an agile mindset, so within this core set of requirements, we were open to changes as we learnt more during the project. All the use cases were made available and kept updated on the intranet. They were also stored to the version control system alongside the code base.

The project was to start in spring 2017 and aim was to be mostly completed by late 2018. We had 1.5 developers and 1.5 support people for specification, documentation and testing. The support people were involved in other projects at the same time, so the we planned schedules so that if one part of the project was blocked, the developers could proceed with a different part.

From the beginning, we realised that the task ahead was large. Solution was to utilise existing work done on the presentation system and to serve the digitisation process more fluently. The goal was to get to the normal mode with books, meaning that, after digitisation, the books would enter into use automatically. As the timeline was short and resources were scarce, we had to be very meticulous with new demands and leave certain things out, if the use case was not strong enough. In order to leave some room for further enhancements, not even all 66 features were developed during the project, as something more important was implemented. All

the features were not of similar sizes, so some features were left for the further development later on.

2.2 User input: the survey and feedback

When it was possible, the new version was put into use in an all-access testing environment. This environment was mentioned in social media and on the front page of digi.kansalliskirjasto.fi, to allow interested parties to use it and give input on the new features as they were being developed. User surveys, most recently the survey in 2018 (Pääkkönen & Kettunen, 2018), contributed greatly to our understanding of the current user outlook on the content. The survey responses helped us to fine-tune the new features in the background. As the overall finding in the survey was that end users were relatively satisfied with the search options and results, the search was kept the same as much as the new material types would allow. We also utilised the Pugh matrix (Pugh, 1991) to evaluate design choices. This method gave us a more objective view of certain features. The Pugh matrix lists core requirements, assigns weights to them and then analyses the current situation with any number of alternatives that relate to the same requirements.

Also, as is core in Agile development (Sutherland, 2014), we welcomed feedback and changed direction if the feedback gave us valuable input about any user needs, which we had missed originally. For example, some usability experts claim, that for usability feedback it can be enough to interview around five people, as typically end users focus on the same things so it is possible to get valuable feedback with limited effort (Krug, 2009).

The long-term development need was that the solutions should be sound and scalable, while utilising the capabilities of internal processes. For example, metadata creation, alongside digitisation, is in the core of the National Library system, so the desire was to utilise the metadata to help in the content search. Even if the content source is the most used feature, expert users can also benefit from seeing the metadata. The metadata can be used to further filter the content-based search results. We also had to improve certain metadata, namely existing publisher and publication place data to make it easier to process and use. The metadata in the presentation system was taken from the long-term preservation package, which contains the metadata as recorded at the time of digitisation. We had to make some adjustments to obtain recent data from national bibliography, which helped to improve the search functions in the presentation system.

2.3 The agile development cycle

The development cycle was quite fast in true Agile manner. We had new versions available in the test environment daily if not hourly. The basic process was for a feature to first go to the alpha environment (separate backend systems), and then to the test environment (full database). When we had to change the internal data model of digi.kansalliskirjasto.fi, database structure or search indexes it was very beneficial to experiment and fine-tune them in the internal quality assurance environment (alpha). Some individual features were also put into the production environment earlier to see how they would be used in real life.

From the use cases, more detailed descriptions were created which were added to our backlog. The most critical items from the backlog were developed. When a feature was developed, it was tested according to the test plan and fixed or changed, if required, based on internal or external use in the public test environment.

3 Presentation and metadata fixes for books

In the simplest form, the goal of the project was to create support for 'books' in the digi.kansalliskirjasto.fi presentation system. In library terminology, books mean monographs, i.e. books, maps, sheet music and catalogues. From the digitisation point of view, the final output of digitisation follows the same principles for all types of materials. Having support for all kinds of monographs was relatively straightforward, as metadata, belonging to different material types, varies in only around 5% of the fields, thanks to the common rules of cataloguing. This meant that once we were able to support one monograph, we were able to add support for new material in the same manner. In the following sections, we explain how two of the most versatile metadata fields: publisher and publication place were normalised, in order to make search fields within the presentation system more user-friendly.

3.1 A publisher with many names

In the creation of the national bibliography, the general work practice has been that the authors, publishers and publication places are stored in the national bibliography as they have been written in the original work. This means that there are differences in how even the same publisher is stored and shown for different works. When the Finnish National Bibliography was used as a research source for

quantitative analysis (Tolonen, Lahti, Roivainen, & Marjanen, 2018), a need was felt for the harmonisation of a few key fields. We wanted to fix this situation in the normalisation of publication places and publishers, of which we tell in the following.

During the long running digitisation effort, as the digitisation methods have evolved, there have also been changes in the rules on how different works are described in the national bibliography. For any information system this can be a problem, because users might not understand all the variations. For example, the user may not know the appropriate organisation name to use, when searching for a particular work. Therefore, we saw the need to harmonise the names when we brought new material to the presentation system. A separate internal tool was made to ease the work, where we have suggestions for normalized names or links to various information sources for finding them. In the tool, the existing description of the work is used as a baseline, and then the publisher names are changed to a harmonised form. In many cases, we can find the name from the Finto.fi ontology system, which also contains all the name variations. This eases the job considerably. Not all names are fully harmonised yet, but work can continue alongside current daily operations. In the long run, we hope to have all names harmonised, which could then be used elsewhere too. We have the variances of the names stored with the final normalised forms, so potentially it could be a data set of its own. By May 2019, we had completed over 10,000 normalisations to publisher data, ranging from correcting simple typos to adding a missing publisher to the metadata of an item.

3.2 Multitude of historical place names

Similar to the publisher names, the publication places were not consistent. One special issue with them was the missing lemmatisation. All words were written as is, so there were a lot of extra data in the place names. For example, ‘painettu Helsingissä’ (meaning ‘printed in Helsinki’) was listed, instead of just the base form Helsinki. Even though this is not a major issue for a user, it is a useful change that we estimated we could fix during the enrichment phase of the digitisation. This was one of the original features that we decided to focus on. The internal tool used for publisher names was extended to include publication places, too.

We hope to receive feedback on this normalisation from the users. There are many things that need to be known for successful normalisation. The original

publication languages vary, and even if we can deduce the information from some, there can be some languages like Latin, which requires knowledge of that language.

Normalisation of place names also requires consideration of what the place name was at the time that the material was written. For example, using all alternatives, like Dorpat, Tartto instead of Tarto would make search fields more complicated. We need to investigate and evaluate how users utilise existing information systems more, when we make further improvements. The Nimisampo (SeCo, 2019) might also be very useful in the future to incorporate the work of researchers and the general public into one system.

3.3 Illustrations within the books

In the digitised materials of Finland, the illustration search is one of most rarely used features. Fewer than 2% of searches utilise this feature. The feature is new and not very granular – just on or off while utilising the text within the same page for enabling the illustration search. Whether the illustration fits the search or not depends on how the layout of the page (text and illustration) was created originally. In the future we hope to make better use of the illustrations, by categorising them. This would make it easier to find usable illustrations. In the case of the British Library, as reported by (Seymour, 2014), digitised books were made free to use by the public to encourage engagement with the content. The British Library has achieved 147 million views and 80,000 tags added by public (Seymour, 2014). This success would support the idea of presenting illustrations in a better way. However, our categorizing method will have less than 10 main categories in the first phase and mainly focusing to the overall category of the item, like ‘ad or ‘photograph’, which we could utilize in determining the copyright status of an illustration.

In the case of the ‘Mechanical Curator’ of the British Library (BL), the overall goal was to go through the digital collections and publish content (OSteen, 2013). The same approach seemed suitable for the NLF as well. For the illustration search, image extraction, as mentioned above, was the first step. The searchability and content itself were also important, so those were the key features we considered. The image extraction process was designed to require the fewest possible human resources and to automate as much as possible.

The first approach to image extraction consisted of four steps: finding the images, extracting the images, tagging them and making them available to the presentation system. Our approach was very similar to the British Library – target

the problem with several small steps and build the whole solution from them. Every library has their own unique process, methods and tools to store digital content. After analysing the code of the Mechanical Curator[3], it became obvious, at least based on the scripts, that the BL and NLF differ in the methods they use to store and access digital collection material in the backend. This meant that it was possible to utilise the BL implementation after customizing it for the environment of NLF.

A major benefit from the post-processing of digitisation, is the page content uses the ALTO (Analysed Layout and Text Object) XML format and METS (Metadata Encoding and Transmission Standard) standards compose the post-processing outputs for all material types. The book content of BL mentioned in O'Steen (2013) utilizes the ALTO standard. This meant that we could take suitable code from the BL embellishments scripts, which targets specific part of the XML, and use it to get the area information for the illustrations. As the image capturing process existed already for newspapers and magazines, adapting that system for books was simple. Most of the changes were mainly about how to control the execution so that it would target the monograph materials and index them accordingly for the search functionalities.

The ALTO file usage made it possible to utilize also the actual image extracting logic from the Mechanical curator. It uses OpenCV[4], an open source computer vision library, which does the "heavy lifting", i.e. extracting the image from the page in question based on the coordinates got from the ALTO XML file. This also means that the quality of the image extraction depends on the quality of the post-processing done in the digitisation. The better the post-processing tool and quality assurance phase, the more accurate the illustration usage is.

4 Discussion

4.1 Content and end-users

In a way images from digital newspapers and journals does not sound so significant. The material could have been found from the digital collections, with enough luck with the search and tenacity. However, in real-life it is quite unfeasible to go

[3] https://github.com/BL-Labs/embellishments
[4] http://opencv.org/

through the over 3 million pages of the collections, which are freely available in the public web service, let alone the full nearly 10 million pages which has includes also the copyrighted material. Especially for the researcher use, the images as a distinctive data set, could be interesting, but even more so for most lay-man users, who browse the materials with related to their work or just for fun. For the National Library of Finland itself both of those user groups are equally important – researchers in the sense of creating long-term impact especially in the humanities, and nowadays digital humanities, which is one of the focus areas. Then again, as the vision of the NLF "national treasures to all" states, the second objective is also to offer all the content to everybody in order to fulfil the societal objectives of NLF.

4.2 Searchability of the content

The findability also relates to the end-user type – some are researchers, some are doing citizen science with their own research goals, there are teachers, students, genealogists, and the end-user ages can range from 11 years to the over 75. There are digital natives who enjoy the search and presentation system as it is, so how does one improve the digital content with something new like images, but still make the search usable for all? After the previous version of the current digital collections of newspaper and journals user interface was retired, there were few but very adamant voices in the feedback, which desired additions to the search. Some feedback has already been incorporated to the presentation system, and some is still in the backlog waiting to be decided whether the wish fits to the overall long-term development plan. A survey for collecting feedback later on can help capture specific things and Feedback functionality is always available.

We believe that best approach is to take small steps of improvement, as it helps us to define the changes needed and it lets us evaluate the feedback we get, and possibly avoid any change resistance (Wodtke, 2013), which is quite common in any information technology project.

5 Conclusions

When implementing new features to support new material types in a digital presentation system, in this case digi.kansalliskirjasto.fi, there are requirements both from the end-user and the technical development perspective, which need to be considered. Given the constraints of the development (resources, time and

desired quality), there is a balance, which needs to be achieved by making compromises along the development journey.

The survey of the end users (Pääkkönen & Kettunen, 2018) was invaluable as it revealed that there are a multitude of different user types, ranging from academic researchers and family researchers to authors and even pure browsers of the content. For all the users, the basic functionality and availability of content from the collections is key. There are also returning users and new users; both groups should get new insight from the contents. The most often needed functions should be easily available. However, even more complicated searches should be possible to accommodate more demanding search requirements. For these reasons, the new search was kept the same as the existing one, with the addition of material-specific fields for books. The new fields are author, keywords and series, and they were obtained from the guidelines of the national bibliography.

In the future, it might become necessary to enrich the metadata even further based on the qualities of the digital object. For example, it would be trivial to add information about page counts, which would enable the calculation of more accurate paper consumption rates for different centuries. In addition, the analysis of illustrations and their creators, could create new, more detailed views of authorship. Digital versions could have different metadata generated from the digital object, which could be offered, for example via the national aggregator, to help users locate the material they are interested in more quickly.

Acknowledgments

We are grateful to the colleagues of National Library of Finland. Part of this work is funded by the Academy of Finland project COMHIS – Computational History and the Transformation of Public Discourse in Finland, 1640–1910, decision number 293341.

References

Biström, A. (2019). Klassikkokirjasto -- Kaikkialla, kaikkien kanssa… och samma på svenska. Retrieved from http://www.doria.fi/handle/10024/168853

Golderman, G., & Connolly, B. (2009). Multimedia multiplied. *Library Journal*, 134(17), 104–112. Retrieved from

http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=44668942&site=ehost-live&scope=site
HS. (2019, May 21). Pääkirjoitus: Helppo tiedonhaku auttaa myös ymmärtämään omaa kulttuuria. Retrieved from: https://www.hs.fi/paakirjoitukset/art-2000006112401.html?share=bbb94a195607c1a01304b5fcacc4e688
Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T., & Kervinen, J. (2014). Analyzing and improving the quality of a historical news collection using language technology and statistical machine learning methods. Paper presented at the *IFLA World Library and Information Congress Proceedings 80th IFLA General Conference and Assembly*.
Krug, S. (2009). *Rocket surgery made easy: The do-it-yourself guide to finding and fixing usability problems* (1 edition). Berkeley, CA: New Riders.
Laine, T. (2019). Humanistisia tietokirjoja ja lähteitä tutkimuksen tueksi. Retrieved from http://www.doria.fi/handle/10024/168850
Lehmikoski-Pessa, T. (2019). Kansalliskirjasto avaa 2000 kirjaa yleisön ja tutkijoiden verkkokäyttöön [Text]. Retrieved from: https://www.kansalliskirjasto.fi/fi/uutiset/kansalliskirjasto-avaa-2000-kirjaa-yleison-ja-tutkijoiden-verkkokayttoon
National Library of Finland. (2018). Kokoelmistamme on digitoitu hyvin pieni osa, n. 2 %. [Tweet]. Retrieved from @NatLibFi website: https://twitter.com/NatLibFi/status/1007511977743339521
National Library of Finland. (2010). The digitisation policy of the national library of finland. Retrieved from http://www.doria.fi/handle/10024/94305
OSteen, B. (2013). *BL-labs/embellishments* British Library Labs Project. Retrieved from https://github.com/BL-Labs/embellishments
Pääkkönen, T., & Kettunen, K. (2018). Kansalliskirjaston sanomalehtiaineistot: Käyttäjät ja tutkijat kesällä 2018. *Informaatiotutkimuksen Päivät 2018*, Retrieved from http://urn.fi/URN:NBN:fi-fe2018110247067
Pugh, S. (1991). *Total design: Integrated methods for successful product engineering*. Wokingham, England; Reading, Mass: Addison-Wesley.
SeCo. (2019). *NameSampo: A linked open data infrastructure and workbench for toponomastic research - semantic computing research group (SeCo)* Retrieved from https://seco.cs.aalto.fi/projects/nimisampo/en/
Seymour, T. (2014, May 2014). Random access memory. *The British Journal of Photography*, 161, 80–81. Retrieved from https://search-proquest-com.libproxy.helsinki.fi/docview/1695234633?accountid=11365

Sutherland, J. (2014). *Scrum: The art of doing twice the work in half the time*. London: RH Business Books.
Tolonen, M., Lahti, L., Roivainen, H., & Marjanen, J. (2018). A quantitative approach to book-printing in sweden and finland 1640–1828. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 0(0), 1–22. doi:10.1080/01615440.2018.1526657.
Whitelaw, M. (2015). Generous interfaces for digital cultural collections.9(1) Retrieved from http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html#p30
Wodtke, C. (2013). *Users don't hate change. they hate you.: The 9x effect applies to redesigns too* Retrieved from https://medium.com/@cwodtke/users-dont-hate-change-they-hate-you-461772fbcac7#.2o6pmlbtp

Suomen viittomakielten korpusta rakentamassa

Juhana Salonen, Anna Puupponen, Ritva Takkinen & Tommi Jantunen
Jyväskylän yliopisto, kieli- ja viestintätieteiden laitos, viittomakielen keskus

Tiivistelmä
Viittomakielikorpusten rakentaminen on lisääntynyt merkittävästi 2000-luvulla: ensimmäiset korpusprojektit käynnistyivät 2000-luvun alussa Australiassa ja Hollannissa, minkä myötä laajoja, koneluettavia aineistokokoelmia on ryhdytty rakentamaan useissa Euroopan maissa 2010-luvulla. Tässä artikkelissa tarkastellaan Suomen viittomakielten, suomalaisen ja suomenruotsalaisen viittomakielen, korpuksen syntyä. Artikkeli esittelee korpuksen rakennusvaiheita eli aineiston keräämistä, käsittelyä, annotointia, pitkäaikaissäilytystä sekä julkaisua tietosuojakysymyksineen. Lisäksi artikkelissa kuvaillaan, miten korpusaineistoa on käytetty ja voidaan hyödyntää viittomakielten tutkimuksessa sekä opetuksessa.

Neljän vuoden mittainen Suomen viittomakielten korpusprojekti käynnistyi Jyväskylän yliopiston viittomakielen keskuksessa vuonna 2014. Projektin aikana kuvattiin keskusteluja ja elisitoituja kertomuksia 91 suomalaista viittomakieltä ja 12 suomenruotsalaista viittomakieltä äidinkielenään käyttävältä, eri puolilla Suomea asuvalta henkilöltä viittomakielisen kuuron projektitutkijan opastuksella. Videomateriaalia kerättiin yhteensä noin 560 tunnin edestä (seitsemästä kamerakulmasta nauhoitetut materiaalit yhteenlaskettuna).

Aineistonkeruun ja editoinnin jälkeen yhteensä 22 suomalaista viittomakieltä äidinkielenään käyttävän kielenoppaan videoaineistoihin on tehty perustason annotaatiot viittoma- ja virketasolla. Annotointivaihe eteni viittomien tunnistamisella, niiden merkitysten erottamisella ja viitotun tekstin ilmauskokonaisuuksien kääntämisellä suomen kielelle. Perusannotointi toteutettiin ELAN-ohjelmalla, jossa viittomia identifioidaan ajallisesti videoon yhteydessä olevien glossien avulla. Annotoinnissa käytettiin lisäksi Suomen Signbank -leksikkotietokantaa, johon ELAN-ohjelman glossit yhdistyvät verkkoyhteyden avulla. Laaja multimodaalinen aineistokokonaisuus täydennettiin metatiedoilla aineiston eri osa-alueista, kuten aineistokokonaisuuden yleisluonteesta, aineistonkeruussa läsnä olleista henkilöistä, videoiden sisällöistä ja video- ja annotaatiotiedostojen muodoista IMDI (ISLE Meta Data Initiative) -standardin mukaisesti. Annotoitu aineisto säilytetään ensisijaisesti Jyväskylän yliopistossa, minkä lisäksi se siirretään maaliskuun 2019 aikana FIN-CLARIN-konsortion

Kielipankkiin pitkäaikaissäilytettäväksi sekä julkaistavaksi kielenoppaiden tutkimussuostumusten ja tietosuoja-asetusten mukaisesti. Kielipankissa julkaistava korpusaineisto sisältää noin 14 tunnin edestä kuudesta kamerakulmasta kuvattua videomateriaalia 21 kielenoppaalta sekä videoihin linkitetyt annotaatiotiedostot ja IMDI-kuvaukset.

Suomen viittomakielten korpuksen luonti kehittää molempien viittomakielten kielellisten ja kulttuuristen piirteiden tutkimusta sekä opetusta. Jyväskylän yliopiston viittomakielen keskuksessa korpusaineiston pohjalta on tehty tähän mennessä useita suomalaiseen viittomakieleen keskittyviä tutkimuksia, minkä lisäksi aineistoa on käytetty myös viittomakieliä vertailevassa tutkimuksessa. Kerätty videoaineisto on ainutlaatuinen kokoelma Suomen viittomakielillä tuotettua kerrontaa ja keskusteluja: materiaali sisältää eri-ikäisten ja eri alueilta tulevien henkilöiden viittomista erilaisissa viestintätilanteissa. Systemaattisen annotoinnin myötä aineisto tulee olemaan merkittävä resurssi tutkimuksen lisäksi viittomakielten opetuksessa, viittomakieliä koskevassa koulutuksessa sekä kielisuunnittelussa.

1 Johdanto

Tässä artikkelissa kuvataan, miten CFINSL-projektissa (Corpus of Finland´s Sign Languages) on ryhdytty rakentamaan Suomen viittomakielten korpusta ja miten ensimmäinen osa korpuksesta on saatettu käytettäväksi. Suomen viittomakielten korpus on multimodaalinen aineistokokonaisuus, joka sisältää videomateriaalia sekä materiaaleihin ajallisesti sidottuja koneluettavia annotaatioita ja metatietoja. Korpusprojektin kuvaukset valmistuivat syksyllä 2017, jolloin oli kuvattu 91 suomalaista viittomakieltä ja 12 suomenruotsalaista viittomakieltä äidinkielenään käyttävän, eri puolilla Suomea asuvan henkilön viittomista. Korpuksen luonti vauhdittaa ja uudentaa molempien viittomakielten kielellisten (sanasto, rakenne, variaatio) ja kulttuuristen (kuurojenyhteisön tavat ja normit) piirteiden tutkimusta sekä kehittää molempien kielten sanakirjatyötä ja opetusta. Suomenruotsalaisen viittomakielen dokumentointi on lisäksi tärkeää kielen elvyttämisen kannalta (ks. De Meulder 2016). Korpusaineistoa taltioidaan FIN-CLARIN-konsortion

Kielipankkiin[1], josta aineistoa voidaan käyttää tutkimus- ja opetustarkoituksiin kielenoppaiden antamien lupien rajoissa.

2 Aineistonkeruu CFINSL-projektissa

Vuonna 2014 käynnistyneen CFINSL-projektin tavoitteena oli kerätä aineistoa yhteensä sadalta suomalaista ja suomenruotsalaista viittomakieltä käyttävältä kielenoppaalta ympäri Suomea. Korpusaineiston keruu tuli ajankohtaiseksi muissa maissa käynnistyneiden projektien johdosta. Ensimmäinen projekti käsitteli australialaista viittomakieltä, jossa materiaalin keruu alkoi vuonna 2004 sadan kielenoppaan voimin (Johnston 2010). Sittemmin ovat seuranneet projektit hollanti- laisesta viittomakielestä (92 kielenopasta vuosina 2006–2008)[2], brittiläisestä viittomakielestä (249 kielenopasta vuosina 2008–2011)[3] ja ruotsalaisesta viittomakielestä (42 kielenopasta vuosina 2009–2011)[4]. Näistä projekteista, erityisesti Ruotsista, saaduista vaikutteista muodostettiin CFINSL-projektin suuntaviivat.

CFINSL-projektin alkuvaiheessa laadittiin suunnitelma siitä, miten etsiä ja tavoittaa kielenoppaita, kuinka tiedottaa viittomakieliselle yhteisölle meneillään olevasta projektista ja miten rakentaa aineiston käsittelyyn vaadittava infrastruktuuri. Korpusprojektin aikana vuosina 2014–2017 aineistonkeruu toteutettiin pääosin Jyväskylän yliopiston AV-studiossa, minkä lisäksi pieni osa aineistosta kuvattiin Helsingissä ja Oulussa. Aineistoa kerättiin seitsemältä pääalueelta, jotka määritettiin Aluehallintoviraston toimialueiden perusteella: Etelä-Suomesta, Lounais-Suomesta, Länsi- ja Sisä-Suomesta, Itä-Suomesta, Pohjois-Suomesta ja Lapista. Lopullinen videoaineisto sisältää 91 suomalaista viittomakieltä (ikäjakauma 18–84 vuotta) ja 12 suomenruotsalaista viittomakieltä (ikäjakauma 27–89 vuotta) äidinkielenään käyttävän kielenoppaan viittomista. Jotta aineisto saatiin edustamaan mahdollisimman kattavasti Suomen viittomakielisen yhteisön keskuudessa käytettyjä kieliä, tehtiin Suomessa asuvista viittojista ennakkokartoitus, jossa selvitettiin muun muassa heidän kuulostatusta (esim. kuuro, huonokuuloinen, kuuleva), syn-

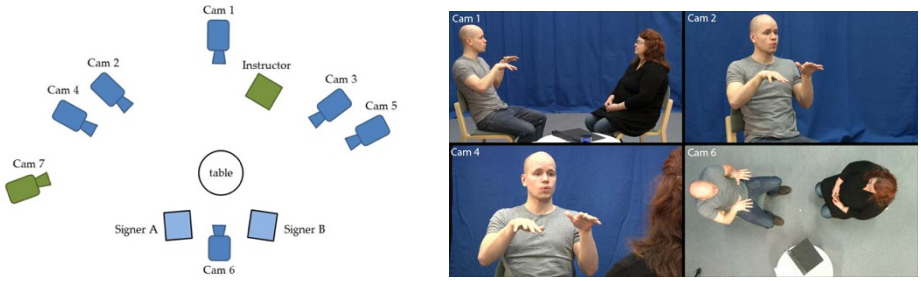
[1] https://www.kielipankki.fi/
[2] https://www.ru.nl/corpusngten/
[3] https://bslcorpusproject.org/project-information/
[4] https://www.ling.su.se/english/research/research-projects/sign-language/swedish-sign-language-corpus-project-1.59270

tymäpaikkaa ja koulutusta. Korpuksen kielenoppaat päätettiin tämän kartoitustyön avulla.

Aineistonkeruuta ja yhteydenpitoa kielenoppaisiin koordinoi viittomakielinen kuuro projektityöntekijä, joka on vastannut tiedottamisen ohella myös kielenoppaiden opastamisesta kuvaussession aikana. Kielenoppaat ovat saaneet valita itselleen oman parin, joka on läheinen tai tuttu esimerkiksi kouluajoilta. Kuvauksiin valituille pareille annettiin etukäteen tietoa yleisistä käytännön asioista, kuten korpusprojektin tavoitteista ja kuvauksille suotuisasta vaatetuksesta, paneutumatta kuitenkaan tarkemmin yksityiskohtiin, jotta voitaisiin välttää etukäteisvalmistelujen vaikuttamista viittojien kielenkäyttöön.

Korpusaineiston kuvauksissa käytettiin seitsemää Full HD-laatuista videokameraa, joista ensimmäinen taltioi yleisnäkymän molemmista viittojista ja toinen sekä neljäs kunkin viittojan yksittäisnäkymän (ks. kuvio 1). Kolmas ja viides ka- mera tallensivat kustakin viittojasta rajatumman lähikuvan ylävartalosta kasvoihin ja kuudes kamera molemmat viittojat lintuperspektiivistä niin, että viittojien pään, vartalon ja käsien syvyyssuuntaisten liikkeiden etäisyyttä on helpompi tarkastella. Seitsemäs kamera tallensi kuvauksissa toimineen viittomakielisen opastajan toimintaa. Videot tallennettiin Material eXhange Format (MXF) -formaattiin ja pakattiin H.264-koodekilla MP4-tiedostoiksi (Puupponen ym. 2014).

Korpusaineiston kuvauksessa kielenoppaat toteuttivat seitsemän eri viestinnällistä tehtävää, joiden yhteenlaskettu kokonaiskesto vaihteli puolestatoista tunnista kahteen tuntiin. Annetut tehtävät muodostuivat keskustelusta ja kerronnasta ja olivat: (1) itsensä esittely, (2) harrastuksesta/työstä kertominen, (3) sarjakuvista viittominen (Ferd´nand-sarjakuvat), (4) videotarinasta viittominen (Mr. Bean sekä Ohukainen ja Paksukainen -videot), (5) kuvakirjasta viittominen (Lumiukko ja Sammakko, missä olet? -kuvakirjat), (6) kuurojen kulttuuriin liittyvästä tapahtumasta keskustelu ja (7) vapaa keskustelu (Salonen ym. 2016). Pareille taattiin keskusteluraha siten, että opastaja vetäytyi tehtävänannon jälkeen taustalle ja AV-teknikot pysyivät erillisessä tarkkaamossa. Tehtävien aikana oli aina mahdollista pyytää tarkennusta viittomakieliseltä opastajalta ja tehtävien välissä pidettiin myös tauko positiivisen ja luonnollisen ilmapiirin saavuttamiseksi.



Kuvio 1. Kuvaustilanne kamera-asetelmineen (a); videoaineistoa eri kuvakulmista (b). (Puupponen ym. 2014; Salonen ym. 2016.)

Kuvausten jälkeen kielenoppailta kerättiin aineiston käyttöön liittyvät suostumukset sekä taustatietoja. Kullekin parille selostettiin tutkimuslupakäytänteet suoma- laisella viittomakielellä ja kielenoppaita pyydettiin täyttämään tutkimuslupa- ja taustatietolomakkeet suomeksi. Tutkimusluvassa tuli vastata joko myöntävästi tai kielteisesti viiteen videoaineiston käyttöön liittyvään kysymykseen, jotka koskivat lupaa (1) käyttää aineistoa tutkimukseen, (2) näyttää aineistosta otteita julkisissa tilaisuuksissa, (3) irrottaa aineistosta kuvia julkaisuja varten, (4) julkaista aineisto kokonaisuudessaan verkossa ja (5) mainita kielenoppaan nimi julkaisuissa. Tausta- tietolomakkeilla kerättiin viittojista tietoa, jonka avulla voidaan verrata muun muassa eri-ikäisten ja eri paikkakunnilla asuvien viittomakielisten käyttämiä kieli- muotoja. Tutkimuslupaa on täydennetty vuonna 2018 lisäkysymyksillä. Näihin sekä metatietojen jatkokäsittelyyn palataan tämän artikkelin luvussa 4.

Projektin aikana videomateriaalia kerättiin yhteensä noin 80 tunnin edestä, mikä tarkoittaa kaiken kaikkiaan noin 560 tuntia eri kamerakulmista kuvattua videoaineistoa. Kunkin parin noin puolentoista tunnin pituinen raakavideo editoitiin eri videotiedostoihin tehtävä- ja kameranumeron mukaisessa järjestyksessä.

3 Aineiston annotointi

Aineistonkeruun ja editoinnin jälkeen videoaineiston käsittely siirtyi annotointivaiheeseen. Jotta puhe- tai multimodaaliset korpukset olisivat koneluettavia, tulee vi-

deoaineistoon tehdä siihen ajallisesti sidottuja merkintöjä eli annotaatioita. Nämä merkinnät mahdollistavat aineistossa navigoinnin, tarkasteltavien kohtien rajaamisen sekä erilaiset haut. Annotaatioiden avulla aineistoon on helpompi myös palata myöhemmin uudestaan. Suurten aineistokokonaisuuksien annotoinnin on tärkeää olla systemaattista, jotta aineisto soveltuisi useiden tutkijoiden käyttöön ja sopisi erilaisiin tutkimustavoitteisiin.

CFINSL-projektissa kuvattu videomateriaali annotoitiin Max Planck -instituutissa Nijmegenissä kehitetyllä ELAN-ohjelmalla (Eudico Linguistic Annotator; Crasborn & Sloetjes 2008)[5]. Videoaineiston perusannotointia eli viittomien ja lauseiden mahdollisimman neutraalia annotointia ryhdyttiin tekemään yhteensä 22 suomalaista viittomakieltä äidinkielenään käyttävän kielenoppaan materiaaleista. Tämä aineisto on kestoltaan yhteensä noin 16 tuntia. Annotointiprosessi alkoi viitomien tunnistamisella, niiden merkitysten erottamisella ja viitotun tekstin ilmaus- kokonaisuuksien kääntämisellä suomen kielelle.

3.1 Glossaus eli viittomatason annotointi

Viittomatason annotointi voidaan toteuttaa eri tavoin. Yleisesti ottaen viittomakielten viittomien merkitsemisessä käytetään glosseja, jotka ovat yleensä suuraakkosin kirjoitettuja puhutun kielen sanoja. Glossiksi valikoituu yleensä sana, jonka merkitystä viittoma vastaa mahdollisimman hyvin (Johnston 2016). Suomalaisen viittomakielen kohdalla käytetään usein suomenkielisiä glosseja, jotka kirjoitetaan perusmuotoisena (esim. Savolainen 2000). Myös CFINSL-projektissa suomalaista viittomakieltä koskeva viittomien perusannotointi on tehty käyttäen suomenkielisiä glosseja (Salonen ym. 2019).

Toimivan korpuksen rakentamisen edellytyksinä ovat yhtenäisyys ja johdonmukaisuus. Tämä tarkoittaa, että yhteisistä periaatteista ja annotointikonventioista tulee sopia kaikkien annotoijien kesken. Annotointikonventioita on kehitelty useissa eri viittomakielten korpusprojekteissa (esim. Johnston 2016 Australia; Crasborn ym. 2015 Hollanti; Wallin & Mesch 2018 Ruotsi). CFINSL-projektissa konventioita kehiteltiin perustason annotointia tehtäessä (ks. Keränen ym. 2016). Ensimmäinen versio annotaatiokonventioista julkaistiin keväällä 2018 ja se sisälsi viittomatason annotointiin liittyvät periaatteet CFINSL-projektissa. Konventioissa

[5] https://tla.mpi.nl/tools/tla-tools/elan/

kuvaillaan viittomiston erilaisten osien, kuten esimerkiksi eriasteisesti leksikaalisten viittomien (ks. Jantunen 2018) muistiinmerkintään liittyviä periaatteita (Salonen ym. 2018). Konventioiden toisessa, helmikuussa 2019 julkaistussa versiossa on kuvattu myös virketason käännöksiin liittyvät periaatteet (Salonen ym. 2019).

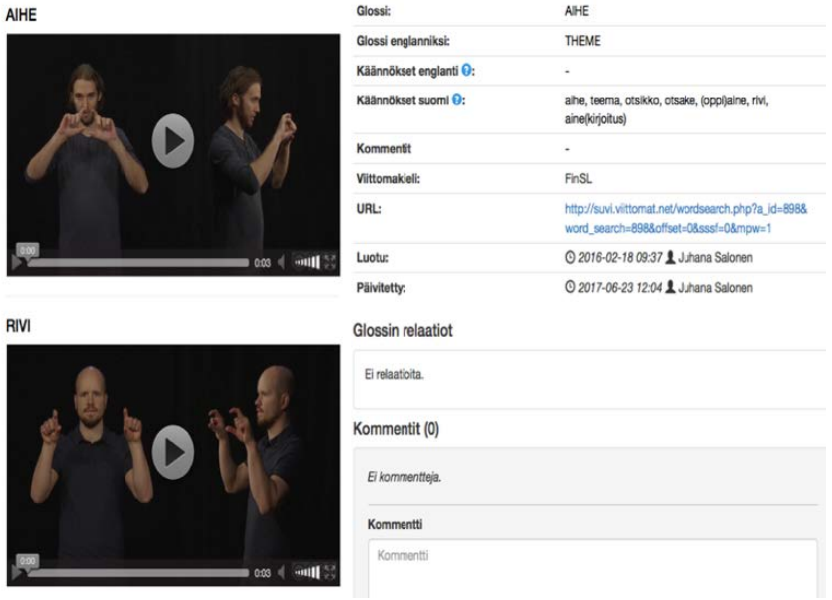
Vuosina 2015 ja 2016 CFINSL-projektissa annotoitiin merkityslähtöisesti siten, että samanmuotoisille viittomaesiintymille nimettiin eri glossi (so. merkitysglossi) lausetason kontekstuaalisen merkityksen perusteella, mikä oli aikaa vievää. Annotoijien kesken oli työlästä löytää yhteistä linjaa paisuvalle merkitysglossijoukolle, saati hallita rakentumassa olevaa korpusta. Esimerkiksi yksi opiskeluun viittaava muoto saatettiin tilanteen mukaan glossata opiskelua, oppimista ja kurssia tarkoittavilla merkitysglosseilla.

Vuonna 2017 päätettiin siirtyä merkitysglosseista ID-glosseihin, jolloin samanmuotoisia (homonyymisiä, polyseemisiä ja foneettisesti varioivia) viittomia alettiin koota yhteen samoin tunnisteglossein. ID-glossilla tarkoitetaan sellaista nimikettä, joka on valittu edustamaan merkitykseltään varioivaa, mutta muodoltaan identtistä viittomaa laajassa korpuksessa (Johnston 2008, 2010). Esimerkiksi suomalaisessa viittomakielessä esiintyy manuaaliselta (käsillä tuotetulta) artikulaatioltaan samanmuotoinen viittoma, joka voi tarkoittaa lauseyhteydestä riippuen arkea, farkkuja, maaseutua, raitista tai oranssia. Aikaisemmin merkitysglossiksi valikoitiin aina jokaisen esiintymän kohdalla kontekstin mukainen merkitys, mutta ID-glossiksi kaikille esiintymille valikoitui ARKI[6]. ID-glossi ei näin ollen ilmaise viittoman merkityskäännöstä, vaan annotoijien kesken sovittua tunnistetta. Useimmiten ID-glossiksi valikoidaan se merkitys, jonka frekvenssi aineistossa on suurin. ID-glossaus mahdollistaa merkitysglossausta tehokkaammat haut aineistosta.

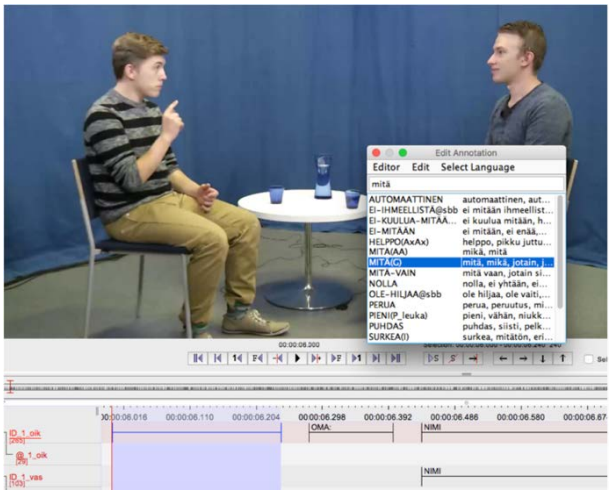
Toteutimme ID-glossauksen kahdella toisiinsa yhteydessä olevalla alustalla: ELAN-ohjelmassa ajallisesti videoon yhteydessä olevat glossit yhdistyvät myös Suomen Signbank -tietokantaan verkkoyhteyden avulla. Suomen Signbank on suomalaiselle ja suomenruotsalaiselle viittomakielelle rakennettu leksikkotietokanta, jonka perustehtävänä on toimia työkaluna viittomakielisten tekstien annotoinnissa[7]. Tietokannassa olevat glossitietueet sisältävät itse glossin lisäksi suomenkieliset käännösvastineet (vrt. merkitysglossit), videon viittomasta, jota glossilla merkitään, sekä tarvittaessa muuta tietoa viittomasta (ks. kuvio 2).

[6] ID-glossin ARKI tietue Signbank-tietokannassa: https://signbank.csc.fi/dictionary/gloss/3564/
[7] https://signbank.csc.fi

ELAN-ohjelma sisältää ominaisuuden, jonka myötä ohjelma voi ottaa annotointeja tehtäessä yhteyden ulkoisella verkkopalvelimella ylläpidettyyn kontrolloituun sanastoon (ECV, external controlled vocabulary). CFINSL-projektissa kontrolloitu sanasto eli lista aikaisemmin luoduista glosseista sijoitettiin Suomen Signbankiin kehittämällä Signbank-alustaa tähän soveltuvaksi. Glossilistan ansi- osta annotoija voi joka annotaatiosolun nimeämisen yhteydessä tarkistaa, löytyykö glossi jo tietokannasta. Jos glossi on jo olemassa, annotoija voi valita sen listalta ja välttää näin ei-automatisoidussa transkriptiossa herkästi syntyviä kirjoitusvirheitä. Jos glossia ei ole vielä luotu kyseiselle viittomalle, tietokantaa voi täydentää luo- malla sinne uuden glossitietueen videoineen, käännöksineen ja muine lisätietoineen. Kuviossa 3 havainnollistetaan, kuinka ELAN-ohjelmassa annotaatiosolun sisältöä luotaessa voidaan hakea Suomen Signbank -tietokannasta sopivaa glossia joko ID- glossin (laatikon vasemmanpuoleinen sarake) tai sen käännösvastineiden (laatikon oikeanpuoleinen sarake) avulla (Salonen ym. 2018). Signbankissa käsin tehdyt glossi- ja käännösvastinemuutokset päivittyvät automaattisesti kaikkiin linkitettyi- hin annotaatiosoluihin jatkuvan ECV-yhteyden myötä.



Kuvio 2. Esimerkkinäkymä Signbankin glossitietueesta.



Kuvio 3. Näkymä annotoinnista ELAN-ohjelmassa käytettäessä Signbankiin sijoitettua kontrolloitua sanastoa.

Suomen Signbankin kehittämistyöstä on vastannut CFINSL-projekti yhdessä Kuurojen Liiton tutkimus- ja sanakirjatyön kanssa. Tietokannan taustalla on alun perin australialaisessa viittomakielityössä kehitelty Auslan Signbank sekä tämän myöhempi sovellus, hollantilainen Signbank-tietokanta. Lähdekoodit kaikista Signbank-tietokannoista ovat saatavissa Github-versionhallintasivustolta. [8] Signbankin kehittämistä on ryhdytty 2010-luvulla tekemään kansainvälisessä yhteistyössä Australiassa, Hollannissa, Suomessa ja Iso-Britanniassa sijaitsevien tutkimusryhmien välillä (Cassidy ym. 2018). Suomen Signbankin kehittelyssä teknisen dokumentoinnin lisäksi tietokannan rakennetta ja ominaisuuksia on dokumentoitu käyttäjälähtöisesti Githubin FinSL-signbank wikiin. Suomen Signbank -tietokannasta löytyvä CFINSL-projektin leksikko perustuu työstettävänä olleeseen noin 16 tunnin pituiseen materiaaliin, ja se julkaistiin huhtikuussa 2018. Tämän lisäksi Suomen Signbank sisältää kesäkuussa 2017 julkaistun Kuurojen Liiton Kipo-korpuksen leksikon, joka pohjautuu noin 2,5

[8] https://github.com/Signbank

tunnin mittaiseen Suomen viittomakielten kielipoliittisen ohjelman annotoituun materiaaliin (Kuurojen Liitto ry 2015).

3.2 Virketason käännöksiä koskeva annotointi

Viittomatason annotoinnin lisäksi CFINSL-projektissa tehtiin annotointia myös virketasolla. Tämä tarkoittaa käytännössä viittomisen kääntämistä suomen kielelle. Käännöksen alkuvaiheessa viitotusta tekstivirrasta eroteltiin kääntäjien intuition perusteella mielekkäitä virkkeitä ilman tarkempaa lauseiden erottelua, mikä on varsinaisen tutkimuksen tehtävä perusannotoinnin jälkeen. Käännös on toteutettu sellaisessa muodossa, joka huomioi lähtökielen tapaa ilmaista asiat niin manuaalisesti (käsillä) kuin ei-manuaalisesti (päällä, keholla ja kasvoilla). Lisäksi käännöksiin on lisätty sulkeiden sisään osia, joita sujuva suomenkielinen teksti edellyttää, mutta jotka tulevat viittomakielisessä tekstissä ilmi edeltävästä diskurssikontekstista tai joita ei välttämättä edellytetä lainkaan (mm. lauseen tekijä, kopula, eräät konjunktiot; ks. esimerkki 1). Käännöksiä koskevat periaatteet kuvataan tarkemmin CFINSL-projektin annotointikonventioissa (ks. Salonen ym. 2019).

(1) KATSOA ULKONA SATAA-LUNTA LUMI SATAA-LUNTA
(Hän) huomaa, (että) ulkona sataa lunta.

Käännös tarjoaa kokonaisvaltaisemman kuvan viitotuista teksteistä, sillä ID-glossauksessa keskitytään pelkästään manuaaliseen artikulaatioon. Käännöksestä voidaan myös tarkistaa, mihin merkitykseen kullakin ID-glossilla on viitattu. Tähän mennessä virketason käännökset on tehty 22 suomalaista viittomakieltä käyttävän kielenoppaan materiaaleista. Kääntäjät ovat myös tehneet Suomen Signbank-tietokannan glossitietueisiin viittomien suomenkielisiä käännösvastineita sen mukaan, millaisia merkityksiä ID-glosseilla merkityillä viittomilla on ilmennyt kääntämisen yhteydessä.

4 Aineiston pitkäaikaissäilytys, julkaisu ja tietosuojakysymykset

CFINSL-projektissa käynnistyneen korpustyön tavoitteena on sekä pitkäaikaissäilyttää aineistoa että julkaista siitä erilaisin käyttöoikeuksin rajattuja osia kielenoppaiden tutkimussuostumusten ja tietosuojalainsäädännön sallimissa puitteissa. Korpusaineistoa rakennettaessa aineistoa säilytetään ensisijaisesti

Jyväskylän yliopistossa, mutta pitkäaikaissäilytystä ja julkaisua varten aineistoa siirretään myös FIN-CLARIN-konsortion Kielipankkiin. Ensimmäinen osakokonaisuus aineistosta siirrettiin Kielipankkiin maaliskuussa 2019. Aineiston nimi on Suomalaisen viittomakielen korpus (Corpus FinSL)[9], ja se sisältää yhteensä noin 14 tuntia videomateriaalia: suomalaisella viittomakielellä viitottuja tarinoita ja keskusteluja yhteensä 21 kielenoppaalta. Corpus FinSL -aineisto on jaettu Kielipankissa käyttöoikeuksien ja videoaineiston sisällön perusteella kahteen osa-aineistoon: Elisitoituihin kertomuksiin (Elicited narratives)[10] ja Keskusteluihin (Conversations)[11]. Elisitoidut kertomukset sisältävät viestinnällisten tehtävien 3–5 materiaaleja (ks. luku 2). Aineiston yhteiskesto on noin 5 tuntia, ja se on julkisesti saatavilla tutkijoiden, kouluttajien sekä laajemman yleisön käyttöön Creative Commons BY NC SA 4.0 -lisenssillä. Keskustelut sisältävät viestinnällisten tehtävien 1, 2, 6 ja 7 aineistoa, joiden yhteis- kesto on noin 9 tuntia. Keskusteluaineiston käyttö edellyttää tutkimussuunnitelmaa sekä henkilökohtaista käyttöoikeutta Kielipankin RES-lisenssin mukaisesti.

Pitkäaikaissäilytykseen ja monipuoliseen tutkimuskäyttöön tarkoitetun laajan, multimodaalisen aineistokokonaisuuden tulee sisältää itse videoaineistot, videoihin synkronoidut annotaatiot sekä riittävän kattavat metatiedot aineiston eri osa-alueista. Metatiedot kuvaavat tässä tapauksessa aineistokokonaisuuden yleisluonnetta, aineistonkeruussa läsnä olleita henkilöitä, videoiden sisältöjä ja video- ja annotaatiotiedostojen muotoja. Kielipankkiin siirrettävän Corpus FinSL -aineiston ano- nymisoidut metatiedot on kuvattu IMDI (ISLE Meta Data Initiative) -standardin mukaisesti. IMDI on Max Planck -instituutissa Nijmegenissä kehitelty kuvausstandardi monimediaisten ja multimodaalisten kieliaineistojen yhdenmukaiseen kuvaukseen[12]. CFINSL-projektissa tuotettiin IMDI-standardien mukaisesti yleiskuvaus aineistosta (Corpus FinSL), sen taustalla olevasta projektista (CFINSL Project) sekä osa-aineistojen sisällöistä (Elicited narratives; Conversations). Osa-aineistojen osalta annettiin myös yleiskuvaukset kustakin viestinnällisestä tehtävästä (1–7, ks. luku 2). Tämän lisäksi jokaisen parin yksittäisistä viestintätehtävistä tehtiin tilannekohtainen kuvaus (Session), joka sisältää tietoja aineistonkeruutilanteen osallistujista (Actors); viestintätilanteen

[9] Suomalaisen viittomakielen korpus: http://urn.fi/urn:nbn:fi:lb-2019012321
[10] Elisitoidut kertomukset: http://urn.fi/urn:nbn:fi:lb-2019012322
[11] Keskustelut: http://urn.fi/urn:nbn:fi:lb-2019012323
[12] https://tla.mpi.nl/imdi-metadata/

laadusta, vuorovaikutuksellisuudesta ja keruu- menetelmästä (Content); videomateriaaleista (MediaFiles) ja annotaatioista (WrittenResources).

Kielenoppaisiin liittyviä taustatietoja kerättiin CFINSL-projektin aineistonkeruun aikana vuosina 2014–2017 hyvin kattavasti. Näistä Kielipankkiin rakennettuun IMDI-kuvaukseen valikoituivat lopulta vain henkilön yksilöivä anonymisoitu koodi, ikä ikäryhmittäin, sukupuoli, asuinalue sekä kätisyys (oikea/vasen). Kaiken kaikkiaan Kielipankkiin siirretty Corpus FinSL -aineisto koostuu 71 viestintätilanteesta ja sisältää yhteensä 343 videotiedostoa (kamerakulmat 1–6), 142 annotaatiotiedostoa (ELANin .eaf- ja .pfsx-tiedostot) sekä IMDI-kuvaukset (taulukko 1).

Taulukko 1. Suomalaisen viittomakielen korpus (Corpus FinSL) Kielipankissa.

Koko aineisto	14 tuntia ja 22 minuuttia
Elisitoidut kertomukset (CC-lisenssi)	5 tuntia ja 4 minuuttia
Keskustelut (RES-lisenssi)	9 tuntia ja 18 minuuttia
Videotiedostot	343 mp4-tiedostoa
Annotaatiotiedostot	142 tiedostoa (eaf + pfsx)
Kielenoppaiden määrä	21 kielenopasta

Vuoden 2018 toukokuussa voimaan tulleen EU:n tietosuoja-asetuksen myötä Corpus FinSL -aineiston lupia oli tarpeen täydentää. Kun alkuperäiset luvat käsittelivät aineiston verkkojulkaisua yhtenä yleisenä kokonaisuutena (ks. luku 2), niin lisäluvat pyytävät kielenoppaiden suostumusta aineiston pitkäaikaissäilytykseen Kielipankissa sekä kunkin viestinnällisen tehtävän vapaaseen julkaisuun samalla alustalla. Tietosuoja-asetuksen mukaisesti kielenoppailta on pyydetty nimenomaista suostumusta myös siihen, että aineisto sisältää heihin liitettäviä erityisiin henkilötietoryhmiin kuuluvia tietoja eli käytännössä kielenoppaiden itsensä kertomaa tietoa oman kuulonsa asteesta. Tarve tämän luvan pyytämiseen kumpuaa erityisesti vuorovaikutuksellisesta tehtävästä 1, jossa kielenoppaat viittomakieliselle kulttuurille luonnollisella tavalla usein identifioivat itsensä kuuroiksi itseään esitellessään. Käytännössä lisälupalomakkeella on pyydetty kielenoppaiden suostumusta myös heidän koko aineistonsa lisensointiin kaupallisen käytön kieltävällä Creative Commons BY NC SA 4.0 -lisenssillä, joskin lopulta Creative Commons -lisenssillä on lisensoitu ainoastaan vapaasti julkaistava Elisitoidut kertomukset -osa-aineisto.

Corpus FinSL -aineiston jako kahteen osa-aineistoon on niin ikään seurausta EU:n tietosuoja-asetuksen mahdollisimman tarkasta noudattamisesta. Kun

Elisitoidut kertomukset -osa-aineisto sisältää ainoastaan temaattisesti rajattuja monologeja, joissa kielenoppaat toisintavat erilaisten kuvamateriaalien tarinoita, niin Keskustelut-osa-aineisto sisältää temaattisesti rajaamattomampia dialogeja, joissa kielenoppaat saattavat ilmaista välillisesti henkilötietoja myös kolmansista osapuolista. Näiden henkilötietojen vapaaseen julkaisemiseen ei ole ollut mahdollista pyytää lupaa suostumuslomakkeella, joten Keskustelut-osa-aineiston saatavuutta on haluttu rajoittaa.

5 Loppusanat

Tässä artikkelissa on esitelty Suomen viittomakielten korpusprojektissa tehtyä viittomakielisen aineiston keruuta, annotointia, pitkäaikaissäilytystä sekä julkaisua. Laaja, elektronisessa muodossa oleva ja tietokoneluettava aineisto tarjoaa uusia mahdollisuuksia Suomessa käytettyjen viittomakielten – suomalaisen ja suomenruotsalaisen viittomakielen – määrälliseen ja laadulliseen tutkimukseen. Kattavasta, usean henkilön viittomista sisältävästä aineistosta voi tutkia esimerkiksi eri-ikäisten tai eri puolilla Suomea asuvien viittomakielisten käyttämää viittomistoa sekä eri tekstilajien välillä olevia eroja viittomistossa ja kielen rakenteessa. Laajat, osin julkisesti saatavilla olevat aineistot mahdollistavat myös aivan uudella tavalla viittomakieliä vertailevan tutkimuksen. Tätä tukee erityisesti eri viittomakielten korpusprojekteissa käytetyt samankaltaiset aineistonkeruumenetelmät.

Viittomakielikorpuksen tuomat mahdollisuudet näkyvät selvästi 2010-luvulla Jyväskylän yliopistossa tehdyssä viittomakielen tutkimuksessa. Korpusaineiston pohjalta on tähän mennessä tehty tutkimusta suomalaisen viittomakielen rakenteesta muuan muassa lauseenjäsenten järjestyksestä intransitiivi- ja transitiivilauseissa (Jantunen 2017), kuvailevista viittomista (Takkinen ym. 2018) sekä ei-manuaalisuudesta pään ja kehon liikkeiden suhteen (Puupponen 2018). Lisäksi aineistoa on käytetty viittomakieliä vertailevissa tutkimuksissa (Jantunen ym. 2016; Puupponen ym. 2016) sekä suomalaisen viittomakielen oppiaineesta valmistuneissa maisterintutkielmissa (esim. Syrjälä 2018; Puhto 2018). Suomenruotsalaisen viittomakielen aineistoa on käytetty Helsingin yliopiston, Kuurojen Liiton ja Humanistisen ammattikorkeakoulun tulkkikoulutuksen yhteisissä projekteissa koskien suomenruotsalaisten kielitietoisuuden kehittämistä ja tulkkikoulutusta. Lisäksi suomenruotsalaisesta viittomakielestä on parhaillaan tekeillä kerättyä aineistoa hyödyntävä väitöskirja Helsingin yliopistossa.

Suomen viittomakielten korpuksella tulee olemaan merkittävä vaikutus Suomen viittomakieliseen yhteisöön sekä viittomakielten yhteiskunnalliseen asemaan. Viittomakielisille se tarjoaa mahdollisuuden kehittää kielitietoisuutta omasta äidin- kielestään, jota ei monille ole opetettu perusopetuksessa. Viittomakieltä vieraana kielenä käyttäville henkilöille – kuten esimerkiksi viittomakielentulkeille – korpus tarjoaa opetusmateriaalia muun muassa kielenkäyttäjien välisten sosiolingvististen erojen tunnistamiseen. Korpusaineistoa tullaan myös hyödyntämään Jyväskylän yliopistossa käynnissä olevassa, opetus- ja kulttuuriministeriön rahoittamassa koulutushankkeessa, jonka tavoitteena on kehittää täydennyskoulutustarjontaa viittomakielen opettajille Suomessa. Opetus- ja koulutussovellusten lisäksi korpus voi tulevaisuudessa toimia myös kielenhuollon ja kielisuunnittelun työvälineenä.

Lähteet

Cassidy, S., Crasborn, O., Nieminen, H., Stoop, W., Hulsbosch, M., Even, S., Komen, E. & John- ston, T. (2018). Signbank: Software to Support Web Based Dictionaries of Sign Language. Proceedings - The 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community, 2359–2364.

Crasborn, O., Bank, R., Zwitserlood, I., Kooij, E., Meijer, A. & Sáfár, A. (2015). Annotation Conventions for The Corpus NGT. Version 3. Radboud University Nijmegen: Centre for Language Studies & Department of Linguistics.

Crasborn, O. & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. Proceedings - The 3rd Workshop on the Representation and Processing of Sign Lan- guages: Construction and Exploitation of Sign Language Corpora, 39–43.

De Meulder, Maartje (2016) Promotion in Times of Endangerment: The Sign Language Act in Finland. Language Policy, 16(2), 189–208.

Jantunen, T. (2017). Fixed and NOT free: Revisiting the order of the main clausal constituents in Finnish Sign Language from a corpus perspective. SKY Journal of Linguistics 30, 137–149.

Jantunen, T. (2018). Viittomakielet hybridisysteemeinä: hämärärajaisuus ja epäkonventionaalisuus osana viittomakielten rakennetta. Puhe ja Kieli 38(3), 109–126.

Jantunen, T.; Mesch, J.; Puupponen, A. & Laaksonen, J. (2016). On the rhythm of head movements in Finnish and Swedish Sign Language sentences. Proceedings - Speech Prosody 2016, 850–853.
Johnston, T. (2008). Corpus linguistics and signed languages: No lemmata, no corpus. Proceedings - The 3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora, 82–87.
Johnston, T. (2010). From archive to corpus: Transcription and annotation in the creation of signed language corpora. International Journal of Corpus Linguistics 15 (1), 106–131.
Johnston, T. (2016). Auslan Corpus Annotation Guidelines. Centre for Languge Sciences, Department of Linguistics, Macquarie University (Sydney) and La Trobe University (Melbourne), Australia.
Keränen, J., Syrjälä, H., Salonen, J. & Takkinen, R. (2016). The Usability of the Annotation. Proceedings - The 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, 111–116.
Kuurojen Liitto ry (2015). Suomen viittomakielten kielipoliittinen ohjelma 2010 - korpus, annotoitu versio. Kielipankki. Saatavilla http://urn.fi/urn:nbn:fi:lb-2014073031
Puhto, J. (2018). Päänpudistuksen käyttötavat ja frekvenssit suomalaisessa viittomakielessä. Suomalaisen viittomakielen pro gradu -tutkielma. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto.
Puupponen, A.; Jantunen, T.; Takkinen, R.; Wainio, T. & Pippuri, O. (2014). Taking non-manuality into account in collecting and analyzing Finnish Sign Language video data. Proceedings - The 6th Workshop on the Representation and Processing of Sign Languages: Beyond the manual channel, 143–148.
Puupponen, A., Jantunen, T. & Mesch, J. (2016). The Alignment of Head Nods with Syntactic Units in Finnish Sign Language and Swedish Sign Language. Proceedings - Speech Prosody 2016, 168–72.
Puupponen, A. (2018). The relationship between the movements and positions of the head and the torso in Finnish Sign Language. Sign Language Studies 18(2), 175–214.
Salonen, J., Takkinen, R., Puupponen, A., Nieminen, H. & Pippuri, O. (2016). Creating Corpora of Finland´s Sign Languages. Proceedings - The 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining, 179–184.

Salonen, J., Wainio, T., Kronqvist, A. & Keränen, J. (2018). Suomen viittomakielten korpus -projektin (CFINSL) annotointiohjeet. 1. versio. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto. http://r.jyu.fi/ygQ
Salonen, J., Wainio, T., Kronqvist, A. & Keränen, J. (2019). Suomen viittomakielten korpus -projektin (CFINSL) annotointiohjeet. 2. versio. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto. http://r.jyu.fi/ygR
Savolainen, L. (2000). Viittomakielten erilaiset muistiinmerkitsemistavat. Teoksessa A. Malm (toim.) Viittomakieliset Suomessa (pp. 189–200). Helsinki: Finn Lectura.
Syrjälä, H. (2018). Hakukysymysviittoman paikka suomalaisessa viittomakielessä. Suomalaisen viittomakielen pro gradu -tutkielma. Kieli- ja viestintätieteiden laitos, Jyväskylän yliopisto.
Takkinen, R., Keränen, J. & Salonen, J. (2018). Depicting Signs and Different Text Genres: Preliminary Observations in the Corpus of Finnish Sign Language. Proceedings – The 8th Workshop on the Representation and Processing of Sign Languages: Involving the Lan- guage Community, 189–194.
Wallin, L. & Mesch, J. (2018). Annoteringskonventioner för teckenspråkstexter. Version 7. Avdelningen för teckenspråk, Institutionen för lingvistik, Stockholms universitet.

II Methods

Guessing a tweet author's political party using weighted n-gram models

Enum Cohrs[1] (University of Eastern Finland) & Wiebke Petersen (University of Düsseldorf)

Abstract
Political parties and their candidates are increasingly using online channels for their electoral campaigns. This was, for instance, observable for the elections for the German parliament (Bundestag) in 2017. But even outside the campaigning time, politicians use Twitter to inform about their work and current topics. For an informed human it is usually easy to guess their political affiliation even if it is not explicitly stated in the tweets. In this paper we present a probabilistic classifier for the political party of a tweet's author and compare different weight configurations, either weighting by word frequency or by part of speech. In opposition to many existing systems that focus on the US and only work with two-party political systems, our model allows an arbitrary amount of parties. For the German election we included 9 political parties into the analysis and our system achieved an accuracy of 72 % when perusing all tweets published by an author in the specified time interval, or 36 % accuracy when using only one single tweet as input. A random guessing baseline system would have an expected accuracy of 11 % in both cases.

1 Introduction

In the recent decade, politicians have entered the World Wide Web and started to use it for electoral campaigns, as well as to keep their supporters motivated. This was observable in the context of many elections, most prominently during the US presidental elections in 2016. An important part of these campaigns are Social Media networks. Among them is Twitter, a so-called microblogging service, where users write small status messages (called tweets) of up to 280 characters. The status messages are then shown to the account's followers. Words prefixed by a number sign (#) are called hashtags, and can be used for search queries.

Whilst many politician and campaign accounts carry party names in their name or biography, not all of them do. For informed humans, who know the political

[1] corresponding author

situation in the country of interest, it is usually still easy to guess the account's party alignment, especially in two-party systems. This is less trivial for computer programs. Still, the party membership information can be useful for applications such as sentiment analysis, social network analysis, to monitor the parties' range of influence or to identify political topics.

In this paper we present a probabilistic classifier for a Twitter account's party membership in a multi-party setting. We focus on a comparison of various classifier con- figurations that either favor frequent or infrequent words or specific parts of speech. Additionally, we investigate whether stemming the tweets enhances the classification results. The classifier has two modes of operation: it can either use a single tweet for classification; or it uses all tweets that an account issued in the relevant time interval. We focus on the political landscape of Germany, where five parties have been part of the federal parliament before the last elections in 2017. For our classifier, we have included four additional parties that are part of the European parliament. Two of the four parties have been elected into the federal government in the 2017 elections.

2 Related Work

Most of the research on political party or conviction focuses on a binary system. For example Pennacchiotti and Popescu [1] use machine learning methods to guess the political affiliation in the binary political system of the US (Democrats or Republicans). They also guess other information about the users, such as ethnicity and sentiment about a large coffeeshop chain. In another study Conover et al. [2] have developed three different SVM-based classifiers for binary political alignment (left or right), based on the tweet text, hashtags or community graphs; they achieve an accuracy of up to 95 %.

Cohen and Ruths [3] evaluate the performance drop of existing binary political classifiers when applied to Twitter users who are not politically active all the time. They observe that most classifiers are trained on and tested against datasets from users that are strongly politically engaged and conclude that the reported accuracies are not rep- resentative for real world applications operating on non-biased data. By choosing only Twitter accounts of well-known politicians for training and testing our classifier, it is very likely that we face this problem as well.

Other approaches make use of the meta-data provided by Twitter. For example, Boutet et al. [4] propose a classification method based on retweet structure, list

membership, self-descriptions and positive affect words. They focused on tweets related to the 2010 UK General Election and used three classes: Labour, Conservative and Liberal Democrats.

Four classes or polarities are used by Pla and Hurtado [5] who have extracted tweets from the TASS2013 corpus that express political sentiment from a Spanish general tweet corpus, and use this subset to guess their political tendency in the categories: Left, right, center and undefined.

One of the first publications about German politics on Twitter is Tumasjan et al. [6]. They have performed a sentiment analysis on tweets about parties and found out that the online sentiments closely relate to the actual election results.

We are not aware of any other publications that approached the problem of classify- ing tweet authors into individual parties in multi-party systems.

3 Data Collection

For our study we have chosen the 9 German political parties listed in Table 1; 5 of the parties have been part of the federal parliament before the 2017 elections, 7 have been elected in the 2017 elections and the other two (Piratenpartei and PARTEI) have representatives in the European parliament.

Table 1. Parties included in the dataset, by alphabet.

official name	abbreviation	alignment	# acc	test dataset
Alternative für Deutschland	AfD	far right	19	6
Bündnis 90/Die Grünen	Grüne	centre left	17	6
Christlich Demokratische Union Deutschlands	CDU	centre right	21	7
Christlich-Soziale Union in Bayern	CSU	right	12	4
Die Linke	Linke	far left	22	7
Freie Demokratische Partei	FDP	econ. liberals	13	4
Partei für Arbeit, Rechtsstaat, Tierschutz, Elitenförderung und basisdemokratische Initiative	PARTEI	satirical	7	2
Piratenpartei	Piraten	centre left	10	3
Sozialdemokratische Partei Deutschlands	SPD	centre left	32	11

For these parties we have chosen 153 Twitter accounts of the parties or fractions, parliamentarians, party functionaries, and other well-known campaigners. Table 1 states the number of accounts chosen per party. From the chosen accounts we have collected the 59 360 tweets that were published between June 1 and September 24, 2017, i.e. the four months preceding the last German federal elections. The tweets were downloaded using the official Twitter API and the Haskell package twitter-conduit. Fifty accounts were randomly chosen for the test dataset, 103 accounts remained for training.

In a preprocessing step common stop words have been excluded. In order to investigate the influence of stemmed input data we apply the Cistem stemmer [7] on the tokenized tweet text. For the part-of-speech-based weight models, we used the ClassifierBasedGermanTagger by Philipp Nolte [8], which was trained on the TIGER corpus [9].

Although all included accounts have in common that they are actively involved in politics, their tweet behaviour is very heterogenous. Some of them use a lot of sharepics and professionally chosen phrases and hashtags (e.g. #fedidwgugl (CDU), #lustauflinks (Linke), #traudichdeutschland (AfD)), while others also write about their regular parliament work, about their personal life, about soccer or about the Octoberfest. These non-political tweets were not removed from the set, and are thus included in the reported accuracies. In addition, many tweets with political content can only be interpreted in context. Here are some examples:

> *Liebe @Piratenlily Du bist wundervoll. Hoffentlich wird nun endlich eine offene und ehrliche Debatte geführt.* – @AnjaHirschel on 2017-06-01 09:55 CEST. (Translation: Dear @Piratenlily, you are wonderful. Hope- fully there will be an open and honest debate now)
>
> *I am still not convinced... #autobahngesellschaft* – @SebRoloff on 2017-06-01 10:24 CEST. (Hashtag translation: highway society)
>
> *Syrien, Nicaragua, USA* – @NielsAnnen on 2017-06-01 21:54 CEST.

All of the above tweets have a political meaning, but are hard to classify even for humans, if the context is not known.

4 Model Description

In this section we present the general architecture of our classifier and the various model configurations that are compared. We will use the following terminology: 'parties' denotes the set of the nine parties given in Table 1, 'tweets' the set of 59 360 tweets posted by the 153 chosen Twitter accounts in the investigated time interval and 'vocabulary' the set of words occurring in the tweets, i.e. vocabulary = $\{w|\exists t \in \text{tweets}, w \sqsubseteq t\}$. For a tweet t we write $w_1 w_2 \ldots w_m \sqsubseteq t$, if and only if $w_1 w_2 \ldots w_m$ is a continuous sequence of words in t. Finally, the function 'party' assigns to a tweet the political party of the account by which it was posted.

Since we cannot know the actual distribution of the a-priori party probability $P(\text{party}(t) = p)$ of a tweet t to belong to an account of party p in the real world, we have chosen to assume a uniform distribution.[2]

We use the relative frequency with add-one smoothing (as described by Koehn [10]) as an estimator for the unigram probability given a specific party:

$$P(w \sqsubseteq t|\text{party}(t) = p) = \frac{|\{t' \in \text{tweets}|w \sqsubseteq t', \text{party}(t',) = p\}|+1}{|\{t' \in \text{tweets}|\text{party}(t') = p\}|+|\text{vocabulary}|}$$

Thus $P(w \sqsubseteq t|\text{party}(t) = p)$ is the probability that a word w occurs in a tweet t given the information that t belongs to party p. Using Bayes' law and the previously established a-priori probability, we can derive the probability for a specific party given a unigram:

$$P(\text{party}(t) = p|w \sqsubseteq t) = \frac{P(w \sqsubseteq t|\text{party}(t) = p) \cdot P(\text{party}(t) = p)}{\sum_{q \in \text{parties}} P(w \sqsubseteq t|\text{party}(t) = q) \cdot P(\text{party}(t) = q)}$$

In order to process entire tweets, we use Markov processes of orders $n = 1$ to 5 for estimation. Thus, our word probability in a given context for a given party is as follows:

[2] Alternatives would have been to choose a distribution given by the strength of the party (members, results in elections, . . .) or by the average amount of posted tweets per party in our example corpus. Both approaches are problematic as well as it is not clear whether the strength of a party correlates with its Twitter activities and whether our corpus is balanced.

$$P(w_k|w_1 \dots w_{k-1}, p) \approx \hat{P}_n(w_k|w_{k-n} \dots w_{k-1}, p)$$
$$= \frac{|\{t' \in \text{tweets}|w_{k-n} \dots w_k \sqsubseteq t', \text{party}(t') = p\}| + 1}{|\{t' \in \text{tweets}|w_{k-n} \dots w_{k-1} \sqsubseteq t', \text{party}(t') = p\}| + |\text{vocabulary}|}$$

And, using Bayes' law again, we can derive the probability of the author supporting a specific party given a string $w_1 \dots w_m$:

$$P(\text{party}(t) = p|w_1 \dots w_m \sqsubseteq t) = \frac{P(w_1 \dots w_m \sqsubseteq t|\text{party}(t) = p) \cdot P(\text{party}(t) = p)}{\sum_{q \in \text{parties}} P(w_1 \dots w_m \sqsubseteq t|\text{party}(t) = q) \cdot P(\text{party}(t) = q)}$$

However, as not all words are similarly meaningful for party classification, we add weights to the word factors. We introduced three kinds of weights:

1. weights by document frequency: words that occur more often in the corpus are considered to be more important (negative α) or less important (positive α)
2. weights by part of speech: words with a specific POS tag are considered more important for the classification
3. uniform weight: as a control weight for comparison, i.e. $\omega 1(w) = 1$

After incorporating the weights, the string probability given a specific party is modeled as follows:

$$P_{\omega,n}(t = w_1 \dots w_m|p) =$$
$$P_1(w_1|p)^{\omega(w_1)} \cdot P_2(w_2|w_1, p)^{\omega(w_2)} \cdot \dots \cdot P_n(w_m|w_{m-n} \dots w_{m-1}, p)^{\omega(w_m)}$$

The document frequency weights use the simplification that a tweet usually does not contain the same word twice. The parameter α controls the influence of the document frequency. A positive α lowers the influence of frequent words, a negative α increases it. The parameter β does not change the party probability order, but it is meant to scale the weights to center around 1. We have tried $\omega_{DF:-1:10}$, $\omega_{DF:-0.1:1.5}$, $\omega_{DF:1:10}$ and $\omega_{DF:0.1:1.5}$.

$$\omega_{DF:\alpha:\beta}(w) = \beta \left(\frac{|\{t \in \text{tweets}|w \sqsubseteq t\}|}{|\text{tweets}|} \right)^{\alpha} \approx \beta \left(\frac{\#\text{occurences of } w}{|\text{tweets}|} \right)^{\alpha}$$

The part-of-speech weight model uses tuples $\nu = (\nu_N, \nu_V, \nu_A, \nu_\#, \nu_@, \nu_X) \in \mathbb{R}^6$, where the word's part-of-speech tag determines which weight to use. Table 2 lists the POS weight models used in our experiment.

Table 2. Four different POS weight models used in the experiment

configuration	nouns	verbs	adjectives	hashtags	mentions	misc
POS_{nouns}	1.5	0.8	0.8	1.0	0.1	0.5
POS_{verbs}	0.8	1.5	0.8	1.0	0.1	0.5
$POS_{ad\ j}$	0.8	0.8	1.5	1.0	0.1	0.5
POS_{htag}	0.8	0.8	0.8	3.0	1.0	0.5

In addition to the problem of classifying single tweets described so far (*single tweet mode*), we aim at a classifier for entire accounts as well (*full-account mode*). To simplify the problem, we pretend that all tweets posted by an account are independent of each other. This is an unrealistic assumption, since often the tweets are about similar topics, and their author still has the same interests, habits and political views. It is still useful, as it substantially simplifies calculations, and allows us to work with our limited training data. Thus we assume that

$$P_\omega(\{t_1, t_2, \ldots, t_m\}|p) = P_\omega(t_1|p) \cdot P_\omega(t_2|p) \cdot \ldots \cdot P_\omega(t_m|p)$$

where $t_1, \ldots, t_m$ are all tweets by the given account in the inspected time interval. From that, we can calculate the party membership probabilities:

$$P_\omega(\text{party}(a) = p|\{t_1, t_2, \ldots, t_m\}) = \frac{P_\omega(\{t_1, t_2, \ldots, t_m\}|p)}{\sum_{q \in \text{parties}} P_\omega(\{t_1, t_2, \ldots, t_m\}|q)}$$

Here, $\text{party}(a)$ denotes the party to which an account from our corpus is assigned.

Table 3. Accuracies of selected model configurations in single-tweet mode.

Configuration	Correct (Test)	Accuracy	Correct (Training)	Accuracy
stemmed:1:df:-0.1:1.5	5908	36.03 %	34164	79.52 %
unstemmed:1:df:-0.1:1.5	5848	35.67 %	35602	82.87 %
unstemmed:1:id	5676	34.62 %	32893	76.56 %
stemmed:1:id	5673	34.60 %	31482	73.28 %
...	...	...	...	...
stemmed:4:pos:adj	2585	15.77 %	37925	88.27 %
stemmed:5:pos:nouns	2568	15.66 %	37978	88.40 %

The source code for the tools used in our experiments is made available.[3]

5 Evaluation

In *single tweet mode* each status message from the test set was classified independently. As observable in Table 3, the unigram models performed best, while the models for larger Markov orders are overfitting.[4] Among the unigram models, only $\omega_{DF:-0.1:1.5}$ (slightly favouring frequent words) performed better than the uniform weight. The best configuration stemmed:1:df:-0.1:1.5 achieved an accuracy of 36 %.

We applied the paired Wilcoxon signed rank test [11] with continuity correction [12] to compare the accuracy of configurations with stemming to those without stemming. It turned out that generally configurations without stemming perform significantly better, at the 0.1 % level, even though the best-performing configuration uses stemming. We then applied the Kruskal-Wallis rank sum test [13] and Dunn's post-hoc test [14] with Holm correction [15] to compare the accuracies of configurations of different Markov orders, and found that there is a significant

[3] 3All software tools developed for this study are free software and published under the GNU Affero General Public License, version 3. The source code is accessible from https://hub.darcs.net/enum/twitbtw, either using Darcs version control, or by downloading the zip archive. Due to copyright limitations we cannot publish the training and test data, but the list of accounts is contained in the plain text file ACCOUNTS.org. The tools are mostly written in Haskell and designed to run on GNU/Linux, although other operating systems may work as well. Happy hacking!

[4] A complete table of all results in single tweet and full-account mode is given at https://hub.darcs.net/enum/twitbtw/browse/evalresults.

difference between $n = 1$ and $n = 2$ at the 5 % level, as well as between $n = 1$ and $n \in \{3, 4, 5\}$ at the 0.1 % level. There were no significant differences in the other pairs. Comparing weight models, the only significant difference was between $\omega_{DF:-0.1:1.5}$ and $\omega_{POS:noun}$, at the 5 % level.

Table 4 shows that there is a considerable bias towards AfD and SPD in the classifier. Hence, there is a high recall for these parties, but a lower precision. The small party PARTEI is almost never guessed.

In the *full-account mode* all tweets from the same author are grouped and there is only one result per account. In this mode, the higher order Markov models performed much better than the unigram models, and the best weight models were $\omega_{DF:-1:10}$ (favouring frequent words) and $\omega_{POS:htag}$ (favouring hashtags). There was no observable difference between weight models favouring specific lexical categories, but all of them performed better than the uniform weight. A possible explanation is that all of them disfavour mentions (cf. Table 2). The best configuration in full-account mode is unstemmed:4:df:-1:10, it achieved an accuracy of 72 % (cf. Table 5 and footnote 4).

Table 4. Actual party distribution in the test dataset vs. distribution among the results with configuration stemmed:1:df:-0.1:1.5, in single-tweet mode.

	actual # of tweets	classif. results	precision	recall	F-measure
AfD	1532	4443	0.26	0.75	0.39
CDU	2687	502	0.58	0.11	0.18
CSU	243	60	0.52	0.13	0.20
FDP	1328	275	0.69	0.14	0.24
Greens	3371	3057	0.46	0.42	0.44
Left	3304	1905	0.32	0.19	0.24
PARTEI	256	4	0.75	0.01	0.02
Pirates	570	521	0.48	0.44	0.46
SPD	3106	5630	0.35	0.63	0.45

In this scenario, the non-stemming configurations performed better as well, but the difference is significant at the 5 % level only. In median, the accuracy difference is 0. According to the Kruskal-Wallis rank sum test there were no significant differences between Markov orders, but there are differences between the weight models: $\omega_{DF:-1:10}$ performed significantly better than all other DF weights and the uniform weight at the 0.1 % level, but not significantly different from any of the

POS weights. $\omega_{POS:htag}$ performed significantly better than the uniform weight and all DF weights except for $\omega_{DF:-1:10}$, at the 1 % level. The weight $\omega_{DF:1:10}$ performed significantly worse than all POS weights at the 1 % level.

Table 5. Words *w* with highest *P*(party(*t*) = *p*|*w* ∈ *t*), for each party *p* (unstemmed and unweighted unigrams)

Configuration	Correct (Test)	Accuracy	Correct (Training)	Accuracy	
unstemmed:4:df:-1:10	36	72.00 %	102	100.00 %	
stemmed:5:df:-1:10	36	72.00 %	102	100.00 %	
...	...	...	...		...
stemmed:1:id	22	44.00 %	64	62.75 %	
unstemmed:1:id	22	44.00 %	66	64.71 %	
...	...	...	...		...
stemmed:1:df:1:10	11	22.00 %	21	20.59 %	
unstemmed:1:df:1:10	11	22.00 %	21	20.59 %	

Table 6. Words w with highest P(party(t) = p|w ∈ t), for each party p (unstemmed and unweighted unigrams).

rank	AfD	CDU	CSU	FDP	Greens
1	#traudichdeutschland	#100hcdu	#fragcsu	cl	#darumgrün
2	#afd	#fedidwgugl	#klarfürunserland	tl	#darumgruen
3	→	@connect17de	#bayernplan	#denkenwirneu	#bdk17
4	@afdberlin	@petertauber	@andischeuer	#bpt17	#ldknds
5	altparteien	angela	#banz17	@danielkolle	katrin

rank	Left	PARTEI	Pirates	SPD
1	#linkebpt	smiley	#piraten	fröhlicher
2	@dietmarbartsch	#diepartei	#freudichaufsneuland	gruss
3	@b_riexinger	#partei	@piratenpartei	traumschön
4	#mahe	#lwa	#copyright	@gabyulm
5	@swagenknecht	#smiley	@anjahirschel	sonnigen

Considering that we have included nine parties in our experiment, a random selection would have produced an accuracy of 11 %. Hence, even the worst model configurations achieved better results than chance.

6 Conclusions

We have presented a probabilistic classifier for party membership using a weighted n-gram model that offers two modes: One mode uses only a single tweet for classification, the other one uses all tweets by an account in the given time interval. In contrast to most existing publications on that topic, we did not assume a binary alignment (left/right or Republicans/Democrats), but instead used nine German parties as classes of different size and popularity.

The results in Table 3 and Table 5 show that the two modes of operations require different settings to achieve their best performance: The single tweet mode performs best with a unigram model, while the full-account mode profits from higher-order Markov chains. Additionally, we have seen that, for German, stemming is counter-productive and leads to a worse performance. However, this is likely to be a language-specific observation. In future research, Finnish tweets will be classified as well. Since Finnish uses a lot more morphology than German does, the stemming might be important there. Furthermore, we found that assigning a higher weight to hashtags and to frequent words improves the result in the full-account mode. We did however not find any differences between lexical categories.

For the single tweet mode we observed an accuracy of 36 % in the best configuration; in the full-account mode, we achieved an accuracy of 72 %. As Cohen and Ruths showed, these results need to be treated carefully. Our system was trained on active political agents and will most likely perform worse on average users. However, it is important to note that we have not excluded tweets that do not comment on political issues. In both modes the baseline approach of a random party guesser with an accuracy of 11.11% is clearly outperformed.

In practice, the classifier could be combined with a filter to exclude unpolitical tweets first. One such filter has e.g. been described by De Mello Araújo and Ebbelaar[16].

References

1. Pennacchiotti M & Popescu AM (2011) A machine learning approach to twitter user classification. In: Fifth International AAAI Conference on Weblogs and Social Media.
2. Conover MD, Gonçalves B, Ratkiewicz J, Flammini A & Menczer F (2011) Predicting the political alignment of twitter users. In: 2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing, pp. 192–199. IEEE.
3. Cohen R & Ruths D (2013) Classifying political orientation on twitter: Its not easy! In: Seventh International AAAI Conference on Weblogs and Social Media.
4. Boutet A, Kim H & Yoneki E (2013) Whats in twitter, I know what parties are popular and who you are supporting now! Social Network Analysis and Mining 3(4): 1379–1391.
5. Pla F & Hurtado LF (2014) Political tendency identification in twitter using sentiment analysis techniques. In: Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical Papers, pp. 183–192.
6. Tumasjan A, Sprenger TO, Sandner PG & Welpe IM (2010) Predicting elections with twitter: What 140 characters reveal about political sentiment. In: Fourth international AAAI conference on weblogs and social media.
7. Weissweiler L & Fraser A (2017) Developing a stemmer for german based on a comparative analysis of publicly available stemmers. In: Proceedings of the International Conference of the German Society for Computational Linguistics and Lan- guage Technology. German Society for Computational Linguistics and Language Technology, Berlin, Germany.
8. Konrad M (2016). Accurate part-of-speech tagging of german texts with nltk. URI: https://datascience.blog.wzb.eu/2016/07/13/ accurate-part-of-speech-tagging-of-german-texts-with-nltk/.
9. Brants S, Dipper S, Eisenberg P, Hansen-Schirra S, König E, Lezius W, Rohrer C, Smith G & Uszkoreit H (2004) Tiger: Linguistic interpretation of a german corpus. Research on Language and Computation 2(4): 597–620.
10. Koehn P (2010) Statistical Machine Translation. Cambridge University Press.
11. Wilcoxon F (1945) Individual comparisons by ranking methods. Biometrics bul- letin 1(6): 80–83.

12. Yates F (1934) Contingency tables involving small numbers and the χ 2 test. Supplement to the Journal of the Royal Statistical Society 1(2): 217–235.
13. Kruskal WH & Wallis WA (1952) Use of ranks in one-criterion variance analysis. Journal of the American statistical Association 47(260): 583–621.
14. Dunn OJ (1964) Multiple comparisons using rank sums. Technometrics 6(3): 241– 252.
15. Holm S (1979) A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics pp. 65–70.
16. de Mello Araújo EF & Ebbelaar D Detecting Dutch Political Tweets: A Classifier based on Voting System using Supervised Learning: In: Proceedings of the 10th International Conference on Agents and Artificial Intelligence, pp. 462–469. SCITEPRESS - Science and Technology Publications.

Optical font family recognition using a neural network

Senka Drobac & Krister Lindén
University of Helsinki

Abstract

Working on OCR (Optical character recognition) of historical newspaper and journal data, we found it beneficial to analyze and evaluate our OCR results based on font family. Our data sets are extracted from a corpus of historical newspapers and magazines (from 1771 until 1874) that have been digitized by the National Library of Finland. Our earlier data is mainly written in Blackletter fonts and later data in Antiqua fonts, while in the transitioning period both font families were used at the same time, even on the same pages. Therefore, in order to make the recognition phase easier and faster, we are building one OCR model, which is able to recognize all fonts represented in the data. In order to make sure that we have sufficient training data for both font families, we need a font family classifier to simplify creation and sampling of training data.

Although there are existing tools for font classification, our problem seems to be overly specific. We only need to distinguish between Blackletter and the other fonts that were printed in Finland between 18th Century and early 20th Century, so the challenge is to find a simple enough font classifier for such a specific task.

Sahare et al., 2017 conveyed a detailed survey of different script identification algorithms. Zramdini et al., 1998 have developed a statistical approach of font recognition based on global typographical features. They report 97 % accuracy of typeface recognition. Brodić et al., 2016 have approached a similar problem as we have, when they do identification of Fraktur and Latin scripts in German historical documents using image texture analysis. The accuracy of their system has been reported to be 98.08 %.

In this work, we build a deep neural network binary font family classifier that for an image of one line of text decides whether it is written in Blackletter or Antiqua typeface. Even with a simple configuration of the network, we get 97.5 % accuracy, leaving space for further improvement.

This font family classifier is specifically created for historical OCR for data printed in Finland. It is useful for collecting and analyzing the data, especially if the OCR is done with line-based software (Ocropy, Kraken, Calamary, Tesseract

4). The font classifier is simple to use, in both the training and prediction phases. It is also easy to change network configurations and parameters.

1 Introduction

Working on OCR (Optical character recognition) of historical newspaper and journal data published in Finland, we found it beneficial to analyze and evaluate our OCR results based on font family. In Drobac et al. (in press) we indicate that recognizing the font family may be more important than recognizing the language of a document as a prerequisite for performing good OCR. This is a largely overlooked problem as many OCR tasks have specialized in documents with just one font, whereas in practice, historical newspapers were printed with several fonts. This is especially true during the transition period from Blackletter to Antiqua in Europe.

From a language technological point of view, the information on font family can be used to ensure that a sufficient amount of training data for machine learning to improve the quality of the OCR of historical newspapers is available. In addition, the font usage in periodicals can be used as to investigate the development of the printing press and the reading preferences during a transition period from 1800 until 1910 when Antiqua had largely replaced the Blackletter.

Our earlier data is mainly printed in Blackletter fonts and later data in Antiqua fonts, while in the transitioning period both font families were used at the same time, even on the same pages. Therefore, in order to make the recognition phase easier and faster, we are building one OCR model, which is able to recognize all fonts represented in the data. In order to make sure that we have sufficient training data for both font families, we need a font family classifier to simplify creation and sampling of training data.

Although there are existing tools for font classification, our problem seems to be overly specific. We only need to distinguish between Blackletter and the other fonts that were printed in Finland between 18th century and early 20th century, so the challenge is to find a simple enough font classifier for such a specific task.

Sahare and Dhok (2017) conveyed a detailed survey of different script identification algorithms. Zramdini and Ingold (1998) have developed a statistical approach of font recognition based on global typographical features. They report 97% accuracy of font family recognition. Brodić et al. (2016) have approached a similar problem as we have, when they do identification of Fraktur and Latin scripts

in German historical documents using image texture analysis. The accuracy of their system has been reported to be 98.08%.

In this work, we build a deep neural network binary font family classifier that for an image of one line of text decides whether it is written in Blackletter or Antiqua. Even with a simple configuration of the network, we get 97.5% accuracy, leaving space for further improvement.

This font family classifier is specifically created for historical OCR for data printed in Finland. It is useful for collecting and analyzing the data, especially if the OCR is done with line-based software (Ocropy,[1] Kraken,[2] Calamary,[3] Tesseract 4[4]). The font classifier is simple to use, in both the training and prediction phases. It is also easy to change network configurations and parameters.

2 Data and resources

In our experiments, we use three data sets: *img-lines, swe-6k* and *fin-7k*, all extracted from a corpus of historical newspapers and magazines that have been digitized by the National Library of Finland. Data ranges from 1771 until 1874 for *img-lines* and *swe-6k* sets, and from 1820 until 1939 for *fin-7k*.

Fig. 1. Training examples: a) Antiqua, b) Blackletter.

The first data set (*img-lines*) was created specifically for this task. The data set was collected from manually classified newspaper and journal pages. First, we randomly picked 307 pages from the entire corpus. We manually checked all the pages and classified them into three classes:

[1] https://github.com/tmbdev/ocropy
[2] http://kraken.re/
[3] https://github.com/Calamari-OCR/calamari
[4] https://github.com/tesseract-ocr/tesseract/wiki/4.0-with-LSTM

- *Antiqua-only* - pages that consist of mostly Antiqua font
- *Blackletter-only* - pages consist of mostly Blackletter font
- *Mixed pages* - pages with both Blackletter and Antiqua fonts

Then we segmented *Antiqua-only* and *Blackletter-only* pages into lines and cleaned the sets of poor quality, miss-segmented lines, or wrong font family. In the end, we were left with total of 6,356 Antiqua lines and 7,205 Blackletter lines.

Figure 1 shows five training examples from this set. On the left there are Antiqua lines and on the right Blackletter lines.

The other two data sets (*swe-6k* and *fin-7k*) were previously used for OCR of historical data. They both consist of randomly picked image lines, *swe-6k* has in total 6,158 image lines of Swedish text and *fin-7k* has 7,000 image lines of Finnish text. These two data sets had previously been divided into Blackletter and Antiqua and were used only for testing purposes.

Keras (Chollet et al., 2015) is a python library for machine learning. It runs on top of Tensorflow, and its simple and user friendly API (application programming interface) together with good quality documentation makes it easy to use.

In additions to neural networks, it also provides functions for image processing which allows dynamical augmentation of training data. This is particularly useful for image classification with small training sets, because in each iteration, the network receives a slightly altered image, and thus it never sees the same training image twice.

3 Method

In this section, we describe the preparation of the training and testing data and the structure of the neural network that we used to train the models. We also describe our evaluation methods.

3.1 Preparing the data

Although the original sizes of *img-lines* data sets were somewhat larger (Antiqua 6,356; Blackletter 7,205), we settled for 6,000 training lines from each category. Additionally, we reserved 200 lines for validation purpose and 100 lines for testing. As noted earlier, we used *swe-6k* and *fin-7k* exclusively for testing to have different test sets from the same corpus.

To load images into the neural network, we use a data generator with augmentation parameters on the training set, described in Table 1.

Table 1. Augmentation parameters used on training set.

rescale=1./255	Rescaling factor
rotation_range=10	Degree range for random rotations
width_shift_range=0.2	Fraction of total width
height_shift_range=0.2	Fraction of total height
shear_range=0.2	Shear Intensity
zoom_range=0.2	Range for random zoom. If a float, [lower, upper] = [1-zoom_range, 1+zoom_range]
fill_mode='nearest'	Points outside the boundaries of the input are filled according to the given mode: 'nearest': aaaaaaaa|abcd|dddddddd

It is also important to note that input image dimensions were set to 100 x 500 pixels for all training, validation and testing.

3.2 Neural network

The neural network consists of three pairs of convolution and pooling layers with a ReLU Activation function, followed by two fully connected layers and a Sigmoid output layer that predicts the probability of the input image belonging to one of the two classes. All convolution layers have a kernel size of 3 x 3 with zero padding. The first and the second layers have 32 filters and the third layer 64 filters. The pooling layers implement MaxPooling with a kernel size and stride of 2 x 2. Each fully connected layer has 64 hidden states, and the first layer has a ReLU Activation function. Between fully connected layers, we apply dropout (Srivastava et al., 2014) with a rate of 0.25 to prevent overfitting. The loss is computed using Binary crossentropy with the RMSProp optimizer. As an input for training, the algorithm expects a list of classified line images stored in respective file folders with corresponding names (i.e. Blackletter and Antiqua). Figure 2 shows a diagram of this neural network.

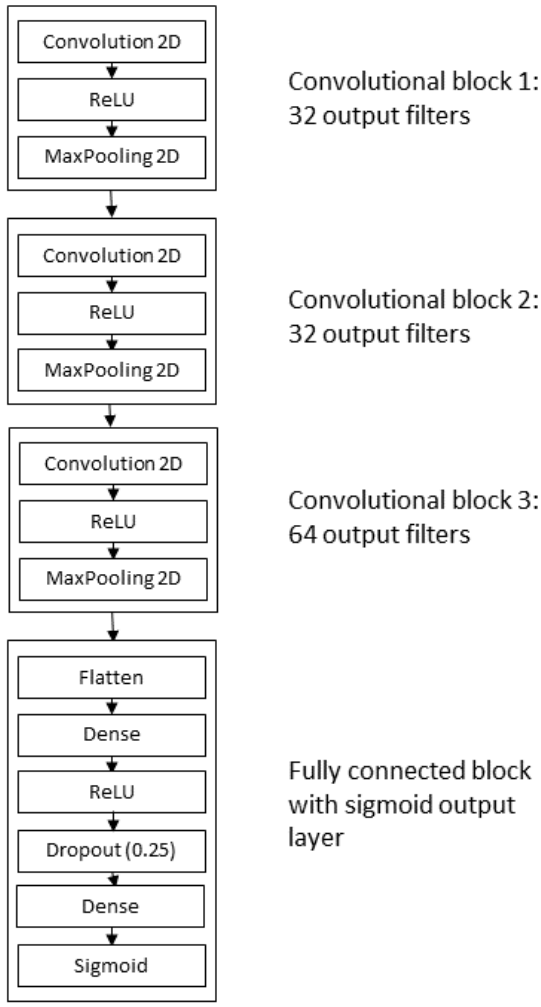


Fig. 2. Diagram of the neural network used for binary classification.

3.3 Evaluation

For evaluation, we used the accuracy as a measure of correctly classified line images per total number of images, expressed in percentage:

$$accuracy = \frac{correct\ nomber}{total\ number} \cdot 100\%$$

To get the best model, we used early stopping on the validation set with patience 2, with the best weights restored.

On the unseen test sets, in addition to accuracy, we also calculate precision, recall and F_1 score. We define True Positive (TP) as correctly predicted Blackletter, False Positive (FP) as incorrectly predicted Blackletter, True Negative (TN) as correctly predicted Antiqua and False Negative (FN) as incorrectly predicted Antiqua. Then we calculate precision and recall as:

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

The F_1 score is defined as the measure that combines precision and recall using the harmonic mean:

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

4 Results

Table 2 shows prediction results on test sets *swe-6k* and *fin-7k* in terms of number of correctly and wrongly predicted line images.

Table 2. Prediction results on unseen data. On the left side, there are results on the test set *swe-6k* and on the right on *fin-7k*. The first row shows the number of predicted results for each font family category from a total of 6,158 test image lines in *swe-6k* and 7,000 images in *fin-7k*. The next two rows show how many were correctly and how many wrongly predicted.

Parameters	*swe-6k*		*fin-7k*	
	Antiqua	Blackletter	Antiqua	Blackletter
Predicted	2842	3316	1296	5704
True	2837	3054	1285	5540
False	5	262	11	164

Table 3 shows Accuracy of all three sets and Precision, Recall and F_1 score on *swe-6k* and *fin-7k*. Since *swe-6k* and *fin-7k* are used in a real world application (training of OCR models), we wanted to evaluate them further.

Table 3. Accuracy, Precision, Recall and F_1 score on different test sets. The first row shows results on the img-lines test set, the second on the *swe-6k* test set and the third on the *fin-7k* test set.

Test set	Accuracy	Precision	Recall	F_1
img-lines	97.5%	-	-	-
swe-6k	95.64%	99.8%	92.1%	95.8%
fin-7k	97.5%	99.8%	97.1%	98.4%

5 Discussion and conclusions

The accuracy results of 97.5% on the classification test set and 95.64% and 97.5% on the real world test sets are quite high considering that we used only a basic setup without any experimentation with different model configurations or parameters. The high F_1 score shows that the model is quite good at balancing precision and recall, especially for *fin-7k*.

It is interesting to see that the accuracy on the *swe-6k* test set is almost 2% lower than the accuracy on *fin-7k*, especially in light of the fact that this set is also more difficult for OCR than the Finnish set. Those two sets also differ on Antiqua and Blackletter ratio, with *swe-6k* having 46% Antiqua and 54% Blackletter fonts, while *fin-7k* only has 18% Antiqua and 82% Blackletter. It is possible that a larger number of Antiqua font lines in *swe-6k* also means a larger variety of fonts, making optical character recognition more challenging. It is possible that more training data and a deeper neural network would solve this problem.

Since this method is developed for a very specific task, it is difficult to make a comparison with other software without actually testing those systems on our data. However, comparing results that they get on their data with our results, we can see that we achieve a similarly high accuracy with a rather standard deep neural network approach.

This is a first approach to automatically classify our data to get enough training data from two font families, and for this purpose, we created a binary classifier. However, with simply changing the output layer to *softmax* and the loss function to *categorical crossentropy*, we could get a multi-class classifier with the same neural network setup.

References

Drobac, S., Kauppinen, P., & Lindén, K. (in press). Improving OCR of historical newspapers and journals published in Finland. In *Proceedings of the 3nd International Conference on Digital Access to Textual Cultural Heritage ACM*.

Sahare, P. & Dhok. S. B. (2017). Script identification algorithms: a survey. *International Journal of Multimedia Information Retrieval* 6(3): 211–232.

Brodić, D., Amelio, A. & Milivojević, Z. N. (2016). Identification of fraktur and latin scripts in german historical documents using image texture analysis. *Applied Artificial Intelligence* 30(5): 379–395.

Zramdini, A. & Ingold, R. (1998). Optical font recognition using typographical features. *IEEE Transactions on pattern analysis and machine intelligence* 20(8): 877–882.

Chollet, F. et al. (2015). Keras. https://keras.io.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* 15(1): 1929–1958.

Distinguishing translations from non-translations and identifying (in)direct translations' source languages[1]

Laura Ivaska
University of Turku

Abstract

The scope of this study is threefold. First, machine learning will be applied to distinguish translated from non-translated Finnish texts. Then, it will attempt to identify the source languages of the translated Finnish texts. Finally, the source language identification will be tested with indirect translations, that is, with translations made from translations. The three underlying research questions are: 1) Can translated Finnish be distinguished from non-translated Finnish? 2) Can the source languages of Finnish translations be identified? 3) If the answer to question 2 is yes, then what happens when the method is applied to indirect translations; will the analysis identify the ultimate source language, the mediating language, or neither?

This study is based on the hypothesis that translated language contains traces of the source language (Toury 1995). The corpus of the study consists of non-translated Finnish prose, Finnish prose literature translations made from English, German, French, Modern Greek, and Swedish, as well as indirect translations from Modern Greek into Finnish via English, German, French, and Swedish. The analyses are based on cluster analysis and support vector machines using the frequencies of the most frequent lemmatized words.

Results show that translated and non-translated Finnish can be distinguished by using machine learning techniques. Support vector machine-based source language identification, however, was only partially successful, while a cluster analysis suggested that there is coherence within a group of texts translated from the same source language and variation between the groups of texts with different source languages. Clustering was further tested with indirect translations, and the results were mixed: six of the thirteen tested indirect translations clustered with

[1] Thank you to Ilmari Ivaska, Jaakko Kankaanpää, Einari Aaltonen, Jonna Joskitt-Pöyry, Arja Kantele, Outi Menna, Jaana Nikula, Raija Rintamäki, Tähti Schmidt, Sirkka-Liisa Sjöblom, Oili Suominen, and Marja Wich. This research was funded by the Kone Foundation.

direct translations from the ultimate source language, two with translations from their mediating languages, and five with neither.

1 Theoretical background

Baker's (1993, 243) suggestion to explore translation universals, or "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" gave rise to corpus-based translation studies in the 1990s. Researchers have since studied what features translations share and what distinguishes translated language from non-translated language (e.g., Mauranen 2004; Baroni and Bernardini 2006). Also, following Toury's (1995) suggestion that translated language contains traces of the source language (SL)—a phenomenon he calls interference—other studies have focused on identifying the SLs of translations and the linguistic features that make the identification possible (e.g., Koppel and Ordan 2011; Lynch and Vogel 2012; Islam and Hoenen 2013).

The main focus of this study, however, is on whether corpus methods can identify (the SLs of) indirect translations (ITrs). In previous research on SL identification, ITrs were not commonly taken into account, which may have led to less accurate findings than what might have otherwise been obtained. Nevertheless, ITrs are interesting because they are the result of a chain of several texts/languages: the ultimate source text/language → mediating text/language → ultimate target text/language (cf. Assis Rosa, Pieta and Maia 2017), for example Greek → French → Finnish. This raises the question of if direct translations contain traces of their SLs, do ITrs contain traces of the ultimate SL, the mediating language, or both?

Recently, Ustaszewski (2018, 173) tackled the question, "Is there an effect of the pivot language on target texts in indirect translation, and is this effect strong enough to discriminate between direct and indirect translations?" However, the results of his study cannot be confirmed because the Europarl corpus that he used lacked metadata on the (in)directness of the translations. In the current study, the (in)directness of the translations and their SLs are known and, therefore, the outcome of the SL identification can be confirmed. The results of this study can, then, be applied to and facilitate further research on ITr: if the SL identification methods detect the mediating languages of ITrs, the methods could be used to uncover ITrs, an arduous task when using the current methods (cf. Ivaska 2018).

2 Materials

The corpus in this study contains different variants of Finnish: 1) non-translated prose literature (Fi–Fi); 2) literary translations from English, French, German, Modern Greek, and Swedish (En–Fi, Fr–Fi, De–Fi, Gr–Fi, and Sv–Fi, respectively); and 3) indirect translations (ITr) of Modern Greek literature translated via English, French, German, and Swedish. The texts included in the corpus are novels and, because direct translations from Modern Greek into Finnish are scarce, there is also one collection of Gr–Fi short stories in the corpus (for the sake of clarity, the latter is considered one text even though it contains texts by several authors and various translators).

The majority of the texts used in this study come from two corpora, the Corpus of Translated Finnish (CTF) (Mauranen 2004) and Intercorp (Cermak and Rosen 2014) (Table 1). Since these two corpora contain only a few translations from languages other than English, further texts were solicited directly from translators. The translations from Modern Greek were scanned and processed into an electronic text format using Adobe Acrobat Pro DC's Optical Character Recognition (OCR). The results of the OCR have not been cleaned, and thus the translations from Greek are likely to contain errors; however, since all of the Gr–Fi and ITr texts went through a similar process, the effect of the eventual errors can be expected to even out.

The texts, except for the ITrs, were divided into subcorpora according to the language variant (De–Fi, En–Fi, Fr–Fi, Fi–Fi, Gr–Fi, Sv–Fi) they represented. Then, these subcorpora were further divided into training and test subcorpora (70% and 30% of the texts, respectively; see Table 2), and, to fade out authorial/translatorial style, texts by one author or translations by one translator in a particular language pair were always included in the same subcorpus (e.g., two novels by J.K. Rowling or three En–Fi translations by Kalevi Nyytajä are all either in training or test subcorpus).

As for the 13 ITrs, their indirectness, as well as their (assumed) SLs, had already been established in an ongoing research project (Ivaska 2016; Ivaska and Paloposki 2018). Some of the ITrs are compilative, meaning that they have been made with the help of support translations, where the translator had more than one language variant of the work (or, several source texts in different languages) at their disposal while composing the translation (cf. Dollerup 2000). However, in this current study, only the primary mediating language of each translation was

considered, as the role of the supporting translations was assumed to be marginal (cf. Ivaska forthcoming).

The texts were lemmatized with UDPipe (Straka and Strakova 2017, 88). Then, as is customary in the field, all the texts in each of the training and test subcorpora were shuffled at the sentence level to fade out features other than those attributable to the SL (e.g., author style; cf. Rabinovich, Nisioi, Ordan and Wintner 2016). Finally, the texts were sliced into chunks of 500 sentences, a number chosen to ensure that their length did not interfere with the analyses (cf. Volansky, Ordan and Wintner 2015). The last slice of each subcorpus was deleted, as these were shorter than 500 sentences and could, therefore, have skewed the results. The ITrs were not divided into training and test subcorpora but only lemmatized because in this study they were studied one by one.

Table 1. The texts in the corpus according to their provenance and language variant.

Language variant	Texts from CTF	Texts from InterCorp	Solicited texts	Scanned texts	Total
De–Fi	2	1	3	0	6
En–Fi	20	16	0	0	36
Fi–Fi	27	25	0	0	52
Fr–Fi	2	1	4	0	7
Gr–Fi	0	0	0	7	7
Sv–Fi	1	3	10	0	14
ITr	0	0	0	13	13
Total	52	46	17	20	135

Table 2. The number of texts and chunks of 500 sentences (lemmatized, shuffled, and sliced) in the training and test subcorpora by language variant.

Subcorpus		No. of texts	No. of chunks of 500 sentences
De–Fi	training	4	38
	test	2	26
En–Fi	training	25	323
	test	11	146
Fi–Fi	training	36	373
	test	16	122
Fr–Fi	training	5	39
	test	2	14
Gr–Fi	training	5	69
	test	2	24
Sv–Fi	training	10	174
	test	4	66

3 Methods

The analyses were done in R using the stylo package (Eder, Rybicki and Kestemont 2016). The main features used in this study included the functions stylo(), with which cluster analysis can be performed, and classify(), which provides supervised methods, such as support vector machines (SVM).

The analyses were based on the frequencies of lemmatized words (most frequent words [MFW]). This means that first, the frequencies of each word (or, in this study, of their lemmatized forms) in the whole (sub)corpus were calculated, and the words were listed from the most to the least frequent. Then, the word frequencies of each individual text were calculated and normalized with z-scores. For example, when a cluster analysis with 100 MFW is run, the first 100 words in the list prepared in the first step are the basis of the analysis: the clustering is based on the frequencies of these 100 words in each individual text. The experiment can also be set to repeat with 30–100 MFW and increases of 10, for example; in this case the test will be done with 1–30 MFW, 1–40 MFW … 1–100 MFW.

An MFW-based analysis is often done by leaving out content words and using only function words in order to fade out topic-specific influences (cf. Grieve 2016; Rabinovich, Nisioi, Ordan and Wintner 2016). There are no widely acknowledged function word lists for Finnish, but a similar effect can be created by using only the words that appear in all the texts; this is done in stylo by setting culling to 100%.

In Finnish, these are words such as *olla*, *ja*, *hän*, *ei*, and *minä* (*to be*, *and*, *s/he*, *no*, and *I*). The MFW function could also be used with other feature sets, such as part-of-speech grams, but because the scanned texts (which include all the Gr–Fi texts and ITrs) have not been cleaned, the accuracy of annotation could distort the results.

Two types of analyses were performed. In the unsupervised cluster analysis, the algorithm clustered the most similar texts together, forming a hierarchical dendrogram that visually illustrated which texts had the most similar MFW profiles. The supervised SVM classifier had two phases. In the training phase, the classifier was given text sets A, B, C … n. It studied their features (in this case, the MFW) and produced a profile for each text set. In the testing phase, the classifier was given text X. It studied its features, produced a profile for it, and compared the profile to those of A, B, C … n to decide which of them was the closest match.

3.1 Distinguishing translations and non-translations

To establish that non-translated Finnish can be distinguished from translated Finnish, a set of three experiments using a SVM classifier was done.

First, experiments were done with 50 chunks of lemmatized, shuffled, and sliced non-translated Finnish and 50 chunks of translated Finnish (consisting of ten chunks of De–Fi, En–Fi, Fr–Fi, Gr–Fi, and Sv–Fi each) in both the training and the test sets. The test run with SVM, with 10–100 MFW at 10-word increases, yielded a general attributive success of 76.4%, meaning that the algorithm correctly identified 76.4% of the chunks as (non)translations. The best result, 97% attributive success, was obtained with 30 MFW; here, the erroneous attributions (three chunks) were non-translated Finnish that were falsely identified as translated Finnish.

Then, to make the experiment as robust as possible, the test was repeated with as many chunks as there were available even if this resulted in the number of chunks ranging from 38 (De–Fi) to 373 (Fi–Fi) in the training set and 14 (Fr–Fi) to 146 (En–Fi) in the test set (see Table 2). The experiment was again based on the SVM, but the number of MFW was narrowed down to 15–50 with increases of 2, as the previous experiment suggested that the best results would be obtained somewhere within that range. The results obtained were slightly stronger, and the best result, 99.2% attributive success, was obtained with 21 MFW. As with the previous experiment setting, the chunks that were wrongly attributed were Fi–Fi chunks misidentified as translated Finnish.

Finally, since the maximum amount of training data seemed to provide the strongest results, the last set of experiments was done, again, using as much training data as possible (Table 2). The test data, however, consisted of the 37 texts in the various test subcorpora (see Table 2) in their full length, which had only been lemmatized (but not shuffled nor sliced) to create a setting resembling real-life conditions, where the method could be applied to full-length texts to verify their (non)translated status. The experiment was done with SVM, with 15–50 MFW in increases of 2. The general attributive success was 81.3%, and the best result, 86.5% attributive success, was obtained with 31 MFW (see Table 3; the one translation erroneously attributed as non-translated Fi–Fi was a Fr–Fi).

Table 3. The confusion matrix of the SVM experiment for distinguishing translations and non-translations with full-length test texts with 31 MFW.

		Attributed		
		Fi–Fi	Tr	Total
Actual	Fi–Fi	12	4	16
	Tr	1	20	21
	Total	13	24	37

3.2 Identifying direct translations' source languages

To take the analysis one step further, a cluster analysis and a SVM analysis to identify the SLs of chunks of translated Finnish were performed. For the cluster analysis, the corpus was tailored to best fit the purpose: all the texts of each language variant (De–Fi, En–Fi, Fr–Fi, Gr–Fi, and Sv–Fi) were put together, lemmatized, shuffled, and sliced into chunks of 500 sentences (once again, the last slices containing less than 500 sentences were deleted) (see Table 4).

Table 4. The number of chunks of 500 sentences (lemmatized, shuffled, and sliced) of translated Finnish by language variant.

Language variant	No. of texts	No. of chunks of 500 sentences
De–Fi	6	64
En–Fi	36	470
Fr–Fi	7	53
Gr–Fi	7	93
Sv–Fi	14	241

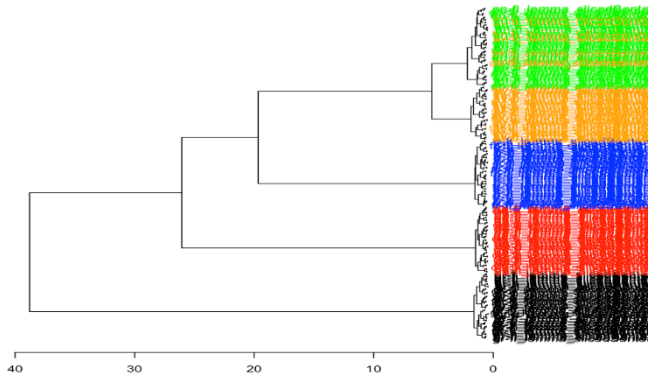


Fig. 1. The dendrogram of the cluster analysis of the translated language variants with 34 MFW; green represents En–Fi, orange Sv–Fi, red De–Fi, blue Fr–Fi, and black Gr–Fi.

In the cluster analysis, 53 chunks (the number of chunks available for the language variant with the smallest number of chunks) of De–Fi, En–Fi, Fr–Fi, Gr–Fi, and Sv–Fi each were used. The clustering was repeated with 2–54 MFW in increments of 2. The dendrogram with 34 MFW (Figure 1) most clearly differentiates the language variants, demonstrating that there is coherence within a group of chunks that were translated from the same SL (they were clustered together in one branch) and variation between the groups of chunks with different SLs (they form different branches), except for Sv–Fi, which paralleled En–Fi to the extent that ten chunks of Sv–Fi were clustered in the En–Fi branch.

After the cluster analysis, a SVM experiment was performed. Here, the pre-manipulated training and test data (Table 2) were used, with the training set consisting of 38 chunks (the number of chunks available for the language variant with the smallest training subcorpus) of each language variant (De–Fi, En–Fi, Fr–Fi, Gr–Fi, Sv–Fi) and the test data of 14 chunks (again, the number of chunks available for the language variant with the smallest test subcorpus) of each language variant. In performing a series of experiments with 2–52 MFW in increments of 2, the general attributive success was 26.3%. The best result, 35.7% attributive success, was gained with 40 MFW. None of the De–Fi nor Fr–Fi chunks were attributed correctly—the Fr–Fi translations were attributed as De–Fi or En–Fi, whereas the De–Fi translations were all attributed as En–Fi; all Gr–Fi

translations and roughly one third of the En–Fi and Sv–Fi translations were attributed correctly; no translations were attributed to Fr–Fi (see Table 5).

Table 5. The confusion matrix of the SVM experiment for attributing SLs with 40 MFW.

		Attributed					
		De–Fi	En–Fi	Fr–Fi	Gr–Fi	Sv–Fi	Total
Actual	De–Fi	0	14	0	0	0	14
	En–Fi	0	10	0	0	4	14
	Fr–Fi	7	7	0	0	0	14
	Gr–Fi	0	0	0	11	3	14
	Sv–Fi	0	10	0	0	4	14
	Total	7	41	0	11	11	70

3.3 Experimenting with indirect translations

In this last set of experiments, the aim was to see how the ITrs cluster. The expectation was that they would cluster either with Gr–Fi chunks or with chunks representing translations from their mediating languages. The former would mean that the interference from the ultimate SL carries over through the chain of ITr, and the latter that the interference from the mediating language overrides that from the ultimate SL. For the purpose of this experiment, the 13 ITrs were each lemmatized in their full length and clustered, one by one, with 53 chunks of each language variant (De–Fi, En–Fi, Fr–Fi, Gr–Fi, and Sv–Fi) (Table 4). Since the most discernible SL clusters were previously formed with 34 MFW (Figure 1), this setting was also used to experiment with the ITrs. In other words, the cluster analysis performed in the previous section was repeated 13 times with a different ITr added to each test.

Six of the ITrs clustered with Gr–Fi, two with chunks that represented translations from their mediating language, and the remaining six with neither Gr–Fi nor their mediating language (Table 6). Interestingly, the Fr–Fi, which had previously been misidentified in the SVM-based SL identification, showed a similar tendency here: none of the five ITrs done via French clustered with Fr–Fi chunks (Table 7).

Table 6. The results of the ITr cluster analysis with 34 MFW.

Result	No. of ITrs
Clustered with Gr–Fi	6
Clustered with mediating language	2
Clustered with neither Gr–Fi nor mediating language	5
Total	13

Table 7. The confusion matrix of the ITr cluster analysis with 34 MFW.

		Clustered					
		De–Fi	En–Fi	Fr–Fi	Gr–Fi	Sv–Fi	Total
Assumed	De–Fi	0	0	0	2	1	3
	En–Fi	0	2	0	0	0	2
	Fr–Fi	1	2	0	1	1	5
	Gr–Fi	0	0	0	0	0	0
	Sv–Fi	0	0	0	3	0	3
	Total	1	4	0	6	2	13

4 Conclusions and discussion

The first set of SVM-based experiments proved that translated Finnish can be distinguished from non-translated Finnish: the highest attributive success was 99.2%. An attempt to identify translations' SLs was also made, and although cluster analysis suggested that there is clear coherence within a group of translations from the same SL as well as clear variation between the groups of translations with different SLs, the best attributive success obtained with a SVM classifier was only 35.7%. The unsupervised cluster analysis yielded better accuracies than SVM because it does not make use of separate training and test sets, and, therefore, the variation within each language variant subcorpus is distributed equally to all the chunks.

The poor results with the SVM-based SL identification may be due to insufficient data, as suggested by the fact that none of the chunks of Fr–Fi and De–Fi, the language variants with the least variegated training and test subcorpora, were correctly attributed. SVM works better with more variegated subcorpora: the more variation there is to make the training profiles robust, the higher the attributive success. The need for a varied corpus is a limitation of supervised machine learning. For example, if the training corpus contained translations from one SL only by

translator X, the translator's style might override the SL features and become the defining element in the profile created by SVM for the training corpus. If, however, the training corpus also contained translations by translators Y and Z, the translators' individual styles would fade out and the common denominator, the same SL, would become the defining feature of the profile. A similar effect may also explain why the only language variant that was perfectly attributed was Gr–Fi; rather than the SVM identifying features caused by interference from a specific SL, the fact that these texts were not cleaned after the OCR may have left behind a feature that immediately distinguished the Gr–Fi translations from all other language variants. However, cleaning the 20 novels translated from Greek manually was not an option due to time constraints. Similarly, if only one text per author/translator had been allowed in the corpus to increase variation, some of the subcorpora would have become too small for the purposes of this study.

Since the cluster analysis could distinguish the SL variants, the last stage of the study, the SL identification with ITr, was also done using cluster analysis. Six of the thirteen ITrs clustered with Gr–Fi, suggesting that the signal of the ultimate SL carries through the chain of translations to the language of ITrs; this conjecture is bolstered by the fact that only two ITrs (which both had English as their mediating language) clustered correctly with their mediating languages. However, five ITrs did not cluster with either Gr–Fi or their mediating language. Perhaps the language of ITrs is actually mixed, containing traces of both the ultimate SL and the mediating language, and making SL identification impossible when using the language of direct translations as reference data. Should this supposition be correct, it might be possible to distinguish the specific language variants of indirect translations (e.g., Gr–De–Fi, Gr–Fr–Fi). However, sometimes translators consult several source texts in different languages, which could lead to the creation of further language varieties. Alternatively, different passages in the translation could contain different language varieties, in which case a windowing procedure, which focuses on one passage at a time, should be performed to identify the SL passage by passage.

It would be interesting to repeat the SVM-based SL identification experiment with a more robust corpus to see if this would yield better attributive success. If so, then the method should also be tested with ITrs. In any case, since the SVM-based classifier can distinguish between translated and non-translated language, it could be tested with pseudotranslations, that is, texts that pretend to be translations although they are not (Du Pont 2005), to see if the method could be used to expose impostor translations. Ultimately, developing a method to identify translations' SLs

could help locate new data for the study of ITr and pseudotranslation. In addition, further studies on these phenomena could provide new information on two specific types of interference: one, in which the interference is fake, as with pseudotranslations, and another where it is mixed, as with compilative ITrs.

References

Assis Rosa, A., Pięta, H. & Bueno Maia, R. (2017). Theoretical, methodological and terminological issues regarding indirect translation: An overview. *Translation Studies 10*(2), 113–132.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair* (pp. 233–250). Amsterdam: John Benjamins.

Baroni, M. & Bernardini S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing 21*(3), 259–274.

Čermák, F. & Rosen, A. (2012). The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics 13*(3), 411–427.

Corpus of Translated Finnish = Käännössuomen korpus. Käännössuomen sähköinen tutkimusaineisto. Käännössuomi ja kääntämisen universaalit -hankkeessa koostanut Joensuun yliopiston kansainvälisen viestinnän laitos 1997–.

Dollerup, C. (2000). Relay and support translations. In A. Chesterman, N. Gallardo San Salvador & Y. Gambier (Eds.), *Translation in Context* (pp. 7–26). Amsterdam: John Benjamins.

Du Pont, O. (2005). Robert Graves's Claudian novels: A case of pseudotranslation. *Target 17*(2), 327–347.

Eder, M., Rybicki, J. & Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. *R Journal 8*(1), 107–121.

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing 22*(3), 251–270.

Islam, Z. & Hoenen, A. (2013). Source and translation classification using most frequent words. *International Joint Conference on Natural Language Processing*, 1299–1305.

Ivaska, L. (forthcoming). The genesis of a compilative translation and its *de facto* source text.

Ivaska, L. (2018). Three methods to uncover the de facto source language(s) of translations. Poster presented at the European Summer University in Digital Humanities, Leipzig, Germany, 17–27 July.
Ivaska, L. (2016). Uncovering the many source texts of indirect translations: Indirect translations of Modern Greek prose literature into Finnish 1952–2004. Poster presented at the 8th European Society for Translation Studies Congress, Aarhus, Denmark, 15–17 September.
Ivaska, L. & Paloposki, O. (2018). Attitudes towards indirect translation in Finland and translators' strategies: compilative and collaborative translation. *Translation Studies 11*(1), 33–46.
Koppel, M. and Ordan, N. (2011). Translationese and its dialects. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, 1318–1326.
Lynch, G. & Vogel, C. (2012). Towards the automatic detection of the source language of a literary translation. *Proceedings of COLING 2012: Posters*, 775–784.
Mauranen, A. (2004). Corpora, universals and interference. In A. Mauranen & P. Kujamäki (Eds.), *Translation Universals: Do they exist?* (pp. 65–82). Amsterdam: John Benjamins.
Rabinonvich, E., Nisioi, S., Ordan, N. & Wintner, S. (2016). On the similarities between native, non-native and translated texts. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-2016), Berlin, Germany*, 1870-1881.
Straka, M. & Strakova. J. (2017). Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Association for Computational Linguistics, Vancouver, Canada*, 88–99.
Toury, G. (1995[/2012]). *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.
Ustaszewski, M. (2018). Tracing the effect of pivot languages in indirect translation. In S. Granger, M.-A. Lefer & L. Aguiar de Souza Penha Marion (Eds.), *Using Corpora in Contrastive and Translation Studies Conference, Louvain-la-Neuve, 12–14 September, 2018. CECL Papers 1* (pp. 174–176). Louvain-la-Neuve: Université catholique de Louvain.
Volansky, V., Ordan, N. & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities 30*(1), 98–118.

From bits and numbers to explanations – doing research on Internet-based big data

Veronika Laippala
University of Turku

Abstract
Internet is a constantly growing source of information that has already brought dramatic changes and possibilities to science. For instance, thanks to the billions of words available online, the quality of many natural language processing (NLP) systems, such as machine translation, has improved tremendously, and people's beliefs, cultural changes and entire nations' mindscapes can be explored on an unprecedented scale (see Tiedemann et al. 2016; Koplenig 2017; Lagus et al. 2018). Importantly, almost anyone can write on the Internet. Therefore, the web provides access to languages, language users, and communication settings that otherwise could not be studied (see Biber and Egbert 2018).

Paradoxically, the Internet's extreme size and diversity also complicate its use as research data. Many Internet-based language resources, such as English Corpus of Global Web-Based English (GloWbE) or the web-crawled Finnish Internet Parsebank developed by our research group, are composed of billions of words. Already searching from these databases requires specific tools, but especially the analysis of the search results may not be straightforward. For instance, the Finnish word *köyhä* 'poor' has 209 609 occurrences in the Finnish Parsebank, and its English correspondant has 312 974 hits in GloWbE. These language resources provide easily bits and numbers, but how to explain them?

In my talk, I will present some of the work we have done in our research group in order to bend Internet-based data collections for research questions in the humanities, where numeric results on frequencies are just the beginning of the analysis. In particular, I will discuss our newly-launched project on improving the usability of Internet-based big data, *A piece of news, an opinion or something else? Different texts and their automatic detection from the multilingual Internet*. In the project, the ultimate objective is to develop a system that could automatically detect different text varieties, or registers (Biber 1988), such as user manuals, news, and encyclopedia articles, from online data. Currently, for instance a Google search can return an overwhelming number of documents from mostly unknown origins and similarly, the origins of the documents in the web-crawled big data language collections are typically unknown. However, in order to explain research results

gotten from these collections, information on the kinds of texts included in the data would be very useful if not mandatory.

Identifying registers from the Internet involves a number of challenges. An essential prerequisite would be information on the registers to be detected. But what kinds of texts is the Internet composed of? A second concern, then, is that online texts do not follow the traditional print media boundaries (see Biber and Egbert 2018). For example, how can one distinguish texts that neutrally report scientific findings from those that use the information to persuade the reader? Additionally, text classification is typically based on manually labeled example documents representing the categories to be detected. However, developing this *training data* is very time-consuming and needs to be done separately for each language. Would it be possible to detect registers without all this manual work?

1 Introduction

The Internet is a constantly growing source of information that has already brought dramatic changes and possibilities to science. For instance, thanks to the billions of words available online, the quality of many natural language processing systems, such as machine translation, has improved tremendously, and people's beliefs and cultural changes can be explored on an unprecedented scale (see Tiedemann et al., 2016; Koplenig, 2017). Almost anyone can write on the Internet. Therefore, the web provides access to languages, language users, and communication settings that otherwise could not be studied (see Biber & Egbert, 2018).

Paradoxically, the Internet's extreme size and diversity also complicate its use as research data. Many Internet-based language resources, such as English Corpus of Global Web-Based English (GloWbE[1]), or the web-crawled Finnish Internet Parsebank,[2] developed by our research group, are composed of billions of words an millions of documents. Searching these databases requires specific tools, and the analysis of the search results may not be straightforward. These language resources can readily provide results in the form of bits and numbers, but how can they be explained?

In this paper, I will present some of the work I have done with our research group and with my collaborators in order to make Internet-based data collections

[1] https://corpus.byu.edu/glowbe/
[2] http://bionlp-www.utu.fi/dep_search/

usable for research questions in the humanities, where numeric results are just the beginning of the analysis.[3] In particular, I will discuss our newly launched project on analyzing and improving the usability of Internet-based big data: "A Piece of News, an Opinion or Something Else? Different Texts and Their Automatic Detection from the Multilingual Internet." In this project, the ultimate objective is to develop a system that could automatically detect from online data different *registers*, i.e. text varieties with specific situational characteristics and communicative purposes, such as user manuals, news, or encyclopedia articles (Biber, 1988)[4]. Currently, for instance, a Google search can return an overwhelming number of documents from mostly unknown origins, and similarly, the registers of the documents in the web-crawled big data language collections are typically unknown. However, ignoring information on the text registers can lead to incorrect conclusions. Therefore, knowing the registers and their characteristics would be very important for all studies aiming at explaining research results obtained from these collections. As a practical outcome, the project applies the developed system to detect registers from Universal Parsebanks (UP), a collection of web corpora we have compiled in our research group by automatically crawling the web. The project focuses on the French, English, and Swedish UP collections, as well as on the Finnish Internet Parsebank, which can be referred to as UP Finnish.

Identifying registers from the Internet involves, however, a number of challenges and an extensive linguistic analysis. An essential prerequisite for a register detection system would be information on the registers to be detected. But what kinds of texts is the Internet composed of? A second concern, then, is that online registers do not follow the traditional print media boundaries (see Biber & Egbert, 2018). For example, how can one distinguish texts that neutrally report scientific findings from those that use information to persuade the reader? Can these different registers be tracked to linguistic characteristics, and can they be identified automatically? Published print media has specific external indicators for register information. For instance, newspapers have sections for news articles and advertisements, and scientific findings are published in recognized journals. On the Internet, however, external indicators are less frequent and they cannot always be

[3] The work is done in collaboration with (in alphabetical order) Douglas Biber, Jesse Egbert, Filip Ginter, Roosa Kyllönen, and Aki-Juhani Kyröläinen.

[4] Following the research tradition established, e.g., in Biber (1988), I use the term 'register' instead of 'genre'. However, in the context of this study, the two notions are essentially interchangeable.

trusted. Therefore, the texts and their characteristics and communicative purposes need to be carefully examined before a detection system can be developed.

Finally, register detection is based on *supervised machine learning* (Manning et al., 2009) and manually labeled example documents representing the register categories to be detected. Our project aims at analyzing online registers in a number of languages: Finnish, English, French, and Swedish. Producing the labeled example documents—training data—is time consuming and needs to be done separately for each language. Thus, it would be useful to develop methods that could group texts to registers without all these manual steps. Would this be possible?

In this article, I present previous and ongoing work that aims at answering these questions. In Section 2, I start by discussing the first challenge: defining the range of registers found on the Internet and thus understanding the composition of the Internet. Furthermore, I present how online registers do not always follow register boundaries similarly to print registers. Second, in Section 3, I present preliminary results on automatically detecting registers from online corpora. Third, in Section 4, I reflect on the possibility of grouping together similar texts using unsupervised machine learning, without training data (Manning et al., 2009).

2 Defining the composition of the Internet

Knowing the text register tells the reader how to interpret it. For instance, phone books, user manuals, company websites, and online discussion forums have different communicative purposes and situational characteristics. Ignoring this information may lead to incorrect conclusions of the text. Therefore, information on the register of a text is essential for everybody reading online texts. Additionally, the linguistic properties of the texts can vary extremely across registers (Biber, 2012). Therefore, register information is also important for large-scale studies that do not necessarily aim at interpreting the texts they use as data, such as the development of part-of-speech taggers or spell checkers (see Zeman et al., 2017).

A first step toward a register detection system is to define the set of possible registers. This means that as a first step, we need to define the composition of the entire Internet in the project languages, i.e., Finnish, English, French, and Swedish. As the basis for this, we use the English online register taxonomy developed by Biber et al. (2015) for Corpus of Online Registers of English (CORE). CORE consists of nearly 50,000 documents and is based on a near-random sample of the English-speaking web. Importantly, CORE is coded for register categories using a

register taxonomy that is developed during the coding process. Therefore, it covers the full range of registers found online and offers a reliable basis for our annotation.

The CORE taxonomy is composed of eight main registers with functional labels, such as *narrative* or *interactive discussion,* and 33 sub-registers, such as *opinion blog* or *news article* (see Biber et al., 2015; Section 3 and Appendix 1). However, during the annotation, the authors found that some documents feature characteristics of more than one register and thus do not fit the taxonomy directly. For these documents, the four annotators' votes were split between several registers. As the combinations of categories were systematic and focused around specific register combinations, the authors concluded that these "hybrid" registers do not reflect a lack of agreement among the annotators but a characteristic of online language use where many of the documents "are not 'pure' instances of a particular register" (Biber & Egbert, 2018: 9).

Our objective is to manually register annotate random samples from the UP collections based on the CORE taxonomy. These samples will be used first to analyze the linguistic variation found in the collections, and then to develop the register detection system. To date, we have annotated 1,380 documents from the Finnish Internet Parsebank. Our annotation focuses on the sub-register level, where the labels are intuitive and correspond to register categories in other corpora (see list of registers in Appendix 2). Based on the documents we have annotated so far, it would seem that the register taxonomy developed for English is relatively easily applicable to Finnish data. During the annotation, we added two register categories that were not present in the original CORE taxonomy: community blog, to mark blogs held, for instance, by different sports associations or political movements, and machine-translated, or generated, texts to signal texts that are not written by humans and that can be excluded from further analysis.

Generally speaking, the annotation of the registers has proceeded smoothly without major problems. However, similarly to Biber et al. (2015) with CORE, the hybrid documents featuring characteristics of several registers needed particular attention. We do not have four annotators per document as Biber et al. (2015) had. Therefore, we have developed two alternative strategies to denote uncertainty and hybrid registers:

- When the document represents the characteristics of two register categories to the extent that just one register cannot be chosen, both categories are marked.

- When the document does not feature clear characteristics of any specific register, a functional label from the eight main register categories defined by Biber et al. (2015) is chosen.

By now, the most frequent documents featuring these hybrid characteristics in our data are combinations of different narrative sub-registers, such as news articles, and persuasive texts, such as description with intent to sell. Table 1 below illustrates these.

Table 1. Example of a hybrid text, originally published in http://jpchenet.fi/etusivu/30-vuotta-suosion-huipulla/.

J.P. Chenet 30 vuotta suosion huipulla!
Kaikki sai alkunsa 1980-luvun alussa erään miehen unelmasta. Joseph Helfrich, nuori ranskalainen viinikauppias, huomasi, etteivät ulkomaalaiset ymmärtäneet ranskalaisia viinejä. […] Joseph aloitti yhteistyön Jean- Paul Chanelin kanssa. […] Loppu onkin historiaa … Ensimmäisen J.P. Chenet -viinin syntymisestä tulee tänä vuonna kuluneeksi 30 vuotta! Juhlavuoden kunniaksi Suomessa esitellään innovatiiviset uutuusviinit sekä uudistuneita tuttuja klassikkoviinejä.
J.P. Chenet: More popular than ever after 30 years!
It all started in the beginning of the 1980s, when a young French wine merchant, Joseph Helfrich, had a dream. He noticed that foreigners did not understand French wines. Joseph started to collaborate with Jean-Paul Chanel. Together, they created easy-going wines that could be enjoyed right away, without conservation.… The rest is history…. This year marks the 30th anniversary of the J.P. Chenet wines! To mark the celebration, innovative new releases and renewed classics will be presented in Finland….

We have annotated the text in Table 1 as a hybrid, combining the registers historical article and description with intent to sell. The text begins with a historical narrative, telling how the J.P. Chenet wines were founded. Based on solely this information, the text could be annotated as historical article. However, the text is published on the J.P. Chenet company's official website. This already changes its communicative purpose, as the objective of a company is to sell their product. Similarly, we think that the ultimate objective of this text is, in addition to telling the story, to advertise the wines. Therefore, the text is annotated as a hybrid.

So far, the results of our annotation experiments have been relatively positive. After the initial stages of agreeing on how to interpret the register categories, even the identification of the hybrid registers has not caused major difficulties. As we have only just started the annotation, we do not yet know how well the registers

will be automatically detected based on these annotations. We have, however, made some initial detection experiments using the CORE data. We will discuss these in the next section.

3 Identifying online registers automatically

In the studies analyzing and developing web-crawled language resources, also called Web-as-Corpus research, the Internet has been described as a jungle because of the wide range of registers and language varieties found in it (Kilgarriff, 2001; Sharoff, 2008). In many of the previous studies developing register-detection systems, one of the main problems has been the lack of a corpus that would be sufficiently large to represent the jungle and thus the registers reliably. Consequently, the systems have targeted topics and other idiosyncrasies presented in the small corpora rather than actual registers (see Sharoff et al., 2010; Petrenz & Webber, 2011). Furthermore, many of the existing corpora do not represent the full range of registers found online; therefore, the results based on them cannot be applied to settings such as our project, where the objective is to classify the entire content of the Internet (e.g., Asheghi et al., 2016).

CORE presented in Section 2 is currently the largest corpus of online documents with manually added information on the register categories. As it is based on a random sample of the English web, it offers a useful scenario for experimenting on how well registers can be detected from the jungle, i.e., a database representing the full range of linguistic variation found on the Internet. In order to test this, we have done classification experiments on the 27 most frequent CORE (sub-) registers (see Egbert et al., 2018; Appendix 2). Altogether, these registers cover 25,178 documents. We have used as a text classification method a support-vector machine (SVM; Vapnik, 1998), a supervised machine-learning method that has been widely applied in previous register detection studies (Sharoff et al., 2010; Petrenz & Webber, 2011; Pritros & Stamatatos, 2018).

In the classification experiments, we used the linear SVM implemented in scikit-learn[5] and an 80/20% data division to train and test sets. We compared the predictive power of three feature sets: lexical information, i.e. words, grammatical information, and the combination of these two. The grammatical information was identified with the Biber tagger that is used extensively in corpus linguistic studies

[5] http://scikit-learn.org

on register characteristics (Biber, 1988; Biber & Egbert, 2015; 2018). The tagger identifies detailed grammatical information associated with the words, and altogether provides 784 different tags (see later this section for examples). Table 2 below presents the results of the experiments.

Table 2. Results from classification experiments on the 27 most frequent CORE registers.

Feature set	Precision	Recall
Combination of lexical and grammatical information	66%	63%
Lexical information	65%	62%
Grammatical information	59%	53%

Table 2 shows that the best classification results are achieved using a combination of lexical and grammatical tags. The classifier scores, 66% for precision and 63% for recall[6], on 27 register classes, are quite encouraging, and prove that registers can be automatically detected from jungle-like data. The results also clearly outperform the previous classification results on the CORE registers reported by Biber & Egbert (2015), with only 32.9% for precision and 41.5% for recall.

In addition to providing high-predictive accuracy, the advantage of SVM is that it estimates the most important features applied in the classification model. Thus, the method allows examining the basis of the classification and interpreting the model. This is very useful for all studies that aim to explain results beyond the numeric classification scores. Specifically, this involves two advantages. First, by analyzing the model and its most important features we can qualitatively evaluate to what extent it targets the intended categories. For instance, using this method, Sharoff et al. (2010) noted that their register-detection model depended on topics rather than registers. Second, the most important features estimated by SVM open up the possibility to examine the properties of the registers or other text classes to be detected. If the model is trained using features that can be interpreted, such as words or structural information, the most important features estimated by the classifier can be considered the linguistic characteristics of the register classes. This

[6] Precision is the fraction of relevant instances in the retrieved instances, and recall is the fraction of relevant retrieved instances among the total amount of relevant instances. F1-score is the balanced and harmonic mean of these two.

is particularly useful when the text classes require further analysis. For instance, in our project, characterizing the register classes and their composition is a crucial step in the analysis to ensure the validity of the classification.

A qualitative analysis of the most important features estimated by the best-performing classifier presented in Table 2 confirms that this model does target registers. Table 3 below presents the 10 most important features estimated for two registers, interviews and sports news.

Table 3. Most important features estimated by the best-performing SVM for two CORE registers.

Interviews	Sports News
nn=interview	nn=fight
wdt+who+whq=what	nn=penalty
wrb+who+whq=how	nn=injury
like=like	nn=win
rn+pl=there	nns=women's
ql+amp=very	np+=jason
vbd+dod+vrb=did	nn=season
vb=play	nn=game
nn=fashion	vprf+xvbnx=won
dt+pdem=that	np=howard

The most important features for interviews and sports news presented in Table 3 reflect various aspects of the two registers. For interviews, the majority of the features denote functional register characteristics, such as question words (*wdt+who-whq=*w*hat, wrb+who+whq=how*), past tense auxiliaries (*vbd+dod+vrb=did*), and amplifiers (*ql+amp=very*). For sports news, the most important features are different nouns (*nn* tag). Their analysis shows that they reflect mostly topical characteristics related to sports, such as *fight*, *penalty*, *injury*. All these properties are already associated with these registers in previous corpus linguistic studies characterizing them (see Biber & Egbert, 2015; 2018). Thus, they confirm that the classifier targets registers instead of topics or other idiosyncrasies. Additionally, these features provide a basis for a detailed linguistic analysis of the texts. In the project, this analysis step is essential prior to the final register detection.

4 Modeling registers without training data

The classification results we presented in the previous section confirmed that registers can be detected automatically. This is very encouraging for our project. However, preparing the example documents with manually coded information on the register they represent—training data—is very time consuming. Furthermore, in our project, this annotation process would need to be repeated for all the project languages. Therefore, it would be useful to develop alternative strategies that would allow to group texts according to registers without this manual work.

In corpus linguistics, Biber (1989) has already grouped similar texts together by using hierarchical clustering, an unsupervised machine-learning method to find previously unknown groupings in the data (see Kaufman & Rousseeuw, 1990). Furthermore, Biber & Egbert (2018) applied the method to CORE registers. The advantage of clustering is that as an unsupervised method, it does not require training data. On the other hand, as registers are defined based on their situational characteristics that do not necessarily show in the linguistic cues, clusters formed on the texts' linguistic similarities do not fully correspond to registers. Furthermore, the featurization of the texts can be difficult, because linguistic data is typically very sparse, where the data include a large number of unique words, of which many occur only in a small number of documents. Without any dimensionality reduction or feature selection, this can be very problematic.

Biber (1989) and Biber & Egbert (2018) solve the sparseness by using similarity metric scores obtained through multi-dimensional analysis (MDA; Berber-Sardinha & Veirano Pinto, 2018). MDA is a quantitative analysis method frequently applied in corpus linguistics to analyze register characteristics. MDA is based on exploratory factor analysis, a dimension reduction method that is used to reduce a large set of correlated variables to a smaller group of factors. In MDA, the correlated variables are typically grammatical forms present in the texts. These are reduced to factors or dimensions that can be functionally interpreted to define the linguistic characteristics associated with the texts.

Biber & Egbert (2018) showed that the clustering of online documents based on dimension scores and grammatical information produces functionally interpretable clusters, such as *oral* + *involved* or *literate* + *narrative.* Consequently, the method does provide a useful alternative for grouping texts in the UP collections. However, as the method builds on the dimension scores, it would require as a pre-processing step an MDA for each of the languages that we want to cluster.

Therefore, we have examined other techniques to reduce the data sparseness that would not depend on such language-specific, time-consuming steps. To this end, we have experimented with word2vecf, a machine-learning method that can be used to represent documents as fixed-size, dense vectors (Levy & Goldberg, 2014).

Word2vecf is an extension of the widely applied word2vec neural algorithm that allows to represent words as dense, predicted vectors, i.e., embeddings (Mikolov et al., 2013). The underlying idea of the method follows the distributional hypothesis stating that semantically similar words occur in similar contexts (Harris, 1954; Firth, 1957). Word embeddings are predicted based on the words the target word co-occurs with, and consequently, semantically similar words are described by nearby vectors. For instance, in the Finnish word2vec model trained on the Finnish Parsebank, the four nearest word vectors to *kissa* 'cat' are *kani* 'rabbit', *kissanpentu* 'kitten', *pentu* 'cub', and *marsu* 'guinea pig'[7]. Word2vecf extends the original word2vec by allowing arbitrary training contexts. Thus, instead of predicting similar words, we can predict similar documents. As contexts, instead of co-occurring words, we can use the text properties.

We have trained document vectors in both English and Finnish. As contexts for the documents, we used lemmas and syntactic information described with Universal Dependencies (UD), a framework for cross-linguistically consistent syntax annotation.[8] This allows us to model different languages as similarly as possible and guarantees that no manual work is needed to code the information. Although we do not yet have quantitative evaluation measures or clustering results for these data, a manual analysis of documents represented by nearby vectors proves that the word2vecf method captures the document's content in a meaningful manner.

To train the document vectors in English, the main source of documents was the aforementioned CORE (see Sections 2 and 3). Additionally, we complemented the dataset with documents from GloWbE.[9] Altogether, the data consisted of 120,000 documents. To generate the syntactic information, we tokenized the data with the UDPipe tool (Straka and Straková, 2017), ran an automatic syntactic analysis with the parser by Dozat et al. (2017), and turned the data into delexicalized syntactic biarcs, i.e., subtrees of dependency analyses with all

[7] http://bionlp-www.utu.fi/wv_demo/
[8] http://universaldependencies.org/
[9] https://corpus.byu.edu/glowbe/

information but the dependency relations deleted, with the tool by Kanerva et al. (2014; for example, see Laippala et al., 2018). We fixed the vector dimensionality to 300. This is thus the final number of variables used to represent the texts.

Table 4 presents the beginning of a CORE document and two other CORE documents that were assigned the nearest document vectors in our word2vecf model. All the documents were coded as sports reports.

Table 4. An example text from CORE and two other texts with nearby vectors. Words reflecting similar topics, such Premier League, player contracts, and financial issues are in bold.

Text 1:
Jelavic Calls for **Old Firm** to Come to the **Premier League**
New **Everton** signing Nikica Jelavic fears **Rangers** and **Celtic** will become second-class **clubs** if they remain in Scottish **football**. The Croatian **striker** moved to **Goodison** Park on **transfer deadline day**, with **Rangers'** manager Ally McCoist powerless to prevent the **6m move**. "**Rangers** and **Celtics** can't compete with **English clubs financially**, so it would be very important to them if they could join the **Premier League** one day."
http://www.sportpulse.net/content/everton-stay-fourth-they-fended-mackems-5370>
Text 2:
The **Premier League** took a big step toward introducing a break-even rule following a meeting in London. There was no formal **agreement** between the 20 chairmen and chief executives over how to introduce **costs controls**, but **clubs** agreed to focus on a model similar to the **Financial** Fair Play **regulations** introduced by UEFA, which require teams to avoid **making losses**.
http://www.talkceltic.net/forum/showthread.php?p=3208981
Text 3:
Didier Drogba has told **Chelsea** he does not want further talks on his future until the end of the season, leaving it ever more likely that the 34-year-old **striker**, who has been such a key figure in the **club's** two semi-final **wins** over the last six days, will leave for nothing in the summer. **Drogba** has **rejected** a one-year **deal** and there have been no **further talks** to resolve the situation […].
http://www.independent.co.uk/sport/football/premier-league/chelsea-left-hanging-by-indemand-and-outofcontract-didier-drogba-7661639.html>

In Table 4, all the three texts discuss sports news. In addition to this broad category, the texts share very specific topics, namely Premier League, player contracts, and financial issues. Thus, the document vectors seem to make sense and capture the contents of the text. Naturally, however, a quantitative evaluation of the vectors is

necessary before final conclusions on their usefulness can be drawn. We are currently investigating this.

5 Conclusion

In this article, I have presented some of the work I have done with our research group and with my collaborators to use large Internet data collections as material for research in the humanities. First, I discussed the different registers the Internet includes, and presented examples of texts that are typical of the Internet but do not fit traditional register categorizations presented in corpus linguistics. Understanding the register categories and knowing their linguistic characteristics would be essential for all studies using online data. I showed that although online registers question the traditional categorizations by combining characteristics of several registers, identifying these hybrid combinations is not necessarily difficult once the annotators are familiar with the particularities of online data.

Second, I presented preliminary text classification experiments on online registers from the CORE corpus. As CORE is based on a random sample of the English web and presents the full range of linguistic variation found online, it offers a useful test scenario for experimenting on to what extent registers can be automatically detected. The results were very encouraging, with 66% precision and 63% recall. Additionally, I presented a short qualitative analysis of the classifier model by examining the most important features it estimated for the register classes. This method allowed us to interpret the model and the register classes in more detail than what the numeric results on the classifier performance can tell. We were able to confirm that the model targets registers instead of specific corpus idiosyncrasies. Furthermore, the features associated with the registers by the classifier were very similar to their linguistic characteristics explored in previous studies.

Third, in the last section I examined the use of machine learning as a pre-processing step to reduce data sparseness. This would be very useful for grouping online documents to similar groups with clustering, an unsupervised method that does not require manually prepared example data. A qualitative analysis of the predicted embeddings representing the documents showed that the method grasps efficiently the document contents. Thus, this line of research is very promising.

To conclude, all three perspectives to online data I presented showed very encouraging outcomes. Specifically, we were able to demonstrate that online data can be successfully analyzed despite the range of linguistic variation and other challenges it features. Importantly, I showed how results based on online data can

be extended beyond numbers to explain and describe the results. This is very promising to all research combining large, textual datasets and research questions in the humanities.

References

Asheghi, N., Sharoff, S., & Markert, K. (2016). Crowdsourcing for web genre annotation. Language Resources and Evaluation, 50(3), 603–641.

Berber-Sardinha, T. B., & Pinto, M. V. (2019). Multi-Dimensional Analysis : Research Methods and Current Issues. London, UK: Bloomsbury Academic.

Biber, D. (1988). Variation across speech and writing. Cambridge: Cambridge University Press.

Biber, D. (1989). A typology of English texts. Linguistics, 27, 3-43.

Biber, D. (2012). Register as a predictor of linguistic variation. Corpus Linguistics and Linguistic Theory, 8, 9-37.

Biber, D., and J. Egbert 2015. Using grammatical features for automatic register identification in an unrestricted corpus of documents from the open web. Journal of Research Design and Statistics in Linguistics and Communication Science 2,3-36.

Biber, D., Egbert, J., & Davies, M. (2015). Exploring the Composition of the Searchable Web: A Corpus-based Taxonomy of Web Registers. Corpora, 10(1), 11-45.

Biber, D., & Egbert, J. (2018). Register variation online. Cambridge: Cambridge University Press.

Dozat, T., Qi, P., Peng, & Manning, C.M. (2017). Stanfords Graph -based Neural Dependency Parser at the CoNLL 2017 Shared Task. Proceedings of the CoNLL conference, 2017.

Egbert, J., Laippala, V., & Biber, D. (2018). Exploring online register variation using machine learning methods. Conference presentation at ICAME 39 - Corpus Linguistics and Changing Society. Tampere Hall Congress Centre 30 May - 3 June 2018.

Firth, J.R. (1957). A synopsis of linguistic theory 1930-1955. Studies in Linguistic Analysis. Oxford: Philological Society, 1–32. Reprinted in F.R. Palmer, ed. 1968. Selected Papers of J.R. Firth 1952-1959. London: Longman.

Harris, Z. (1954). Distributional structure. Word, 10(23), 146–162.

Kanerva, Jenna, Luotolahti, Juhani, Laippala, Veronika, and Ginter, Filip. 2014. Syntactic N-gram Collection from a Large-Scale Corpus of Internet Finnish. In Proceedings of the Sixth International Conference Baltic HLT.
Kaufman, L., & Rousseeuw, P.J. (1990). Finding groups in data: An introduction to cluster analysis. New York: John Wiley.
Kilgarriff, A. (2001). The web as corpus. Proceedings of Corpus Linguistics, Lancaster University.
Koplenig, A. (2017). The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets—Reconstructing the composition of the German corpus in times of WWII. Digital Scholarship in the Humanities, 32(1), 169–188.
Laippala, V., Kyröläinen, A-J., Kanerva, J., & Ginter, F. (2018 / Aop). Dependency profiles in the large-scale analysis of discourse connectives. Corpus linguistics and linguistic theory.
Levy, O., & Goldberg, Y. (2014). Dependency-based word embeddings. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 203-208.
Linzen, T., Chrupała, G., & Alishah, A. (2018). Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Association for Computational Linguistics.
Manning, C.M., Raghavan, P., Schütze, H. (2009). An Introduction to Information Retrieval. Cambridge: Cambridge University Press.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26, 3111–3119.
Petrenz, P., & Webber, B. (2011). Stable classification of text genres. Computational Linguistics, 37(2), 385-393.
Pritsos, D., & Stamatatos, E. (2018). Open Set Evaluation in Web Genre Identification. Language Resources and Evaluation, 52(4), 949–968.
Sharoff, S. (2008). In the garden and in the jungle: comparing genres in the BNC and Internet. Genres on the Web, 149-166.
Sharoff, S., Wu, Z., & Markert, K. (2010). The web library of babel: evaluating genre collections. Proceedings of the Seventh Conference on International Language Resources and Evaluation, 3063–3070.
Straka, M., & Straková, J. (2017). Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. Proceedings of the CoNLL 2017 Shared Task:

Multilingual Parsing from Raw Text to Universal Dependencies, Vancouver, Canada, August 2017.
Tiedemann, J., Cap, F., Kanerva, J., Ginter, F., Stymne, S., Östling, R., &Weller-Di Marco, M. (2016). Phrase-based SMT for Finnish with more data, better models and alternative alignment and translation tools. Proceedings of the First Conference on Machine Translation, Volume 2: Shared Task Papers. Berlin, Germany. Association for Computational Linguistics, 391–398.
Vapnik, Vladimir N. (1998). Statistical learning theory. New York: Wiley Interscience.
Zeman, D., M. Popel, M. Straka, J. Hajic, J. Nivre, F. Ginter, J. Luotolahti, S. Pyysalo, S. Petrov, M. Potthast, F.M. Tyers, E. Badmaeva, M. Gokirmak, A. Nedoluzhko, S. Cinková, J. Hajic Jr., J. Hlavácová, V. Kettnerová, Z. Uresová, J. Kanerva, S. Ojala, A. Missilä, C.D. Manning, S. Schuster, S. Reddy, D. Taji, N. Habash, H. Leung, M.-C. de Marneffe, M. Sanguinetti, M. Simi, H. Kanayama, V. de Paiva, K. Droganova, H. Martínez Alonso, Ç. Çöltekin, U. Sulubacak, H. Uszkoreit, V. Macketanz, A. Burchardt, K. Harris, K. Marheinecke, G. Rehm, T. Kayadelen, M. Attia, A. El-Kahky, Z. Yu, E. Pitler, S. Lertpradit, M. Mandl, J. Kirchner, H. Fernandez Alcalde, J. Strnadová, E. Banerjee, R. Manurung, A. Stella, A. Shimada, S. Kwak, G. Mendonça, T. Lando, R. Nitisaroj, J. Li (2017). CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. CoNLL Shared Task 2, 1–19.

Appendix 1

Registers and subregisters used in the Finnish register annotation.

Narrative
News reports/News blogs
Sorts reports
Personal blog
Historical article
Short story / Fiction
Travel blog
Community blog
Online article
Informational Description
Description of a thing
Encyclopedia articles
Research articles
Description of a person
Information blogs
FAQs
Course materials
Legal terms / conditions
Report
Opinion
Reviews
Personal opinion blogs
Religious blogs/sermons
Advice
Interactive discussion
Discussion forums
Question-Answer forums
How-to/instructional
How-to/instructions
Recipes
Informational persuasion
Description with intent to sell
News+Opinion blogs/Editorials

Lyrical
Songs
Poems
Spoken
Interviews
Formal speeches
TV transcripts
Other
Machine-translated/generated texts

Appendix 2

CORE subregisters used in the experiment:

Advice
Discussion forum
Description of a person
Description with intent to sell
Description of a thing
Encyclopedia article
FAQ about information
Formal speech
Historical article
How-to
Informational blog
Interview
News article / news blog
Opinion blog
Poem
Personal blog
Question / answer
Research article
Recipe
Religious blog / sermon
Review
Song lyrics
Sports report
Short story
Travel blog
Tv subscripts

The extent of similarity: comparing texts by their frequency lists

Mikhail Mikhailov
Tampere University

Abstract
Measuring distances between texts can be useful in document search, automated classification, detecting plagiarism, etc. (Hoad, T. C. and Zobel, J., 2003, Gomaa & Fahmy 2013). One of the possible ways to do it is to compare frequency lists of the texts (Kilgarriff 1997, Piperski 2018). In this paper, two methods of such comparison are discussed: the first one is based on top X items from normalized frequency lists and the second compares frequency lists of N random samples of X running words per text.

The research data were literary texts in Russian and in Finnish, original texts and translations. The data included different authors, different genres and translations by different translators. During the first experiment the top 1000 words from lemmatized frequency lists were compared. The frequency lists were merged into a single data frame, a distance matrix was calculated, and finally the cluster analysis was run on the matrix. For the second experiment, 50 random samples of 3000 running words were drawn from each text and the frequency lists of these samples were processed in the similar way as in the first experiment.

The results for both methods were quite satisfactory. The texts by the same authors and texts devoted to the same topics were often clustered together. All retranslations of the same texts were clustered to the same groups. However, the time period, the translator and the source language did not seem to influence the result much. The terminal clusters were more interesting than the upper-level classification. Obviously, a human would have defined the larger groups in a different way. At the same time, this computer-made classification might bring new insights for the researchers.

1 Introduction

"How similar are these two texts?" This question arises in different situations: when we wish to group texts by genre or topic, when we look for other texts on similar issues, when we desire different versions of the same text, and, last but not least,

when we suspect plagiarism. In each case, it would be very useful to be able to use technology to check large amounts of data without having to read the texts.

In corpus linguistics, the automated comparison of large numbers of texts has become very important for many reasons. Any large corpus may contain duplicate texts: the same text copied twice by accident or published by different publishers, as well as different versions of the same text (e.g. the same news distributed by different news agencies). Although a corpus will contain similar texts (e.g. many weather forecasts by different news companies in different seasons in different places), the duplicate use of a text (such as a BBC weather forecast for Great Britain in Christmas 2018) must be avoided. Another important issue for corpus studies is the grouping of texts into subcorpora. The automated clustering of texts of similar style and vocabulary can validate grouping by external criteria (author, place, time, topic, purpose), provide an additional check on the balancedness and proportionality of the data, and give additional insights into the variation of language. Thus, comparing texts can improve the quality of the data and open new functionalities for researchers.

Earlier studies in text comparison were connected with plagiarism detection (Lyon et al 2004, Hoad and Zobel 2003, Chong et al 2010, Abdel-Hamid et al 2009), developing language technologies (Islam et al 2012, Gomaa and Fahmy 2013) and the classification of texts in corpora (Kilgarriff 1997 and 2001, Piperski 2017 and 2018). There exist two approaches: comparing whole texts and comparing frequency lists compiled from texts.

The automated comparison of texts is not a straightforward task unless we are looking for a perfect match. Any pair of documents written in the same language will have a certain degree of similarity. It may seem that the easiest way to check the extent of this similarity is to compare texts sentence by sentence. However, this method is neither the most effective nor the fastest, requiring sophisticated techniques of fuzzy matching. Although the first draft of a document and the same document after heavy editing might look very different, any human reader would notice that they are two versions of the same text. Moreover, one can imagine texts on the same topic without matching sentences as well as texts on different topics with similar sentences. There exist many ways and approaches for evaluating text similarity by comparing whole texts (see e.g. Abdel-Hamid et al 2009, Chong et al 2010, Gomaa and Fahmy 2013, Islam et al 2012 and Lyon et al 2004).

An alternative approach is to compute the similarity of texts by comparing not whole texts but individual words, the very bricks of which texts are made.[1] Many researchers develop methods of measuring the distance between texts based on comparison of their frequency lists (Kilgarriff 2001, Piperski 2018). It is obvious that the more similar texts are, the closer the resemblance will be between their frequency lists. Identical texts would produce identical frequency lists, while frequency lists of entirely different texts would lack any matching items. Various methods are available for distance measurements depending on the purpose of the comparison: chi-square testing, mutual information; Euclidean, Manhattan, and Canberra distance, and so on. Measurements can be performed on either complete or truncated frequency lists. For details, see for example Kilgarriff 1997, 2001, and 2009.

Corpus linguistics researchers are mostly attracted to informative texts with relative homogeneity of size, composition, structure, and vocabulary. In this paper, I shall try to examine whether comparing frequency lists is a fruitful approach with literary texts. The problem of literary texts lies in their non-homogeneity: they can differ in size, consist of different types of microtexts (narration, argumentation, description, etc.), and may contain lexis from different registers. Even texts written by the same author may vary considerably across different topics, purposes, and time periods. Therefore, methods effective for documents and articles might perform poorly with literary works.

The research data for this study will include both original literary texts and literary translations. Fiction texts vary by author, genre, or topic. Translations are especially interesting from the point of view of similarity, as there may be multiple of the same work. There can be translations of the works of one author performed by the same or different translators, or translations of works by different authors performed by the same translator. Archaic language may distinguish very old translations from those recently published.

Are all such issues visible in frequency lists? What factors make frequency lists more similar to each other? Another question that arises when working with language data is whether the suggested methods are equally effective for different languages. This paper will deal with data in two strongly dissimilar languages:

[1] It is also possible to check units smaller than words: some researchers have developed methods of comparing texts based on letter N-grams (see e.g. Islam et al 2012, Piperski 2017).

Russian and Finnish. The research data will include original works of fiction in Russian and literary translations into both Russian and Finnish.

2 Observing frequency lists

A text in any natural language is composed of words (or rather, *tokens*: strings of characters delimited by spaces and punctuation marks), some of which occurring many times and others only once (so-called *hapax legomena*) or twice (*hapax dislegomena*). The number of tokens in a text is usually almost twice the number of different tokens (*types*), and a lemmatized frequency list (with inflected forms of the same lexemes grouped under single headwords) – especially for languages with rich morphologies – will be many times shorter than the text. One corollary of Zipf's law is that, in any frequency list, most types are *hapax legomena* and *hapax dislegomena* (see e.g. Mikhailov and Cooper 2016: 9-10). As a result, at the tops of frequency lists drawn from texts in the same language, the same high-frequency words will occur, while at the bottoms, there will be almost no overlaps. This phenomenon is easy to demonstrate in retranslations of the same literary work: in Table 1, only four words (in bold) do not occur in all three top-20 frequency lists of three translations of Astrid Lindgren's *Lillebror och Karlsson på taket* from Swedish into Russian (by Lilija Lungina, Ljudmila Braude, and Eduard Uspenski respectively). Other words differ only in rank position and frequency.

Table 1. Top 20 words of the lemmatized word lists[2] from three different Russian translations of the same source text.

Lungina, 1957			Braude, 1997			Uspenski, 2008		
Lemma	F	ipt[3]	Lemma	F	ipt	Lemma	F	ipt
i	935	41.59	i	785	38.23	i	931	41.33
malyš	608	27.05	on	631	30.73	on	703	31.21
on	551	24.51	malyš	493	24.01	karlson	487	21.62
karlson	488	21.71	karlsson	446	21.72	v	468	20.77
v	465	20.68	ne	443	21.57	ne	452	20.06
ne	450	20.02	v	431	20.99	â	442	19.62
â	426	18.95	â	410	19.96	byt'	398	17.67
na	379	16.86	na	388	18.89	malyš	380	16.87
čto	369	16.41	čto	286	13.93	skazat'	380	16.87
byt'	336	14.95	byt'	285	13.88	čto	373	16.56
ty	323	14.37	ty	270	13.15	na	346	15.36
skazat'	265	11.79	s	233	11.35	ty	292	12.96
s	225	10.01	a	225	10.96	èto	282	12.52
no	194	8.63	skazat'	213	10.37	s	218	9.68
a	188	8.36	oni	188	9.15	oni	197	8.74
mama	187	8.32	no	176	8.57	no	189	8.39
èto	182	8.10	mama	172	8.38	že	174	7.72
oni	164	7.30	èto	167	8.13	mama	174	7.72
u	158	7.03	u	156	7.60	tak	167	7.41

In Table 2, which displays the last 20 words from the frequency lists of the same translations, the opposite occurs: only three words (in bold) occur in all three lists, while the existence of any overlap can be explained by all three texts' being translations of the same source text. Obviously, high- and low-frequency words do not affect the similarity of texts, but rather words of medium frequency.

[2] Russian words are transliterated using ISO-9 standard.
3 ipt = items per thousand (Frequency / Size of the corpus X 1000).

Table 2. The last 20 words of the lemmatized word lists from three different Russian translations of the same source text.

Lungina, 1957			Braude, 1997			Uspenski, 2008		
Lemma	F	ipt	Lemma	F	ipt	Lemma	F	ipt
šuršanie	1	0.04	šmel'	1	0.05	šuher	1	0.04
šutit'	1	0.04	šnyrât'	1	0.05	è-gej	1	0.04
šutka	1	0.04	štany	1	0.05	èkzemplâr	1	0.04
šutnik	1	0.04	štora	1	0.05	èlegantno	1	0.04
šenka	1	0.04	šuršat'	1	0.05	èlegantnyj	1	0.04
šenka-pudelâ	1	0.04	šutit'	1	0.05	èmigrant	1	0.04
šenoček	1	0.04	šš-šš	1	0.05	èrovyj	1	0.04
šečka	1	0.04	šel'	1	0.05	èskil'stuna	1	0.04
èkonomka	1	0.04	šečka	1	0.05	èstetičnyj	1	0.04
èlektričeskij	1	0.04	è	1	0.05	ûnyj	1	0.04
èskil'stun	1	0.04	èlegantno	1	0.05	ûrknut'	1	0.04
ûnyj	1	0.04	èskil'stun	1	0.05	âbedničat'	1	0.04
ûrknut'	1	0.04	èstermal'm	1	0.05	âvit'sâ	1	0.04
âvstvennyj	1	0.04	ûrknut'	1	0.05	âvlât'	1	0.04
âzvitel'no	1	0.04	âbeda	1	0.05	âzvitel'no	1	0.04
âzyk	1	0.04	âičnica	1	0.05	âzyk	1	0.04
âzyčok	1	0.04	âmočka	1	0.05	âjco	1	0.04
âmočka	1	0.04	ârkij	1	0.05	âmočka	1	0.04
ârostno	1	0.04	âstreb	1	0.05	âstreb	1	0.04
âstreb	1	0.04	â-to	1	0.05	âšiček	1	0.04

Related texts, that is, different versions of the same document (e.g. drafts and the final version), different renderings of the same story (e.g. *Cinderella* rendered by both Perrault and by the Brothers Grimm), and retranslations of the same work (e.g. six different Finnish translations of Dostoyevsky's *Brothers Karamazov*) will show similar frequency lists. Texts devoted to similar matters – politics, philosophy, nature, education – should also yield frequency lists with some frequencies in common. It is more difficult to say whether frequency lists reflect the individual style of a writer or translator.

In corpus linguistics, two types of frequency word list are used: unlemmatized and lemmatized. In unlemmatized or types lists, different forms of the same lexemes are registered as separate items. In lemmatized lists, different forms of the same lexemes are merged and, as a result, the list better reflects vocabulary (see e.g. Mikhailov and Cooper 2016: 51-54). Lemmatized lists are also shorter, especially for languages with rich morphologies. On the other hand, certain grammatical forms can dominate unlemmatized lists. At the same time, lemmatization performed by modern parsers is of low quality, with lemmatized lists containing numerous mistakes. Thus, it is difficult to determine whether lemmatized or unlemmatized lists are more useful for comparing texts.

3 The research data

As already mentioned, this study was performed on Russian and Finnish literary texts – both originals and translations from other languages. As it was necessary for my case study to have heterogeneous data, common-sense criteria for selection could include the following:

- long and short texts
- texts of different genres and topics
- original texts by the same author
- original texts by different authors
- translations from different source languages
- translations of works by one author performed by the same translator
- translations of works by one author performed by different translators
- translations of works by different authors performed by the same translator
- multiple translations of the same works by different translators.

The texts selected for this pilot study, which provide a modest amount of data, fulfil most of these criteria. Three subcorpora were formed: Russian originals (RuOrig) comprising 32 texts of 2.1 million running words, Russian translated data (RuTr) comprising 38 texts of 2.3 million running words, and Finnish translated data (FiTr) comprising 44 texts of 2 million running words. The texts belong to different genres and chronological period, and had different translator. RuTr is composed of translations from English, French, and Swedish. FiTr consists of only translations from Russian into Finnish. Retranslations, and especially multiple translations, received particular attention. Examples include five translations into Finnish of

Nikolai Gogol's *Taras Bulba* and six translations into Russian of George Orwell's *Animal Farm*.

4 The experiments

I carried out a series of experiments with cluster analysis on unlemmatized and lemmatized frequency lists. These lists were drawn from the corpus databases using original scripts of my own written in PHP programming language. The lists were then processed using R Studio (packages cluster, data.table, party, smacof).

The frequency lists of texts of each subcorpus were loaded into R and merged into a single data frame (for this purpose, a simple script in R was written; the merging of data frames via console commands is also possible, but is time consuming). A merge with the outer join was then performed (any word that occurred in at least one list was preserved in the merged table). Thus, the combined table included a complete lemmatized word list of test data as row headers, the codes of all texts as column headers, and the IPT (items per thousand) frequencies of each word in each text as table cells. To make the table suitable for calculating distance measures, it was transposed (columns became rows and rows became columns). The next steps were to calculate and run cluster analysis on the distance matrix.

Initially, I worked with complete frequency lists for whole texts. However, in spite of normalized frequencies enabling the comparison of frequencies in texts of different lengths, this method yielded unsatisfactory results. The main reason for this was the great variation in list length. For example, Dostoevsky's novel *Crime and Punishment* contains 170,000 running words, resulting in a lemmatized frequency word list of 13,500 items. In contrast, Gogol's novella *Overcoat* contains 10,000 running words; its lemmatized frequency list has only 2500 items. Even if all the items from the *Overcoat* list match well with those of the *Crime and Punishment* list, over 10,000 items would remain unmatched. As a result, the cluster analysis run on the complete frequency lists produced no reasonable patterns. Explainably and as expected, only retranslations were clustered together in the great majority of cases.

To minimize the effect of differences in text length, I tested two methods, both of which producing positive results. The first was to take the first *n* items from the lists, sorted by frequency in descending order with normalized frequencies. Although this allowed comparison, it did not completely overcome the problem,

because even the top of the frequency list of a very long text will be more diverse. The second method was to draw random samples from the texts. Although using one sample would be too arbitrary, especially with fiction texts, drawing multiple samples would do the trick, even if some samples for shorter texts would overlap.

4.1 Comparing the top 1000 words

What should the length of a truncated list be to provide reliable results? Very short lists emphasize high frequency words. Very long lists emphasize words making thus all texts different and, furthermore, precluding the comparison of short texts. It might seem a good idea to draw *n* words from the middle of the list, but how can we find the same mid-range for different texts? After some consideration, I decided to try the top 1000 words. For this purpose, each frequency list was truncated prior to the merging procedure described in the previous section.

For cluster analysis, I tried both the Euclidean and Manhattan distance measures, the latter of which producing better results. This can be explained by the tendency of the Euclidean distance measure to emphasize words with higher frequencies; as has already been mentioned above, high-frequency words are of little importance for comparing texts. I undertook clustering on both unlemmatized and lemmatized word lists. Although, as expected, the comparison performed best on lemmatized word lists, the clustering of unlemmatized lists also produced acceptable results.

Results are displayed in the Appendix (Figs. 1-3). When clustering original texts, works by the same author were usually, though not always, clustered together. For example, although Pushkin's novellas form a single cluster, his novel *The Captain's Daughter* is grouped with works by Tolstoy and Lermontov (Fig. 1). Obviously, works by the same author belong to the same cluster if they also belong to the same genre and were written in the same period, as for example detective stories by Marinina and novels by Trifonov (Fig. 1). In general, the terminal clusters of the dendrograms are more interesting than the upper-level groupings. Time period does not seem to influence clustering at all, with nineteenth- and twentieth-century works often seen in the same clusters. Obviously, the key issue is the similarity of the topics and of the literary school tradition, as a kind of "apprenticeship".

As expected, the experiments with the translated data show that retranslations of the same works are the "closest" texts in both the Russian and the Finnish data;

all were clustered together without a single mistake (Figs. 2 and 3). Translations of works by the same authors were also grouped together successfully with few exceptions. As for the individual style of translators, it is evidently a less important factor: the multiple retranslations of Astrid Lindgren's stories by different translators are grouped first by story and then by author (Fig. 3). Although the translations of three Lindgren stories by Nora Gal are indeed clustered together, her translations of the works of other authors are clustered separately.

Interestingly, the results of clustering the original works in RuOrig and their translations in FinTr often match quite well. For example, works by Solzhenitsyn, Belov and Baklanov inhabit the same cluster in both RuOrig and FinTr. On the other hand, some relations found between originals do not persist in translations; for example, Erofeev and Aksenov inhabit the same cluster in RuOrig but not in FinTr (Fig. 2).

4.2 Comparing 3000-word samples

The alternative method of comparing texts of different lengths is to draw random samples. The size of a sample should be large enough to allow for grasping the specific features of the text. At the same time, very large samples might incorporate differing types of text. Moreover, it would be impossible to measure the distance between short texts. For this study, I set the size of the sample to 3000 words, having already tried smaller and larger samples. As it has already been mentioned, one sample is unable to cover the potential diversity of a very long text. The way to overcome this problem is to draw multiple samples. Evidently, it is possible to draw samples continuously, updating the distances until the values stabilize. In this particular case, I decided to simply set the number of samples to 50.

A PHP script was used to generate frequency lists for fifty random samples taken from each text of a subcorpus. These lists were then processed with an R-script that loaded data into fifty data frames, calculated fifty distance matrices on their basis, and subsequently created a distance matrix with means of fifty values for each cell. Finally, the cluster analysis was run on the resulting distance matrix. As in the previous case, I used the Manhattan distance measure. In contrast with the previous experiment, the frequency lists were not lemmatized: for short extracts, lemmatization would have affected the results too strongly. Results are shown in Figures 4-6.

The effectiveness of this method is mainly confirmed by all retranslations clustering together. The only exception (which also proves the rule) is that although all translations of Astrid Lindgren's stories into Russian form one large cluster, only the first two translations are clustered pairwise by book. Conversely, the translations of all three books by Eduard Uspenski form a distinct cluster together (see Fig. 6).

In many respects, the results of clustering are similar to those of the previous experiment. Works by the same authors were often clustered together, especially within the same genre and when written in the same time period, as with two detective stories by Marinina and three works by Dostoevsky, respectively. Three novellas by Pushkin are in the same cluster; however, the novel *The Captain's Daughter* and the novella *The Shot* occupy a separate cluster together with Lermontov's *Hero of Our Time*. All three texts were written in the same time span (1830-1840) as well as deal with such topics as the lives of aristocrats, travel, love, duels, and war. In the same way, Gogol's *Overcoat* is closer to Dostoevsky (*Crime and Punishment*, *Notes from the Underground*, *The Brothers Karamazov*) than to other works by Gogol (*Dead Souls*, *Taras Bulba*). The reason must be the same: *The Overcoat* is closer by topic to Dostoevsky's works than to *Dead Souls* and *Taras Bulba*. As in the previous experiment, retranslations of the same works are the first to appear in the subcorpora of translations. The next most important factor is the author: texts written by the same author are often in the same terminal cluster. However, as with the clustering of the original texts, works by the same authors in different genres or covering different topics might belong to different clusters, as with George Orwell's *Animal Farm* and *1984* or Saint Exupery's *Little Prince* and *Wind, Sand and Stars*. Interestingly, the Finnish translation of the works of Gogol and of Pushkin all clustered together successfully. Could this mean that authorial style differs less in translation than in original works?

5 Discussion

I could not decide which of the two methods was preferable. Although similar in many respects, results were not identical. Each method shows distinct strengths and weaknesses.

The top 1000 words method is easy to use, with a short processing time. At the same time, taking only the top of the frequency list may exclude words of the same frequency band, which can affect the comparison of some pairs of texts.

The 3000-word samples method is more difficult to use, and processing multiple lists requires more time. Working with samples may produce different results for each run of the programme.

The results of testing both methods of comparing texts are satisfactory. Each method makes it possible to find similar texts even when their lengths vary. It is important to note that these methods are not suitable for comparing very short texts. Each method is language independent.

In most cases, the terminal groups were reasonable. However, high-level clustering did not yield a usable classification: the human classification performed by a literary scholar would have been entirely different. Nonetheless, such automated clustering may provide interesting insights for literary scholars. For example, what have the works of Venedikt Erofeev and Dostoevsky – or of Tolstoy, Bulgakov, Troepolski and Trifonov – in common apart from their lexicon? Such unexpected results might still be caused by the small amount of research data: with more texts compared, there would be fewer surprises, and data would be clustered by author and topic.

Each method may be useful for both corpus linguistics and translation studies. Corpus linguists can check whether external criteria really work and the corpus is balanced; the researchers of translation can measure such distances as those between translations of works of the same author, between translations performed by the same translator, and between retranslations.

References

Abdel-Hamid, O., Behzadi, B., Christoph, S., and Henzinger, M., 2009. Detecting the Origin of Text Segments Efficiently. In *Proceedings of the 18th International Conference on World Wide Web*, 61–70, Madrid.

Chong, M., Specia, L., and Mitkov, R., 2010. Using Natural Language Processing for Automatic Detection of Plagiarism. In *Proceedings of the 4th International Plagiarism Conference*. Newcastle upon Tyne, UK.

Gomaa, W.H. and Fahmy, A.A., 2013. A Survey of Text Similarity Approaches. *International Journal of Computer Applications* (0975 –8887). Volume 68–No.13. <https://research.ijcaonline.org/volume68/number13/pxc3887118.pdf>

Hoad, T. C. and Zobel, J., 2003. Methods for Identifying Versioned and Plagiarized Documents. *Journal of the American Society of Information Science and Technology*, 54(3): 203–215.

Islam, A., Milios, E., and Keselj, V., 2012. Text Similarity Using Google Tri-Grams. In Proceedings of the 25th Canadian Conference on Artificial Intelligence, pages 312–317, Toronto, Canada.

Kilgarriff A., 1997. Using word frequency lists to measure corpus homogeneity and similarity between corpora. In *Information Technology Research Institute Technical Report Series.* 97-07. <http://aclweb.org/anthology/W97-0122>.

Kilgarriff A., 2001. Comparing corpora. *International Journal of Corpus Linguistics.* 6(1), pp. 97–133. <https://www.sketchengine.eu/wp-content/uploads/comparing_corpora_ 2001.pdf>

Kilgarriff A., 2009. Simple maths for keywords. In *Proceedings of Corpus Linguistics Conference CL2009*, University of Liverpool. <https://www.sketchengine.eu/wp-content/uploads/2015/04/2009-Simple-maths-for-keywords.pdf>.

Lyon, C., Barrett, R., and Malcolm, J., 2004. A Theoretical Basis to the Automated Detection of Copying Between Texts, and its Practical Implementation in the Ferret Plagiarism and Collusion Detector. In *Proceedings of the Conference on Plagiarism: Prevention, Practice and Policies*. Newcastle upon Tyne, UK.

Mikhailov M. and Cooper R., 2016. *Corpus linguistics for Translation and Contrastive Studies. A guide for research.* Routledge: London and New York.

Piperski, A. Ch., 2018. Corpus size and the robustness of measures of corpus distance. In *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018"*. Moscow, May 30—June 2, 2018. < http://www.dialog-21.ru/media/4327/piperskiach.pdf>

Piperski A. Ch., 2017. Sravnenie korpusov meroj χ^2: simvoly, slova, lemmy ili časterečnye pomety? [Comparing corpora with χ^2: characters, words, lemmata, or PoS tags?], In *Korpusnaja lingvistika–2017* [Corpus Linguistics–2017]. Saint Petersburg, Saint Petersburg State University, pp. 282–286.

Appendix. Results of the cluster analysis

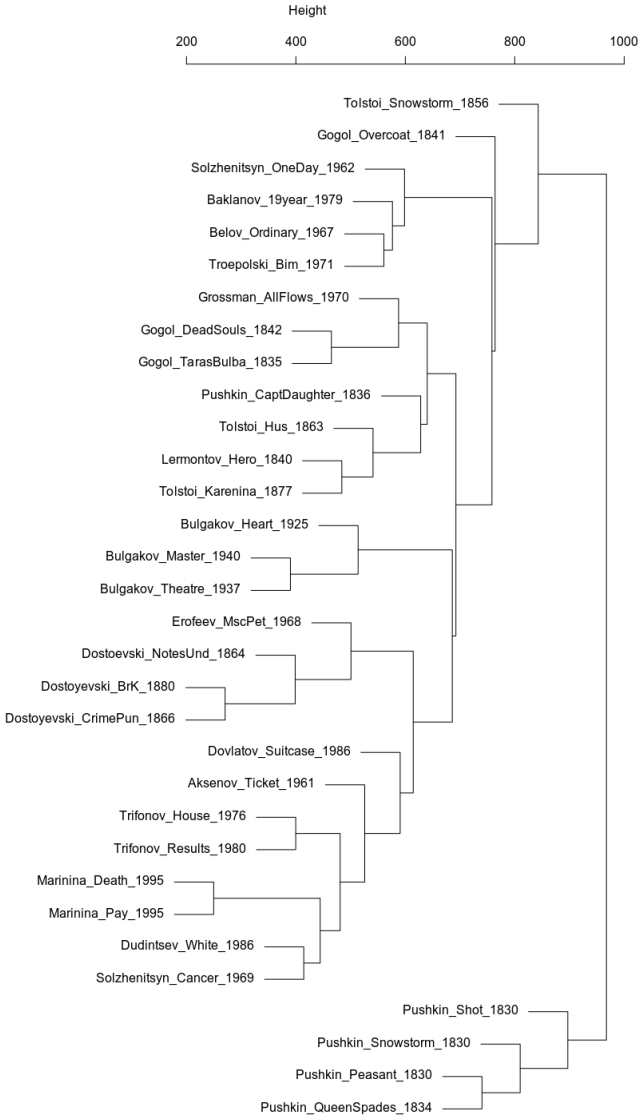


Fig. 1. Whole texts, RuOrig.

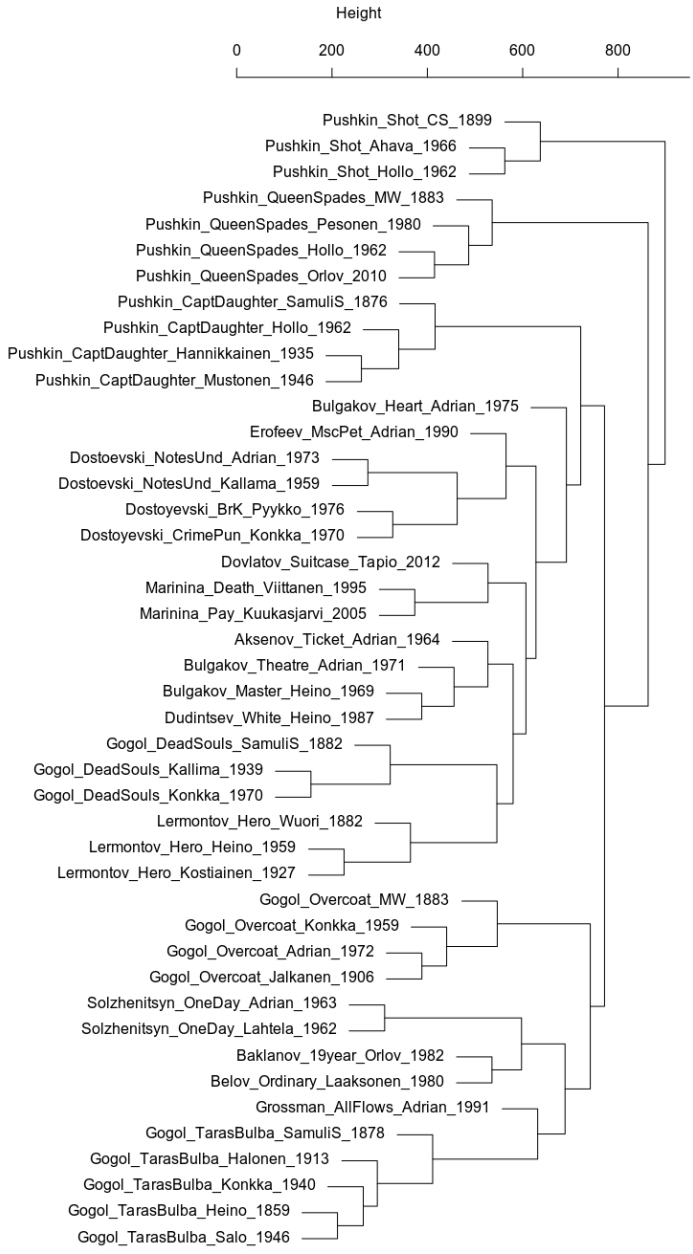


Fig. 2. Whole texts, FiTr.

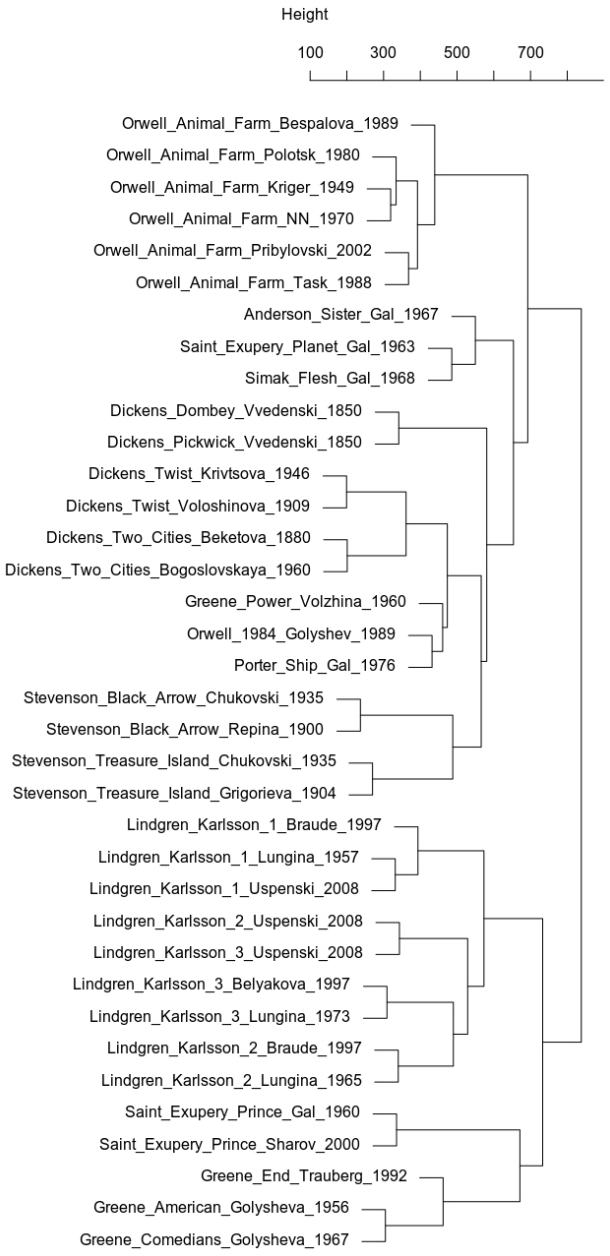


Fig. 3. Whole texts, RuTr.

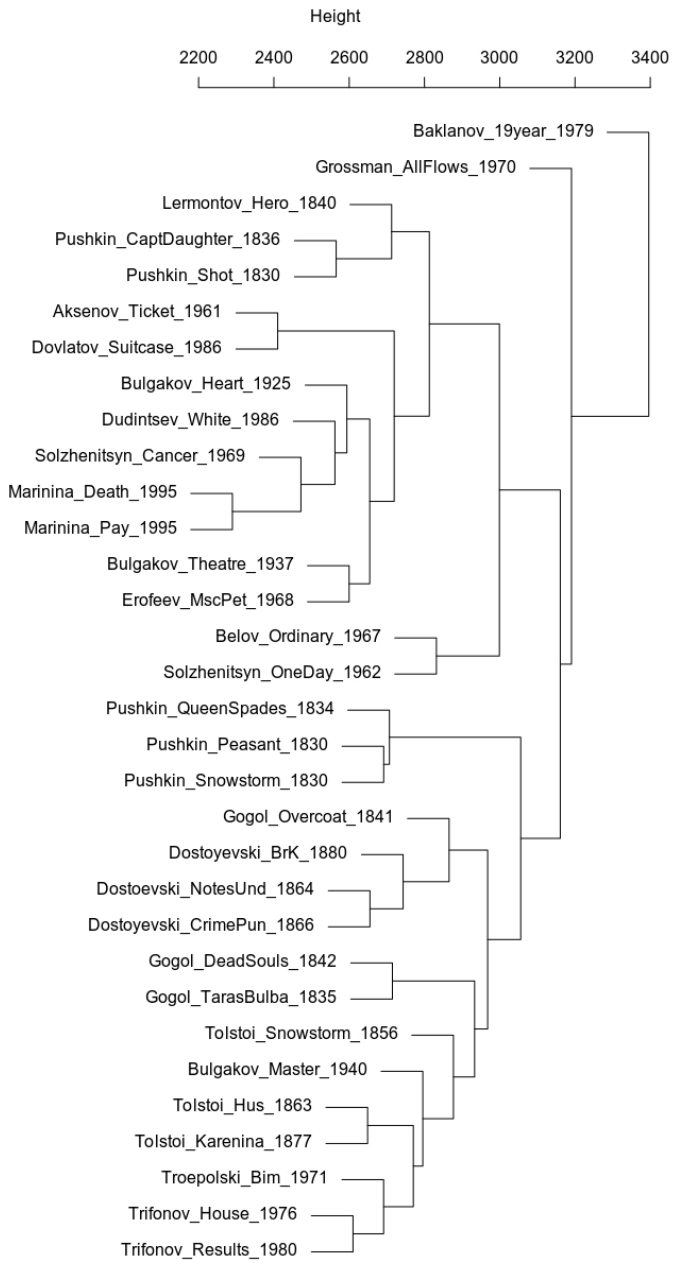


Fig. 4. Samples 3000 tokens (50 samples), RuOrig.

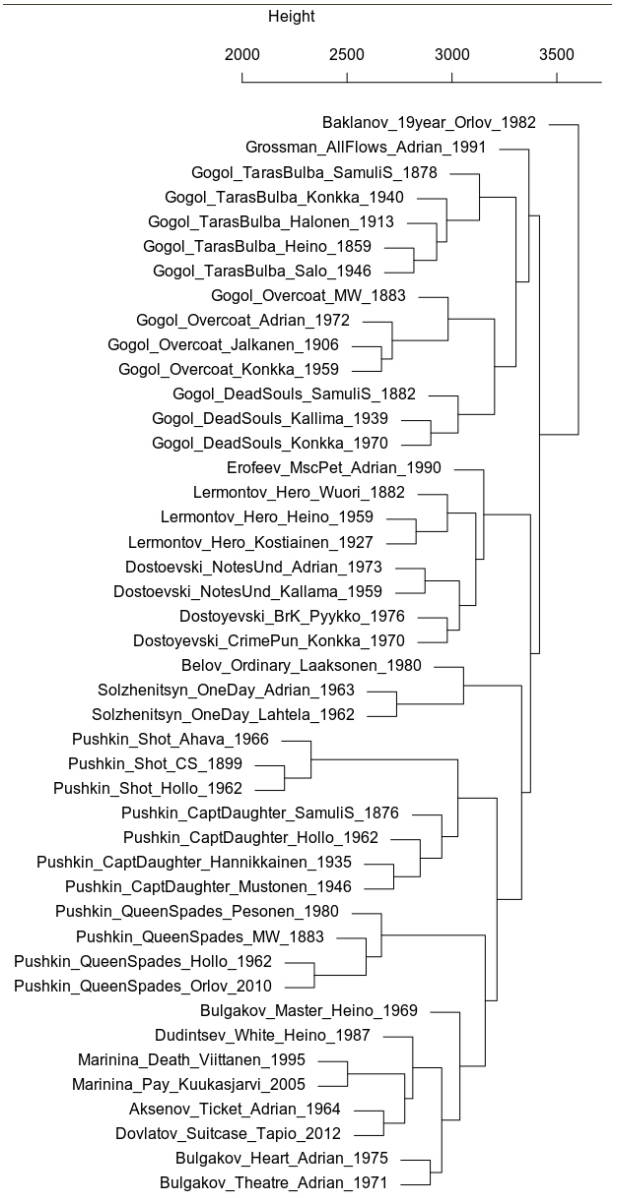


Fig. 5. Samples 3000 tokens (50 samples), FiTr.

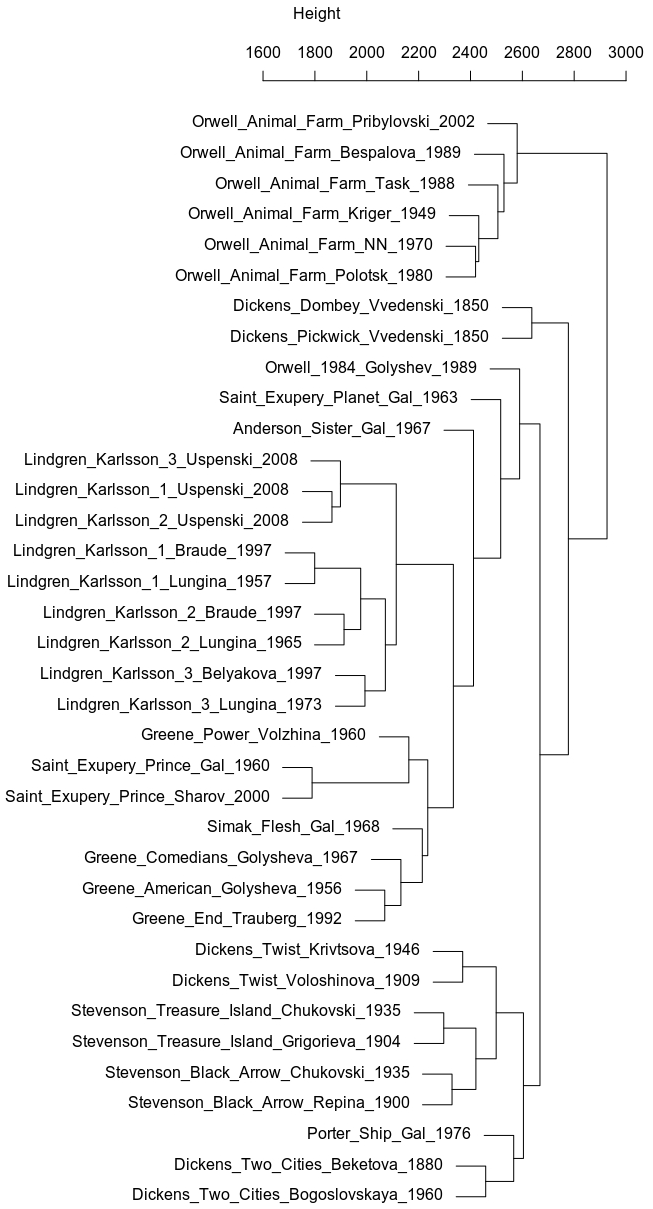


Fig. 6. Samples 3000 tokens (50 samples), RuTr.

Search options used in digitized serial publications – observational user data and future challenges

Tuula Pääkkönen, Kimmo Kettunen, Jukka Kervinen
The National Library of Finland, DH Projects

Abstract

Easy access to digital data resources is one of the key components of successful data intensive Digital Humanities research. Despite increased use of programming languages, web data services and different digital tools, there is increasing demand for researcher friendly tools that are close to the actual materials.

In the digitized newspaper collection of the National Library of Finland (digi.nationallibrary.fi) there has been continuous efforts to incorporate the internal needs to the functionalities, which are offered to end-users. Therefore during the last development project we utilized user survey and log data to help us to design features for the future.

This paper discusses the key findings of both a user survey and gathered user log data of a digitized historical material Web collection. The aim of the discussion is to pinpoint the most salient features of an interface that users of the digitized historical newspaper and journal collection use. Inclusion of new type of data, digitized books, will bring new challenges to the planning and design of a user interface so that it would serve Digital humanities researchers as well as possible.

1 Introduction

One tool for data intensive Digital Humanities research, is a search engine attached to the representation system of a digital collection. The key functional feature of the Web presentation system of digitized newspapers and journals developed by the National Library of Finland (NLF)[1] is keyword search from all the contents. As the full amount of the material is over 15 million pages from 1771–1929, the only way to get a grasp of the material is to be able to perform an exhaustive search. Full content keyword search is the major feature, but also additional search options, like

[1] digi.nationallibrary.fi

publication dates, place of publication, and even page number, are useful as search operators.

In this paper we wanted to approach two research questions: 1) how the search is currently used and 2) how the queries are formed. These questions emerged, as in an earlier user survey with 140 respondents an overwhelming majority of the answerers stated that the most used feature of the Web service of NLF was free text search and its different operators (Pääkkönen & Kettunen, 2018). The questionnaire did not go into details of which search operators were used. Fortunately, we can improve our insight of the user query behaviour with anonymous search data that has been collected for almost two years in order to help The National Library to develop the search functionalities of the Web service. These functionalities are universal and could also be applicable in other presentation systems of similar types. In addition, via looking at the technical features, we can also evaluate the overall researcher needs and services within the NLF. Needs of users are complex and variable, and the search features of the Web system should enable both novice and expert users to find the content they need (Fuhr et al., 2007). The usability goals of the Digi service overall are as Rubin (1994) states: usefulness, effectiveness (ease of use), learnability and attitude (likeability), (Rubin, 1994). This study was also needed, because new material types bring new challenges – we need to both hold onto existing users, but also be able to incorporate new search functionalities in an easy-to-use way to the search.

This paper discusses the key findings of both a user survey and gathered user log data of a digitized historical material Web collection. The aim of the discussion is to pinpoint the most salient features of an interface that users of the digitized historical newspaper and journal collection use. Inclusion of new type of data, digitized books, will bring new challenges to the planning and design of a user interface so that it would serve Digital humanities researchers as well as possible.

2 Background: the search of 2018

Information retrieval from the digitized materials of The NLF starts from content-based search. Metadata is used to filter down the amount of the content. A long-standing philosophy of the full-text search interface of Digi has been that different material types – newspapers, journals and small prints – have been provided separately, although with the same search form. Thus each search is limited to a specific material type. Search fields Title, Place of Publication, and Publisher have

an internal filtering within the field. In time range field the user can type the dates or pick them from the calendar. In the middle of the form is the essential Search words field, where the search terms are keyed. In the bottom there are some options to control the search – require all search terms or possibility to use fuzzy search to overcome text recognition errors. An illustration of the present search form from the early November 2018 is shown in Figure 1.

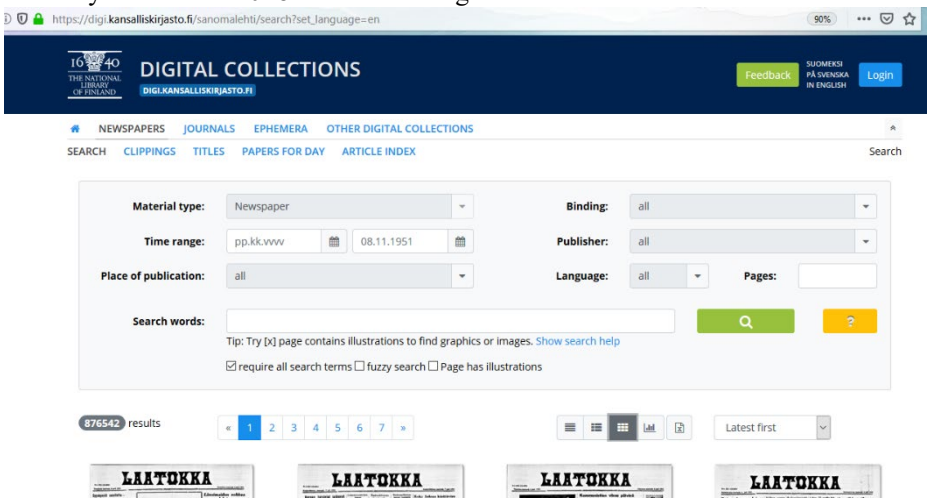


Fig. 1. Digi.nationallibrary.fi - search form as of November 2018

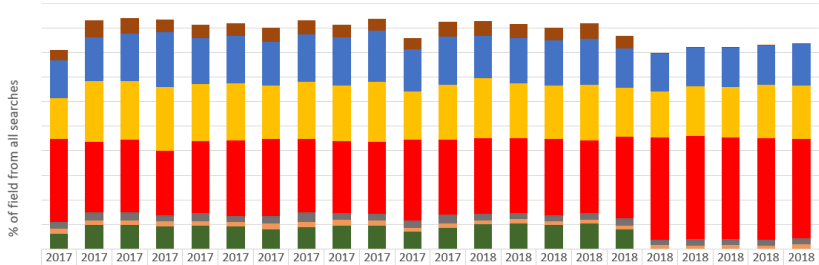
3 Data of the search logs

In order to improve our insight into user behavior of our search system we have gathered anonymized search logs. Our data covers time period of almost two years from January 2017 until October 2018. The top-level research question of how search is currently used was divided into two further questions. Firstly which search fields are used generally based on log data and secondly how different search operators are used in queries. With this question we wanted to see how to develop the search interface further so that it would fulfill different user needs. The problem, even today, is how to have an easy-to-use search interface, which still has powerful enough capabilities (Shneiderman, 1997). Answers to these questions help development of the service and the user interface. When we have data of the most used fields in the search form, we can rearrange the search fields based on the actual

usage. The data, together with concept-decision frameworks like Pugh matrix (Bailey & Lee, 2016), enable us to have fact-based development.

The most used search operators of the interface from January 2017 till October 2018 are depicted in Figure 2. Whenever user makes a search in Digi.nationallibrary.fi, the search operators are stored to the database alongside with a timestamp and number of returned results. Originally this data was collected to ease monitoring of the service, statistics and customer service, e.g. to analyze why certain searches were not giving the results the users were expecting. Collecting of the data is mentioned in the data privacy statement, and is anonymous.

The most used search fields according to the log data are the free-text search field, publishing dates, bindings (i.e. name of the newspaper or magazine), and publishing place. After these come fuzzy search and language selection options. Page numbers or illustrations are used only in very few searches. This result correlates quite nicely with the generic feedback got from our user survey of summer 2018 (Pääkkönen & Kettunen, 2018). 90% of answerers used text search, but only 58% used advanced search options.



	2017-01	2017-02	2017-03	2017-04	2017-05	2017-06	2017-07	2017-08	2017-09	2017-10	2017-11	2017-12	2018-01	2018-02	2018-03	2018-04	2018-05	2018-06	2018-07	2018-08	2018-09	2018-10
PublishingPlaces%	8,8	13,7	12,4	10,1	10,8	10,2	11,1	11,2	10,2	9,8	9,4	11,9	12,1	11,1	10,0	12,8	10,4	0,0	0,1	0,1	0,1	0,1
End_publishingdate%	30,4	35,7	38,6	44,9	38,0	39,0	35,7	39,0	39,0	41,9	34,4	39,5	34,6	36,9	36,9	37,5	31,9	30,8	31,4	32,3	32,2	34,1
Start_publishingdate	33,4	50,1	48,2	52,0	46,2	46,6	43,9	46,6	45,8	48,5	39,3	44,6	49,2	45,1	43,9	45,8	40,0	38,0	40,2	40,9	43,6	43,8
TextSearch%	89,9	87,0	88,9	80,2	88,1	88,3	89,5	89,6	87,7	87,5	89,1	89,2	90,1	90,2	89,6	88,3	91,7	90,7	92,3	91,1	90,4	89,4
Fuzzy%	5,3	6,3	6,6	4,8	6,4	4,4	5,8	7,7	5,3	5,7	5,7	7,0	5,7	4,9	4,4	5,6	6,0	4,4	5,1	4,8	4,7	5,2
Language%	4,5	4,0	4,1	3,9	4,0	3,6	4,6	4,4	4,5	4,3	3,4	3,7	3,3	3,9	3,3	3,2	3,0	3,0	2,6	3,1	2,8	3,5
Bindings%	12,1	19,2	19,2	18,4	18,7	18,5	16,1	17,5	19,1	18,7	13,8	17,2	19,8	20,6	19,5	20,5	15,8	0,0	0,0	0,0	0,0	0,0

Fig. 2. Different search operators used – log data analysis

4 Results: An Improved Search User Interface

Our presentation system has so far included only newspapers, journals, and small prints. A development project at the NLF brought new material types to the Web

service digi.nationallibrary.fi in February 2019. Digitized books are our first focus, and they bring new kinds of search needs for the users. Our constant development challenge has been where to fit the new search fields that the monograph material brings, namely: material selection, series, keywords, and last but not least the collection concept, which brings new opportunities in grouping material across material types. A constant discussion during the development concerned "how do we fit these fields to the screen", which led to this analysis of the actual search field use

4.1 Search across different material types

While adding monograph support to the search, it is also a logical step to lift the traditional model where there was a search barrier between newspapers and journals. New idea enables end-users to search across all the material easily.

Also certain structural changes were now possible. The basic search bar, as seen in Figure 3, is brought to the front page – this was earlier one click further down. As the basic text search is the most used option both according to our user survey and search logs, this change will make search easier to use. On-site customer service in NLF stated that 'serial publications' are the most used feature when end-users query them on legal deposit library terminals, so that also has now a more prominent location.

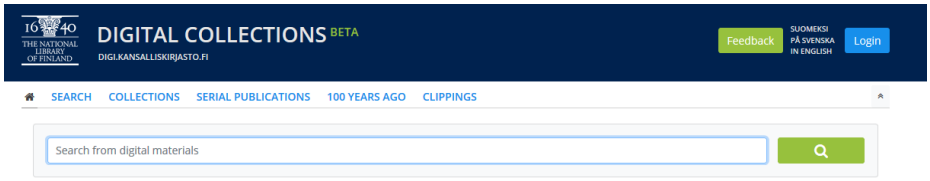


Fig. 3. New simple search bar in the front page

When the user has made the initial search with the new search form, all the options of the search become available on the advanced search form, which can be seen in Figure 4. The so far the collected log data from searches has been utilized to organize the most used search fields on the top of the page, and the less used options are hidden until the user needs them. (Nielsen, 1993) recommends that statistics can guide development to focus on the most used functionalities, which in this case are the search fields themselves. We aimed to follow the ideology of "making the

easy jobs easy, without making the hard jobs impossible" (Christiansen, Schwartz, & Wall, 1996). Therefore the insight from the actual usage of the search fields was very valuable. We will keep experimenting with expanding the search fields to full use if it seems that the additional fields are not used as much as compared to the earlier situation.

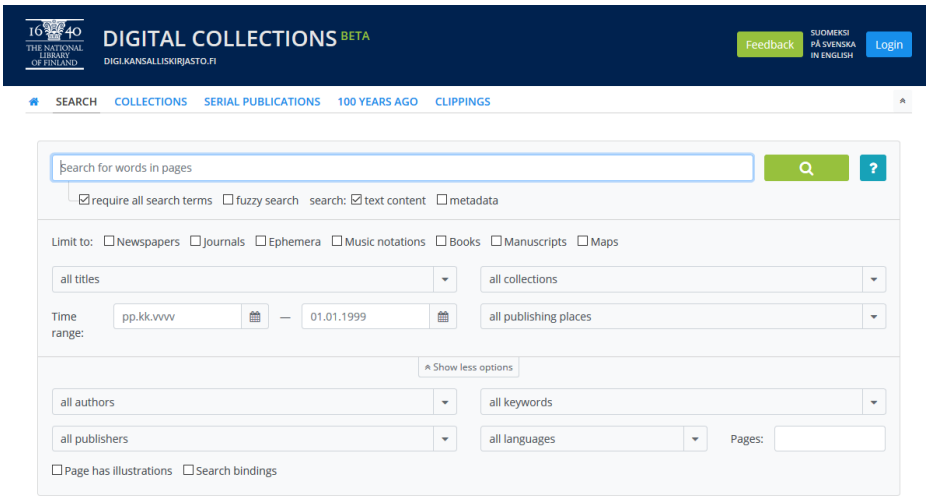


Fig. 4. New full search form with fields organized according to the usage data

4.2 Beyond the search – factors that affect use of the materials

Direct search from the digi.nationallibrary.fi is not the only way to get access to the data. Users can end up to the various content pages of our Web service also from general search engines, social media, national aggregator of all content, or from the calendar of the newspaper, and therefore the search log does not tell every detail, even though it is quite revealing. The amount of material users view is also directly related to the time span that is accessible for the users. For example, when NLF announced in early 2018 that the years 1918–1929 are accessible for users, the number of searches on the day was the fourth largest during the time the data has been collected.

Search functionalities are only one factor that impacts the end-user. The quality of text recognition, for example, has an effect on number of pages returned for the user. OCR error correction could lead to more returned relevant pages for the end

user (Traub, Samar, Van Ossenbruggen, & Hardman, 2018; Järvelin, Keskustalo, Sormunen, Saastamoinen, & Kettunen, 2016). Search functionalities, search engine capabilities and the quality of the material need all to be considered when responding to the end-user needs. Data improvement is on our focus with a new OCR pipeline that should improve the quality of the texts clearly (Kettunen & Koistinen, 2019).

5 Discussion

Surveys and user log data give together good insight on the usage patterns of digitized libraries (Fuhr et al., 2007). This information helps us to be closer to the end-user and enables us to develop better services. In the future, however, it would be advantageous to add a series of user labs, with e.g. researcher groups of different fields, to our service development tool box. User labs could give us quick feedback on existing or possibly missing features of the interface. Improved services for researchers will soon be even more important than before in the evolving open science spirit (Lilja & Hakkarainen, 2018). Different kinds of researchers can have different needs from the service, which needs to be taken into account.

Easy data access with search and browsing tools, such as different visualizations, can give a digital humanities researcher a starting point to think about which topic to start to research, when a phenomenon has appeared or when it might have disappeared. Tools integrated to the digital presentation system can be more approachable than large open data sets. Integrated tools, datasets and e.g. comprehensive instructions to the researchers can provide a powerful tool set, which libraries can offer even with limited resources.

Acknowledgments

Fig. 5. This work was funded by the EU commission through its European Regional Development Fund and the program Leverage from the EU 2014-2020.

References

Bailey, B. D., & Lee, J. (2016). Decide and CONQUER. *Quality Progress*; Milwaukee, 49(4), 30-37. Retrieved from http://search.proquest.com/docview/1783899539/abstract/865BB1B2DE404BD4PQ/1

Christiansen, T., Schwartz, R. L., & Wall, L. (1996). *Programming perl* (Second edition ed.). Sebastopol, CA: OReilly Media.

Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Kals, C.-P., Kovács, L., Landoni, M., Micsik. A., Papatheodorou C., Peters, C. & Sølvberg, I. (2007). Evaluation of digital libraries. *International Journal on Digital Libraries*, 8(1), 21. Retrieved from http://search.ebscohost.com.libproxy.helsinki.fi/login.aspx?direct=true&db=a9h&AN=27201179&site=ehost-live&scope=site

Järvelin, A., Keskustalo, H., Sormunen, E., Saastamoinen, M., & Kettunen, K. (2016). Information retrieval from historical newspaper collections in highly inflectional languages: A query expansion approach. *Journal of the Association for Information Science and Technology*, 67(12), 2928-2946. doi:10.1002/asi.23379

Kettunen, K., & Koistinen, M. (2019). Open source tesseract in re-OCR of finnish fraktur from 19th and early 20th century newspapers and journals – collected notes on quality improvement. Paper presented at the *Digital Humanities in the Nordic Countries* 2019. Retrieved from https://cst.dk/DHN2019Pro/papers/dhnocrsplnproc1703.pdf

Lilja, J., & Hakkarainen, J. (2018). Re-defining our services - national librarys new initiatives to support open science. Paper presented at the *HELDIG Summit 2018*,

Nielsen, J. (1993). *Usability Engineering*. Boston: Morgan Kaufmann.

Pääkkönen, T., & Kettunen, K. (2018). Kansalliskirjaston sanomalehtiaineistot: Käyttäjät ja tutkijat kesällä 2018. *Informaatiotutkimus*, 37(3), 15–19. doi:10.23978/inf.76067

Rubin, J. (1994). *Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests*. New York: Wiley.

Shneiderman, B. (1997*). Designing the user interface* Reading: Addison Wesley.

Traub, M. C., Samar, T., van Ossenbruggen, J., & Hardman, L. (2018). Impact of Crowdsourcing OCR Improvements on Retrievability Bias. *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, 29–36. https://doi.org/10.1145/3197026.3197046

Border crossing and trespassing? Expanding digital humanities research to developing peripheries with the novel digital technologies

Toni Ryynänen and Torsti Hyyryläinen
University of Helsinki, Ruralia Institute, Finland

Abstract
Definitions of and perspectives to Digital Humanities (DH) research tend to deviate amongst the disciplines involved. Typically, DH refers to the application of novel technology and methods in the humanities and social sciences (HSS) research: the usage of data in the context of computational science, collaborative effort combining the expertise from humanities and data sciences as well as examination of digitalisation as a cultural and social phenomenon. We propose an expansion for DH research by discussing an on-going research project funded by the EU's Northern Periphery and Arctic 2014−2020 programme. The Emergreen project (2018–2021) is utilised here as an illustrative case: is there a role for development-oriented research based on the local needs aiming at producing practical and transnational technological solutions for the stakeholders' (end-users', consumers', companies', the public sector actors') real needs? The article expands the current method discussions about DH research emphasising a need for the future oriented and practically relevant research and development methods. In this context, the digitalisation of the society is analysed from the humanistic perspective aiming at understanding the needs of the public services development. In the conclusions, we propose a concept of "practical digital humanities" for describing research utilising a humanist approach to practical problem solving with digital technology development in the DH context.

Keywords: Case study*,* Digital Humanities, Humanities and social sciences, Novel technologies, Practical digital humanities.

1 Introduction

1.1 What digital humanities (DH) research is?

DH research refers to the use of data science within the realm of Social Sciences and Humanities research (SSH). DH research defined both as the utilisation of computerised methods and as *the exploration of the digitalisation in societies and cultures* is a liquid research area that has expanded significantly during the last ten years. Growing interest towards this novel compilation of various disciplines has raised the question how DH research should be defined. Both strict restrictions and open approach to embrace all that define themselves as digital humanists have been surfaced.

The computational methods have been used in humanities (e.g. linguistics) from 1950s onwards. In the Digital Humanities Manifesto 2.0[1], DH research is divided in recent and subsequent waves. Research in the first wave in the 1990s and beginning of the 2000s was "quantitative, mobilizing the search and retrieval powers of the database, automating corpus linguistics, stacking hypercards in critical arrays". The second wave emphasises born digital phenomena and is "qualitative, interpretive, experiential, emotive, generative in character. It harnesses digital toolkits in the service of the Humanities' core methodological strengths: attention to complexity, medium specificity, historical context, analytical depth, critique and interpretation." The first tendency emphasises programming and productive activities whereas the second stresses understanding about digitalisation and digitalised culture or theorising things. However, it is not yet evident that even the first wave utilising quantitative research is changing humanities research traditions (Matres, Oiva, & Tolonen, 2018).

1.2 Purpose of the Article – Border crossing and trespassing the field of DH research

Digitalisation is currently advancing probably faster than ever, and new digital products and cultures are created. It is justified to ask, how DH researchers participate in this creation of digital worlds. Should we study past worlds and

[1] http://www.humanitiesblast.com/manifesto/Manifesto_V2.pdf (retrieved 28.10.2018)

historical documents converted into a digital format with the advanced computational methods, or is there any role for DH researchers in creating just, participatory and equal digital world around us?

As a qualitative case study, we will contextualise and analyse an on-going Emergreen project that does not exactly fit to the current DH research "paradigm". However, it is an example of "the other DH research": *the exploration of the digitalisation in societies and cultures* as it develops a participatory digital application for the citizens to engage in the sustainable development of their living environment. Our goal is to open up a case-based discussion about whether the development of novel digital technologies for answering the concrete needs of several stakeholders that develop effective public services for remote and periphery areas could have a home in the DH research community. There has been a transition from reading to doing in DH research. The research can also be constructive and creative in a sense that something never seen before (a digital application for citizen participation) is realised during a research process.

DH research concerns not only the use of data science in SSH research, but also the study of the digital world (Matres, Oiva, & Tolonen, 2018). The New Media research is an informative example of the examination of the digital world in this context. If we accept this claim that DH research has something to do with the examination of digitalisation as a cultural and social phenomenon, DH research should consider the drivers of digitalisation and evaluate the outcomes or applications of these processes. However, the creation of practical solutions via digital technologies – an activity that drives digitalisation in contemporary societies – is not considered yet in its full potential. This entails bringing different stakeholders together in close collaboration that facilitates the creation of something new by the digital means. Could this be a novel expansion for DH research or just traditional technology development?

2 The Post-digital humanities and the latest turns

Some theorists are already talking about the post-digital humanities (Berry, 2014). Instead of waves, Berry (2012) discusses about moments or overlapping layers, which can be seen as perspectives to DH research. Placing attention to these "digital components", DH research is shifting to the third moment or layer. In that case,

research interests are linked to the epistemic changes and transformations of digitalisation.

Berry (2014) suggests that there is a shift to the post-digital humanities research phase because people might no longer talk about digital versus analog but instead about the modulations of the digital or different intensities of the computational. The post-digital humanists would study the increasingly computational world and culture, the ways in which culture is materialised and fixed in forms specific to digital material culture and how culture is created, for example, in the technical devices, recording systems and databases (Berry, 2014).

These discussions have an impact on the human conceptions. Digitalisation of the society shapes and has an inevitable effect on the current idea of man, the so called homo connectus, defined as people available online and living in a digital ecosystem. They are gregarious and hyper social, use cyber language, are involved in social activism, are trapped in the net and always learning and sharing (Llamas & Belk, 2013).

Haverinen and Suominen (2015) have illustrated the development of DH research in examining how research themes are weighted in relation to digital: is digital technology emphasised instrumentally when doing, developing and making things or is the weight put on theoretical understanding about the digital technology itself.

Critical tones towards digitalisation in general are common, since novel practices and rapidly advancing technologies have provoked both fear and awe throughout the history (Peters, 2013). For example, the general utilisation of digital tools and research on digital artefacts, new media and contemporary culture, if these take place in the relation to physical products and technology or media history, are sometimes not considered as DH research (Burdick, Drucker, Lunenfeld, Presner, & Schnapp, 2012). Critical voices warn also of being too enthusiastic and uncritical towards the new analysis methods, which can obscure the humanist perspective and the role for humanistic disciplines in understanding humanity in general (Evans & Rees, 2012).

DH research has faced resistance from the outside of the DH community but also from the traditional humanities perspectives. In addition, the DH research community has formed "tribes" that emphasise different aspects inside the scholarship as the majority of the practitioners embrace method-driven approaches. It has been emphasised that DH researchers should be technologically perceptive and know how to produce computer code and program software. It is also stressed

that a digital humanist should be "a real humanist", an expert with a strong knowledge on the substance of the study. Voices calling for broader perspective to DH research have intensified (Schreibman, Siemens, & Unsworth, 2016). DH research does not consist of two separate components, digitalisation and humanism, but it is their fusion (Burdick et al., 2012).

How people actively create digitalisation in their daily practices, what are the expected impacts on the societal level and what kind of role people have in these processes could be relevant questions in the future. In addition, DH community faces a challenge how to improve the means of cooperation as the work is increasingly conducted in expert networks, with several partner organisations, and in various scientific and geographical fields. There are also challenges when research projects join researchers, technology developers and the local actors together. The characteristics of DH research work will be different in comparison with traditional humanistic research conducted by a single individual.

3 In between? A case for "practical digital humanities"

3.1 Overview of the Emergreen project

The EU's Northern Periphery and Arctic 2014−2020 programme funds the *Emergreen – Emerging technologies for greener communities* -project (10/2018–9/2021), which aims at developing several public services with the novel digital technologies. Altogether seven partners from five NPA-regions will cooperate following an open and innovative approach underpinned by the introduction of novel emerging technologies and the exploration of new business models for public services provision in remote areas.

The Emergreen project develops and produces concrete public services with the aid of digital applications. Northern peripheries and remote locations share similar challenges that can be facilitated with the novel technologies. The common territorial challenge tackled by the project is how to deliver quality and sustainable public services in remote areas to overcome factors such as long distances, high service delivery costs due to low demand aggregation, shortages in human and material resources and lack of access to the latest innovations. The project is 1) establishing a wider range of user-friendly channels for the community to access and participate in the co-production process, 2) exploring new models based on

shared solutions at regional and transnational levels to ensure quality and sustainable services, and 3) offering an opportunity to introduce new emerging technologies assisting the public services provision and test their impact and viability within this changing landscape.

3.2 The Emergreen project and DH research

It is suggested that DH research should concentrate on communal problems rather than emphasise technical expertise (Fitzpatrick, 2010). DH researchers should also focus on developing institutions and advance novel academic thinking, which is enabled and reinforced by the on-going digitalisation processes. Along with studying technological methods and their possibilities, DH researchers could examine how making of technology and new digitalisation related phenomena will have an impact on societies (Drucker, 2012).

The participating communities and various stakeholders will be put at the centre of the Emergreen project; they are empowered to build up capacities by facilitating the adoption of innovative solutions. The approach revolves around "practical digital humanities", where concrete needs of the communities are canvassed, suitable solutions are mapped, novel technological solutions are developed, digital products are tested and lastly the process is evaluated in practice. A variety of quantitative and qualitative research materials are collected spanning from clarifying the stakeholders needs to possible digital solutions to be developed. Big data sets are not utilised and therefore, the latest computational methods are not needed. However, the project utilises novel technologies and digitalisation, has a strong humanist approach to solve practical problems the communities are facing and develop concrete digital applications. In other words, the project operates on the fringes of DH research with very practical notions.

Spiro (2012) utilises opposites when defining digital humanists. These are at least making vs. theorising, computation vs. communication and practice vs. theory. In the Emergeen project making, communicating and production of practical solutions are emphasised. The project introduces new emerging technologies in the provision of green type of public services, which include the solutions of marine litter and zero waste circular management, green growth advisory services, intelligent green participation and real time visualization of climate data. Technologies such as virtual reality, data visualization, gamification, drone technology and mobile participative tools will be tested. The technological

solutions will be also utilised in seeking new business models for public services provision in remote areas: a transnational platform of technology-led services will be set up and serve to test the effectiveness of these new models.

Kirschenbaum (2010) describes digital humanism as a social contract, which enables cooperation between various disciplines, to form novel research projects and create justified debates about the themes of digitalisation. DH research is therefore a multi- and interdisciplinary concept that compiles computational and humanistic research agendas and methods together. An essential part of DH research is an expert network created by the researchers, as it is increasingly difficult to realise this kind of research alone.

The main reasons why the Emergreen project requires transnational cooperation to achieve the expected results are a need for bringing a significant amount of different knowledge and experiences impossible to be only found at regional or local level. In this sense, the involved partners present a mix of knowledge and experience required ranging from emerging technologies, operational models for remote areas, experience in shared and green-growth public services, an engagement with communities, crowdsourced GIS systems and transnational cooperation management. There is also a need for testing new approaches in different contexts and target groups. The high number of challenges that the provision of public services in remote areas present requires a joint and open approach where organisations from different contexts are committed to test innovative solutions on different target groups and share their findings and achievements with others. Transnational cooperation brings a multiplier effect when it comes to find solutions adapted to very different demands. In addition, the shared services or adaptation of existing ones is one of the key areas to sustain quality public services in remote areas.

3.3 Expected contributions of the Emergreen project

In the first instance, the approach presented will bring a change in relation to the introduction of innovation processes for the provision of public services. It presents a completely new scenario for the organisations responsible for the provision of the public services from a threefold dimension: technology-wise, where new emerging technologies will be tested; methodology-wise, reflected in an open innovation approach involving all the relevant stakeholders; and business-wise, where new business models based on sharing solutions will be explored. This will lead into an

increased openness by the public authorities to apply these new approaches in their task of providing public services.

The Emergreen project will also produce an increased awareness in the users about the key role they can play as part of the solution and a subsequent behavioural change in the way they currently interact to receive these services and adopting an active role. The project will bring the enhanced capacity of the communities to effectively manage their resources and develop in a sustainable way. The new participatory services that will be piloted in the project are oriented to actively change the behaviour of citizens and business providing them with enhanced capacities and prepare them to take their communities into a greener status. Furthermore, the project will seek suitable models that permit the delivery of quality and sustainable public services adapted to the reality of remote areas. The transnational cooperation of a balanced partnership with the required knowledge, experience and competence in the fields tackled, will permit to test new solutions aimed to overcome factors such as the scarcity of human and the material resources of the public sector.

4 Conclusions

Digital humanism is a discipline or a field of study actively created and maintained by researchers. It orients towards interesting research topics and develops together with applied technologies. However, the definitions of DH research are under constant debate.

Based on the definitions of DH research presented in this article and the introduced case study about the Emergreen project, we propose an extension to the existing definitions. This extension could capture studies that aim at the practical development of involved communities from the humanist perspective. We call this kind of research as “practical digital humanities”.

The shortly introduced Emergreen project is an example of practical digital humanities research with the distinctive characteristics: the project aims at 1) resolving concrete practical problems faced by the remote areas, 2) with novel digital technologies, 3) in the collaboration with local and key stakeholders, and 4) in transnational cooperation with research organisations. We asked how this kind of research fits in the current domain of DH research and what kinds of novelties it could bring to this liquid and developing research field. Practical digital humanities research not only expand the current field of DH research, but it could feed novel

research ideas to the traditional humanistic research and offer novel opportunities for data scientists.

General DH research has focused on two directions: to purely humanistic data processing and software design, and to humanistic research that applies technological solutions. In other words, the dominant question in defining DH research is whether the digital technology is the subject of research or is it a tool for research? The third option outlined in this article is practical digital humanities where central factors are a strong humanistic approach, practical problem solving and digital technology development.

References

Berry, D. M. (2014). Post-digital humanities: computation and cultural critique in the arts and humanities. Educause, 49, 22–26. http://sro.sussex.ac.uk/id/eprint/49324

Berry, D. M. (2012). Introduction: Understanding the Digital Humanities. In David M., Berry (ed.) Understanding Digital Humanities (pp. 1–20). New York: Palgrave Macmillan.

Burdick, A., Drucker, J., Lunenfeld, P., Presner, T. & Schnapp, J. (2012). Digital Humanities. Cambridge (MA): The MIT Press.

Drucker, J. (2012). Humanistic Theory and Digital Scholarship. In Gold, M. K. (ed.) Debates in the Digital Humanities. Minnesota: University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates, retrieved 28.10.2018.

Evans, L. & Rees, S. (2012). An Interpretation of Digital Humanities. In Berry, D. M. (ed.) Understanding Digital Humanities (pp. 21–41). New York: Palgrave Macmillan.

Fitzpatrick, K. (2010). Reporting from the Digital Humanities 2010 Conference. The Chronicle of Higher Education, 13.7.2010. http://chronicle.com/blogs/profhacker/reporting-from-the-digital-humanities-2010-conference/25473, retrieved 28.10.2018.

Haverinen, A. & Suominen, J. (2015). Koodaamisen ja kirjoittamisen vuoropuhelu? – mitä on digitaalinen humanistinen tutkimus. Ennen ja Nyt, retrieved 28.10.2018. http://www.ennenjanyt.net/2015/02/koodaamisen-ja-kirjoittamisen-vuoropuhelu-mita-on-digitaalinen-humanistinen-tutkimus/

Kirschenbaum, A. (2010). What Is Digital Humanities and What's It Doing in English Departments? ADE Bulletin, No. 150, 55–61.

http://mkirschenbaum.files.wordpress.com/2011/03/ade-final.pdf, retrieved 28.10.2018.
Llamas, R. & Belk, R. (2013). Living in a digital world. In Belk, R. & Llamas, R. (eds.) The Routledge Companion to Digital Consumption (pp. 3–12). New York: Routledge.
Matres, I, Oiva, M. & Tolonen, M. (2018). In Between Research Cultures – The State of Digital Humanities in Finland. Informaatiotutkimus 2(37), 37–61.
Peters, O. (2013). Against the Tide Critics of Digitalisation Warners, Sceptics, Scaremongers, Apocalypticists 20 Portraits. Oldenburg: BIS-Verlag.
Schreibman, S., Siemens, R. & Unsworth, J. (eds). (2016). A New Companion to Digital Humanities. West Sussex: John Wiley & Sons.
Spiro, L. (2012). ”This is why we fight”: Defining the Values of Digital Humanities. In Gold, M. K. (ed.) Debates in the Digital Humanities. University of Minnesota Press: Minnesota. http://dhdebates.gc.cuny.edu/debates, retrieved 28.10.2018.

III Tools

Tutkimusaineiston kerääminen ja analysointi monipuolisia digitaalisia keinoja hyödyntäen. Esimerkkinä Tunne-etsivät-tutkimushankekokonaisuus

Kerttu Huttunen
Oulun yliopisto, humanistinen tiedekunta, logopedian tutkimusyksikkö ja Lapsenkielen tutkimuskeskus

Tiivistelmä
Digitaaliset pelit kuuluvat nykyään kiinteänä osana lasten elämään – niiden pelaaminen on yksi nykylasten leikin muoto. Sähköisiä, vuorovaikutteisia pelejä käytetään yhä enemmän myös kommunikointihäiriöisten lasten kuntoutuksessa, sillä mielenkiintoisten pelien avulla saadaan ylläpidettyä lasten harjoittelumotivaatiota ja harjoitteluun voi sisällyttää loputtoman määrän toistoja. Digitaaliset pelit soveltuvat erityisen hyvin autismikirjon lasten kuntoutukseen.

Monitieteisessä neljän suomalaisen yliopiston hankkeena toteutettavassa Tunne-etsivät-tutkimuskokonaisuudessa tuetaan lasten kielellisiä ja sosioemotionaalisia taitoja, erityisesti lasten kykyä tunnistaa tunteita kasvoilta, puheesta ja sosiaalisista vuorovaikutustilanteista. Tukeminen tapahtuu hankkeessa laaditun, verkossa pelattavan Tunne-etsivät-pelin avulla. Kyseinen peli on tähän saakka laajin ja monipuolisin Suomessa tunteiden tunnistustaitojen harjaannuttamiseen laadittu sähköinen materiaali.

Kommunikointihäiriöisiä lapsia koskevassa hankkeen osassa tutkitaan, voidaanko digitaalisen pelin avulla vahvistaa lasten kielellisiä ja sosioemotionaalisia taitoja ja erityisesti, voidaanko pelin avulla parantaa lasten kykyä tunnistaa tunteita. Jyväskylän yliopistossa väitöskirjahankkeena toteutuvassa osassa taas keskitytään siihen, miten peli toimii yhteisöllisen oppimisen välineenä: millaista on yhteisöllinen oppiminen ja millaista vuorovaikutusta tyypillisesti kehittyvien lasten kesken syntyy, kun he pelaavat peliä internetissä yhdessä toisen lapsen kanssa.

Tässä artikkelissa kuvataan digitaalisen Tunne-etsivät-verkkopelin kahden tutkimusversion ominaisuuksia. Nämä tutkimusversiot kehitettiin hankekokonaisuuden kahden osahankkeen aineistonkeruuta varten. Lisäksi kuvataan näiden osahankkeiden digitaalista aineistonkeruuta ja analysointia.

Hankkeen tutkimusaineistona on lasten suoriutuminen kielellisiä, kognitiivisia ja tunteiden tunnistamisen ja tuoton taitoja kartoittavissa tehtävissä ja testeissä. Tutkittavana oli 55 lasta, joilla oli joko autismikirjon häiriö, ADHD, kehityksellinen kielihäiriö tai kuulovika. Näiltä lapsilta tutkittiin mm. tuottava sanavarasto, lyhytaikainen muisti, mielen teorian taidot sekä kerrontataidot. Lisäksi tutkittiin keskittymiskykyä ja reaktionopeutta sekä tunteiden tunnistus- ja tuottokykyä. Aineistonkeruussa hyödynnettiin näillä alueilla keskeisinä ärsykkeinä digitaalisia materiaaleja. Lapset mm. tekivät hankkeessa laaditun, tietokoneella tehtävän reaktioaikatestin sekä nimesivät ja tuottivat ilmeitä ja äänensävyjä heille tietokoneella esitettyjen valokuvien, videoleikkeiden ja äänitiedostojen pohjalta. Testaushetket videoitiin ja myös tallennettiin Zoom-äänitallentimella äänitiedostoiksi ennen pelaamisinterventiota ja sen jälkeen. Tunne-etsivät-pelin lokitiedostoista kerättiin tieto kunkin lapsen harjoitusmääristä eli kertyneestä pelaamisajasta ja myös eri tehtävissä onnistumisesta ja kunkin tehtävän suorittamiseen kuluneesta ajasta.

Tausta-aineistoksi kerättiin 109 tyypillisesti kehittyvän lapsen suoriutumistulokset edellä kuvatuissa mielen teorian ja tunteiden tunnistuskyvyn testeissä ja tehtävissä. Väitöskirjahankkeessa puolestaan videoitiin vuorovaikutuksen ja mm. paripelaamistyyppien ja niiden muutosten analysointia varten 16 tyypillisesti kehittyvän lapsen Tunne-etsivät-pelin pelaamista toisen lapsen kanssa parina.

Tutkimusaineisto on nyt kerätty ja se on analyysivaiheessa. Osa tuloksista on jo raportoitu. Videoituja testaustilanteita on hyödynnetty eri testien ja tehtävien pisteytyksessä tai pisteytyksen tarkistamisessa. Niiden avulla on myös litteroitu lasten kuvasarjakerrontatehtävässä tuottamat kertomukset ja lasten sanallinen ja ei-sanallinen vuorovaikutus paripelaamishetkissä. Videoista havainnoidaan myös eleiden käyttöä kuvasarjakerronnassa. Äänitiedostoista puolestaan analysoidaan kuvasarjakerrontatehtävässä sitä, miten lapset käyttävät prosodiikkaa tunteiden ja muiden mielentilailmausten korostamiseen.

Edellä kuvatut tutkimushankkeet edustavat monitieteisiä ja moniammatillisia tutkimuskokonaisuuksia, joissa aineistoja kerätään digitaalisesti, ja aineistot myös analysoidaan monipuolisia sähköisiä työkaluja käyttäen. Olennaista on myös eri tieteenaloja edustavien tutkijoiden yhteistyö, toisilta oppiminen ja toisten tieteenalojen tutkimustyökaluihin tutustuminen ja niiden hyödyntäminen. Lopuksi artikkelissa kuvataan laajojen digitaalisten aineistojen keräämiseen, analysointiin ja säilyttämiseen liittyviä kokemuksia.

Asiasanat: digitaaliset ihmistieteet, digitaaliset pelit, digitaaliset tallenteet, kielelliset häiriöt, kuntoutus, pelit, tietokoneavusteisuus, tunteet, tunnetaidot, verkkopelit, viestintä

1 Tutkimusaineiston digitaalinen kerääminen ja analysointi

Seuraavassa kuvataan tutkimushankekokonaisuutta, jonka lähtökohtana on lasten tunteiden tunnistustaitojen ja muiden sosioemotionaalisten taitojen tukemiseen tarkoitettu Tunne-etsivät-peli ja siihen liittyvä tutkimustyö.

1.1 Digitaalisuuden hyödyntäminen aineistonkeruussa

Lasten sosioemotionaalisia taitoja tutkittaessa Tunne-etsivät-hankkeessa kerättiin aineistoa erilaisin kielenkehitystä, mielen teorian sekä tunnetaitoja mittaavin testein ja tehtävin. Lisäksi hyödynnettiin Tunne-etsivät-pelin lokitiedostoja.

Kommunikointihäiriöisiä lapsia koskevassa tutkimushankkeen osassa yhteensä 35 lasta pelasi noin tunnin viikossa kahden kuukauden ajan hankkeessa laadittua, Opetushallituksen tuottamaa ja sen verkkosivuilla pelattavaa Tunne-etsivät-peliä (Huttunen, Hyvärinen, Laakso, Parkas & Waaramaa, 2015). Esimerkki pelin Tunne-etsivien toimistosta on kuvassa 1.

Kuva 1. Tunne-etsivät-pelin toimisto, josta pelaajat voivat valita ns. toimisto- ja kenttätehtäviä. Myös Tunnetutka-hahmo neuvoo seuraavan pelattavaksi aukeavan tehtävän.

Pelin kaikkiin osiin pääsee keskitetysti verkko-osoitteen https://www.edu.fi/verkko_oppimateriaalit/tunne_etsivat kautta. Pelistä erillisinä pelattaviksi omaan valikkoonsa irrotetut 23 minipeliä löytyvät sen lisäksi myös verkko-osoitteesta https://tunneetsivat3.oph.oodles.fi/minipelit.html ja Tarinatuubi-sarjakuvakone puolestaan erillisenä osoitteesta https://tarinatuubi.oodles.fi/. Tunne-etsivät-peli on sarjakuvapohjainen, mutta vuorovaikutteinen siten, että pelaaja saa runsaasti kannustusta ja myös palautetta tekemistään valinnoista (esimerkiksi, oliko vastaus oikein vai väärin).

Jokaiselle peliä pelaamaan pyydetylle eli koeryhmään kuuluvalle lapselle annettiin yksilölliset tunnukset peliin kirjautumiseksi. Näiden pelaajatunnusten käyttäminen johti pelaajan automaattisesti pelaamaan pelin tutkimusversiota, joka muokattiin tutkimusaineiston keräämistä varten tietyiltä taustaominaisuuksiltaan erilaiseksi Opetushallituksen tuottamaan Tunne-etsivät-peliin verrattuna. Vanhempia kehotettiin olemaan läsnä pelaamishetkissä ja kytkemään pelin sisältöjä keskustelemalla lapsen arjen tapahtumiin. Vanhemmille ja tarvittaessa lapsen puheterapeutille tai esimerkiksi opettajalle annettiin seurantatunnus, jonka avulla aikuisen oli mahdollista nähdä pelin tutkimusversion analytiikkasivustolta reaaliaikaisesti, kauanko lapsi oli kullakin kahdeksasta peliviikosta siihen mennessä peliä pelannut. Kunkin yksilöllisen seurantatunnuksen avulla oli mahdollista nähdä vain yhden tietyn lapsen pelaamistilastot. Tavoitteena oli, että

lapset olisivat pelanneet peliä vähintään tunnin, mutta enintään kaksi tuntia viikossa. Näin tavoiteltiin tiettyä kuntoutuksen määrää eli ”annosta”.

Tutkijoilla oli puolestaan käytössään kirjautumistunnukset, joiden avulla he pystyivät seuraamaan samoin reaaliajassa jokaisen koeryhmän lapsen kertynyttä peliaikaa. Mikäli tunnin minimipelaamisaika ei ollut vielä täyttynyt perjantaina, lapsen vanhemmille lähetettiin asiasta tekstiviestillä kehotus, että he pyytäisivät lasta vielä pelaamaan tai pelaisivat lapsen kanssa Tunne-etsivät-peliä viikonloppuna niin, että pelin minimipelaamisaika tulisi täyteen viimeistään sunnuntai-iltana. Mikäli lapsi pelasi jollakin viikolla peliä enemmän kuin oli tavoitteena, peli katkesi automaattisesti silloin, kun peliaikaa oli kertynyt kaksi tuntia. Lapsi sai kuitenkin pelata ajan täyttyessä pelattavana olleen pelin osan loppuun. Pelaamisajat ja eri tehtävissä onnistuminen (vastaukset ja vastaamiseen kulunut aika) kirjautuivat automaattisesti pelin lokitiedostoihin niin, että pelin tutkimusversion tuottamat Excel-tiedostot olivat nimettyinä kunkin pelaajatunnuksen mukaisesti. Excel-tiedostoista tieto on helposti siirrettävissä esimerkiksi SPSS-tilasto-ohjelmaan tarkempia analyysejä varten.

Kommunikointihäiriöisiä lapsia oli tutkimuksessa mukana yhteensä 55. Peliä pelanneista lapsista 30:lla oli jokin neurokehityksellinen häiriö: autismikirjon diagnoosi (useimmiten Aspergerin oireyhtymä), ADHD, kehityksellinen kielihäiriö tai joillakin jopa kaikki nämä diagnoosit. Heidän vertailuryhmänään toimi 20 lasta, joilla oli samoin yksi tai useampi edellä mainituista diagnooseista. Vertailuryhmän jäsenet eivät kuitenkaan vielä tutkimusaineiston keruuaikana pelanneet peliä, vaan heille tehtiin vain alku- ja loppumittaukset. Heille kerrottiin, että tutkimusaineiston keräämisen jälkeen peliä sai pelata vapaasti.

Vertailuryhmän avulla pyrittiin selvittämään, kuinka suuri rooli testioppimisella ja lapsen taitojen luonnollisella kypsymisellä voi olla lasten eri mittauksissa saamissa tunteiden tunnistustuloksissa. Lisäksi peliä pelasi viisi kuulovammaista lasta, joista kaksi käytti kuulokojeita ja kolme sisäkorvaistutteita. Kaikki edellä mainitut 55 lasta olivat iältään 6–10-vuotiaita.

Vertailukohdaksi kerättiin tiedot myös 109:n saman ikäisen, tyypillisesti kehittyvän lapsen tunteiden tunnistus- ja tuottokyvystä, jotta tiedettäisiin, poikkesivatko kommunikointihäiriöisten lasten tunnetaidot heidän tyypillisesti kehittyvien ikätovereidensa taidoista.

Yhteensä 55 kommunikointihäiriöiseltä lapselta ja suurimmalta osalta tyypillisesti kehittyvistä ikäverrokeista tutkittiin (pelaamisesta erillisessä testaustilanteessa) keskittymistä, toiminnanohjaustaitoja ja reaktionopeutta noin

kaksi minuuttia kestävän reaktioaikatestin avulla. Se ohjelmoitiin tutkimushankkeessa ns. two-choice reaction-time task -tyyppiseksi tehtäväksi niin, että tietokoneen näytölle ilmaantui satunnaisesti yhdestä kolmeen sekunnin välein vasempaan reunaan suurikokoinen numero yksi tai oikeaan reunaan numero kaksi (kuva 2). Numeron ilmaannuttua tietokoneen näytölle lapsen tehtävänä oli painaa näppäimistöllä vastaavia eli vasemmalle tai oikealle osoittavia nuolinäppäimiä niin nopeasti kuin mahdollista. Ohjelma laski tulokset eli tunnistustarkkuuden ja vastausnopeuden keskiarvon ja keskihajonnan.

Kuva 2. Lapsi tekemässä reaktioaikatestiä.

Reaktioaikatestin tekemisen lisäksi lapsia pyydettiin tekemään vielä monia muitakin digitaalisuuteen perustuvia tehtäviä. Tunteiden tunnistustaitojen kartoittamiseen kasvoilta eli ilmeiden tunnistamiseen käytettiin digitaalisen FEFA 2 -testin Kasvot-osatestiä (Bölte, Ollikainen, Feineis-Matthews & Poustka, 2013). Testin kolme lisenssiä ostettiin aineistonkeruuta varten Karoliinisesta Instituutista Tukholmasta. Ohjelma laski automaattisesti oikeiden vastausten yhteismäärän, kaikkien mukana olleiden seitsemän eri tunteen oikeiden vastausten määrän ja lisäksi kunkin tunteen tunnistamiseen kuluneen ajan. Lapset myös nimesivät ja tuottivat ilmeitä ja äänensävyjä heille tietokoneella esitettyjen valokuvien, videoleikkeiden ja äänitiedostojen pohjalta.

Koska kommunikoinnissa tarvitaan sekä vastaanoton että tuoton taitoja, lapsilta ei tutkittu pelkästään kielellisiä taitoja ja tunteiden tunnistustaitoja (ks. kuva 3), vaan myös kyky tuottaa ilmeitä ja tunteisiin liittyviä äänensävyjä. Erityisesti autismikirjon lasten on usein vaikea tuottaa sosiaaliseen

vuorovaikutustilanteeseen sopivia ilmeitä (Rodgers ym., 2015) ja äänensävyjä. Lasten tuottamia ilmeitä kerättäessä imitoitavat kuvat näytettiin useimmiten paperivalokuvista videokameran takaa, sillä siten lapsi saatiin katsomaan suoraan kameraan ja hänen ilmeistään saatiin näin hyvä tallenne. Kaikki testaushetket nimittäin videoitiin, ja lisäksi ne tallennettiin Zoom H2N -äänitallentimella äänitiedostoiksi myöhempiä määrällisiä ja laadullisia analyysivaiheita varten. Äänitallennin sijoitettiin lapsen eteen pöydälle noin 30 cm:n päähän lapsesta. Zoom-tallentimessa on viisi sisäistä mikrofonia. Niiden suuntakuvioksi valittiin eteenpäin suuntautunut suuntakeila. Äänitiedostot tallennettiin pakkaamattomassa aaltomuodossa (.wav) bittisyvyydellä 16 ja 44 100 Hz:n näytteenottotaajuudella.

Kuva 3. Esimerkki tunteiden tunnistustehtävissä käytetyistä ärsykekuvista. Lapsen piti monivalintatehtävässä kertoa, mikä neljästä valintavaihtoehdoksi annetusta tunteesta on kulloinkin kyseessä.

Keskeiset Tunne-etsivät pelin vaikuttavuutta koskevat tulokset on julkaistu laajassa, avoimesti saatavilla olevassa tutkimusraportissa (Huttunen, Kosonen, Waaramaa & Laakso, 2018), ja aiheesta on parhaillaan työstettävänä useita tutkimusartikkeleita (mm. Löytömäki, Ohtonen, Laakso & Huttunen, käsikirjoitus arvioitavana). Riippuvien otosten t-testi osoitti, että peliä pelanneiden 30:n lapsen tunteiden tunnistuskyky parani tilastollisesti merkitsevästi kaikilla kuudella mitatulla alueella. Vertailuryhmässä, joka ei peliä pelannut, tulokset olivat kahden kuukauden seurantajakson jälkeen tilastollisesti merkitsevästi paremmat ainoastaan äänensävyjen tunnistamisessa merkityksettömistä sanoista.

Lasten keskinäisen vuorovaikutuksen piirteisiin keskittyvän väitöskirjatutkimuksen aineisto puolestaan koostuu 16:n tyypillisesti kehittyvän, 5–6-vuotiaan lapsen videoiduista pelaamishetkistä (Lipponen, Koivula, Huttunen, Turja & Laakso, 2018). Aineistosta on valmistunut myös kirjan luku (Koivula,

Huttunen, Mustola, Lipponen & Laakso, 2017) sekä erityispedagogiikan pro gradu -tutkielma (Vallenius, 2018).

Kun haluttiin selvittää, millaisen digitaalisen oppimisympäristön Tunne-etsivät-peli tarjoaa lasten sosioemotionaalisen kehityksen tukijana ja millä eri tavoin lapset toimivat keskenään vuorovaikutuksessa pelin äärellä, pyydettiin näitä edellä mainittuja, tyypillisesti kehittyviä 5–6-vuotiaita lapsia pelaamaan peliä parinsa kanssa päiväkotipäiviensä aikana (Lipponen ym., 2018). Koska lapset olivat näin nuoria, tätä aineistonkeruuta varten Tunne-etsivät-pelistä laadittiin vielä toinen tutkimusversio, jossa oli tavanomaista rajallisempi määrä osioita. Lasten ei tarvinnut näitä tehtäviä tehdäkseen osata vielä lukea, sillä samoin kuin ensimmäisessäkin tutkimuspeliversiossa, kaikki pelissä näkyvät tekstit oli äänitetty ja lapset pystyivät kuuntelemaan kaikki ohjeet, valintavaihtoehdot ja kaikkien puhekuplien tekstit. Pelissä esiintyvien hahmojen ääniä oli tuottamassa useita lapsia ja aikuisia. Osa aikuisista oli ammattinäyttelijöitä. Äänet kuuluivat pelissä automaattisesti, samoin vaimea taustamusiikki, mutta taustamusiikin sai halutessaan myös kytkeä pois päältä.

Kahden kuukauden intervention aikana kullekin kahdeksasta pelaajaparista kertyi yhdessä pelaamista yhteensä keskimäärin kaksi tuntia. Pelaaminen videoitiin GoPro-videokameroilla. Aluksi lapset hiukan ujostelivat pientä, kenttätallennuksiin käytettävää kameraa, joka oli kiinnitetty sen pöydän reunalle, jota lapset käyttivät Tunne-etsivät-peliä pelatessaan. Pian kameran uutuusarvo kuitenkin laimeni, eivätkä lapset enää juuri kiinnittäneet siihen huomiota. Kaikki pelaajaparit kuitenkin ilmeilivät tai muilla tavoin hassuttelivat ja esiintyivät kameralle – yleensä kuitenkin vain silloin, kun aikuinen ei ollut läsnä samassa huoneessa. Samalla lapset pohtivat, näkeekö kamera heidät nyt varmasti, ja osalla pelaajapareista kameran ”valvovan silmän” läsnäolo ryhdisti yhdessä toimimista (Lipponen ym., 2018).

Kaikille päiväkodeille lainattiin aineistonkeruun ajaksi langaton laajakaistalaite ja kannettava tietokone, jotta lapset olisivat voineet pelata niitä käyttäen. Kaikissa kannettavissa tietokoneissa oli levyhiiri. Sen käyttö ei ollut useimmille lapsille tuttua, sillä he olivat tottuneet käyttämään kotona pääasiassa kosketusnäytöllisiä laitteita (tablettitietokoneita). Levyhiiren käyttö ei ollut alussa sen vaatimien hyvien hienomotoristen taitojen takia kaikille kovin helppoa, mutta videotallenteissa välillä näkyikin, kuinka parista taitavampi osoitti prososiaalista käytöstä eli tarjoutui auttamaan pelin käytössä sitä parin jäsentä, jolla oli tietokoneen käytössä vaikeuksia.

1.2 Digitaalisuuden hyödyntäminen tutkimusaineistoa analysoitaessa

Kun Tunne-etsivät-peliä pelanneiden lasten tuloksia tarkasteltiin pelin Excel-muodossa olevista lokitiedostoista, voitiin niistä kerätä tieto kunkin lapsen harjoitusmääristä eli kertyneestä pelaamisajasta ja myös eri tehtävissä onnistumisesta ja kuhunkin tehtävään vastaamiseen kuluneesta ajasta. Esimerkki pelin tutkimusversion lokitiedostosta on taulukossa 1.

Taulukko 2. Esimerkki Tunne-etsivät-pelin tutkimusversion lokitiedostoista.

Pelaaja	Aika	Kysymys	Vastaus	Vastausaika
Tutkimus-ryhma057	2016-03-19 18:06:02	Minkä tunteet uskot pojalla olevan päällimmäisenä juuri nyt? Hän on...	kiusaantunut - oikein	21
Tutkimus-ryhma057	2016-03-19 18:37:19	Miltä uskot Viljamista tuntuvan juuri nyt? Hän on...	rohkea - väärin	19
Tutkimus-ryhma057	2016-03-19 18:37:53	Miltä uskot Viljamista tuntuvan juuri nyt? Hän on...	surullinen - oikein	29

Lokitiedostojen data on helposti siirrettävissä tilasto-ohjelmaan muuttujien jatkotyöstämistä ja tulosten analysointia varten.

Lasten pelaamistilanteista erillisiä, videoituja testaushetkiä on käytetty kielellisiä ja kognitiivisia taitoja kartoittavien testien pistemäärien laskemisessa ja tarkistamisessa. Lasten kuvasarjakerronnasta tehtyjen video- ja äänitallenteiden avulla selvitetään parhaillaan kahdessa pro gradu -tutkielmassa, millä tavalla lapset käyttävät eleitä ja äänensävyjä, kun he kertovat kuvasarjojen tapahtumista. Erityisen huomion kohteena on se, esiintyykö eleitä ja käyttävätkö lapset prosodiikkaa jollakin tietyllä tavalla korostamaan tunteisiin viittaavia sanoja ja mielentilailmauksia eli täydentääkö ei-kielellinen tai ekstraverbaalinen informaatio kielen avulla välitettyä tietoa. Äänitiedostojen editointiin käytetään Praat- ja Audacity-ohjelmia ja prosodiikan analysointiin Praat-ohjelmaa (Boersma, 2001). Prosodiikassa huomion kohde on erityisesti F0-mittauksissa.

Lapsilta kerättiin myös valokuvissa esiintyvien ilmeiden imitointiin perustuvat ilmetuotokset. Ne editoitiin Lightworks- ja iMovie-videoeditointiohjelmilla kukin kahden sekunnin mittaiseksi ja upotettiin PowerPoint-dioihin. Viiden puheterapeutin tai puheterapeutiksi tai psykologiksi opiskelevan muodostamat arvioijapaneelit katsoivat lasten ilmetuotokset tietokoneen ruudulta ja merkitsivät,

mikä annetuista valintavaihtoehdoista eli ilmeistä kulloinkin oli kyseessä. Näitä arvioijapaneelien tuloksia verrataan tyypillisesti kehittyvien lasten aineiston osalta konenäön eli tekoälyalgoritmien tunnistustuloksiin (Huttunen, Huang & Zhao, käsikirjoitus valmisteilla).

Selvitettäessä tyypillisesti kehittyvien 5–6-vuotiaiden lasten kahdeksan pelaajaparin keskinäistä vuorovaikutusta pelihetkien aikana tulokset analysoitiin videotallenteista ensin litteroimalla lasten käymät keskustelut. Litteraateissa kuvattiin myös lasten ei-kielellinen käyttäytyminen. Sellaista oli esimerkiksi se, kun pelaajapari kiisteli kannettavan tietokoneen käyttövuorosta. Toinen lapsi saattoi kurottaa kädellään ulottuakseen kannettavan tietokoneen levyhiirelle ja voidakseen sitä käyttäen valita pelissä haluamansa vastauksen pelin esittämään kysymykseen. Tietokoneen äärellä oleva lapsi saattoi taas puolestaan koettaa suojella näppäimistöä ja levyhiirtä omalla kädellään niin, ettei parin toinen jäsen olisi päässyt tietokonetta käyttämään. Tällainen ei-kielellisen vuorovaikutuksen tarkastelu auttoi osaltaan vuorovaikutuksen piirteiden ja paripelityyppien luokittelussa (Lipponen ym., 2018).

2 Kokemuksia työskentelystä digitaalisten tutkimusaineistojen parissa

Datan hallinta on osoittautunut Tunne-etsivät-osahankkeissa välillä haasteelliseksi, koska suuren aineiston keräämiseen tarvittiin paljon aineistonkerääjiä. Osalla aineistonkerääjistä käytössä oli videokamera, joka tallensi kuvan ja äänen suoraan kameran kiintolevylle. Siinä tapauksessa videot piti siirtää kamerasta tietokoneelle tai muistitikulle tarkoitukseen käytettävissä olevalla ohjelmalla. Muissa videokameratyypeissä aineisto tallentui pienille SDHC-muistikorteille, joita kertyikin sitten suuri määrä. Suuri onni on ollut se, että muistivälineiden hinta tippui dramaattisesti vajaat 10 vuotta sitten. Lisäksi tulivat ZOOM H2N

-äänitallentimilla erikseen taltioidut äänitiedostot, jotka niin ikään olivat SDHC-muistikorteilla. Niiden tallessa pitäminen ja niiltä aineiston siirtäminen ulkoisille kovalevyille vaati suurta huolellisuutta. Myös varakopioiden tekeminen suurella vaivalla kerätyn aineiston säilyvyyden takaamiseksi vaati runsaasti aikaa ja suunnitelmallisuutta. Tärkeintä oli siirtää aineisto muistitikuilta ulkoisille kovalevyille mahdollisimman pian, sillä muistitikku on aineistojen säilyttämiseen kaikkein haavoittuvaisin muistiväline. Koska aineistoja ei voinut tietosuojan

turvaamisen vuoksi postittaa eri tutkijoiden välillä, niitä siirrettiin kädestä käteen aina tavatessa.

Videoita arvioijaryhmille segmentoitaessa ja myöhempiä tekoälyanalyysejä varten valmisteltaessa eri videokameroiden tallennusformaatit tuottivat välillä hankaluuksia. Koska aineistoa kerättiin eri puolilla Suomea osin mm. opinnäytetöihin liittyvinä oppimistehtävinä ja yliopistoissa kulloinkin käytettävissä olevilla, lähinnä kuluttajakäyttöön tarkoitetuilla videokameroilla (merkkeinä mm. JVC, Canon ja Panasonic), poikkesivat niiden tallennusmuodot usein toisistaan. Käytössä olivat mm. resoluutioltaan huippulaatuinen teräväpiirtovideon tallennusmuoto .mts (AVCHD; Advanced Video Codec High Definition), resoluutioltaan vaatimaton .mod ja lisäksi .mov-tiedostomuoto, jota QuickTime Movie -ohjelma käyttää. Eri tallennemuodoista .mod-tiedostomuodon kanssa on ollut kaikkein eniten hankaluuksia. Tekoälyohjelmistot edellyttävät analysoitavalta aineistolta riittävän suurta resoluutiota. Heikomman resoluution tallennemuotoja sisältävät videotiedostot on jouduttu sen vuoksi analysoimaan erilaisia algoritmeja käytettäessä erikseen (Huttunen, Huang & Xhao, valmisteilla).

Kaikista näistä edellä mainituista tallennusmuodoista lasten ilmeitä sisältävät jaksot videoista tuli yhteensopivuuden varmistamiseksi muuttaa eri muunto-ohjelmilla .mp4-formaattiin, jotta ilmeet olisi ollut mahdollista upottaa arvioijaryhmien myöhemmin katsomiin PowerPoint-esityksiin. Videoidut ilmeet editoitiin Lightworks- ja iMovie -videoeditoriohjelmilla niin, että lopputuloksena aineistoon sisältyvien, videoitujen lasten kasvot olivat keskenään suunnilleen samankokoisia ja kunkin ilmeen kesto oli tasan kaksi sekuntia. Lightworks-ohjelmankin ilmaisversio ehti muuttua kesken analyysien, ja analyysien tekoa aloittaville uusille opinnäytetöiden tekijöille oli laadittava taas uudet, hyvin yksityiskohtaiset ohjeet. Editointiin vaikutti myös se, oliko kulloinkin käytössä Macintosh- vai Windows-tietokone ja kuinka paljon siinä oli käytettävissä välimuistia. Videoiden muokkaaminen vaatii tietokoneilta yllättävän paljon kapasiteettia.

Tutkimuksen suunnittelun, aineiston keräämisen ja analysoinnin ja tulosten raportoinnin aikana on ollut mahdollista tutustua moniin uusiin digitaalisiin tutkimuksentekovälineisiin ja oppia uutta eri tieteenaloja edustavilta yhteistyökumppaneilta. Digitaalisuus on tuonut tutkimuksen tekemiseen paljon päänvaivaa, mutta ehdottomasti kuitenkin sitä enemmän iloa ja hyötyä.

Lähteet

Boersma, Paul (2001). Praat, a system for doing phonetics by computer. Glot International, 5(9–10), 341–345.

Bölte, S., Ollikainen, R., Feineis-Matthews, S. & Poustka, F., 2013, Frankfurtin mallin mukainen tunteiden tunnistamisen testi ja harjoitteluohjelma. Tukholma: Karolinska Institutet, Center of Neurodevelopmental Disorders.

Huttunen, K., Huang, X. & Zhao, G. (käsikirjoitus valmisteilla). A new children's dynamic facial expressions' database and its baseline results based on human recognition and computer vision technology.

Huttunen, K., Hyvärinen, H., Laakso, M.-L., Parkas, R., & Waaramaa, T. (2015). Tunne-etsivät. Tietokoneohjelma tunteiden tunnistamisen harjoitteluun. Helsinki: Opetushallitus. Haettu: http://www.edu.fi/verkko_oppimateriaalit/tunne_etsivat sekä minipelit haettu: http://emo.oph.oodles.fi/minipelit.html

Huttunen, K., Kosonen, J., Waaramaa, T., & Laakso, M-L. (2018). Tunne-etsivät-pelin vaikuttavuus lasten sosioemotionaalisen kehityksen tukemisessa. Kelan tutkimusosaston julkaisusarja, Sosiaali- ja terveysturvan raportteja 2018:8. Helsinki: Kela. Haettu: https://helda.helsinki.fi/handle/10138/233957

Koivula, M., Huttunen, K., Mustola, M., Lipponen, S., & Laakso, M-L. (2017). The Emotion Detectives game: Supporting the social-emotional competence of young children. In M. Ma & A. Oikonomou (Eds.), *Serious Games and Edutainment Applications II* (pp. 29–53). Cham: Springer International Publishing.

Lipponen, S., Koivula, M., Huttunen, K., Turja, L., & Laakso, M-L. (2018). Children's peer interaction while playing the digital Emotion Detectives game. Journal of Early Childhood Education Research (JECER), 7(2), 282–309. Haettu: https://jecer.org/fi/wp-content/uploads/2018/12/Lipponen-Koivula-Huttunen-Turja-Laakso-issue7-2.pdf

Löytömäki, J., Ohtonen, P., Laakso, M-L. & Huttunen, K. (käsikirjoitus arvioitavana). The role of linguistic and cognitive factors in emotion recognition difficulties in children with ASD, ADHD or DLD.

Rodgers, J. D., Thomeer, M. L., Lopata, C., Volker, M. A., Lee, G. K., McDonald, C. A., Smith, R. A. & Biscotto, A. A. (2015). RCT of a psychosocial treatment for children with high-functioning ASD: Supplemental analyses of treatment effects of facial emotion encoding. Journal of Developmental and Physical Disabilities 27(2), 207–221. DOI 10.1007/s10882-014-9409-x

Vallenius, N. (2018). ”Hei sä voit kuljettaa ja mä painan tästä”. Lasten yhteistoiminnallinen vertaistyöskentely Tunne-etsivät-pelin aikana. Erityispedagogiikan pro gradu -tutkielma. Kasvatustieteiden laitos. Jyväskylä: Jyväskylän yliopisto. Haettu: https://jyx.jyu.fi/handle/123456789/58516

Kirjoitetun nykysuomen automaattisesta semanttisesta merkitsemisestä

Kimmo Kettunen
Kansalliskirjasto

Tiivistelmä
Tässä julkaisussa esitellään FiST, työn alla oleva kirjoitetun nykysuomen kokotekstien semanttinen merkitsin. FiSTin ensimmäinen versio perustuu vapaasti saatavilla oleviin osiin: 46 226 sanan semanttiseen leksikkoon (Löfberg, 2017; Multilingual USAS) sekä morfologisen analyysin ohjelmiin Omorfi ja FinnPos (Silfverberg ja kumppanit, 2016). FiSTin tämänhetkistä versiota on testattu systemaattisesti erilaisilla suomen aineistoilla, joista laajin on 45 miljoonan sanan elokuva- ja tv-tekstitysten OpenSubtitlesin osa-aineisto. FiST merkitsee teksteihin sanojen semanttisia luokkia noin 82–91 %:n sanastollisella kattavuudella (Kettunen, 2019). Toistaiseksi ohjelmasta puuttuu kaksi merkittävää osaa: semanttisesti monitulkintaisten sanojen käsittely (Navigli, 2009; Ragnato ja kumppanit, 2017; Robertson, 2019) sekä semanttisesta sanastosta puuttuvien yhdyssanojen kattava käsittely. Nykytilassaankin ohjelma tarjoaa toimivan työvälineen laajojen digitaalisten tekstiaineistojen tietokoneavusteiseen analyysiin sekä digitaalisten ihmistieteiden tutkijoille että muille tekstien sisältöä analysoiville.

1 Johdanto

Kirjoitetun nykysuomen kieliteknologisten analyysivälineiden saatavuutta voi pitää yleisesti ottaen hyvänä. Yksi oleellinen tekstiaineistojen analyysitapa on kuitenkin jäänyt vähälle huomiolle: suomea varten ei ole olemassa yleisesti saatavaa semanttisen eli kielen merkitystason analyysiohjelmaa. Suomen kielen semanttisten digitaalisten resurssien puutteesta huomautettiin jo META NET -raportissa (Koskenniemi ja kumppanit, 2012). Tilanne ei ole juurikaan parantunut sen jälkeen, vaikka joitakin semanttisia sanastoja on julkaistu.[1]

[1] En käsittele muita semanttisia leksikoita tai kielivarantoja tarkemmin, mutta on syytä kirjata, että ne sisältävät ainakin seuraavat: FinnWordnet (Linden ja Carlson, 2010; Linden ja Niemi, 2014), FrameNet (Linden ja kumppanit, 2017). Yleinen suomalainen ontologia (YSO, Hyvönen ja kumppanit, 2008) ei ole varsinainen semanttinen sanasto, mutta sillä voidaan myös tehdä tekstisisältöjen merkintää ja luokittelua karkeammin kuin semanttisella sanastolla (Hirst, 2004).

Vapaasti saatavia suomen kielen keskeisiä kieliteknologisia ohjelmia on olemassa tällä hetkellä hyvin morfologiseen ja syntaktiseen analyysiin, esimerkiksi Omorfi, [2] Voikko [3] ja FinnPos [4] morfologiaan ja Finnish depedency parser [5] lauseenjäsennykseen. FiNER-ohjelmistolla[6] voidaan tunnistaa ja merkitä erisnimiä ja muita niiden kaltaisia elementtejä, ja se on joukon ainoa semanttisesti suuntautunut työvälineohjelma. Toistaiseksi ei kuitenkaan ole olemassa ainoatakaan vapaasti saatavaa suomenkielisten kokotekstien kattavaa semanttista merkintää tekevää ohjelmaa, semanttista taggeria. Voikin todeta, että suomen kielen automaattiseen semanttiseen käsittelyyn on jäänyt jos ei aivan tyhjiö, niin kuitenkin suuri aukko.

2 Semanttinen merkintä

Tekstien semanttinen merkintä (tai annotaatio) määritellään tässä yhteydessä prosessiksi, jossa annetun tekstin sanojen merkitys tunnistetaan ja merkitään jonkin semanttisen luokitusmallin mukaan. Semanttista merkintää voidaan tehdä eri tavoin, esimerkiksi koneoppimiseen perustuen tai laajoja merkityssanastoja hyväksi käyttäen. FiST käyttää semanttiseen merkintään laajaa suomen kielen merkityssanastoa, joka on julkaistu vuonna 2016 (Löfberg, 2017; Multilingual USAS).

Suomen kielen julkaistu semanttinen sanasto pohjautuu työhön, jota on tehty Lancasterin yliopistossa englannin kielen semanttisen merkitsimen ja sen sanaston kanssa. Yliopiston korpustutkimuskeskuksessa (UCREL) kehitettiin 1990-luvulla Longman Lexicon of Contemporary English -sanakirjan (McArthur, 1981) pohjalta oma semanttinen luokitusjärjestelmä, UCREL Semantic Analysis System (USAS).[7] Tämä luokitus perustuu semanttisten kenttien ajatukselle: sanasto jaetaan

---

Haverinen (2014) käsittelee suomen kielen lauseiden semanttisten roolien merkintää. BabelNet (https://babelnet.org/, Navigli ja Ponzetto, 2012) puolestaan on laaja monikielinen ensyklopedinen tietokanta, jossa on mukana myös suomen kieli. Euroopan parlamentin monikielisiin rinnakkaisteksteihin perustuen on myös julkaistu Eurosense-kielivaranto, johon on merkitty sanojen merkitykset (Bovi ja kumppanit, 2017).

[2] https://github.com/flammie/omorfi

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

[4] https://github.com/mpsilfve/FinnPos

[5] https://github.com/TurkuNLP/Turku-neural-parser-pipeline

[6] https://korp.csc.fi/download/finnish-tagtools/v1.3/

[7] http://ucrel.lancs.ac.uk/usas/

aihealueisiin tai kenttiin, joiden alle sanojen merkitysluokat sijoitetaan. USASissa on 21 ylemmän tason semanttista luokkaa, jotka sisältävät 232 semanttista luokkaa tai kategoriaa. Taulukossa 1 on esitetty esimerkkinä yhdestä merkitysluokasta luokka I, *raha ja liiketoiminta* (Löfberg, 2017; Multilingual USAS).

Taulukko 1. USASin semanttinen luokka Raha ja liiketoiminta.

I Raha & liiketoiminta
I1 Raha (yleiskategoria)
I1.1 Raha: varakkuus
I1.2 Raha: velka
I1.3 Raha: hinta
I2 Liiketoiminta
I2.1 Liiketoiminta (yleiskategoria)
I2.1 Liiketoiminta: myyminen
I3 Työ ja työllisyys (yleiskategoria)
I3.1 Työ ja työllisyys: ammattimaisuus
I4 Teollisuus

Esimerkiksi kaikki valuutat on luokiteltu sanastossa luokkaan I1: dinaari Noun I1, dirhami Noun I1, dobra Noun I1, dollari Noun I1, dong Noun I1, drakma Noun I1, dram Noun I1, ecu Noun I1.

Aakkosnumeeriset lyhenteet merkitysluokkien edessä ovat sanaston käyttämiä varsinaisia semanttisia merkitsimiä. Piaon ja kumppaneiden (2005) mukaan semanttisten luokkien syvyys on rajoitettu kolmeen, koska tämän on havaittu olevan parhaiten toimiva.

Semanttista kenttää voi pitää teoreettisena konstruktiona, joka yhdistää sanoja, jotka ovat merkitykseltään jollain lailla samankaltaisia, liittyvät yhteiseen käsitteeseen (Wilson ja Thomas, 1997). Hieman toisin sanoen niitä yhdistää jokin yhteinen semanttinen komponentti (Dullieva, 2017; Geeraerts, 2010; Lutzeier, 2006).

USASin käyttämät 21 semanttista luokkaa[8] on esitelty taulukossa 2.

[8] https://github.com/UCREL/Multilingual-USAS/tree/master/Finnish

Taulukko 2. USASin semanttiset luokat.

A YLEISET & ABSTRAKTIT SANAT
B KEHO & IHMINEN
C TAIDE & KÄSITYÖ
E TUNNE-ELÄMÄ & MIELENTILAT
F RAVINTO & MAATALOUS
G HALLINTO, POLITIIKKA & LAKI
H ARKKITEHTUURI, RAKENNUKSET & KOTI
I RAHA & LIIKETOIMINTA
K VIIHDE, URHEILU & PELIT
L ELÄMÄ & ELOLLINEN LUONTO
M LIIKKUMINEN, SIJAINTI, MATKAILU & KULJETUS
N NUMEROT & MITTAAMINEN
O AINEET, ESINEET & TARVIKKEET
P KOULUTUS
Q KIELELLISET TOIMINNOT & PROSESSIT
S SOSIAALISET TOIMINNOT & PROSESSIT
T AIKA
W MAAILMA & YMPÄRISTÖ
X PSYKOLOGISET TOIMINNOT, TILAT & PROSESSIT
Y TIEDE & TEKNOLOGIA
Z NIMET & KIELIOPILLISET SANAT

2.1 Suomen kielen semanttisen sanaston koostumus

Suomen kielen semanttisen sanaston koostaminen on aloitettu 2000-luvun alkuvuosina EU:n rahoittamassa Benedict-projektissa (Löfberg ja kumppanit, 2005). Löfberg (2017) esittelee väitöskirjassaan yksityiskohtaisesti suomen kielen semanttisen sanaston koostamisen periaatteet, kuvaa sanastoa sekä evaluoi Kielikoneen Benedict-projektissa tuottamaa semanttista merkintäohjelmaa (FST). Merkityssanastossa on 46 226 lekseemiä tai leksikaalista kuvausta. Sanoista noin 58 % on substantiiveja, 7 % verbejä, 17 % erisnimiä, 7 % adjektiiveja ja 7 % adverbeja. Loppu osa sanoista kuuluu pieniin suljettuihin sanaluokkiin. Sanojen jakaantuminen eri semanttisiin luokkiin on kuvattu Löfbergin teoksen sivulla 139 taulukossa 7 (Löfberg, 2017). Tässä riittää todeta, että sanaston semanttinen jakauma vaikuttaa tasapuoliselta. Selvää on, että tämäntyyppisestä sanastosta puuttuu sanoja tai jonkin merkitysluokan rakenne tai sisältö voi olla sopimaton tai puutteellinen tiettyyn käyttötarkoitukseen, mutta yleisenä suomen kielen semanttisena sanastona sanasto on kattava.

USAS-tyyppisiä sanastoja on tähän mennessä julkaistu 12 eri kielelle (Piao ja kumppanit, 2016).[9] Suomen sanastoa voi pitää yhtenä parhaista: se on koostettu leksikografisena käsityönä käyttäen eri lähteitä. Useiden muiden kielten vastaavat sanastot on tuotettu ainakin osin kääntämällä englanninkielinen USAS-sanasto automaattisesti tai puoliautomaattisesti.

2.2 FiSTin rakenne ja analyysin tulos

FiSTin tämänhetkisen version rakenne on yksinkertainen. Kokonaisuus käyttää vapaasti saatavia suomen kielen morfologisia ohjelmistoja Omorfia ja FinnPosia tekstien sanojen perusmuotoistamiseen ja morfologisesti yksitulkintaisen muodon valintaan (ks. kaavakuvaa Kettunen, 2019). Morfologisen vaiheen jälkeen sanat ovat perusmuotoisia ja niitä voidaan etsiä semanttisesta leksikosta. Jos sana löytyy semanttisesta leksikosta, se merkitään sanaston antamilla leimoilla: sanaluokalla sekä yhdellä tai useammalla semanttisella merkitsimellä. Jos sanaa ei ole semanttisessa leksikossa, sille annetaan merkintä Z99 (tuntematon) ja sanaan liitetään sen morfologinen tieto, jos se on saatavilla. Pentti Saarikosken esikoisrunokokoelman *Runoja* avausrunon ensimmäinen säkeistö analysoituisi taulukossa 3 kuvatusti.

[9] Kielten lista on seuraava: arabia, espanja, hollanti, italia, kiina, malaiji, portugali, tšekki, venäjä, urdu ja wales (kymri). Leksikoiden koot vaihtelevat 1 800:sta 64 800 sanaan. Suomen kielen sanasto on kolmanneksi laajin, sitä suurempia ovat malaijin ja kiinan sanastot. Kahdeksalle kielistä on olemassa semanttinen merkintäohjelma. Tiettävästi myös muille kielille on tekeillä sanastoja.

Taulukko 3. FiSTin analyysi Saarikosken runon säkeistöstä.

Syöte	FiSTin tuotos	Selitys
Taivas	taivas Noun W1 S9 Z4	Substantiivi, jolla on kolme semanttista merkintää. Ensimmäinen, maailman-kaikkeuteen viittaava, on oikea.
on	olla Verb A3+ A1.1.1 M6 Z5	Verbi, jolla on neljä semanttista merkintää. Olemassaoloon viittaava ensimmäinen merkitsin on oikea.
paperia	paperi Noun O1.1 Q1.2 B4 P1/Q1.2	Substantiivi, jolla on neljä semanttista merkintää. Kiinteään aineeseen viittaava ensimmäinen merkitsin olisi paras valinta.
,	PUNCT	Välimerkki
paperia	paperi Noun O1.1 Q1.2 B4 P1/Q1.2	Substantiivi, jolla on neljä semanttista merkintää. Kiinteään aineeseen viittaava ensimmäinen merkitsin olisi paras valinta.
maa	maa Noun M7	Alueisiin viittaava yksikäsitteinen merkitys.
.	PUNCT	Välimerkki

2.3 Semanttisesti merkityn tuotoksen evaluaatio

Suomen kielen semanttisen merkinnän evaluoimiseen ei ole olemassa yhtään standardiaineistoa,[10] johon FiSTin merkintöjä voisi verrata. Niinpä tässä vaiheessa voi vain analysoida FiSTin sanaston kattavuutta analysoimalla sillä erilaisia aineistoja ja laskemalla sanastollisen kattavuuden analyyseista. Taulukossa 4 on esitetty kahdeksan erilaisen nykysuomen aineiston analyysin kattavuustulokset. Vertailun vuoksi mukana on myös morfologisen tunnistuksen antama kattavuus.

[10] Kieliteknologiassa puhutaan yleensä kultakorpuksesta tai kultastandardista englanniksi. Tällä tarkoitetaan aineistoa, johon on tehty jonkin kielen tason merkintä, joka on huolellisesti tarkastettu. Uusien merkintäohjelmistojen toimintaa voidaan evaluoida suhteessa tällaiseen aineistoon (vrt. esim. Kilgariff, 1998). Bovi ja kumppanit (2017) ovat tuottaneet monikielisen EuroSense-kielivarannon Euroopan parlamentin istuntopöytäkirjoista. Merkitysten kuvaus tässä kielivarannossa ei kuitenkaan ole yhteensopiva FiSTin käyttämän USAS-tyylisen merkityskuvauksen kanssa.

Käytetyssä morfologisessa analyysiohjelmassa Omorfissa on melkein kymmenen kertaa semanttista sanastoa suurempi sanasto, 424 259 sanaa (Pirinen, 2015).[11]

Taulukko 4. FiSTin ja Omorfin tuntemien sanojen määrät kahdeksassa eri aineistossa.

Aineisto	Tekstilaji	Saneita	Omorfin tunnistusprosentti	FiSTin tunnistusprosentti
Suomi24-aineistoa	Nettikeskustelua	494 000	92 %	81,9 %
Miljoona lausetta Leipzigin korpuksen uutisaineistoa[12]	Netistä automaattisesti koottua uutisaineistoa	7 295 230	95,9 %	84,9 %
Elokuvien ja tv-ohjelmien tekstityksiä[13]	Tekstitys	45 204 076	96,5 %	90,9 %
Europarl v6[14]	EU-parlamentin pöytäkirjat	28 600 000	---------	90,9 %
Finnish treebank[15]	Ison suomen kieliopin esimerkkilauseet	138 949	99,2 %	90,2 %

Taulukosta 4 ilmenee, että FiSTin sanastollinen kattavuus erilaisilla kirjoitetun nykysuomen aineistoilla on hyvä, 82:n ja 91 prosentin välillä. Matalimman sanastollisen kattavuuden saavat kaksi ensimmäistä aineistoa. Niistä Suomi24-aineisto on tyypiltään erilaista kuin muut: se muodostuu verkkokeskusteluista, joissa käytetään paljon epämuodollisempaa kieltä ja enemmän vieraskielisiä sanoja. Leipzigin korpuksen lehtiaineistot on koottu verkosta automaattisesti, joten ne sisältävät myös enemmän hälyä (Quasthoff ja kumppanit, 2006). Loput kolme aineistoa ovat laadullisesti parempia, joten niiden saama sanastollinen kattavuuskin on parempi.

Aiemmin olen esittänyt myös tuloksia FiSTin peruskäyttöalueen ulkopuolisen aineiston analyyseista (Kettunen, 2019). Analysoin muun muassa 1800-luvun

[11] Laskentakaava FiSTin kattavuudelle on seuraava: (100* (1-(puuttuva merkitsin/(NR-välimerkki-numero)))). Puuttuvat merkitsimet ovat FiSTille tuntemattomia sanoja (Z99). Välimerkit ja numerot vähennetään sanojen kokonaismäärästä (NR). Evaluaation syöte on yksi merkitty sana rivillä, eikä aineistossa ole tyhjiä rivejä.
[12] http://wortschatz.uni-leipzig.de/en/download/
[13] http://opus.nlpl.eu/OpenSubtitles2016.php; Lison & Tiedeman (2016).
[14] http://www.statmt.org/europarl/archives.html
[15] https://universaldependencies.org/

lopun lehtitekstejä, suomalaisten kaunokirjailijoiden tuotantoa 1800- ja 1900-luvun vaihteesta sekä Raamatun käännöstä vuodelta 1938. Sekä Raamatun käännös että kaunokirjalliset tekstit saivat korkean sanastollisen kattavuuden (noin 90 %), ja vanhojen lehtitekstien kattavuus vaihteli 69:n ja 84 prosentin välillä. Lisäksi olen analysoinut eduskunnan julkaisemaa asiakirja-aineistoa vuosilta 1907–2000[16] sekä tweet-aineistoa (Laitinen ja kumppanit, 2018). Niissä sanastollinen kattavuus jää matalammaksi. Eduskunta-aineistoissa saavutetaan eri vuosikymmenillä vaihtelevasti noin 55–78 prosentin sanastollinen kattavuus, tweeteissä noin 65 prosentin kattavuus pienellä noin 33 000 sanan otoksella. Eduskunnan aineistoissa matala kattavuus johtuu monesta tekijästä: optisen luvun tekstiin tuomista virheistä, yhdyssanoista, runsaista erisnimistä, katkenneista sanoista sekä lyhenteistä (Kettunen ja La Mela, 2019). Tweeteissä tunnistuksen mataluus johtuu pääasiassa kielenkäytön erilaisesta luonteesta.

2.4 Semanttisen merkinnän käyttökohteet

Englannin kielen semanttista merkintäohjelmaa on käytetty muun muassa tyylintutkimuksessa, korpusanalyyseissa, erilaisten diskurssien tutkimuksessa, nettikeskustelujen analyysissa, sentimenttianalyysissa, ontologioiden oppimisessa, poliittisen kielen tutkimuksessa jne.,[17] mikä kertoo sen, että tämäntyyppisen analyysivälineen käyttötarkoitukset voivat olla hyvin moninaisia. Ensimmäinen FiSTiä käyttävä digitaalisten ihmistieteiden tutkimus, Kettunen ja La Mela (2019), tutkii eduskunnan valtiopäiväasiakirjoja ja niistä erityisesti jokamiehen oikeuksia käsittelevää keskustelua kolmella vuosikymmenellä. Tästä tutkimuksesta saatujen kokemusten mukaan FiST on osoittautunut hyödylliseksi analyysivälineeksi. Tutkimus on tehty melko pienellä, noin 44 000 saneen aineistolla, joka on poimittu ja koostettu ensin käsin eduskunnan pöytäkirjojen jokamiehen oikeuksia käsittelevästä keskustelusta tekstihaun osumien perusteella eri vuosikymmeniltä. Tämän koosteaineiston lisäksi on tutkittu FiSTillä merkittyjä eduskunnan pöytäkirjoja kolme vuotta ennen ja jälkeen jokamiehen oikeuksista käydyn

[16] http://avoindata.eduskunta.fi/digitoidut/download. La Melan (2019) morfologisten analyysien perusteella näyttäisi siltä, että eduskunta-aineiston digitoinnin laatutaso olisi korkeampi. Erot analyyseissa johtuvat siitä, että morfologinen tunnistus tunnistaa myös erilaisia merkkijonoja, jotka eivät ole kokonaisia sanoja. Lisäksi morfologinen tunnistus kattaa yhdyssanat ja erisnimet paremmin.
[17] http://ucrel.lancs.ac.uk/wmatrix/#apps

keskustelun 1940-, 1970- ja 1990-luvuilla. Tutkimuksessa pystyttiin havaitsemaan jokamies-käsitettä käytetyn 1940-luvulla lähinnä kalastusoikeuksiin liittyvässä keskustelussa. 1970-luvulla käsitettä käytettiin yleisemmin luonnon virkistyskäyttöön liittyvässä keskustelussa, ja 1990-luvulla käsitettä käytettiin yhä laajemmin myös luontoteeman ulkopuolella, esimerkiksi Suomen EC-jäsenyydestä keskusteltaessa.

3 Kehitystarpeet

Kettunen (2019) käsittelee FiSTin tämänhetkisiä puutteita kattavasti, joten puutteet listataan tässä vain lyhyesti. Kaksi oleellista puutosta FiSTissä on yhdyssanojen osien käsittely ja monimerkityksisten sanojen merkityksen yksikäsitteistämisen puuttuminen (disambiguointi, Edmonds, 2006; Navigli, 2009; Robertson, 2019). Yhdyssanojen muodostaminen on suomen kielessä runsasta, ja mikään sanasto ei voi sisältää kaikkia yhdyssanoja. Semanttisessa sanastossa on tuhansia keskeisiä yhdyssanoja, mutta sanastoon sisältymättömät yhdyssanat tulisi analysoida niiden osien kautta. FinnPos ei kuitenkaan palauta morfologisesti analysoituja sanoja yhdyssanojen osat eroteltuina,[18] joten FiST ei toistaiseksi kykene analysoimaan semanttisesta sanastosta puuttuvia yhdyssanoja. Muut morfologiset analyysiohjelmat, kuten Omorfi ja Voikko, osittavat yhdyssanat osiinsa, mutta eivät tee morfologista disambiguointia, joten niidenkään käyttö ei auta. Suomen kielen dependenssianalyysia tekevä jäsennin osaa palauttaa osan yhdyssanoista osiinsa pilkottuina, mutta se kattavuus yhdyssanojen osittamisessa on melko vajavainen.

Sanojen merkityksen monitulkintaisuuden käsittelyn puuttuminen on toinen FiSTin puutos. Suomen kielen semanttinen sanasto kuvaa monitulkintaiset sanat yksinkertaisesti. Esimerkiksi sanalle *huone* annetaan kuvaus *huone Noun H2 S9.* Sana on substantiivi, jolla on kaksi merkitystulkintaa: rakennuksen osa tai joissain asiayhteyksissä, erityisesti astrologiassa, yliluonnollinen tulkinta. Semanttisen sanaston periaate on merkitä sanan yleisin merkitys ensimmäiseksi (Löfberg, 2017: 74). Lisäksi semanttisessa sanastossa käytetään toistakin tapaa merkitä sanan kuuluminen useaan mahdolliseen merkitysluokkaan. Taulukossa 3 annettiin sanalle *paperi* seuraava merkitystulkinta: *paperi Noun O1.1 Q1.2 B4 P1/Q1.2* Viimeisessä merkityksessä käytetään vinoviivaa (slash tag, portmanteau tag) kertomaan, että

[18] Kommentti perustuu aineistojen ajojen aikaiseen tilanteeseen syksyllä 2018 ja alkutalvesta 2019. FinnPosin uusimmassa Myllyssä toimivassa versiossa on osittainen yhdyssanojen osien merkintä toukokuussa 2019.

sana voi kuulua useaan luokkaan: tässä tapauksessa koulutukseen sekä kirjallisiin dokumentteihin tai kirjoittamiseen. Semanttinen sanasto sisältää tällä hetkellä 7 791 sanaa, jotka ovat monimerkityksisiä. 10 556 sanaa on merkitty vinoviivalla. Kaksi merkitystä on 70,3 prosentilla monimerkityksisistä sanoista, kolme merkitystä 18,6 prosentilla. Vinoviivalla merkityistä sanoista 85 prosentilla on yksi lisämerkitys.

Ilman semanttista disambiguointia joudumme siis toistaiseksi olettamaan, että FiSTin ensimmäinen semanttinen merkintä olisi oikea. Tätä pidetään semanttisen monitulkintaisuuden kirjallisuudessa ns. yleisimmän merkityksen lähtötilanteena (baseline), johon erilaisia disambiguointistrategioita voidaan verrata. Erilaisia merkityksen yksikäsitteistämisen tapoja arvioitaessa on todettu, että sanan yleisimmän merkityksen antama lähtötilanne on melko vaikeasti ylitettävä, joten FiSTin kannalta ensimmäisen merkityksen käyttäminen lähtökohtana on hyvin perusteltu. (Ragnato ja kumppanit, 2019; Robertson, 2019.)

Voimme tutkia FiSTin ensimmäisen semanttisen merkitsimen oikeellisuutta pienellä vertailulla. Koska suomelle ei ole olemassa sopivaa semanttista evaluaatiokokoelmaa, joudumme tekemään vertailun kahden ohjelman tekemien merkintöjen välillä. FiSTin rinnalla toisena ohjelmana käytämme Kielikoneen semanttista merkitsintä, jossa on kevyt semanttinen disambiguointi (Löfberg, 2017: 180). Se käyttää sanaluokkamerkitsimen sanaluokkatietoa sekä semanttisen sanaston ensimmäistä merkintää. Löfberg on käynyt lävitse Kielikoneen merkitsimen merkinnöistä noin 10 000 sanaa viidestä eri aineistosta. Analyysin tuloksen mukaan Kielikoneen merkitsin on merkinnyt oikean semanttisen merkitsimen noin 79,5–83,5 prosenttiin sanoista.

Tämä tieto lähtökohtana voimme verrata Kielikoneen merkitsimen tekemää merkintää FiSTin merkintään ja analysoida eroja. Analysoitavaksi valittiin Finnish treebankin aineisto. Siitä analysoitiin 102 050 sanaa molemmilla ohjelmilla. Analyysin mukaan FiST analysoi 85,1 prosenttia aineistosta samalla tavoin kuin Kielikoneen semanttinen merkitsin. Eroja analyyseissa oli 15 176 sanassa. Löydetyistä eroista 5 812 (38,3 %) on sellaisia, joissa FiST merkitsee sanaan leiman Z99, tuntematon, ja Kielikoneen merkitsin tunnistaa sanan. Eroavien analyysien joukon molempien tunnistamasta 9 364 sanasta FiST merkitsee 1 674:ään (17,9 %) sanaan ensimmäisen semanttisen merkitsimen samoin kuin Kielikoneen merkitsin, mutta lopuilta osin analyysit voivat erota.

FiSTille tuntemattomista sanoista 4 173 (71,8 %) on yhdyssanoja, joita FiST ei ole kyennyt analysoimaan. Loppujen erojen analyysi on hankalaa, mutta näyttää

siltä, että paikoittain FiSTin käyttämä ajantasaisempi morfologinen analyysi tuottaa tuloksia ja FiST analysoi sanoja, jotka jäävät Kielikoneen merkitsimeltä tunnistamatta. Kielikoneen merkitsimeltä jää tunnistamatta 801 sanaa, joille FiST antaa analyysin. Näin ollen noin viisi prosenttia analyysien eroista on mahdollisia positiivisia osumia FiSTille.

Analyysi on tehty vain yhdellä melko pienellä aineistolla, joten sitä voi pitää vain suuntaa antavana. Vaikuttaa kuitenkin siltä, että yhdyssanojen puutteellinen analyysi vaikuttaa FiSTin analyysiin jo melko paljon: 27,5 prosenttia analyysitta jääneistä sanoista oli yhdyssanoja. Semanttisen monimerkityksisyyden vaikutuksen arviointi analyysien eroihin vaatisi yksityiskohtaista perehtymistä loppuihin analyysien eroihin.

Yhteenvetona voi sanoa, että FiSTin sanastollinen kattavuus erilaisten aineistojen analyysissa on hyvä. Semanttinen merkintä toimii erityyppisissä teksteissä, myös vanhemmassa aineistossa, jota varten sanastoa ei ole luotu. Ohjelma vaatii kehittämistä, mutta sen tämänhetkisellä versiolla voi analysoida tekstejä, ja analysoitujen tekstien käyttötarkoitus voi olla hyvin monenlainen. FiSTin käyttökelpoisuuden lisäselvittäminen olisi paikallaan myös muilla aineistoilla. Ohjelman jatkokehitykselle olisi oleellista, että suomen kielen semanttista merkintää varten tuotettaisiin vertailukelpoisia evaluaatioaineistoja, joilla tekstien merkinnän oikeellisuutta voisi arvioida systemaattisesti.

Lähdeluettelo

Bovi, D. C., Camacho-Collados, J., Raganato, A. & Navigli, R. (2017). EUROSENSE: Automatic harvesting of multilingual sense annotations from parallel text. Teoksessa *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics Volume 2: Short Papers*, 594–600.

Dullieva. K. (2017). Semantic Fields: Formal Modelling and Interlanguage Comparison. *Journal of Quantitative Linguistics*, 24:1, 1–15. DOI: 10.1080/09296174.2016.1239400.

Edmonds, P. (2006). Disambiguation. Teoksessa Allan, K. (ed.), *Concise Encyclopedia of Semantics*, 223–239. Oxford: Elsevier.

Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Haverinen, K. (2014). *Natural Language Processing Resources for Finnish Corpus Development in the General and Clinical Domains.* TUCS Dissertations No 179. https://www.utupub.fi/bitstream/handle/10024/98608/TUCSD179Dissertation.pdf?sequence=2&isAllowed=y.

Hirst, G. (2004). Ontology and the lexicon. Teoksessa Staab, S., Studer, R. (eds.), *Handbook on Ontologies.* International Handbooks on Information Systems. Springer, Berlin.

Hyvönen, E., Viljanen, K., Tuominen, J. & Seppälä, K. (2008). Building a National Semantic Web Ontology and Ontology Service Infrastructure – The FinnONTO Approach. Teoksessa Bechhofer, S., Hauswirth, M., Hoffmann, J., Koubarakis, M. (eds.), *The Semantic Web: Research and Applications. ESWC 2008. Lecture Notes in Computer Science, vol 5021.* Springer: Berlin.

Kettunen, K. (2019). FiST – towards a Free Semantic Tagger of Modern Standard Finnish. IWCLUL2019, http://aclweb.org/anthology/W19-0306.

Kettunen, K. & La Mela, M. (2019). Digging Deeper into Finnish Parliamentary Protocols 1907-2000 – Using a Lexical Semantic Tagger for Analysis of Parliamentary Protocols' Discussion of Everyman's Rights (arvioitavana).

Kilgarriff, A. (1998). Gold standard datasets for evaluating word sense disambiguation programs. *Computer Speech & Language,* 12(4), 453–472.

Koskenniemi, K. et al. (2012). The Finnish Language in the Digital Age. META NET White paper series. http://www.meta-net.eu/whitepapers/e-book/finnish.pdf/view?searchterm=Finnish.

La Mela, M. (2019). Tracing the Emergence of Nordic Allemansrätten through Digitized Parliamentary Sources. Teoksessa Fridlund, M., Paju, P., Oiva, M. (eds.), *Digital, Computational and Distant Readings of History: Emergent approaches within the new digital history* (ilmestyy).

Laitinen, M., Lundberg, J., Levin, M. & Martins, R. (2018). The Nordic tweet stream: A dynamic real-Time monitor corpus of big and rich language data. Teoksessa *3rd Conference on Digital Humanities in the Nordic Countries.*

Lindén, K. & Carlson, L. (2010). FinnWordNet – WordNet på finska via översättning. *LexicoNordica – Nordic Journal of Lexicography*, 17, 119–140.

Lindén, K. & Niemi, J. (2014). Is it possible to create a very large wordnet in 100 days? An evaluation. *Language Resources and Evaluation,* 48(2), 191–201.

Lindén, K., Haltia, H., Luukkonen, J., Laine, A. O., Roivainen, H., & Väisänen, N. (2017). FinnFN 1.0: The Finnish frame semantic database. *Nordic Journal of Linguistics*, 40(3), 287–311.

Lison, P. & Tiedemann, J. (2016). OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. Teoksessa *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*.
Lutzeier, P. R. (2006). Lexical fields. Teoksessa Allan, K. (ed.), *Concise Encyclopedia of Semantics*, 470–473. Oxford: Elsevier.
Löfberg, L. (2017). Creating large semantic lexical resources for the Finnish language. Lancaster University. http://www.research.lancs.ac.uk/portal/en/publications/creating-large-semantic-lexical-resources-for-the-finnish-language(cc08322c-f6a4-4c2b-8c43-e447f3d1201a)/export.html.
Löfberg, L., Piao, S., Rayson, P., Juntunen, J.-P., Nykänen, A. & Varantola., K. (2005). A semantic tagger for the Finnish language. http://eprints.lancs.ac.uk/12685/1/cl2005_fst.pdf.
McArthur, T. (1981). *Longman Lexicon of Contemporary English*. Longman, London.
Multilingual USAS. https://github.com/UCREL/Multilingual-USAS.
Navigli, R. (2009). Word Sense Disambiguation: A Survey. *ACM Computing Surveys*, 41, 10–69.
Navigli, R. & Ponzetto, S. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193, 217–250.
Piao, S., Archer, D., Mudraya, O., Rayson, P., Garside, R., McEnery, T. & Wilson, A. (2005). A Large Semantic Lexicon for Corpus Annotation. Teoksessa *Proceedings of the Corpus Linguistics 2005 conference, July 14-17, Birmingham, UK*. Proceedings from the Corpus Linguistics Conference Series on-line e-journal, Vol. 1, no. 1.
Piao, S. et al. (2016). Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages. Teoksessa *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC2016), Portoroz, Slovenia*, 2614–2619.
Pirinen, T. A. (2015). 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,* vol 28, 381–393. http://www.linguistics.fi/julkaisut/SKY2015/SKYJoL28_Pirinen.pdf.
Quasthoff, U., Richter, M. & Biemann, C. (2006). Corpus portal for search in monolingual corpora. Teoksessa *Proceedings of the fifth international conference on Language Resources and Evaluation, LREC 2006, Genoa*, 1799–1802.

Ragnato, A., Collados, J.-C. & Navigli, R. (2017). Word Sense Disambiguation: a Unified Evaluation Framework and Empirical Comparison. Teoksessa *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 99–110, Valencia, Spain, April 3–7, 2017.

Robertson, F. (2019). A Contrastive Evaluation of Word Sense Disambiguation Systems for Finnish. IWCLUL2019. http://aclweb.org/anthology/W19-0304.

Silfverberg, M., Ruokolainen, T., Lindén, K. & Kurimo, M. (2016). FinnPos: an open-source morphological tagging and lemmatization toolkit for Finnish. *Lang Resources & Evaluation*, 50, 863–878. https://doi.org/10.1007/s10579-015-9326-3.

Wilson, A. & Thomas, J. (1997). Semantic annotation. Teoksessa Garside, R., Leech, G., McEnery, T. (eds.), *Corpus annotation: Linguistic information from computer text corpora*, 53–65. Longman: New York.

Oulun yliopisto • Humanistinen tiedekunta
University of Oulu • Faculty of Humanities

ISBN 978-952-62-2320-9
ISSN 1796-4725

KANSI RAIMO AHONEN