La Mela, Matti; Tamper, Minna ; Kettunen, Kimmo

Finding Nineteenth-century Berry Spots

Published in:
DHN 2019 - Digital Humanities in the Nordic Countries

Published: 01/01/2019

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

Please cite the original version:
La Mela, M., Tamper, M., & Kettunen, K. (2019). Finding Nineteenth-century Berry Spots: Recognizing and Linking Place Names in a Historical Newspaper Berry-picking Corpus. In C. Navarretta, M. Agirrezabal, & B. Maegaard (Eds.), *DHN 2019 - Digital Humanities in the Nordic Countries : Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, Copenhagen, Denmark, March 5-8, 2019* (pp. 295-307). (CEUR Workshop Proceedings; Vol. 2364). CEUR. http://ceur-ws.org/Vol-2364/

Finding Nineteenth-century Berry Spots: Recognizing and Linking Place Names in a Historical Newspaper Berry-picking Corpus

Matti La Mela[1[0000-0003-0340-9269]], Minna Tamper[1[0000-0002-3301-1705]], Kimmo Kettunen[2[0000-0003-2747-1382]]

[1] Semantic Computing Research Group (SeCo), Aalto University, Finland
firstname.secondname@aalto.fi
[2] The National Library of Finland
firstname.secondname@helsinki.fi

Abstract. The paper studies and improves methods of named entity recognition (NER) and linking (NEL) for facilitating historical research, which uses digitized newspaper texts. The specific focus is on a study about historical process of commodification. The named entity detection pipeline is discussed in three steps. First, the paper presents the corpus, which consists of newspaper articles on wild berry picking from the late nineteenth century. Second, the paper compares two named entity recognition tools: the trainable Stanford NER and the rule-based FiNER. Third, the linking and disambiguation of the recognized places is explored. In the linking process, information about the newspaper publication place is used to improve the identification of small places.

The paper concludes that the pipeline performs well for mapping the commodification, and that specific problems relate to the recognition of place names (among named entities). It is shown how Stanford NER performs better in the task (F-score of 0.83) than the FiNER tool (F-score of 0.68). Concerning the linking of places, the use of newspaper metadata appears useful for disambiguation between small places. However, the historical language (with its OCR errors) recognized by the Stanford model poses challenges for the linking tool. The paper proposes that other information, for instance about the reuse of the newspaper articles, could be used to further improve the recognition and linking quality.

Keywords: Historical newspapers, Named Entity Recognition, Named Entity Linking, Berry picking, Commodification

1 Introduction

Berry picking has been a common pastime in the Nordic countryside for centuries. Wild berries have been picked for personal consumption, but also for local trade and for the national exporting industries. The locations of good berry spots are something foragers keep to their own knowledge. In this paper, we want to identify place names in a historical

nineteenth-century newspaper corpus, which does not only regard concrete berry spots, but a wide range of locations from export destinations to local market places. The aim of the paper is to test and improve methods of named entity recognition and linking to discover these locations from a large text corpus.

In the paper, we compare two named entity recognition tools—the trainable Stanford NER[1] and the rule-based FiNER[2]—, and link the recognized place names by using the ARPA linking tool [1] and newspaper metadata. The method pipeline is being developed for an actual research case, which uses Finnish historical newspaper articles and studies the commodification of nature during an export boom of lingonberries in the late nineteenth century [2]. The research case employs place names for studying the developing export networks and the geography of local conflicts concerning wild berries. Automated named entity recognition and linking is very useful, while the newspaper material about berry picking is large and it is not possible to go through it manually. Moreover, the linking will enable to derive relevant information from other databases, for instance, about the recognized places' geographic location.

At the same time, the historical research case helps to understand what the methodological challenges concerning named entities, their recognition and linking are. The paper presents a method pipeline where place names are identified in a historical newspaper research corpus. The named entity recognition tools have been previously evaluated with the Finnish historical newspaper data [3], and the results we obtain are comparable to studies with similar French and Dutch data (analyzed with Stanford NER) [4]. Moreover, we use the ARPA tool in the paper to link named entities in historical newspapers, and enhance the disambiguation of potential links with our solution to make use of geographic ontology hierarchies and newspapers' publication place information.

In the paper, we will present and discuss the three steps of the pipeline. In section two, the paper presents the berry corpus and named entity recognition that has been done for the historical newspaper data. The paper shows how the quality of recognition remains adequate with the recognition methods included in the pipeline. In the third section, the focus is on named entity linking. The aim is to show how well the identified place names can be linked to other databases, for example, to retrieve

[1] https://nlp.stanford.edu/software/CRF-NER.shtml
[2] https://korp.csc.fi/download/finnish-tagtools/v1.1/

coordination information. Finally, in the last section, we will discuss the results from the perspective of the research project.

2 Recognizing Place Names in a Corpus of Nineteenth Century Newspaper Articles

Our berry-picking corpus has been collected from the digital historical newspaper corpus of The National Library of Finland, known also as Digi[3]. This collection contains over 14 million digitized pages of newspapers and journals published in Finland since 1771. The open part of the corpus, 1771-1929, consists of ca. 7.45 million pages mainly in Finnish and Swedish.

The berry-picking corpus consists of a total of 303 historical newspaper articles (42 179 word tokens) from the late nineteenth century.[4] The articles include local, national and international news about wild berry picking: children lost in berry woods, exports of wild berries, industrial visions or reports from local market places. In the late nineteenth century, a lingonberry boom developed in Finland and the Nordic countries that initiated in the 1870s with the growing demand of lingonberries in Western Europe. News about Swedish exports were read in the newspapers in Finland, where the "red gold fever" led to initiatives for export and commercial use of wild berries [2]. Moreover, this berry boom led to conflicts in the local woods about their ownership, when the demand for the red berries intensified and the prices rose [5].

The articles were handpicked by conducting key word searches about wild berries, their foraging, economic use and trade in the online interface of the Digi-collection. Manual work was preferred at this stage, to be able to control closely the quality of the search results and to code the articles based on their content for the purposes of the historical research (eg. commercial, non-commercial news). Even though the newspapers have been optically character read, it is not possible to extract automatically complete articles based on the search results. The article structure has not been recognized well in the OCR-process, and, thus, the articles in the corpus were collected by copying the text layer by hand.

[3] https://digi.kansalliskirjasto.fi/etusivu?set_language=en

[4] The articles in the corpus are from the years 1880-1881, 1885-1886, 1890 and 1895.

2.1 Named Entity Recognition

We spotted first names of locations in the manually prepared berry picking corpus with named entity recognition software. Named Entity Recognition (NER), search, classification and tagging of names and name like frequent informational elements in texts, has become a standard information extraction procedure for textual data. NER has been applied to many types of texts and different types of entities: newspapers, fiction, historical records, persons, locations, chemical compounds, protein families, animals etc. Performance of a NER system is usually heavily genre and domain dependent. Entity categories used in NER may also vary. The most used set of named entity categories is usually some version of three partite categorization of locations, persons and organizations [6]. In this study, we are only interested in names of locations.

The names in the berry corpus were recognized with two NE tools: Stanford NER and FiNER. Stanford NER is a standard trainable named entity recognition tool that is based on conditional random fields [7]. Stanford NER models have been trained for several languages, e.g. for English, German, Dutch, French [4], Chinese[5] and Finnish [3]. FiNER, on the other hand, is a rule-based named entity recognizer that has been produced solely for Finnish names in the Fin-CLARIN consortium [8].

FiNER has earlier been evaluated with OCRed Finnish newspaper data along with other modern Finnish NER tools. Results with low quality OCRed 19th century Finnish were not very good: FiNER was able to achieve F-score of 0.57 with locations in the data [8]. Ruokolainen and Kettunen [3] describe creation of a Stanford NER model for 19th century Finnish using training data of ca. 380 000 words that were annotated with names of locations and persons manually and semi-manually. They were able to achieve F-score of 0.79 with locations in an improved quality OCR of a subpart of the Finnish newspaper collection. Considering the quality of the OCR, these NER results are quite good. Better results are not easily achieved without the use of more training data for Stanford NER, better quality OCR, or some other NER system.

Both of the taggers are used for recognizing Finnish language named-entities, and the berry corpus contains texts only from newspapers in Finnish. We estimated the word level quality of the berry-picking corpus by running it through a morphological analyzer Omorfi[6]. 79.1% of the

[5] https://nlp.stanford.edu/software/CRF-NER.shtml
[6] https://github.com/jiemakel/omorfi

words in the corpus were recognized by Omorfi. This quality is slightly better than the quality of NER evaluation collection used in Kettunen et al. [8]. Anyhow the quality is not very high, but of typical OCRed historical newspaper data level.

The result differences between the two taggers are clear. As shown in Table 1, Stanford NER outperforms FiNER in both precision and recall: Stanford receives an F-score of 0.83 and FiNER a clearly lower score of 0.68. It is seen clearly how a trained tagger works much better with data that includes historical language use, and which has been OCRed. The Stanford NER results are also better—although not directly comparable—than the previous evaluations of named entity recognition using historical Finnish newspaper data [8].

Table 1. Performance of the two taggers tested with the berry-picking corpus

	Stanford NER	**FiNER (Mylly[7])**	**Manual**
Place names tagged, all (n)	672	551	691
Manually verified place names (n)	567	425	
Erroneous place names	15.6 %	22.9 %	
Precision	0.84	0.77	
Recall	0.82	0.62	
F-score	**0.83**	**0.68**	

To be able to pinpoint some of the problems of our OCRed newspaper data for the NE taggers, we performed first error analysis of the output of the Stanford tagger in the NER evaluation data of Ruokolainen and Kettunen [3]. The parallel data has available both manually corrected ground truth (GT) and a reasonably good quality new OCR version with Tesseract 3.04.01.

Ehrmann et al. [9] suggest that application of NE tools on historical texts faces three challenges: i) noisy input texts, ii) lack of coverage in linguistic resources, and iii) dynamics of language. Lack of coverage in linguistic resources can be e.g. be missing old names in the lexicons of the NER tools. With dynamics of language Ehrmann et al. refer to different rules and conventions for the use of written language in different

[7] https://www.kielipankki.fi/support/mylly/

times. In this respect, late 19th century Finnish is not that different from current Finnish, but obviously also this can affect the results.

In an earlier historical newspaper data NER evaluation [8] especially Ehrman's first point, noisy input, was the obvious reason for low performance of evaluated NER tools. Now that we have available a good quality ground truth evaluation collection along with a lower quality re-OCRed version of the same data, we can see more clearly effects of OCR quality on the results. We performed a detailed error analysis on results of locations in GT and OCR evaluation data to pinpoint problems of OCRed data and Stanford NER's performance in it. We found 437 misclassifications in the results of locations in the GT evaluation data. In OCR evaluation data there were 491 errors (+14% units). Error classes and their counts are shown in Table 2.

Table 2. Error amounts in tagged data

Error	**Amount in GT data**	**Amount in Tesseract OCR data**
LOC missed	224	204
NULL marked as LOC	106	162
LOC marked as PER	58	76
PER marked as LOC	40	46
Confused beginnings and endings of LOC	9	3
	437	**491**

As the two first content rows in the table show, about 75% of the errors in both data are either missing entity tags or marked entities in case, where there should be none. Locations and persons do not get confused to each other as much, although this is usually a common error. It seems also that lower quality data provokes Stanford NER to mark common words more as locations. Common possible causes for errors are the following:

- spelling variants of words (variant/common): *Itaalia*/*Italia, Buda-Pestiä/Budapestiä*, *Amsterdami/Amsterdam*, *Tukholmi/Tukholma, Kiöpenhawni/Köpenhamina, Kalefornia*/*Kalifornia*
- spelling errors or erroneous OCR (*Vulgarian* pro *Bulgarian*, *Insbuckissä* pro *Innsbruckissa*)

- broken lines (e.g. *Hel- sinki* broken to two separate lines)
- *Stynnyrin, Viinakaupan* (initial upper case letter in a common word)

2.2 Analysis of Errors in the Berry-picking Data

The locations of the berry-picking corpus have been extracted manually in an Excel sheet for P/R counting, but their comparative analysis is difficult, as right and wrong markings are not separated in the entity data, only counts. We can anyhow make some observations between differences of Stanford NER's location markings and those of FiNER.

Stanford has marked 783 words as locations in 672 entities. Out of the word tokens marked as entities 73.56% are recognized by Omorfi. FiNER has marked 551 word tokens as locations, and 88.38% of the words are recognized by Omorfi. It seems, thus, that Stanford NER is clearly more robust in tagging of named entities, as out of its entities more are misspelled but still better marked correctly as entities.

Some of the erroneous word forms that Stanford NER model gets right are shown below:

Lcppämirran pitäjään	(pro Leppävirran)
Cyslöjärmen kylässä	(pro Syslöjärven)
Uustaarlcbyyssä	(pro Uuskaarleby, Uusikaarlepyy)
Hinvcnsalon saarella	(pro Hirvensalon)
Ccderhwarfin tilalle	(pro Cederhwarfin)
Smeitsin	(pro Sveitsin)
Ruotiin	(pro Ruotsiin)
Länsi-*Cuomessa*	(pro Länsi-Suomessa)
Iymäskylän	(pro Jyväskylän)

These examples contain usually 1-3 character errors. FiNER marks also some of them correctly as locations, but Stanford's ability to mark misspellings correctly is clearly better.

Both taggers mark false strings as locations. A common error for both is marking of a word with initial upper case character as a location. Some examples are *Stynnyrin, Viinakaupan, Viinan, Vähemmissä, Vapaasta, Väkijuomakaupasta, Vähemmin, Viinaliikkeen, Vuosittain.*

Another important feature, which separates the two tools is the ability of Stanford NER to recognize named entities with multiple terms. For

instance, with Stanford NER, we were able to detect *Mikkelin kaupunki* and *Mikkelin lääni*, which are the *town* of Mikkeli and the Mikkeli *province*. Moreover, we are able to qualify some locations as *rautatiepysäkki*, *railway station*, which is of particular interest when studying processes of commodification and exports. As we will see below, the ontologies that we are using enable linking to these more specific spatial categories. At the same, this poses even more acutely the question of the historical dimensions of the places contained in the ontologies.

3 The Linking of Recognized Place Names for Creating Structured Data

After the tests about recognizing the named-entities, we continued the study only with the results of the Stanford NER, as it performed clearly better than FiNER. We used the complete list of place names recognized by Stanford NER, and did not remove the wrong locations of the results to keep the process as "genuine" and automated as possible. The next aim was to link the recognized place names to ontologies (i.e. controlled vocabularies), which would provide more detailed location information about the places. In the linking, we took use of the information about the newspaper publication places that is available in the newspaper metadata.

Named-entity linking (NEL) [10–11] refers to the task of determining the identity of named entities mentioned in a text, by linking found named entity mentions to strongly identified entries in ontologies. NEL process consists of NER, entity linking (EL) and named entity disambiguation (NED). In this case, the Stanford NER's results are used to search matching entities from ontologies, which cover historical Finnish and contemporary place names: WarSampo's Karelian places[8], Finto's YSO places[9], and Finnish Geographic Places ontology[10]. The NED determines the correct identity for the entity from a pool of entities extracted from ontologies. Each ontology contained or was linked to other ontologies that contained coordinates for places.

For the linking of the entities, we use ARPA [1], which is a NER and EL tool that queries matches from controlled vocabularies. For this paper, ARPA tool has been configured to link only extracted entities or n-

[8] https://www.ldf.fi/dataset/warsa
[9] https://finto.fi/yso-paikat/en/
[10] http://www.ldf.fi/dataset/pnr/

grams that start with a capital letter, are nouns, or proper nouns. The NED uses newspaper metadata and information provided by the ontologies about the linked targets to determine the correct identity. In our case of historical newspapers, additional newspaper metadata was previously manually enriched with publication place's coordinates. The disambiguation and identification of the places was done in three steps in relation to their position in the ontology hierarchies. Our solution is to use the newspaper publication place for delineating the area or group of potential places.

First, if the newspaper place name referred to a foreign country or their cities, towns, and villages, these were preferred. For example, when "Russia" is mentioned it is linked to a small place in Finland and to the country Russia. It is far more likely in such corpus that when a country is mentioned, the place should be preferably linked to it rather than a Finnish town or village. In these cases, thus, the countries and continent names are prioritized.

Second, for national towns and smaller places of the same name, we prioritize the larger one. Third, the most problematic to identify were the "local" place names in the hierarchy (villages and farm houses), which can be found with similar place names around the country. An example is the place *Niinimäki*, to which 11 different targets were linked, all in the lowest hierarchy classified in the ontology as village, town quartier or neighbourhood. In such cases, we have used the coordinates of the newspaper publication place and the linked targets to determine, which target was the nearest to the publication place. The idea is that smaller places received publicity foremost in the newspapers of the region.

The results of the linking is evaluated in two steps: concerning the linking on the one hand, and the disambiguation on the other. The errors encountered can be divided into five groups: OCR errors, NER errors, Linking tool errors (ARPA/LAS error), ontology errors, and place not found from selected ontologies. In earlier work [12] similar errors such as OCR errors, tool errors, and ontology related errors were encountered. The OCR'd input text contains errors that impact the entity linking as they reduce the amount of produced entity links. The OCR application may incorrectly identify certain words and letters due to poor quality of the newspaper.

The NER errors are produced by the Stanford NER whereas the linking tool errors are produced by ARPA and the tools it uses. The ARPA tool [1] (that is used in the linking process) uses LAS for lexical analysis

to lemmatize and inflect the words. In case of some place names the tools cannot always find the original base form or inflected form to correctly match the names into ontologies. This leads to loss of links. In addition, in some cases the ontologies do not contain all place names in Finnish or all required information for the algorithm to function properly (for example missing coordinates).

In the linking, 388 of the 672 places (of which 567 were correct places) recognized by Stanford NER were linked to an ontology. The result is explained mainly by two factors. First, the Stanford model recognized also false positives, which the link tool, then, could not identify. Second, the trained Stanford tagger could recognize also correct places with OCR-errors, which could not be handled in the linking. Moreover, some Linking tool errors were encountered, which regard the inflected word forms.

The linking process found 809 linked targets, which were identified in the NED: 33 locations of the places linked were not correctly identified, that is, in all 355 of the 672 Stanford recognized places were linked correctly. Seven errors were generated by a false positive recognized by Stanford NER, three errors were created by the linking tool, one error was related to an OCR mistake, and one (historical) location was not found in the used ontologies. The rest of the errors (21) were related to problems of disambiguation part of our method: either caused by the hierarchical identification or the demarcation by newspaper publication coordinates. There are cases where the demarcation helps to locate the ambiguous small place correctly nearby the newspaper's home town. At the same time, due to the reuse of articles by other newspapers, several small places in reproduced articles were identified wrongly. It is notable, however, that in most cases the first newspaper to publish an article gave the right geographic context to the local places described in the article, which supports our idea of using ontology hierarchies.

4 Conclusion

This paper has built and evaluated the functioning of named entity recognition and linking in historical research, which uses location information in nineteenth century historical newspaper data. We started our inquiry with a manually generated corpus consisting of 303 newspaper articles on wild berries, their foraging, economic use and trade. The aim was to evaluate the quality and problems related to an automated named entity

recognition and linking pipeline that we built. From the 303 articles, we generated 672 automatically tagged locations (691 locations were tagged manually in the corpus), of which 567 were correct. These Stanford NER tagged locations resulted further into 388 locations, which were identified in the linking, and of these 355 were linked to correctly.

We have shown in this paper that a Stanford NER model developed with nineteenth-century newspaper data outperforms clearly a rule-based NER software FiNER in location analysis of OCRed newspaper corpus containing news related broadly to berry-picking. Although the corpus is smallish, differences in performance are clear. Despite the low quality of the OCR in the berry-picking corpus, NER analysis of locations provided by the Stanford model are useful and give also a good basis for larger data analysis, if more data is gathered.

The paper has highlighted, how there are challenges related to the linking of the historical places due to the discrepancy between the linking tool and the trained Stanford NER, which is able to detect places with considerable spelling mistakes. One solution would be to process the recognized named entities to a more consistent and modern written form before the linking. At the same time, the linking tool improves the results to some extent, as it is able to drop out almost all false positives recognized by Stanford NER.

From the perspective of the historical research, the pipeline produces adequate level results. The quality of the named entity recognition of the locations is good. The NER results—manually read—show how the share of European place names, such as *Sweden*, *(North) Germany*, *Stettin*, *Hamburg*, *Lübeck*, but also *Saint Petersburg*, increase in the berry corpus towards the end of the century. This supports one of the research case's hypotheses that wild berries became discussed and viewed in relation to the expanding western European market. Moreover, if we look at the corpus texts that were coded as being about exports, we can pinpoint actual export links. Especially notable is the appearance of the Swedish *Moheda* station in the recognition results, as the station was one known link in the Swedish berry exports of the late nineteenth century. Also, the town of *Vaasa* on the west coast of Finland stands out as a surprisingly central link and is the most cited place in the export texts.

The linking offers interesting results already at this point. The method for detecting smaller places enables to map the developments regionally and inside the country. However, to improve the recognition quality and the depth of the historical and statistical analysis, more attention should

be paid to the uniqueness of the events in the texts, on the one hand, and the virality or reuse of the texts, on the other. In the berry corpus, for example, the most reproduced text was about a small girl who handed wild berries as a gift to the Empress, during the summer trip of the Imperial family in the Finnish archipelago in 1886[11]. Adding a text reuse detection tool to the pipeline, like the tool developed for historical newspapers by the COMHIS consortium [13], would enable to control for the geographic over-representation of single events, and to improve the identification of the linked targets.

Acknowledgements

The third author's work is part of the project Computational History and the Transformation of Public Discourse in Finland 1640–1910 (COMHIS) funded by the Academy of Finland. We would like to thank the anonymous referees, and Jouni Tuominen and Esko Ikkala (Semantic Computing Research Group) for their comments.

References

1. Mäkelä, E.: Combining a REST Lexical Analysis Web Service with SPARQL for Mashup Semantic Annotation from Text. In: Valentina Presutti et al. (eds.) The Semantic Web: ESWC 2014 Satellite Events, ESWC 2014, Vol. 8798, pp. 424–428, Springer, Cham (2014).
2. La Mela, M.: The Politics of property in a European periphery : The ownership of books, berries, and patents in the Grand Duchy of Finland 1850-1910. PhD Thesis, European University Institute (2016), pp. 257–268. http://dx.doi.org/10.2870/604750
3. Ruokolainen, T., Kettunen, K.: À la recherche du nom perdu – searching for named entities with Stanford NER in a Finnish historical newspaper and journal collection. In: 13th IAPR International Workshop on Document Analysis Systems (2018).
4. Neudecker, C.: An Open Corpus for Named Entity Recognition in Historic Newspapers. In: Proceedings of Tenth International Conference on Language Resources and Evaluation, LREC 2016, pp. 4348–4352 (2016).
5. La Mela, M.: Property rights in conflict: wild berry-picking and the Nordic tradition of allemansrätt. Scandinavian Economic History Review 62(3), pp. 266–289 (2014). https://doi.org/10.1080/03585522.2013.876928
6. Nadeau, D., Sekine, S.: A Survey of Named Entity Recognition and Classification. Linguisticae Investigationes 30(1), pp. 3–26 (2007).
7. Finkel, J.R., Grenager, T., Manning, C.: Incorporating non-local information into information extraction systems by Gibbs sampling. In: Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics, ACL 2005, pp. 363–370 (2005).

[11] Two different versions of this event appeared in the corpus 15 times.

8. Kettunen, K., Mäkelä, E., Ruokolainen, T., Kuokkala, J., Löfberg, L: Old Content and Modern Tools – Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910. Digital Humanities Quarterly 11(3), (2017).
9. Ehrmann, M., Colavizza, G., Rochat, Y., Kaplan, F.: Diachronic Evaluation of NER Systems on Old Newspapers. In: Proceedings of the 13th Conference on Natural Language Processing, KONVENS 2016, pp. 97–107 (2016). https://www.linguistics.rub.de/konvens16/pub/13_konvensproc.pdf (accessed on 8 February 2019).
10. Hachey, B., Radford, W., Nothman, J., Honnibal, M., Curran, J.R: Evaluating entity linking with Wikipedia. Artificial intelligence, 194, pp. 130–150 (2013).
11. Bunescu, R., Paşca, M.: Using encyclopedic knowledge for named entity disambiguation. In: Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, pp. 9–16 (2006).
12. Tamper, M., Leskinen, P., Ikkala, E., Oksanen, A., Mäkelä, E., Heino, E., Tuominen, J., Koho, M., Hyvönen E.,: AATOS – a Configurable Tool for Automatic Annotation. In: Gracia J. et al. (eds.) Language, Data, and Knowledge. LDK 2017, vol. 10318, pp. 276–289, Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59888-8_24
13. Vesanto, A., Nivala A., Rantala, H., Salakoski, T., Salmi, H., Ginter, F.: Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910. In: Proceedings of the 21st Nordic Conference of Computational Linguistics, NoDaLiDa 2017, pp. 54–58 (2017). http://www.ep.liu.se/ecp/133/010/ecp17133010.pdf (accessed on 8 February 2019).