Kettunen, Kimmo; Mäkelä, Eetu; Kuokkala, Juha; Ruokolainen, Teemu; Niemi, Jyrki

Modern tools for old content-in search of named entities in a finnish ocred historical newspaper collection 1771-1910

Published in:
Lernen, Wissen, Daten, Analysen 2016

Published: 01/01/2016

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

Please cite the original version:
Kettunen, K., Mäkelä, E., Kuokkala, J., Ruokolainen, T., & Niemi, J. (2016). Modern tools for old content-in search of named entities in a finnish ocred historical newspaper collection 1771-1910. In R. Krestel, D. Mottin, & E. Müller (Eds.), Lernen, Wissen, Daten, Analysen 2016: Proceedings of the Conference "Lernen, Wissen, Daten, Analysen", Potsdam, Germany, September 12-14, 2016 (pp. 124-135). (CEUR Workshop Proceedings; Vol. 1670). CEUR. CEUR WORKSHOP PROCEEDINGS

Modern Tools for Old Content – in Search of Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910

Kimmo Kettunen[1], Eetu Mäkelä[2], Juha Kuokkala[3],Teemu Ruokolainen[4], Jyrki Niemi[5]

[1] National Library of Finland, Centre for Preservation and Digitization, Mikkeli, Finland
kimmo.kettunen@helsinki.fi
[2] Aalto University, Semantic Computing Research Group, Espoo, Finland
eetu.makela@aalto.fi
[3] University of Helsinki, Department of Modern Languages, Helsinki, Finland
juha.kuokkala@helsinki.fi
[4] Aalto University, Department of Signal Processing and Acoustics, Espoo, Finland
teemu.ruokolainen@aalto.fi
[5]University of Helsinki, Department of Modern Languages, Helsinki, Finland
jyrki.niemi@helsinki.fi

Abstract. 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. In general a NER system's performance is genre and domain dependent and also used entity categories vary [1]. The most general set of named entities is usually some version of three partite categorization of locations, persons and organizations. In this paper we report first trials and evaluation of NER with data out of a digitized Finnish historical newspaper collection Digi. Digi collection contains 1,960,921 pages of newspaper material from years 1771–1910 both in Finnish and Swedish. We use only material of Finnish documents in our evaluation. The OCRed newspaper collection has lots of OCR errors; its estimated word level correctness is about 74–75 % [2]. Our principal NER tagger is a rule-based tagger of Finnish, FiNER, provided by the FIN-CLARIN consortium. We show also results of limited category semantic tagging with tools of the Semantic Computing Research Group (SeCo) of the Aalto University. FiNER is able to achieve up to 60.0 F-score with named entities in the evaluation data. Seco's tools achieve 30.0–60.0 F-score with locations and persons**.** Performance of FiNER and SeCo's tools with the data shows that at best about half of named entities can be recognized even in a quite erroneous OCRed text.

Keywords: named entity recognition, historical newspaper collections, Finnish

1 Introduction

The National Library of Finland has digitized a large proportion of the historical newspapers published in Finland between 1771 and 1910 [3, 4]. This collection contains 1,960,921 million pages in Finnish and Swedish. Finnish part of the collection consists of about 2.39 billion words. The National Library's Digital Collections are offered via the *digi.kansalliskirjasto.fi* web service, also known as Digi. Part of the newspaper material (years 1771–1874) is freely downloadable in The Language Bank of Finland provided by the FIN-CLARIN consortium[1]. The collection can also be accessed through the Korp[2] environment that has been developed by Språkbanken at the University of Gothenburg and extended by FIN-CLARIN team at the University of Helsinki to provide concordances of text resources. A Cranfield style information retrieval test collection has been produced out of a small part of the Digi newspaper material at the University of Tampere [5].

The web service digi.kansalliskirjasto.fi is used, for example, by genealogists, heritage societies, researchers, and history enthusiast laymen. There is also an increasing desire to offer the material more widely for educational use. In 2015 the service had about 14 million page loads. User statistics of 2014 showed that about 88.5 % of the usage of the Digi came from Finland, but an 11.5 % share of use was coming outside of Finland.

Named entity recognition has become one of the basic techniques for information extraction of texts. In its initial form NER was used to find and mark semantic entities like person, location and organization in texts to enable information extraction related to these kinds of entities. Later on other types of extractable entities, like time, artefact, event and measure/numerical, have been added to the repertoires of NER software [1], [6].

Our aim with usage of NER is to provide users of Digi better means for searching and browsing of the historical newspapers. Different types of names, especially person names and names of locations are used frequently as search terms in different newspaper collections [7]. They can provide also browsing assistance to collections, if the names are recognized and tagged in the newspaper data and put into the index [8]. A fine example of usage of name recognition with historical newspapers is La Stampa's historical newspaper collection[3]. After basic keyword search users can browse or filter the search results by using three basic NER categories of person (authors of articles or persons mentioned in the articles), location (countries and cities mentioned in the articles) and organization. Thus entity annotations of newspaper text allow a more semantically-oriented exploration of content of the large archive. A large scale (152 M articles) NER analysis and usage examples of the Australian historical newspaper collection Trove is described in Mac Kim and Cassidy [9].

[1] https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/KielipankkiAineistotDigilibPub
[2] https://korp.csc.fi/
[3] http://www.archiviolastampa.it/

2 NER Software and Evaluation

For recognition and labelling of named entities we use principally FiNER software. SeCo's ARPA is of different type, it is mainly used for Semantic Web tagging and linking of entities [10][4], but it could be adapted for basic NER, too. Before choosing FiNER we also tried a commonly used trainable free tagger, Stanford NER[5], but were not able to get reasonable performance out of it for our purposes.

FiNER is a rule-based named-entity tagger, which in addition to surface text forms utilizes grammatical and lexical information from a morphological analyzer (Omorfi[6]). FiNER pre-processes the input text with a morphological tagger derived from Omorfi. The tagger disambiguates Omorfi's output by selecting the statistically most probable morphological analysis for each word token, and for tokens not recognized by the analyzer, guesses an analysis by analogy of word-forms with similar ending in the morphological dictionary. The use of morphological pre-processing is crucial in performing NER with a morphologically rich language such as Finnish, where a single lexeme may theoretically have thousands of different inflectional forms.

The focus of FiNER is in recognizing different types of proper names. Additionally, it can identify the majority of Finnish expressions of time and e.g. sums of money. FiNER uses multiple strategies in its recognition task:

1) Pre-defined gazetteer information of known names of certain types. This information is mainly stored in the morphological lexicon as additional data tags of the lexemes in question. In the case of names consisting of multiple words, FiNER rules incorporate a list of known names not caught by the more general rules.

2) Several kinds of pattern rules are being used to recognize both single- and multiple-word names based on their internal structure. This typically involves (strings of) capitalized words ending with a characteristic suffix such as Inc, Corp, Institute etc. Morphological information is also utilized in avoiding erroneously long matches, since in most cases only the last part of a multi-word name is inflected, while the other words stay in the nominative (or genitive) case. Thus, preceeding capitalized words in other case forms should be left out of a multi-word name match.

3) Context rules are based on lexical collocations, i.e. certain words which typically or exclusively appear next to certain types of names in text. For example, a string of capitalized words can be inferred to be a corporation/organization if it is followed by a verb such as *tuottaa* ('produce'), *työllistää* ('employ') or *lanseerata* ('launch' [a product]), or a personal name if it is followed by a comma- or parenthesis-separated numerical age or an abbreviation for a political party member.

The pattern-matching engine that FiNER uses, HFST Pmatch, marks leftmost longest non-overlapping matches satisfying the rule set (basically a large set of disjuncted patterns) [11, 12]. In the case of two or more rules matching the exact same passage in the text, the choice of the matching rule is undefined. Therefore, more

[4] An older demo version of the tool is available at http://demo.seco.tkk.fi/sarpa/#/
[5] http://nlp.stanford.edu/software/CRF-NER.shtml
[6] https://github.com/flammie/omorfi

control is needed in some cases. Since HFST Pmatch did not contain a rule weighing mechanism at the time of designing the first release of FiNER, the problem was solved by applying two runs of distinct Pmatch rulesets in succession. This solves for instance the frequent case of Finnish place names used as family names: in the first phase, words tagged lexically as place names but matching a personal name context pattern are tagged as personal names, and the remaining place name candidates are tagged as places in the second phase. FiNER annotates 15 different entities that belong to five categories: location, person, organization, measure and time [12].

SeCo's ARPA [10] is not actually a NER tool, but instead a dynamic, configurable entity linker. In effect, ARPA is not interested in locating all entities of a particular type in a text, but instead locating all entities that can be linked to strong identifiers elsewhere. Through these, it is then for example possible to source coordinates for identified places, or associate different name variants and spellings to a single individual. For the pure entity recognition task presented in this paper, ARPA is thus at a disadvantage. However, we wanted to see how it would fare in comparison to FiNER.

The core benefits of the ARPA system lie in its dynamic, configurable nature. In processing, ARPA combines a separate lexical processing step with a configurable SPARQL-query -based lookup against an entity lexicon stored at a Linked Data endpoint. Lexical processing for Finnish is done with a modified version of Omorfi[7], which supports historical morphological variants, as well as lemma guessing for out of vocabulary words. This separation of concerns allows the system to be speedily configured for both new reference vocabularies as well as the particular dataset to be processed.

2.1 Evaluation Data

As evaluation data for FiNER we used samples from the Digi collection. Kettunen and Pääkkönen [2] calculated among other things number of words in the data for different decades. It turned out that most of the newspaper data was published in 1870–1910, and beginning and mid of the 19th century had much less published material. About 95 % of the material was printed in 1870–1910, and most of it, 82.7 %, in two decades of 1890–1910.

We aimed at an evaluation collection of 150,000 words. To emphasize the importance of the 1870–1910 material we took 50 K of words from time period 1900–1910, 10 K from 1890–1899, 10 K from 1880–1889, and 10 K from 1870–1879. Rest 70 K of the material was picked from time period of 1820–1869. Thus the collection reflects most of the data from the century but is also weighed to the end of the 19th century and beginning of 20th century.

The final manually tagged evaluation data consists of 75,931 lines, each line having one word or other character data. The word accuracy of the evaluation sample is on the same level as the whole newspaper collection's word level quality: about 73 % of the words can be recognized by a modern Finnish morphological analyzer [2]. 71

[7] https://github.com/jiemakel/omorfi

% of the tagger's input snippets have five or more words, the rest have fewer than five words in the text snippet.

FiNER's 15 tags for different types of entities is too fine a distinction for our purposes. Our first aim was to concentrate only on locations and person names, because they are mostly used in searches of the Digi collection, as was detected in an earlier log analysis [4]. After reviewing some of the FiNER tagged material, we included also three other tags, as they seemed important and were occurring frequently enough in the material. The final chosen eight tags are shown and explained below.

	Entity/tag	Meaning
1.	<EnamexPrsHum>	person
2.	<EnamexLocXxx>	general location
3.	<EnamexLocGpl>	geographical location
4.	<EnamexLocPpl>	political location (state, city etc.)
5.	<EnamexLocStr>	street, road, street address
6.	<EnamexOrgEdu>	educational organization
7.	<EnamexOrgCrp>	company, society, union etc.
8.	<TimexTmeDat>	expression of time

The final entities show that our interest is mainly in the three most used semantic NER categories: persons, locations and organizations. With locations we use two subcategories and with organizations one. Temporal expressions were included in the tag set due to their general interest in the newspaper material.

Manual tagging of the evaluation material was done by the fourth author, who had previous experience of tagging modern Finnish with tags of the FiNER tagger. Tagging took one month, and quality of the tagging and its principles were discussed before starting based on a sample of 2000 lines of evaluation data. It was agreed, for example, that words that are misspelled but are recognizable for the human tagger as named entities would be tagged (cf. 50 % character correctness rule in Packer et al. [15]). If orthography of the word was following 19th century spelling rules, but the word was identifiable as a named entity, it would be tagged, too.

2.2 Results of the Evaluation

We evaluated performance of FiNER and SeCo's ARPA using the *conlleval*[8] script used in Conference on Computational Natural Language Learning (CONLL). Evaluation is based on "exact-match evaluation" [1], [16]. In this type of evaluation NER system is evaluated based on the micro-averaged F-measure (MAF) where precision is the percentage of correct named entities found by the NER software; recall is the percentage of correct named entities present in the tagged evaluation corpus that are found by the NER system. A named entity is considered correct only if it is an exact

[8] http://www.cnts.ua.ac.be/conll2002/ner/bin/conlleval.txt, author ErikTjong Kim Sang, version 2004-01-26

match of the corresponding entity in the tagged evaluation corpus: “a result is considered correct only if the boundaries and classification are exactly as annotated” [17]. Thus the evaluation criteria are strict, especially for multipart entities.

Detailed results of the evaluation of FiNER are shown in Table 1. Entities <ent/> consist of one word token, <ent> are part of a multiword entity and </ent> are last parts of multiword entities.

Label	**P**	**R**	**F-score**	**Number of tags found**	**Number of tags in the evaluation data**
<EnamexLocGpl/>	6.96	9.41	8.00	115	85
<EnamexLocPpl/>	89.50	8.46	15.46	181	1920
<EnamexLocStr/>	23.33	50.00	31.82	30	14
<EnamexLocStr>	100.00	13.83	24.30	13	94
</EnamexLocStr>	100.00	18.31	30.95	13	71
<EnamexOrgCrp/>	2.39	6.62	3.52	376	155
<EnamexOrgCrp>	44.74	25.99	32.88	190	338
</EnamexOrgCrp>	40.74	31.95	35.81	189	250
<EnamexOrgEdu>	48.28	40.00	43.75	29	35
</EnamexOrgEdu>	55.17	64.00	59.26	29	25
<EnamexPrsHum/>	16.38	52.93	25.02	1819	564
<EnamexPrsHum>	87.44	26.67	40.88	438	1436
</EnamexPrsHum>	82.88	31.62	45.78	438	1150
<TimexTmeDat/>	5.45	14.75	7.96	495	183
<TimexTmeDat>	68.54	2.14	4.14	89	2857
</TimexTmeDat>	20.22	2.00	3.65	89	898

Table 1. Evaluation results of FiNER with strict CONLL evaluation criteria. Data with zero P/R is not included in the table. These include categories <EnamexLocGpl>, </EnamexLocGpl>, <EnamexLocPpl>, </EnamexLocPpl>, <EnamexLocXxx>, <EnamexLocXxx/>, </EnamexLocXxx>, and <EnamexOrgEdu/>. Most of these have very few entities in the data, only <EnamexLocXxx> is frequent with over 1200 occurences

Results of the evaluation show that named entities are not recognized very well, which is not surprising, as the quality of the text data is quite low. Especially recognition of multipart entities is very low. Some part of the entities may be recognized, but rest is not. Out of multiword entities corporations and educational organizations are recognized best. Names of persons are the most frequent category. Recall of one part person names is best, but its precision is low. Multipart person names have a more balanced recall and precision, even if their overall recognition is not high.

In a looser evaluation the categories were treated so that any correct marking of an entity regardless its boundaries was considered a hit. Four different location categories were joined to two: general location <*EnamexLocXxx*> and that of street names. End result was six different categories instead of eight. Table 2 shows evaluation results with loose evaluation. Recall and precision of the most frequent categories of person and location was now clearly higher, but still not very good.

Label	**P**	**R**	**F-score**	**Number of tags**
<EnamexPrsHum>	63.30	53.69	58.10	2681
<EnamexLocXxx>	69.05	49.21	57.47	1541
<EnamexLocStr>	83.64	25.56	39.15	55
<EnamemOrgEdu>	51.72	47.62	49.59	58
<EnamemOrgCrp>	30.27	32.02	31.12	750
<TimexTmeDat>	73.85	12.62	21.56	673

Table 2. Evaluation results of FiNER with loose criteria and six categories

Our third evaluation was performed for a limited tag set with tools of the SeCo's ARPA. First only places were identified so that one location, *EnamexLocPpl*, was recognized. For this task, ARPA was first configured for the task of identifying place names in the data. As a first iteration, only the Finnish Place Name Registry[9] was used. After examining raw results from the test run, three issues were identified for further improvement. First, PNR contains only modern Finnish place names. To improve recall, three registries containing historical place names were added: 1) the Finnish spatiotemporal ontology SAPO [18] containing names of historic municipalities, 2) a repository of old Finnish maps and associated places from the 19th and early 20th Century, and 3) a name registry of places inside historic Karelia, which does not appear in PNR due to being ceded by Finland to the Soviet Union at the end of the Second World War [19]. To account for international place names, the names were also queried against the Geonames database[10] as well as Wikidata[11]. The contributions of each of these resources to the number of places identified in the final runs are shown in Table 3. Note that a single place name can be, and often was found in multiple of these sources.

Source	**Matches**	**Fuzzy matches**
Karelian places	461	951
Old maps	685	789
Geonames	1036	1265
SAPO	1467	1610
Wikidata	1877	2186
PNR	2232	2978

[9] http://www.ldf.fi/dataset/pnr/
[10] http://geonames.org/
[11] http://wikidata.org/

Table 3. Number of distinct place names identified using each source

Table 4 describes the results of location recognition with ARPA. Without one exception (*New York*), only one word entities were discovered by the software

Label	P	R	F-score	Number of tags
<EnamexLocPpl/>	39.02	53.24	45.03	2673
</EnamexLocPpl>	100.00	5.26	10.00	1
<EnamexLocPpl>	100.00	4.76	9.09	1

Table 4. Basic evaluation results for ARPA

A second improvement to the ARPA process arose from the observation that while recall in the first test run was high, precision was low. Analysis revealed this to be due to many names being both person names as well as places. Thus, a filtering step was added, that removed 1) hits identified as person names by the morphological analyzer and 2) hits that matched regular expressions catching common person name patterns found in the data (I. Lastname and FirstName LastName). However, sometimes this was too aggressive, ending up for example in filtering out also big cities like Tampere and Helsinki. Thus, in the final configuration, this filtering was made conditional on the size of the identified place, as stated in the structured data sources matched against.

Finally, as the amount of OCR errors in the target dataset was identified to be a major hurdle in accurate recognition, experiments were made with sacrificing precision in favor of recall through enabling various levels of Levenshtein distance matching against the place name registries. In this test, the fuzzy matching was done in the query phase after lexical processing. This was easy to do, but doing the fuzzy matching during lexical processing would probably be more optimal, as currently lemma guessing (which is needed because OCR errors are out of the lemmatizer's vocabulary) is extremely sensitive to OCR errors particularly in the suffix parts of words.

After the place recognition pipeline was finalized, a further test was done to test if the ARPA pipeline could be used for also person name recognition. Here, as a lexicon of names, the Virtual International Authority File was used, as it contains 33 million names for 20 million people. In the first run, the query simply matched all uppercase words against both first and last names in this database, while allowing for any number of initials to also precede such names matched. This way, the found names can't actually be always any more linked to strong identifiers, but for a pure NER task, recall is improved.

Table 5 shows results of this evaluation without fuzzy matching of names and Table 6 with fuzzy matching.

Label	P	R	F-score	Number of tags
<EnamexLocPpl/>	58.90	55.59	57.20	1849
</EnamexLocPpl>	1.49	10.53	2.61	134
<EnamexLocPpl>	1.63	14.29	2.93	184
<EnamexPrsHum/>	30.42	27.03	28.63	2242
</EnamexPersHum>	83.08	47.39	60.35	656
<EnamePersHum>	85.23	43.80	57.87	738

Table 5. Evaluation results for ARPA: no fuzzy matching

Label	P	R	F-score	Number of tags
<EnamexLocPpl/>	47.38	61.82	53.64	2556
</EnamexLocPpl>	1.63	15.79	2.96	184
<EnamexLocPpl>	1.55	14.29	2.80	193
<EnamexPrsHum/>	9.86	66.79	17.18	3815
</EnamexPersHum>	63.07	62.97	63.01	1148
<EnamePersHum>	62.25	61.77	62.01	1425

Table 6. Evaluation results for ARPA: fuzzy matching

Recall of recognition increases markedly in fuzzy matching, but precision deteriorates. More multipart location names are also recognized with fuzzy matching.

Loose evaluation without fuzzy matching gave 44.02 % precision, 64.58 % recall and 52.35 F-score for locations with 2933 found tags. For persons it gave precision of 63.61%, recall of 45.27% and F-score of 52.90 with 3636 found tags.

Loose evaluation with fuzzy matching gave 44.02 % precision, 64.58 % recall and 52.35 F-score for locations. Number of found tags was 2933. For persons it gave precision of 34.49, recall of 78.09 and F-score of 51.57 with 6388 found tags.

3 Discussion

We have shown in this paper first evaluation results of NER for historical Finnish newspaper material from the 19th and early 20th century with two different tools, FiNER and SeCo's ARPA. Word level correctness of the digitized newspaper archive is approximately 70–75 %; the evaluation corpus had a word level correctness of about 73 %. Regarding this and the fact that FiNER and ARPA were developed for modern Finnish, the newspaper material makes a very difficult test for named entity recognition. It is obvious that the main obstacle of high class NER in this material is bad quality of the text. Also historical spelling variation has some effect, but it should not be that high.

Evaluation results in this phase were not very good, best basic F-scores were ranging from 30 to 60 in the basic evaluation, and slightly better in a looser evaluation. We have ongoing trials for improving word quality of our material, which may yield also better NER results. We made some unofficial tests with three versions of a 500,000 word text material that is different from our NER evaluation material but

derives from the 19th century newspapers as well. One version was manually corrected OCR, another old OCRed version and third a new OCRed version. Besides character level errors also word order errors have been corrected in the two new versions. For these texts we did not have a ground truth tagged version, so we could only count marking of NER tags. With FiNER total number of tags increased from 23,918 to 26,674 (+11.5 % units) in the manually corrected text version. Number of tags increased to 26,424 tags (+10.5 % units) in the new OCRed text version. Most notable increase in the number of tags was in categories *EnamexLocStr* and *EnamexOrgEdu*. With ARPA results were even slightly better. ARPA recognized 10 853 places in the old OCR, 11,847 in the new OCR (+ 9.2 % units) and 13,080 (+20.5 % units) in the ground truth version of the text. There is about a 10–20 % unit overall increase in the number of NER tags in both of the new better quality text versions in comparison to the old OCRed text with both taggers.

NER experiments with OCRed data in other languages show usually some improvement of NER when the quality of the OCRed data has been improved from very poor to somehow better [15, 16]. Results of Alex and Burns [18] imply that with lower level OCR quality (below 70 % correctness) name recognition is harmed clearly. Packer et al. [15] report partial correlation of Word Error Rate of the text and achieved NER result; their experiments imply that word order errors are more significant than character errors. On the other hand, results of Rodriquez et al. [17] show, that manual correction of OCRed material that has 88–92 % word accuracy does not increase performance of four different NER tools significantly. As the word accuracy of our material is low, it would be expectable, that somehow better recognition results would be achieved, if the word accuracy was round 80–90 % instead of 70–75 %. Our informal test with different quality texts suggests this, too. Our material has also quite a lot of word order errors which may affect results.

Another option for better recognition results is that we can use more historical language sensitive NER software. Such may become available, if the historically more sensitive version of morphological recognizer Omorfi can be merged with FiNER. A third possibility is to train a statistical name tagger described by Silfverberg [11] with labeled historical newspaper material.

Other causes for poor performance are probably due to 19th century Finnish spelling variation and perhaps also due to different writing conventions of the era. It is possible, for example, that the genre of 19th century newspaper writing differs from modern newspaper writing in some crucial aspects. Considering that both FiNER and ARPA are made for modern Finnish, our evaluation data is heavily out of their main scope [19], even if ARPA uses historical Finnish aware Omorfi.

In our case extraction of names is primarily a tool for improving access to the Digi collection. After getting the recognition rate of the NER tool to acceptable level, we need to decide, how we are going to use extracted names in Digi. Some exemplary suggestions are provided by archive of La Stampa and Trove Names [9]. La Stampa style usage of names provides informational filters after a basic search has been conducted. You can further look for persons, locations and organizations mentioned in the article results. This kind of approach enables browsing access to the collection and possibly also entity linking [20, 21, 22]. Trove Names's name search takes the oppo-

site approach: you first search for names and then you get articles where the names occur. We believe that the La Stampa style of usage of names in the GUI of the newspaper collection is more informative and useful for users, as the Trove style can be already obtained with the normal search function in the GUI of the newspaper collection. If we consider possible uses of NER in Digi, FiNER does so far only basic identifying and classification of names. ARPA is basically not a NER software, but a semantic entity linking system, and thus of broader use. Our main emphasis with NER will be on the use of the names with the newspaper collection as a means to improve browsing and general informational usability of the collection. A good enough coverage of the names with NER needs to be achieved also for this use, of course. A good balance of P/R should be found for this purpose [15], but also other capabilities of the software need to be considered. These remain to be seen later, if we are able to connect some type of functional NER to our historical newspaper collection.

Acknowledgements

First author is funded by the EU Commission through its European Regional Development Fund, and the program Leverage from the EU 2014–2020.

References

1. Nadeau, D., Sekine, S.: A Survey of Named Entity Recognition and Classification. Linguisticae Investigationes 30(1):3–26 (2007)
2. Kettunen, K., Pääkkönen, T.: Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means. Accepted for LREC 2016. http://lrec2016.lrec-conf.org/en/conference-programme/accepted-papers/ (2016).
3. Bremer-Laamanen, M-L.: In the Spotlight for Crowdsourcing. Scandinavian Librarian Quarterly, 1, 18–21 (2014)
4. Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T., Kervinen, J.: Analyzing and Improving the Quality of a Historical News Collection using Language Technology and Statistical Machine Learning Methods. In: Proceedings of IFLA 2014, Lyon (2014) http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-honkela-en.pdf
5. Järvelin, A., Keskustalo, H., Sormunen, E., Saastamoinen, M. and Kettunen, K.: Information Retrieval from Historical Newspaper Collections in Highly Inflectional Languages: A Query Expansion Approach. *Journal of the Association for Information Science and Technology* doi: http://onlinelibrary.wiley.com/doi/10.1002/asi.23379/epdf (2015)
6. Kokkinakis, D., Niemi, J., Hardwick, S., Lindén, K., Borin. L.: HFST-SweNER – a New NER Resource for Swedish. In: Proceedings of LREC 2014, http://www.lrec-conf.org/proceedings/lrec2014/pdf/391_Paper.pdf (2014)
7. Crane, G., Jones, A.: The Challenge of Virginia Banks: An Evaluation of Named Entity Analysis in a 19th-Century Newspaper Collection. In *Proceedings of JCDL'06*, June 11–15, 2006, Chapel Hill, North Carolina, USA. http://repository01.lib.tufts.edu:8080/fedora/get/tufts:PB.001.001.00007/Archival.pdf (2006)
8. Neudecker, C., Wilms, L., Faber, W. J., van Veen, T.: Large-scale Refinement of Digital Historic Newspapers with Named Entity Recognition.

http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-neudecker_faber_wilms-en.pdf (2014)
9. Mac Kim, S., Cassidy, S.: Finding Names in Trove: Named Entity Recognition for Australian. In: Proceedings of Australasian Language Technology Association Workshop, pp. 57–65, https://aclweb.org/anthology/U/U15/U15-1007.pdf (2015)
10. Mäkelä, E.: Combining a REST Lexical Analysis Web Service with SPARQL for Mashup Semantic Annotation from Text. In Presutti, V. et al. (eds.), The Semantic Web: ESWC 2014 Satellite Events. Lecture Notes in Computer Science, vol. 8798, pp. 424–428 (2014)
11. Lindén, K., Axelson, E., Drobac, S., Hardwick, S., Kuokkala, J., Niemi, J., Pirinen, T.A., Silfverberg, M.:. HFST—a System for Creating NLP Tools. In Mahlow, C., Piotrowski, M. (eds.) Systems and Frameworks for Computational Morphology. Third International Workshop, SFCM 2013, Berlin, Germany, September 6, 2013 Proceedings, pp. 53–71 (2013).
12. Silfverberg, M.: Reverse Engineering a Rule-Based Finnish Named Entity Recognizer. https://kitwiki.csc.fi/twiki/pub/FinCLARIN/KielipankkiEventNERWorkshop2015/Silfverberg_presentation.pdf (2015)
13. Packer, T., Lutes, J., Stewart, A., Embley, D., Ringger, E., Seppi, K., Jensen, L. S.: Extracting Person Names from Diverse and Noisy OCR Text. In: Proceedings of the fourth workshop on Analytics for noisy unstructured text data. Toronto, ON, Canada: ACM. (2010)
14. Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., Gómez-Berbís, J.M.: Named Entity Recognition: Fallacies, challenges and opportunities. Computer Standards & Interfaces 35, 482–489 (2013)
15. Rodrigues, K.J., Bryant, M., Blanke, T., Luszczynska, M.: Comparison of Named Entity Recognition Tools for raw OCR text. In: Proceedings of KONVENS 2012 (LThist 2012 wordshop), Vienna September 21, pp. 410–414 (2012)
16. Alex, B., Burns, J.: Estimating and Rating the Quality of Optically Character Recognised Text. In: DATeCH '14 Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage, pp. 97–102. http://dl.acm.org/citation.cfm?id=2595214 (2014)
17. Poibeau, T., Kosseim, L.: Proper Name Extraction from Non-Journalistic Texts. Language and Computers, 37, pp. 144-157 (2001)
18. Hyvönen, E., Tuominen, J., Kauppinen T, Väätäinen, J: Representing and Utilizing Changing Historical Places as an Ontology Time Series. In Ashish, N. and Sheth, V. (eds.) Geospatial Semantics and Semantic Web: Foundations, Algorithms, and Applications, Springer-Verlag, (2011)
19. Ikkala, E., Tuominen, J., Hyvönen, E.: Contextualizing Historical Places in a Gazetteer by Using Historical Maps and Linked Data. In: Proceedings of Digital Humanities 2016, short papers, Kraków, Poland (2016)
20. Bates, M.: What is Browsing – really? A Model Drawing from Behavioural Science Research. Information Research 12. http://www.informationr.net/ir/12-4/paper330.html (2007)
21. Toms, E.G.: Understanding and Facilitating the Browsing of Electronic Text. International Journal of Human-Computer Studies, 52(3), 423–452 (2000)
22. McNamee, P., Mayfield, J.C., Piatko, C.D.: Processing Named Entities in Text. Johns Hopkins APL Technical Digest, 30(1), pp. 31–40. (2011)