DHQ: Digital Humanities Quarterly
2017
Volume 11 Number 3

Old Content and Modern Tools – Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910

Kimmo Kettunen <kimmo_dot_kettunen_at_helsinki_dot_fi>, National Library of Finland, Mikkeli, Finland
Eetu Mäkelä <eetu_dot_makela_at_aalto_dot_fi>, University of Helsinki, Helsinki Centre for Digital Humanities
Teemu Ruokolainen <teemu_dot_ruokolainen_at_helsinki_dot_fi>, National Library of Finland, Mikkeli, Finland
Juha Kuokkala <juha_dot_kuokkala_at_helsinki_dot_fi>, University of Helsinki, Department of Modern Languages, Helsinki, Finland
Laura Löfberg <l_dot_lofberg_at_lancaster_dot_ac_dot_uk>, Department of Linguistics and English Language, Lancaster University, UK

Abstract

Named Entity Recognition (NER), search, classification and tagging of names and name-like 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, the performance of a NER system is genre- and domain-dependent and also used entity categories vary [Nadeau and Sekine 2007]. The most general set of named entities is usually some version of a tripartite categorization of locations, persons, and organizations. In this paper we report trials and evaluation of NER with data from a digitized Finnish historical newspaper collection (Digi). Experiments, results, and discussion of this research serve development of the web collection of historical Finnish newspapers.

Digi collection contains 1,960,921 pages of newspaper material from 1771–1910 in both 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 70–75 % [Kettunen and Pääkkönen 2016]. Our principal NE tagger is a rule-based tagger of Finnish, *FiNER*, provided by the FIN-CLARIN consortium. We also show results of limited category semantic tagging with tools of the Semantic Computing Research Group (SeCo) of the Aalto University. Three other tools are also evaluated briefly.

This paper reports the first large scale results of NER in a historical Finnish OCRed newspaper collection. Results of this research supplement NER results of other languages with similar noisy data. As the results are also achieved with a small and morphologically rich language, they illuminate the relatively well-researched area of Named Entity Recognition from a new perspective.

1. Introduction

The National Library of Finland has digitized a large proportion of the historical newspapers published in Finland between 1771 and 1910[1] [Bremer-Laamanen 2014] [Kettunen et al. 2014]. This collection contains 1,960,921 million pages in Finnish and Swedish. The Finnish part of the collection consists of about 2.4 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.[2] The collection can also be accessed through the Korp[3] environment, 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[4] style information retrieval test collection has been produced out of a small part of the Digi newspaper material at the University of Tampere [Järvelin et al. 2016]. An open data package of the

whole collection was released in the first quarter of 2017 [Pääkkönen et al. 2016].

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 2016 the service had about 18 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. 2

Digi is part of the growing global network of digitized newspapers and journals, and historical newspapers are considered more and more as an important source of historical knowledge. As the amount of digitized data accumulates, tools for harvesting the data are needed to gather information and to add structure to the unstructured mass [Marrero et al. 2013]. Named Entity Recognition has become one of the basic techniques for information extraction of texts since mid-1990s [Nadeau and Sekine 2007]. In its initial form, NER was used to find and mark semantic entities like persons, locations, and organizations 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 [Nadeau and Sekine 2007][Kokkinakis et al. 2014]. 3

Our goal using NER is to provide users of Digi better means for searching and browsing the historical newspapers (i.e. new ways to structure, access, and, possibly, enhance information). Different types of names, especially person names and names of locations, are frequently used as search terms in different newspaper collections [Crane and Jones 2006]. They can also provide browsing assistance to collections if the names are recognized, tagged in the newspaper data, and put into the index [Neudecker et al. 2014]. A fine example of applying name recognition to historical newspapers is La Stampa's historical newspaper collection[5]. After basic keyword searches, 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 named entity annotations of newspaper text allow a more semantically-oriented exploration of content of the larger archive. Another large scale (152M articles) NER analysis of the Australian historical newspaper collection (Trove) with usage examples is described in Mac Kim and Cassidy [Mac Kim and Cassidy 2015]. 4

Named Entity Recognition is a tool that needs to be used for some useful purpose. In our case, extraction of person and place names is primarily a tool for improving access to the Digi collection. After getting the recognition rate of the NER tool to an acceptable level, we need to decide how we are going to use extracted names in Digi. Some exemplary suggestions are provided by the archives of La Stampa and Trove Names [Mac Kim and Cassidy 2015]. La Stampa's usage of names provides informational filters after a basic search has been conducted. User can further look for persons, locations, and organizations mentioned in the article results. This kind of approach expands browsing access to the collection and possibly enables entity linking [Bates 2007] [Toms 2000] [McNamee et al. 2011]. Trove Names' name search takes the opposite approach: users first search for names and the search reveals articles where the names occur. We believe that the La Stampa's use of names in the GUI of the newspaper collection is more informative and useful for users, as the Trove style can be achieved with the normal search function in the GUI of the newspaper collection. 5

If we consider possible uses of presently evaluated NER tools–FiNER, the FST, and Connexor's tagger–for our newspaper collection, they only perform basic recognition and classification of names, which is the first stage of entity handling [McNamee et al. 2011]. To be of practical use names would need both intra-document reference entity linking, as well as multiple document reference entity linking [McNamee et al. 2011] [Ehrmann et al. 2016b]. This means that all the occurrences of equal names either in a single document or several documents, would be marked as occurrences of the same name. ARPA's semantic entity linking is of broader use, and entity linking to external knowledge sources, such as Wikipedia, has been used, for example, in the Europeana newspaper collection with names [Neudecker et al. 2014] [Hallo et al. 2016]. 6

One more possible use for NER involves tagging and classifying images published in the newspapers. Most of the images (photos) have short title texts. It seems that many of the images represent locations and persons, with names of the objects mentioned in the image title. As image recognition and classifying low-quality print images may not be very 7

feasible, image texts may offer a way to classify at least a reasonable part of the images. Along with NER, topic detection could also be done to the image titles. Image content tagging thus could be one clear application for NER.

Our main research question in this article is, how well or poorly names can be recognized in an OCRed historical Finnish newspaper collection with readily available software tools. The task has many pitfalls that will affect the results. First, the word level quality of the material is quite low [Kettunen and Pääkkönen 2016], which is common for OCRed historical newspaper collections. Second, the language technology tools that are available are made for modern Finnish. Third, there is neither comparable NER data of historical Finnish, nor a standard evaluation corpus. Thus our results form a first baseline for NER of historical Finnish. One would be right to expect that results will not be very good, but they will give us a realistic empirical perspective on NER's usability with our data [Kettunen and Ruokolainen 2017]. 8

We are using five readily available tagging tools for our task. By using a set of different types of tools we are able to pinpoint common failures in NE tagging our type of material. Observations of error analysis of the tools will help us possibly to improve further NE tagging of historical OCRed material of Finnish. Differences in tagging between these tools also help us to analyze further the nature of our material. 9

We will not provide a review of basic NER literature; those who are interested in getting an overall picture of the topic, can start with Nadeau and Sekine [Nadeau and Sekine 2007] and Marrero et al. [Marrero et al. 2013], who offer historical and methodological basics as well as critical discussion of the theme. References in Nadeau and Sekine and Marrero et al. as well as references in the current paper allow further familiarization. Specific problems related to historical language and OCR problems are discussed in the paper in relation to our data. 10

The structure of the paper is following: first we introduce our NER tools, our evaluation data and the tag set. Then we will show results of evaluations, analyze errors of different tools, and finally discuss the results and our plans for using NER with the online newspaper collection. 11

2. NER Software and Evaluation

For recognizing and labelling named entities in our evaluation we use FiNER software as a baseline NER tool. Our second main tool, SeCo's ARPA, is a different type of tool, mainly used for Semantic Web tagging and linking entities [Mäkelä 2014][6], but it could be adapted for basic NER, too. Besides these two tools, three others were also evaluated briefly. Connexor[7] has NER for modern Finnish, which is commercial software. The multilingual package Polyglot[8] also works for Finnish and recognizes persons, places, and organizations. A semantic tagger for Finnish [Löfberg et al. 2005] also recognizes the three types of names. 12

All our taggers have been implemented as analysers of modern Finnish, although ARPA's morphological engine is able to deal with 19th century Finnish, too. As far as we know there is no NE tagger for historical Finnish available. Before choosing FiNER and ARPA we also tried a commonly used trainable free statistical tagger, Stanford NER, but were not able to get reasonable performance out of it for our purposes although the software has been used successfully for other languages than English. Dutch, French and German named entity recognition with the Stanford NER tool has been reported in the Europeana historical newspaper project, and the results have been good [Neudecker et al. 2014] [Neudecker 2016]. However, we have been able to use Stanford NER with our material after this evaluation was finished, but results of this are not available yet. 13

As far as we now, besides the five tools evaluated in this paper, there are not many other existing tools to do NER analysis for Finnish[9]. FiNER has an updated version, but it does not seem to perform any better than the version we are using in our evaluation. In our discussion section we will get back to possibility of using a more historically aware statistical NE tagger. It should also be noted that all the tools except ARPA were used *as is*. That is, we did not enhance their performance with any additional tools or modifications, and the tools were treated as black boxes. 14

2.1. FiNER

FiNER[10] is a rule-based named-entity tagger, which in addition to surface text forms utilizes grammatical and lexical 15

information from a morphological analyzer (Omorfi[11]). 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 (and e.g. Estonian, [Tkachenko et al. 2013]), where a single lexeme may theoretically have thousands of different inflectional forms.

16 The focus of FiNER is in recognizing different types of proper names. Additionally, it can identify the majority of Finnish expressions of time and, for example, 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 remain in the nominative (or genitive) case. Thus, preceding 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.

17 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) [Lindén et al. 2013] [Silfverberg 2015]. 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 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 fifteen different entities that belong to five semantic categories: location, person, organization, measure, and time [Silfverberg 2015].

2.2. ARPA

18 SeCo's ARPA [Mäkelä 2014] 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.

19 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[12], which supports historical morphological variants, as well as lemma guessing for words outside the standard vocabulary. 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.

2.3. Evaluation Data

As there was no evaluation collection for Named Entity Recognition of 19th century Finnish, we needed to create one first. As evaluation data we used samples from different decades out of the Digi collection. Kettunen and Pääkkönen calculated, among other things, number of words in the data for different decades [Kettunen and Pääkkönen 2016]. It turned out that most of the newspaper data was published in 1870–1910, and beginning and middle 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 the two decades of 1890–1910. 20

We aimed for an evaluation collection of 150,000 words. To emphasize the importance of the 1870–1910 material we took 50K of data from time period 1900–1910, 10K from 1890–1899, 10K from 1880–1889, and 10K from 1870–1879. The remaining 70K 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. Decade-by-decade word recognition rates in Kettunen and Pääkkönen show that word recognition rate during the whole 19th century is quite even, variation being maximally 10% units for the whole century [Kettunen and Pääkkönen 2016]. In the latter part of the 19th century variation of recognition of words is maximally 4% units. Thus we believe that temporal dimension of the data should not bring great variation to the NER results. It may be possible, however, that older data has old names that are out of the scope of our tools.[13] 21

The final manually tagged evaluation data consists of 75,931 lines, each line having one word or other character data. By character data we mean here that the line contains misrecognized words that have a variable amount of OCR errors. 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 in the evaluation collection can be recognized by a modern Finnish morphological analyzer. The recognition rate in the whole index of the newspaper collection is estimated to be in the range of 70–75% [Kettunen and Pääkkönen 2016]. Evaluation data was input to FiNER as small textual snippets. 71% of the tagger's input snippets have five or more words, the rest have fewer than five words in the text snippet. Thus the amount of context the tagger can use in recognition is varying. 22

FiNER uses fifteen tags for different types of entities, which 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 where 80% of the ca. 149,000 occurrences of top 1000 search term types consisted of first and last names of persons and place names [Kettunen et al. 2014]. This kind of search term use is very common especially in the humanities information seeking [Crane and Jones 2006]. 23

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 eight final chosen tags are shown and explained below (Entity followed by Tag Meaning). 24

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 commonly used semantic NE categories: persons, locations, and organizations. In locations we have four different categories and with organizations two. Temporal expressions were included in the tag set due to their general interest in the newspaper material. Especially persons and locations fulfill content validity condition of an experimental unit [Marrero et al. 2013] [Urbano 2011] with our material, as locations and persons are most sought for by users in the newspapers according to our log studies. 25

Manual tagging of the evaluation corpus was done by the third author, who had previous experience in tagging modern Finnish with tags of the FiNER tagger. Tagging took one month, and quality of the tagging and its principles were 26

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 2010]). 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. Thus no spelling corrections or normalizations were made in the evaluation data.

All the evaluation runs were performed with the tagged 75K evaluation set. This set was not used in configuration of the tools. 27

2.3.1 Testing Lexical Coverage of FiNER

To get an idea how well FiNER recognizes names in general, we evaluated it with a separate list of 75,980 names of locations and persons. We included in the list modern first names and surnames, old first names from the 19th century, names of municipalities, and names of villages and houses. The list also contains names in Swedish, as Swedish was the dominant language in Finland during most of the 19th century[14]. The list has been compiled from independent open sources that include, for example, the Institute for the Languages of Finland, the National Land Survey of Finland, and the Genealogical Society of Finland, among others. All the names were given to FiNER as part of a predicative pseudo sentence *X on mukava juttu* ("X is a nice thing") so that the tagger had some context to work with, not just a list of names. 28

FiNER recognized 55,430 names out of the list, which is 72.96%. Out of these 8,904 were tagged as persons, 35,733 as *LocXxxs*, and 10,408 as *LocGpls*. The rest were tagged as organizations, streets, time, and titles. Among locations, FiNER favors general locations (*LocXxxs*). As *LocGpls* it tags locations that have some clear mention of a natural geographical entity as part of the name (lake, pond, river, hill, rapids, etc.), but this is not clear cut, as some names of this type seem to get tag of *LocXxx*. It would be reasonable to use only one location tag with FiNER, as the differences between location categories are not very significant. 29

Among the names that FiNER does not recognize are foreign names, mostly Swedish (also in Sami), names that can also be common nouns, different compound names, and old names. Variation of *w/v*, one the most salient differences of 19th century Finnish and modern Finnish, does not impair FiNER's tagging, although it has a clear impact on general recognizability of 19th century Finnish [Kettunen and Pääkkönen 2016]. Some other differing morphological features of 19th century Finnish [Järvelin et al. 2016] (cf. Table 1) may affect recognition of names with FiNER. Most of the differences can be considered as spelling variations, and some of them could affect results of NER too. Also writing of compounds was unstable in 19th century Finnish. Compounds could be spelled either together, with hyphen, or with whitespace, for example *diakonissalaitos*, *diakonissa-laitos* or *diakonissa laitos.* This may also affect spelling of names in the texts. 30

2.4. Results of the Evaluation

We evaluated performance of the NER tools using the *conlleval*[15] script used in Conference on Computational Natural Language Learning (CONLL). *Conlleval* uses standard measures of precision, recall and F-score–one defined as 2PR/(R+P)–where P is precision and R recall (cf. [Manning and Schütze 1999, 269]). Evaluation is based on "exact-match evaluation" [Nadeau and Sekine 2007]. 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 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" [Poibeau and Kosseim 2001]. Thus the evaluation criteria are strict, especially for multipart entities. Strict evaluation was possible only for FiNER and ARPA, which marked the boundaries of the entities. 31

We performed also a looser evaluation for all the taggers. In a looser evaluation the categories were treated so that any correct marking of an entity regardless its boundaries was considered a hit. 32

2.5. Results of FiNER

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. An example of a multipart name would be *G. E. Jansson*, with two initials. A proper tagging for this would be *G. <EnamexPrsHum> E. <EnamexPrsHum> Jansson</EnamexPrsHum>*. If only the surname *Jansson* would appear in the text, a proper tagging for this would be *<EnamexPrsHum/>*. 33

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 recognized quite badly by FiNER, which is not surprising as the quality of the text data is quite low. Recognition of multipart entities is mostly very low. Some part of the entities may be recognized, but rest is not. Out of multiword entities person names 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, and their F-score is 40–45. If the three different locations (*LocGpl*, *LocPpl* and *LocXxx*) are joined in strict evaluation as one general location, *LocXxx*, one part locations get precision of 65.69, recall of 50.27, and F-score of 56.96 with 1533 tags. Multipart locations are found poorly even then. FiNER seems to have a tendency to tag most of the *LocPpl*s as *LocXxx*s. *LocGpl*s are also favored instead of *LocPpl*s. On the other hand, only one general location like LocXxx could be enough for our purposes, and these results are reasonably good 34

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. The 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. 35

Label	P	R	F-score	Number of tags found
<EnamexPrsHum>	63.30	53.69	58.10	2681
<EnamexLocXxx>	69.05	49.21	57.47	1541
<EnamexLocStr>	83.64	25.56	39.15	55
<EnamexOrgEdu>	51.72	47.62	49.59	58
<EnamexOrgCrp>	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

2.6. Results of ARPA

Our third evaluation was performed for a limited tag set with tools of the SeCo's ARPA. We first analyzed ARPA's lexical coverage with the same word list that was used with FiNER. ARPA recognized in the recognition word list (of 75,980 tokens) 74,068 as either locations or persons (97.4 %). 67,046 were recognized as locations, and 37,456 as persons. 30,434 names were tagged as both persons and locations. Among the 1912 names that were not recognized by ARPA were the same kind of foreign names that were left unrecognized by FiNER. 13% of the unrecognized names were compounds with hyphen, such as Esa-Juha, Esa-Juhani. This type could be easily handled by ARPA with minor modifications to configuration. In general, the test showed that ARPA's lexicons are more comprehensive than those of FiNER. 36

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[16] 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 [Hyvönen et al. 2011] 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 [Ikkala et al. 2016]. To account for international place names, the names were also queried against the Geonames database[17] as well as Wikidata[18]. 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. 37

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

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

Table 4 describes the results of location recognition with ARPA. With one exception (*New York*), only one word entities were discovered by the software. 38

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. 39

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 lemma guessing (which is needed because OCR errors are out of the lemmatizer's vocabulary) is currently extremely sensitive to OCR errors–particularly in the suffix parts of words. 40

After the place recognition pipeline was finalized, a further test was done to see if the ARPA pipeline could be used for also person name recognition. The Virtual International Authority File was used as a lexicon of names 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 cannot always be linked to strong identifiers, but for a pure NER task, recall is improved. 41

Table 5 shows results of this evaluation without fuzzy matching of names and Table 6 with fuzzy matching. Table 7 shows evaluation results with loose criteria without fuzzy matching and Table 8 loose evaluation with fuzzy matching. 42

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
</EnamexPrsHum>	63.07	62.97	63.01	1148
<EnamePrsHum>	62.25	61.77	62.01	1425

Table 6. Evaluation results for ARPA: fuzzy matching

Label	P	R	F-score	Number of tags
<EnamexPrsHum>	63.61	45.27	52.90	3636
<EnamexLocXxx>	44.02	64.58	52.35	2933

Table 7. Evaluation results for ARPA with loose criteria: no fuzzy matching

Label	P	R	F-score	Number of tags
<EnamexPrsHum>	34.39	78.09	51.57	6388
<EnamexLocXxx>	44.02	64.58	52.35	2933

Table 8. Evaluation results for ARPA with loose criteria: fuzzy matching

Recall of recognition increases markedly in fuzzy matching, but precision deteriorates. More multipart location names are also recognized with fuzzy matching. In loose evaluation more tags are found but precision is not very good and thus the overall F-score is a bit lower than in the strict evaluation. 43

2.7. Comparison of Results of FiNER and ARPA

To sum up results of our two main tools, we show one more table where the main comparable results of FiNER and ARPA are shown in parallel. These are results of loose evaluations from Tables 2 and 7. 44

	P FiNER	P ARPA	R FiNER	R ARPA	F-score FiNER	F-score ARPA
<EnamexLocXxx>	69.05	63.61	49.21	45.27	57.47	52.90
<EnamexPrsHum>	63.30	44.02	53.69	64.58	58.10	52.35

Table 9. Comparative evaluation results of FiNER and ARPA with loose criteria

As one can see, FiNER perfoms slightly better with locations and persons than ARPA. The difference in F-scores is about 5 percentage units. 45

2.8. Results of Other Systems

Here we report briefly results of three other systems that we evaluated. These are Polyglot, a Finnish semantic tagger [Löfberg et al. 2005] and Connexor's NER. 46

Polyglot[19] is a natural language pipeline that supports multilingual applications. NER is among Polyglot's tools. According to Polyglot Website, the NER models of Polyglot were trained on datasets extracted automatically from Wikipedia. Polyglot's NER supports currently 40 major languages. 47

Results of Polylot's performance in a loose evaluation with three categories are shown in table 10. 48

Label	P	R	F-score	Number of tags found
<EnamexPrsHum>	75.99	34.60	47.55	1433
<EnamexLocXxx>	83.56	32.28	46.57	821
<EnamemOrgCrp>	5.77	1.70	2.63	208

Table 10. Evaluation results of Polyglot with loose criteria and three categories

As can be seen from the figures, Polyglot has high precision with persons and locations, but quite bad recall, and F-scores are thus about 10% units below FiNER's performance and clearly below ARPA's performance. With corporations Polyglot performs very poorly. The reason for this is probably the fact that names of companies have changed and organizations taken out of Wikipedia do not contain old company names. 49

2.8.1. Results of a Semantic Tagger of Finnish

Our fourth tool is a general semantic tagger tool for Finnish. The Finnish Semantic Tagger (FST) has its origins in Benedict, the EU-funded language technology project, the aim of which was to discover an optimal way of catering to the needs of dictionary users in modern electronic dictionaries by utilizing state-of-the-art language technology. Semantic tagging in its rule-oriented form (vs. statistical learning) can be briefly defined as a dictionary-based process of identifying and labeling the meaning of words in a given text according to some classification. FST is not a NER tool as such; it has first and foremost been developed for the analysis of full text. 50

The Finnish semantic tagger was developed using the English Semantic Tagger as a model. This semantic tagger was developed at the University Centre for Corpus Research on Language (UCREL) at Lancaster University as part of the UCREL Semantic Analysis System (USAS) framework [Rayson et al. 2004], and both these equivalent semantic taggers were utilized in the Benedict project in the creation of a context-sensitive search tool for a new intelligent dictionary. In different evaluations the FST has been shown to be capable of dealing with most general domains which appear in a modern standard Finnish text. Furthermore, although the semantic lexical resources were originally developed for the analysis of general modern standard Finnish, evaluation results have shown that the lexical resources are also applicable to the analysis of both older Finnish text and the more informal type of writing found on the Web. In addition, the semantic lexical resources can be tailored for various domain-specific tasks thanks to the flexible USAS category system. The semantically categorized single word lexicon of the FST contains 46,225 entries and the multiword expression lexicon contains 4,422 entries [Piao et al. 2016], representing all parts of speech. There are plans to expand the semantic lexical resources for the FST by adding different types of proper names in the near future in order to tailor them for NER tasks. 51

FST tags three different types of names: personal names, geographical names, and other proper names. These are tagged with tags Z1, Z2, and Z3, respectively [Löfberg et al. 2005]. FST does not distinguish first names and surnames, but it is able to tag first names of persons with male and female sub tags. As Z3 is a slightly vague category with names of organizations among others, we evaluate only categories Z1 and Z2–persons and locations. 52

FST tagged the list of 75,980 names as follows: it marked 5,569 names with tags Z1-Z3. Out of these 3,473 were tagged as persons, 2,010 as locations and rest as other names. It tagged 47,218 words with the tag Z99, which is a mark for lexically unknown words. Rest of the words, 23,193, were tagged with tags of common nouns. Thus FST's recall with the name list is quite low compared to FiNER and ARPA. 53

In Table 11 we show results of FST's tagging of locations and persons in our evaluation data. As the tagger does not distinguish multipart names only loose evaluation was performed. We performed two evaluations: one with the words as they are, and the other with w to v substitution. 54

Label	P	R	F-score	Number of tags found
<EnamexPrsHum>	76.48	22.48	34.75	897
<EnamexLocXxx>	67.11	47.72	55.78	1420
<EnamexPrsHum> w/v	76.10	23.06	35.39	908
<EnamexLocXxx> w/v	69.66	51.34	59.12	1536

Table 11. Evaluation of FST tagger with loose criteria and two categories. W/v stands for w to v substitution in words.

Substitution of *w* with *v* decreased number of unknown words to FST with about 3% units and has a noticeable effect on detection of locations and a small effect on persons. Overall locations are recognized better; their recognition with w/v substitution is slightly better than FiNER's and better than ARPA's overall. FST's recognition of persons is clearly inferior to that of FiNER and ARPA. 55

2.8.2 Results of Connexor's NER

Connexor Ltd. has provided different language technology tools, and among them is name recognition[20]. There is no documentation related to the software, but Connexor states on their Web pages that "using linguistic and heuristic methods, the names in the text can be tagged accurately". Software's name type repertoire is large; at least 31 different types of names are recognized. These are part of 9 larger categories like NAME.PER (persons), NAME.PRODUCT (products), NAME.GROUP (organizations), NAME.GPE (locations) etc. Boundaries of names are not tagged, so we perform only a loose evaluation. 56

As earlier, our interest is mainly in persons and locations. Connexor's tags NAME.GPE, NAME.GPE.City, NAME.GPE.Nation, NAME.GEO.Land and NAME.GEO.Water were all treated as <EnamexLocXxx>. NAME.PER, NAME.PER.LAW, NAME.PER.GPE, NAME.PER.Leader, NAME.PER.MED, NAME.PER.TEO and NAME.PER.Title were all treated as <EnamexPrsHum>. All other tags were discarded. Results of Connexor's tagger are shown in Table 12. 57

Label	P	R	F-score	Number of tags found
<EnamexPrsHum>	44.86	76.02	56.40	5321
<EnamexLocXxx>	66.76	55.93	60.87	1802

Table 12. Evaluation of Connexor's tagger with loose criteria and two categories

Results show that Connexor's NE tagger is better with locations–achieving the best overall F-score of all the tools–but also persons are found well. Recall with persons is high, but low precision hurts overall performance. Data inspection shows that Connexor's tagger has a tendency to tag words beginning with upper case as persons. Locations are also mixed with persons many times. 58

2.9. Results overall

If we consider results of FiNER and ARPA overall, we can make the following observations. They both seem to find two part person names best, most of which consist of first name and last name. In strict evaluation ARPA appears better with locations than FiNER, but this is due to the fact that FiNER has a more fine-grained location tagging. With one location tag FiNER performs equally well as ARPA. In loose evaluation they both seem to find almost equally well locations and persons, but FiNER gets slightly better results. FiNER finds educational organizations best, although they are scarce in the data. Corporations are also found relatively well, even though this category is prone to historical changes. FiNER is precise in finding two part street names, but recall in street name tagging is low. High precision is most likely due to common part *–katu* in street names: they are easy to recognize, if they are spelled right in the data. Low recall indicates bad OCR in street names. 59

Out of the other three tools we evaluated, the FST was able to recognize locations slightly better than FiNER or ARPA in loose evaluation when w/v variation was neutralized. Connexor's tagger performed at the same level as FINER and ARPA in loose evaluation. Its F-score with locations was the best performance overall. Polyglot performed worst of all the systems. 60

We evaluated lexical coverage of three of our tools with a wordlist that contained 75,980 names. ARPA's lexical coverage of the names was by far the best as it was able to recognize 97.4% of the names. FiNER recognized 73% of the names in this list and the FST recognized only about 7% of them as names. It marked about 62% of the names as unknown. Thus it seems that very high lexical coverage of names may not be the key issue in NER, as all three tools performed tagging of locations at the same level. The FST performed worst with persons although it had clearly more person names than locations in its lexicon. 61

One more caveat of performance is in order, especially with FiNER. After we had achieved our evaluation results, we evaluated FiNER's context sensitivity with a small test. Table 13 shows effect of different contexts on FiNER's tagging for 320 names of municipalities. In the leftmost column are results, where only a name list was given to FiNER. In the three remaining columns, names of the municipality was changed from the beginning of a clause to middle and end. Results imply that there is context sensitivity in FiNER's tagging. With no context at all results are worst, and when the location is at the beginning of the sentence, FiNER also misses more tags than in the other two positions. Overall it tags about two thirds of the municipality names as locations (*LocXxx* and *LocGpl*) in all the three context positions. The high number of municipalities tagged as persons is partly understandable as names are ambiguous, but in many cases interpretation as a person is not well grounded. This phenomenon derives clearly from FiNER's tagging strategy that was explained at the end of section 2.1. At the beginning of the clause locations are not confused as much as persons, but this comes with a cost of more untagged names. 62

No context, list of names	With context 1: location at the beginning	With context 2: location in the middle	With context 3: location at the end
111 LocXxx	151 LocXxx	158 LocXxx	159 LocXxx
84 PrsHum	66 PrsHum	80 PrsHum	80 PrsHum
7 LocGpl	56 LocGpl	54 LocGpl	54 LocGpl
12 OrgCrp	10 OrgCrp	12 OrgCrp	11 OrgCrp
2 OrgTvr	2 OrgTvr	2 OrgTvr	2 OrgTvr
102 no tag	35 no tag	14 no tag	14 no tag

Table 13. FiNER's tagging for 320 names of municipalities with different positional context for the name

Same setting was tested further with 15,480 last names in three different clause positions. Positional effect with last name tagging was almost nonexistent, but amount of both untagged names and locative interpretations is high. 39% of last names are tagged as PrsHum, 19.5% are tagged as LocXxx, and about 34.6% get no tag at all. The rest 7% are in varying categories. Tagging of last names would probably be better if first names were given together with last names. Isolated last names are more ambiguous. 63

We did not test effects of contextualization with other taggers, but it may have had a minor effect on all our results, as input text snippets were of different sizes (see section 2.3). Especially if first and last names are separated to different input snippets identification of person names may suffer. 64

2.10. Error Analysis of Results

Ehrmann et al. [Ehrmann et al. 2016a] suggest that application of NER tools on historical texts faces three challenges: 1) noisy input texts, 2) lack of coverage in linguistic resources, and 3) dynamics of language. In our case the first obstacle is the most obvious, which will be shown in more detail. Lack of coverage in linguistic resources (e.g. in the form of missing old names in the lexicons of the NE tools) is also a considerable source of errors in our case, as our 65

tools are made for modern Finnish. With dynamics of language Ehrmann et al. refer to different rules and conventions for the use of written language in different times. In this respect late 19th century Finnish is not that different from current Finnish, but this can certainly affect the results and should be studied more thoroughly. Considering that all our NE tools are made for modern Finnish, our evaluation data is heavily out of their main scope [Poibeau and Kosseim 2001], even if ARPA uses historical Finnish aware Omorfi and FiNER is able to guess unrecognized word forms.

66 To be able to estimate the effect of bad OCR on the results, we made some additional trials with improved OCR material. We made 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 an old OCRed version, and third a new OCRed version. Besides character-level errors, word order errors have been corrected in the two new versions. For these texts we did not have a ground truth NE tagged version, and thus we could only count number of NE tags in different texts. 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 were increases in the number of tags in the 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. Thus there is about a 10–20 % unit overall increase in the number of NE tags in both of the new better quality text versions in comparison to the old OCRed text with both taggers.

67 Another clear indication of effect of the OCR quality on the NER results is the following observation: when the words in all the correctly tagged FiNER *Enamex*es of the evaluation data are analyzed with Omorfi, only 14.3% of them are unrecognized. With wrongly tagged FiNER *Enamex*es 26.3% of the words are unrecognized by Omorfi. On tag level the difference is even clearer, as can be seen in recognition figures of Fig. 1 with words of locations and persons of FiNER, ARPA, FST and Connexor analyses (FiNER's analysis was reduced to a single location class). Thus improvement in OCR quality will most probably bring forth a clear improvement in NER of the material.

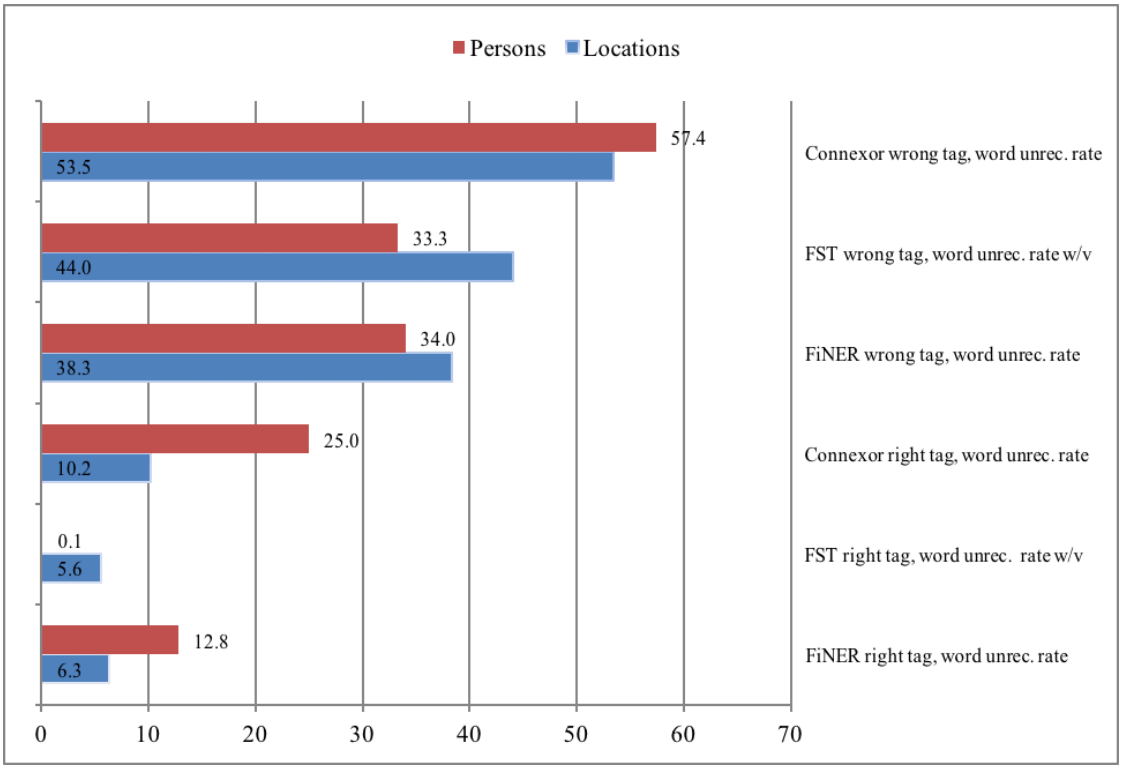

Figure 1. Word unrecognition percentages with rightly and wrongly tagged locations and persons – recognition with Omorfi 0.3

68 We also performed tagger specific error analysis for our tools. This analysis is not systematic because sources for the errors often remain unclear, but it shows some general tendencies of our tools. Besides general and tagger errors it also

reveals some general characteristics of our data. The errors reported here can also be seen as common improvement goals for better NE tagging of our newspaper data.

69 Connexor's NER tool seems to get misspelled person names many times right, even though percentage of morphologically unrecognized words among the locations is quite high in Fig. 1. Some of the rightly tagged but clearly misspelled examples are *Elwira4w, Aletsanterinp, Söderhslm, Tuomaanp, Hcikinp, AleNander, Hartifaincu, Schvvindt* and *NyleniUß.* On the other hand, the tagger also tags common nouns with upper case initial as person names. Examples of these are words such as *Prospekteja, Diskontto, Telefooni.* Without proper documentation of the tagger we are inclined to believe that upper case detection is one of the cues for person name detection in the tagger. This favours detection of misspelled person names even without any handling of typographical errors as such. It can also bring quite odd errors that we listed – these words have nothing in common with person names. Especially odd is the error with *Telefooni* ("telephone") which is many times followed by a number, and the whole is presenting a phone number. Both the word *Telefooni* and the number are tagged as person names. Also some title-like words before person names are tagged as persons, although they are not names of persons. With regards to locations, Connexor seems to be more cautious: most of its locations are correct spellings of words. Only a few misspellings like *Suomuzsalmi* and *Siräiöniemi* are analyzed right as locations. In this case it seems that the correctly spelled locative parts–*salmi* ("strait") and *-niemi* ("cape") are cues for the tagger.

70 FiNER analyses some misspelled person names right, but previously mentioned *Söderhslm* and *AleNander* are not analysed as person names. The FST analyses only correctly-spelled person and location names correctly as it has no means to handle spelling errors or variation in names. It could benefit from some form of fuzzy matching to be able to handle misspellings and variations, although this can also be counterproductive, as was seen with ARPA's fuzzy matching results in Tables 5–8: fuzzy matching increased recall and lowered precision, and the overall effect in F-score could be either slightly positive or negative.

71 Closer examination of FiNER's street name results shows that problems in street name recognition are due to three main reasons: OCR errors in street names, abbreviated street names, and multipart street names with numbers as part of the name. In principle streets are easy to recognize in Finnish, while they have most of the time common part *-katu* ("street") as last part of their name, which is usually a compound word or a phrase. Common use of abbreviations with street names seems to be characteristic for the data and it can be considered as a style of writing for the era. Street name like *Aleksanterinkatu* would be written as *Aleksanterink.*

72 Another similar case are first name initials which are used a lot in 19th century Finnish newspaper texts. Names like *O. Junttila*, *Y. Koskinen* (first name initial with surname, there can also be more initials) are common. FiNER gets initials right if they are preceded by a title, but otherwise not, and thus it gets these mostly wrong. The FST is not able to analyse these at all. Connexor's tagger gets one initial right if it is followed by surname, but out of two initials it only analyses the second one. This kind of writing of names belongs also to Ehrmann et al's dynamics of language type of challenges [Ehrmann et al. 2016a]. Usage of initials only for first names was common in the 19th century Finnish newspapers, but it is not used in modern newspaper writing.

73 One common source of errors for all NE taggers originates from ambiguity of some name types. Many Finnish surnames can be also names of locations, either names of municipalities, villages or houses. These kind of names are e.g. *Marttila, Liuhala,* and *Ketola.* All our taggers make errors with these, as only contextual cues specify whether the name is used as a person name or as a name of location. Proper handling of these names would need some extra disambiguation effort from the taggers. Although FiNER has a pattern matching mechanism for these (cf. 2.1) it also makes errors with ambiguous names.

3. Discussion

74 We have shown in this paper evaluation results of NER for historical Finnish newspaper material from the 19th and early 20th century with two main tools, FiNER and SeCo's ARPA. Besides these two tools, we briefly evaluated three other tools: a Finnish semantic tagger, Polyglot's NER and Connexor's NER. We were not able to train Stanford NER for Finnish. As far as we know, the tools we have evaluated constitute a comprehensive selection of tools that are capable

of named entity recognition for Finnish although not all of them are dedicated NER taggers.

75 Word level correctness of the whole digitized newspaper archive is approximately 70–75% [Kettunen and Pääkkönen 2016]; the evaluation corpus had a word level correctness of about 73%. Regarding this and the fact that FiNER and ARPA and other tools 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 the poor quality of the text. Also historical spelling variation had some effect, but it should not be that high, as late 19th century Finnish is not too far from modern Finnish and can be analyzed reasonably well with modern morphological tools [Kettunen and Pääkkönen 2016]. Morphological analyzers used in both FiNER and ARPA seem to be flexible and are able to analyze our low quality OCRed texts with a guessing mechanism too. The FST and Connexor's NER also performed quite well with morphology.

76 NER experiments with OCRed data in other languages usually show some improvement of NER when the quality of the OCRed data has been improved from very poor to slightly better [Packer et al 2010] [Marrero et al. 2013] [Miller et al. 2000]. Results of Alex and Burns (2014) imply that with lower level OCR quality (below 70% word level correctness) name recognition is harmed clearly [Alex and Burns 2014]. Packer et al. (2010) 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 level errors [Packer et al 2010]. Miller et al. (2000) show that rate of achieved NER performance of a statistical trainable tagger degraded linearly as a function of word error rates [Miller et al. 2000]. On the other hand, results of Rodriguez et al. (2012) show that manual correction of OCRed material that has 88–92% word accuracy does not increase performance of four different NER tools significantly [Rodriguez 2012].

77 As the word accuracy of our material is low, it would be expected that better recognition results would be achieved if the word accuracy was around 80–90% instead of 70–75%. Our tests with different quality texts suggest this too, as do the distinctly different unrecognition rates with correctly and incorrectly tagged words.

78 Better quality for our texts may be achievable in the near future. Promising results in post-correction of the Finnish historical newspaper data have been reported recently: two different correction algorithms developed in the FIN-CLARIN consortium achieved correction rate of 20–35 % [Silfverberg et al. 2016]. We are also progressing in re-OCRing tests of the newspaper data with open source OCR engine, Tesseract[21], and may be able to improve the OCR quality of our data [Kettunen et al. 2016] [Koistinen et al. 2017]. Together improved OCR and post-correction may yield 80+% word level recognition for our data. Besides character level errors our material also has quite a lot of word order errors which may affect negatively the NER results [Packer et al 2010]. Word order of the material may be improved in later processing of the XML ALTO and METS data, and this may also improve NER results. It would also be important that word splits due to hyphenation could be corrected in the data [Packer et al 2010].

79 Four of the five taggers employed in the experiments are rule-based systems utilizing different kinds of morphological analysis, gazetteers, and pattern and context rules. The only exception was Polyglot. However, while there has been some recent work on rule-based systems for NER [Kokkinakis et al. 2014], the most prominent research on NER has focused on statistical machine learning methodology [Nadeau and Sekine 2007]. Therefore, we are currently developing a statistical NE tagger for historical Finnish text. For training and evaluating the statistical system, we are manually annotating newspaper and magazine text from the years 1862–1910 with classes for person, organization, and location. The text contains approximately 650,000 word tokens. After annotation we can utilize freely available toolkits, such as the Stanford Named Entity Recognizer, for teaching the NE tagger. We expect that the rich feature sets enabled by statistical learning will alleviate the effect of poor OCR quality on the recognition accuracy of NEs. Preliminary unpublished results comparing FiNER and adjusted Stanford NER for modern Finnish business newspaper data show that Stanford NER outperforms FiNER with about 30% units in combined results of five name categories. For recent work on successful statistical learning of NE taggers for historical data, see [Neudecker 2016].

80 On general level, there are a few lessons to be learned from our experiments for those that are working with other small languages that do not have a well-established repertoire of NER tools available. Some of them are well known, but worth repeating, too. First and most self-evidently, bad OCR was found to be the main obstacle for good quality NER

once again. This was shown clearly in section 2.10. The implications of this are clear: better quality data is needed to make NER working well enough to be useful. Second, slightly surprisingly, we noticed that differences in lexical coverage of the tools will not show that much in the NER results. ARPA had clearly the best lexical coverage of the tools and FST the worst coverage, but their NER performance with locations are quite equal. This could imply that very large lexicons for a NE tagger are not necessary and a good basic coverage lexicon is enough, but this could also be language specific. Third, we showed that historical language can be processed to a reasonable extent with tools that are made for modern language if nothing else is available. However, if best possible results need to be achieved, more historical data oriented tools need to be used. It is also possible that the quite short time frame of our material enhances performance of our tools. Fourth, results of The Finnish Semantic Tagger showed that NER does not need to be only a task for dedicated NER tools. This has also been shown with modern Finnish [Kettunen and Löfberg 2017]. FST performed almost as well in a modern Finnish NER evaluation task as FiNER. Thus, at least in some cases, one can have available a suitable tool for doing NER even if the tool does not look like a NE tagger. Fifth, results of Polyglot hint that general multilingual packages do not work very well with data of small languages if their training data was not suitable for the task at hand. A lesson that we learned too was, that too detailed named entity classification (original FiNER labeling) is probably not a good solution for our purposes. In the future we shall stick to the use of persons, locations, and organizations.

Our main emphasis with NER will be to use the names with the newspaper collection as a means to improve structuring, browsing, and general informational usability of the collection. A good enough coverage of the names with NER also needs to be achieved for this use, of course. A reasonable balance of P/R should be found for this purpose, but also other capabilities of the software need to be considered. These lingering questions must be addressed if we are to connect some type of functional NER to our historical newspaper collection's user interface. 81

Acknowledgements

First and third author were funded by the Academy of Finland, project Computational History and the Transformation of Public Discourse in Finland 1640–1910 (COMHIS), decision number 293 341. 82

Thanks to Heikki Kantola and Connexor Ltd. for providing the evaluation data of Connexor's NE tagger. 83

Corresponding author: Kimmo Kettunen, ORCID: 0000-0003-2747-1382 84

Notes

[1] Ten more years, 1911–1920, were opened for free use on Feb. 1, 2017. Our trials have been carried out with the material of 1771–1910.

[2] https://kitwiki.csc.fi/twiki/bin/view/FinCLARIN/KielipankkiAineistotDigilibPub

[3] https://korp.csc.fi/

[4] Cranfield style test collection is a prototypical evaluation model for information retrieval systems. It consists of a set of predefined topics/queries, a set of documents to be searched for and relevance assessments for documents. First such collections were used by Cyril W. Cleverdon at Cranfield University in the early 1960s to evaluate efficiency of indexing systems.

[5] http://www.archiviolastampa.it/

[6] An older demo version of the tool is available at http://demo.seco.tkk.fi/sarpa/#/

[7] https://www.connexor.com/nlplib/?q=technology/name-recognition

[8] http://polyglot.readthedocs.io/en/latest/NamedEntityRecognition.html

[9] Lingpipe's NER tool could probably be taught for Finnish, too, but we did not try it.

[10] http://www.helsinki.fi/~jkuokkal/finer_dist/

[11] https://github.com/flammie/omorfi

[12] Mäkelä, Eetu (2016). “LAS: an integrated language analysis tool for multiple languages.” *The Journal of Open Source Software.* http://joss.theoj.org/papers/10.21105/joss.00035; https://github.com/jiemakel/omorfi

[13] Ehrmann et al. show irregular time-based variance for Swiss French, but their data consists of almost 180 years between 1804 and 1981 [Ehrmann et al. 2016a].

[14] Finnish became the dominant language in newspapers from the beginning of 1890s [Kettunen et al. 2016].

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

[16] http://www.ldf.fi/dataset/pnr/

[17] http://geonames.org/

[18] http://wikidata.org/

[19] http://polyglot.readthedocs.io/en/latest/index.html

[20] https://www.connexor.com/nlplib/?q=technology/name-recognition

[21] https://github.com/tesseract-ocr

Works Cited

Alex and Burns 2014 Alex, B. and Burns, J. (2014), “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*, available at: http://dl.acm.org/citation.cfm?id=2595214 (accessed 10 October 2015).

Bates 2007 Bates, M. (2007), “What is Browsing – really? A Model Drawing from Behavioural Science Research”, *Information Research* 12, available at: http://www.informationr.net/ir/12-4/paper330.html(accessed 1 June 2016).

Bremer-Laamanen 2014 Bremer-Laamanen, M-L. (2014), “In the Spotlight for Crowdsourcing”, *Scandinavian Librarian Quarterly*, Vol. 46, No. 1, pp. 18–21.

Crane and Jones 2006 Crane, G. and Jones, A. (2006), “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, available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.6257&rep=rep1&type=pdf (accessed 1 June 2016).

Ehrmann et al. 2016a Ehrmann, M., Colavizza, G., Rochat, Y. and Kaplan, F. (2016a). “Diachronic Evaluation of NER Systems on Old Newspapers.” In *Proceedings of the 13th Conferenceon Natural Language Processing* (KONVENS 2016), 97–107. https://www.linguistics.rub.de/konvens16/pub/13_konvensproc.pdf (accessed March 28 2017).

Ehrmann et al. 2016b Ehrmann, M., Nouvel, D. and Rosset, S. (2016b), “Named Entity Resources – Overview and Outlook”, in *LREC 2016, Tenth International Conference on Language Resources and Evaluation*, available at http://www.lrec-conf.org/proceedings/lrec2016/pdf/987_Paper.pdf (accessed June 15 2016).

Hallo et al. 2016 Hallo, M., Luján-Mora, S., Maté, A. and Trujillo, J. (2016), “Current State of Linked Data in Digital Libraries”, *Journal of Information Science*, Vol. 42, No. 2, pp. 117–127.

Hyvönen et al. 2011 Hyvönen, E., Tuominen, J., Kauppinen T. and Väätäinen, J. (2011), “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 US, pp. 1–25.

Ikkala et al. 2016 Ikkala, E., Tuominen, J. and Hyvönen, E. (2016), “Contextualizing Historical Places in a Gazetteer by Using Historical Maps and Linked Data”, *in Digital Humanities 2016: Conference Abstracts*, Jagiellonian University & Pedagogical University, Kraków, pp. 573-577, available at: http://dh2016.adho.org/abstracts/39(accessed October 1 2016).

Järvelin et al. 2016 Järvelin, A., Keskustalo, H., Sormunen, E., Saastamoinen, M. and 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.

Kettunen and Löfberg 2017 Kettunen, K. and Löfberg, L. (2017). “Tagging Named Entities in 19th Century and Modern Finnish Newspaper Collection with a Finnish Semantic Tagger.” In *Nodalida 2017*, http://www.ep.liu.se/ecp/131/Title_Pages.pdf(accessed August 7 2017).

Kettunen and Pääkkönen 2016 Kettunen, K. and Pääkkönen, T. (2016), “Measuring Lexical Quality of a Historical Finnish Newspaper Collection – Analysis of Garbled OCR Data with Basic Language Technology Tools and Means”, in *LREC 2016, Tenth International Conference on Language Resources and Evaluation*, available at http://www.lrec-conf.org/proceedings/lrec2016/pdf/17_Paper.pdf(accessed 15 June 2016).

Kettunen and Ruokolainen 2017 Kettunen, K. and Ruokolainen, T. (2017). “Names, Right or Wrong: Named Entities in an OCRed Historical Finnish Newspaper Collection.” In *DATeCH 2017*, http://dl.acm.org/citation.cfm?id=3078084(accessed Augut 7, 2017).

Kettunen et al. 2014 Kettunen, K., Honkela, T., Lindén, K., Kauppinen, P., Pääkkönen, T. and Kervinen, J. (2014), “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), available at: http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-honkela-en.pdf (accessed March 15 2015).

Kettunen et al. 2016 Kettunen, K., Pääkkönen, T. and Koistinen, M. (2016), “Between Diachrony and Synchrony: Evaluation of Lexical Quality of a Digitized Historical Finnish Newspaper and Journal Collection with Morphological Analyzers”, in: Skadiņa, I. and Rozis, R. (Eds.), *Human Language Technologies – The Baltic Perspective*, IOS Press, pp. 122–129. Available at: http://ebooks.iospress.nl/volumearticle/45525 (accessed October 12 2016).

Koistinen et al. 2017 Koistinen, M., Kettunen, K. and Pääkkönen, T. (2017). “Improving Optical Character Recognition of Finnish Historical Newspapers with a Combination of Fraktur & Antiqua Models and Image Preprocessing.” In *NoDaLiDa 2017*, http://www.ep.liu.se/ecp/131/Title_Pages.pdf(accessed August 7 2017).

Kokkinakis et al. 2014 Kokkinakis, D., Niemi, J., Hardwick, S., Lindén, K., and Borin. L. (2014), “HFST-SweNER – a New NER Resource for Swedish”. in: *Proceedings of LREC 2014*, available at: http://www.lrec-conf.org/proceedings/lrec2014/pdf/391_Paper.pdf(accessed 15 June 2016).

Lindén et al. 2013 Lindén, K., Axelson, E., Drobac, S., Hardwick, S., Kuokkala, J., Niemi, J., Pirinen, T.A. and Silfverberg, M. (2013) “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.

Lopresti 2009 Lopresti, D. (2009), “Optical character recognition errors and their effects on natural language processing”, *International Journal on Document Analysis and Recognition,* Vol. 12, No. 3, pp. 141–151.

Löfberg et al. 2005 Löfberg, L., Piao, S., Rayson, P., Juntunen, J-P, Nykänen, A. and Varantola, K. (2005), “A semantic tagger for the Finnish language”, available at http://eprints.lancs.ac.uk/12685/1/cl2005_fst.pdf(accessed 15 June 2016).

Mac Kim and Cassidy 2015 Mac Kim, S. and Cassidy, S. (2015), “Finding Names in Trove: Named Entity Recognition for Australian”, in *Proceedings of Australasian Language Technology Association Workshop*, available at: https://aclweb.org/anthology/U/U15/U15-1007.pdf (accessed August 10 2016).

Manning and Schütze 1999 Manning, C. D., Schütze, H. (1999) *Foundations of Statistical Language Processing*. The MIT Press, Cambridge, Massachusetts.

Marrero et al. 2013 Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J. and Gómez-Berbís, J.M. (2013), “Named Entity Recognition: Fallacies, challenges and opportunities”, *Computer Standards & Interfaces* Vol. 35 No. 5, pp. 482–489.

McNamee et al. 2011 McNamee, P., Mayfield, J.C., and Piatko, C.D. (2011), “Processing Named Entities in Text”, *Johns Hopkins APL Technical Digest*, Vol. 30 No. 1, pp. 31–40.

Miller et al. 2000 Miller, D., Boisen, S., Schwartz, R. Stone, R. and Weischedel, R. (2000), “Named entity extraction from noisy input: Speech and OCR”, in *Proceedings of the 6th Applied Natural Language Processing Conference*, 316–324, Seattle, WA, available at: http://www.anthology.aclweb.org/A/A00/A00-1044.pdf (accessed 10 August 2016).

Mäkelä 2014 Mäkelä, E. (2014), “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, Springer, pp. 424–428.

Nadeau and Sekine 2007 Nadeau, D., and Sekine, S. (2007), “A Survey of Named Entity Recognition and Classification”, *Linguisticae Investigationes,* Vol. 30 No. 1, pp. 3–26.

Neudecker 2016 Neudecker, C. (2016), “An Open Corpus for Named Entity Recognition in Historic Newspapers”, in *LREC 2016, Tenth International Conference on Language Resources and Evaluation*, available at http://www.lrec-conf.org/proceedings/lrec2016/pdf/110_Paper.pdf (accessed June 17 2016).

Neudecker et al. 2014 Neudecker, C., Wilms, L., Faber, W. J., and van Veen, T. (2014), “Large-scale Refinement of Digital Historic Newspapers with Named Entity Recognition”, In *Proceedings of IFLA 2014,* available at: http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-neudecker_faber_wilms-en.pdf (accessed June 10 2016).

Packer et al 2010 Packer, T., Lutes, J., Stewart, A., Embley, D., Ringger, E., Seppi, K. and Jensen, L. S. (2010), “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, available at: http://dl.acm.org/citation.cfm?id=1871845 (accessed May 10 2016).

Piao et al. 2016 Piao, S., Rayson, P., Archer, D., Bianchi, F., Dayrell, C., El-Haj, M., Jiménez, R.-M., Knight, D., Kren, M., Löfberg, L., Nawab, R.M.A., Shafi, J., The, P.L. and Mudraya, O. (2016), “Lexical Coverage Evaluation of Large-scale Multilingual Semantic Lexicons for Twelve Languages”, in *LREC 2016, Tenth International Conference on Language Resources and Evaluation*, available at: http://www.lrec-conf.org/proceedings/lrec2016/pdf/257_Paper.pdf (accessed August 10 2016).

Poibeau and Kosseim 2001 Poibeau, T. and Kosseim, L. (2001), “Proper Name Extraction from Non-Journalistic Texts”, *Language and Computers,* Vol. 37 No. 1, pp. 144–157.

Pääkkönen et al. 2016 Pääkkönen, T., Kervinen, J., Nivala, A., Kettunen, K. and Mäkelä E. (2016), “Exporting Finnish Digitized Historical Newspaper Contents for Offline Use”, *D-Lib Magazine*, July/August, available at http://www.dlib.org/dlib/july16/paakkonen/07paakkonen.html(accessed August 15 2016).

Rayson et al. 2004 Rayson, P., Archer, D., Piao, S. L. and McEnery, T . (2004), “The UCREL semantic analysis system” in *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks* in association with *4th International Conference on Language Resources and Evaluation (LREC 2004)*, 25th May 2004, Lisbon, Portugal, pp. 7-12. Available at: http://www.lancaster.ac.uk/staff/rayson/publications/usas_lrec04ws.pdf (accessed August 10 2016).

Rodriguez 2012 Rodriguez, K.J., Bryant, M., Blanke, T. and Luszczynska, M. (2012), “Comparison of Named Entity Recognition Tools for raw OCR text”, in: *Proceedings of KONVENS 2012* (LThist 2012 wordshop), Vienna September 21, pp. 410–414.

Silfverberg 2015 Silfverberg, M. (2015), “Reverse Engineering a Rule-Based Finnish Named Entity Recognizer”, paper presented at Named Entity Recognition in Digital Humanities Workshop, June 15, Helsinki available at: https://kitwiki.csc.fi/twiki/pub/FinCLARIN/KielipankkiEventNERWorkshop2015/Silfverberg_presentation.pdf(accessed April 5 2016).

Silfverberg et al. 2016 Silfverberg, M., Kauppinen, P., and Linden, K. (2016), “Data-Driven Spelling Correction Using Weighted Finite-State Methods”, in *Proceedings of the ACL Workshop on Statistical NLP and Weighted Automata,* pp. 51–59, available at: https://aclweb.org/anthology/W/W16/W16-2406.pdf (accessed August 20 2016).

Tkachenko et al. 2013 Tkachenko, A., Petmanson, T., and Laur, S. (2013), “Named Entity Recognition in Estonian”, in *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing,* pp. 78–83, available at: http://aclweb.org/anthology/W13-24 (accessed May 10 2016).

Toms 2000 Toms, E.G. (2000), “Understanding and Facilitating the Browsing of Electronic Text”, *International Journal of Human-Computer Studies*, Vol. 52 No. 3, pp. 423–452.

Urbano 2011 Urbano, J. (2011). “Information Retrieval Meta-Evaluation: Challenges and Opportunities in the Music Domain.” International Society for Music Information Retrieval Conference, 609-614. https://pdfs.semanticscholar.org/df87/1a4c635d8b21fc68f2de0bd58ca32fa557ae.pdf