Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature

Lars Borin, Dimitrios Kokkinakis, Leif-Jöran Olsson
Litteraturbanken and Språkdata/Språkbanken
Department of Swedish Language, Göteborg University
Sweden
{first.last}@svenska.gu.se

Abstract

This paper provides a description and evaluation of a generic named-entity recognition (NER) system for Swedish applied to electronic versions of Swedish literary classics from the 19th century. We discuss the challenges posed by these texts and the necessary adaptations introduced into the NER system in order to achieve accurate results, useful both for metadata generation, but also for the enhancement of the searching and browsing capabilities of *Litteraturbanken*, the Swedish Literature Bank, an ongoing cultural heritage project which aims to digitize significant works of Swedish literature.

1 Introduction

In this paper we investigate generic named entity recognition (NER) technology and the necessary adaptation required in order to automatically annotate electronic versions of a number of Swedish literary works of fiction from the 19th century. Both the genre and language variety are markedly different from the text types that our NER system was originally developed to annotate. This presents a challenge, posing both specific and more generic problems that need to be dealt with.

In section 2 we present briefly the background and motivation for the present work, and section 3 gives some information on related work. In section 4 we provide a description of the named entity recognition system used in this work, its entity taxonomy, including the animacy recognition component and the labeled consistency approach that is explored. Problems faced in the literary texts and the kinds of adaptations performed in the recognition system as well as evaluation and error analysis are given in section 5. Finally, section 6 summarizes the work and provides some thoughts for future work.

2 Background

Litteraturbanken <http://litteraturbanken.se/> (the Swedish Literature Bank) is a cultural heritage project financed by the Swedish Academy[1]. Litteraturbanken has as its aim to make available online the full text of significant works of Swedish literature, old and new, in critical editions suitable for literary research and for the teaching of literature. There is also abundant ancillary material on the website, such as author presentations, bibliographies, thematic essays about authorships, genres or periods, written by experts in each field.

Similarly to many other literature digitization initiatives, most of the works in Litteraturbanken are such for which copyright has expired (i.e., at least 70 years have passed since the death of the author); at present the bulk of the texts are from the 18th, 19th and early 20th century. However, there is also an agreement with the organizations representing authors' intellectual property rights, allowing the inclusion of modern works according to a uniform royalty payment scheme. At present, Litteraturbanken holds about 150 works – mainly novels – by about 50 different authors. The text collection is slated to grow by 80–100 novel-length works (appr. 4–6 million words) annually.

1 The present permanent version of Litteraturbanken was preceded by a two-year pilot project by the same name, funded by *the Bank of Sweden Tercentenary Foundation*.

Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007), pages 1–8,
Prague, 28 June 2007.

Even at outset of the Litteraturbanken project, it was decided to design the technical solutions with language technology in mind. The rationale for this was that we saw these literary texts not only as representing Sweden's literary heritage, but also as high-grade empirical data for linguistic investigations, i.e. as corpus components. Hence, we wanted to build an infrastructure for Litteraturbanken which would allow this intended dual purpose of the material to be realized to the fullest.[2] However, we soon started to think about how the kinds of annotations that language technology could provide could be of use to others than linguists, e.g. literary scholars, historians and researchers in other fields in the humanities and social sciences.

Here, we will focus on one of these annotation types, namely NER and entity annotation. Combined with suitable interfaces for displaying, searching, selecting, correlating and browsing named entities, we believe that the recognition and annotation of named entities in Litteraturbanken will facilitate more advanced research on literature (particularly in the field of literary onomastics; see Dalen-Oskam and Zundert, 2004), but also, e.g., historians could find this facility useful, insofar as these fictional narratives also contain, e.g. descriptions of real locations, characterizations of real contemporary public figures, etc. Flanders et al. (1998: 285) argue that references to people in historical sources are of intrinsic interest since they may reveal "networks of friendship, enmity, and collaboration; familial relationships; and political alliances […] class position, intellectual affiliations, and literary bent of the author".

3 Related Work

The presented work is naturally related to research on NER, particularly as applied to diachronic/historical corpora. The technology itself has been applied to various domains and genres over the last couple of decades such as financial news and biomedicine, with performance rates difficult to compare since the task is usually tied to particular domains/genres and applications. For a concise overview of the technology see Borthwick, (1999). Even though this technology is widely used in a number of domains, studies dealing with historical corpora are mostly comparatively recent (see for instance the recent workshop on historical text mining; <http://ucrel.lancs.ac.uk/events/htm06/>).

Shoemaker (2005) reports on how the *Old Bailey Proceedings*, which contain accounts of trials that took place at the Old Bailey, the primary criminal court in London, between 1674 and 1834, was marked up for a number of semantic categories, including the crime date and location, the defendant's gender, the victim's name etc. Most of the work was done manually while support was provided for automatic person name[3] identification (*cf.* Bontcheva et al., 2002). The author mentions future plans to take advantage of the structured nature of the Proceedings and to use the lists of persons, locations and occupations that have already been compiled for annotating new texts.

Crane and Jones (2006) discuss the evaluation of the extraction of 10 named entity classes (personal names, locations, dates, products, organizations, streets, newspapers, ships, regiments and railroads) from a 19th century newspaper. The quality of their results vary for different entity types, from 99.3% precision for *Streets* to 57.5% precision for *Products*. The authors suggest the kinds of knowledge that digital libraries need to assemble as part of their machine readable reference collections in order to support entity identification as a core service, namely, the need for bigger authority lists, more refined rule sets and rich knowledge sources as training data.

At least least two projects are also relevant in the context of NER and historical text processing, namely NORA <http://www.noraproject.org/> and ARMADILLO <http://www.hrionline.ac.uk/armadillo/>. The goal of the first is to produce text mining software for discovering, visualizing, and exploring significant patterns across large collections of full-text humanities resources in existing digital libraries. The goal of the latter is to evaluate the benefits of automated mining techniques (including information extraction) on a set of online resources in eighteenth-century British social history.

2 This precluded the use of ready-made digital library or CMS solutions, as we wanted to be compatible with emerging standards for language resources and tools, e.g. TEI/(X)CES and ISO TC37/SC07, which to our knowledge has never been a consideration in the design of digital library or CM systems.

3 By using the General Architecture for Text Engineering (GATE) platform; <http://gate.ac.uk>.

4 Named Entity Recognition

Named entity recognition (NER) or entity identification/extraction, is an important supporting technology with numerous applications in a number of human language technologies. The system we use originates from the work conducted in the *Nomen Nescio* project; for details see Johannessen et al. (2005). In brief, the Swedish system is a multi-purpose NER system, comprised by a number of modules applied in a pipeline fash-ion. Six major components can be distinguished, making a clear separation between lexical, gram-matical and processing resources. The six compo-nents are:

- lists of **multiword names**, taken from various Internet sites or extracted from various corpora, running directly over the tokenised text being processed;
- a rule-based, **shallow parsing** component that uses finite-state grammars, one grammar for each type of entity recognized;
- a module that uses **the annotations produced by the previous two components**, which have a high rate in precision, in order to make decisions regarding other un-annotated entities. This module is further discussed in Section 4.2;
- lists of **single names** (approx. 100,000);
- **name similarity,** this module is further discussed in Section 4.3;
- a **theory revision and refinement** module, which makes a final control of an annotated document, in order to detect and resolve possible errors and assign new annotations based on existing ones, for instance by applying name similarity or by combining various annotation fragments.

4.1 Named-Entity Taxonomy

The nature and type of named entities vary depending on the task under investigation or the target application. In any case, *personal names*, *location* and *organization names* are considered “generic”. Since semantic annotation is not as well understood as grammatical annotation, there is no consensus on a standard tagset and content to be generally applicable. Recently, however, there have been attempts to define and apply richer name hierarchies for various tasks, both specific (Fleischman and Hovy, 2002) and generic (Sekine, 2004). Our current system implements a rather fine-grained named entity taxonomy with 8 main named entitiy types as well as 57 subtypes. Details can be found in Johannessen et al., 2005, and Kokkinakis, 2004. The eight main categories are:

- **Person** (PRS): people names (forenames, surnames), groups of people, animal/pet names, mythological, theonyms;
- **Location** (LOC): functional locations, geographical, geo-political, astrological;
- **Organization** (ORG): political, athletic, media, military, etc.;
- **Artifact** (OBJ): food/wine products, prizes, communic. means (vehicles) etc.;
- **Work&Art** (WRK): printed material, names of films and novels, sculptures etc.;
- **Event** (EVN): religious, athletic, scientific, cultural etc.;
- **Measure/Numerical** (MSR): volume, age, index, dosage, web-related, speed etc.;
- **Temporal** (TME).

Time expressions are important since they allow temporal reasoning about complex events as well as time-line visualization of the story developed in a text. The temporal expressions recognized include both relative (*nästa vecka* ‘next week’) and absolute expressions (*klockan 8 på morgonen i dag* ‘8 o’clock in the morning today’), and sets or sequences of time points or stretches of time (*varje dag* ‘every day’).

4.2 Animacy Recognition

The rule-based component of the person-name recognition grammar is based on a large set of designator words and a group of phrases and verbal predicates that most probably require an animate subject (e.g. *berätta* ‘to tell’, *fundera* ‘to think’, *tröttna* ‘to become tired’). These are used in conjunction with orthographic markers in the text, such as capitalization, for the recognition of personal names. In this work, we consider the first group (designators) as relevant knowledge to be extracted from the person name recognizer, which is explored for the annotation of animate instances in the literary texts. The designators are implemented as a separate module in the current pipeline, and constitute a piece of information which is considered important for a wide range of tasks (*cf.* Orasan and Evans, 2001).

The designators are divided into four groups: designators that denote the nationality or the ethnic/racial group of a person (e.g. *tysken* 'the German [person]'); designators that denote a profession (e.g. *läkaren* 'the doctor'); those that denote family ties and relationships (e.g. *svärson* 'son in law'); and finally a group that indicates a human individual but cannot be unambiguously categorized into any of the three other groups (e.g. *patienten* 'the patient'). Apart from this grouping, inherent qualities, for at least a large group of the designators, (internal evidence/morphological cues) also indicate referent (natural) gender. In this way, the animacy annotation is further specified for male, female or unknown gender; unknown in this context means unresolved or ambiguous, such as *barn* 'child'.

Swedish is a compounding language and compound words are written as a single orthographic unit (i.e. solid compounds). This fact makes the recognition of animacy straightforward with minimal resources and feasible by the use of a set of suitable headwords, and by capturing modifiers by simple regular expressions. Approximately 25 patterns are enough to identify the vast majority of animate entities in a text; patterns such as "inna/innan/innor", "man/mannen/män/männen", "log/logen/loger", "ktör/ktören/ktörer" and "iker/ikern/ikerna". For instance, the pattern in (1) consists of a reliable suffix "inna" which is a typical designator for female individuals, preceded by a set of obligatory strings and an optional regular expression which captures a long list of compounds (2).

(1)
```
[a-zåäö]*(kv|älskar|man|grev|…)inna
```

(2)
```
taleskvinna, yrkeskvinna, idrotts-
kvinna, ungkvinna, Stockholmskvin-
na, Dalakvinna, samboälskarinna,
lyxälskarinna, ex-älskarinna, sam-
largrevinna, exälskarinna, markgre-
vinna, majgrevinna, änkegrevinna,…
```

Examples of animacy annotations are given in (3). The attribute value *FAM* stands for *FAmily* relation and *Male*; *PRM* for *PRofession* and *Male*; *FAF* for *FAmily* relation and *Female* and finally *UNF* for *UNknown* and *Female*.

(3)
```
[…]<ENAMEX TYPE="FAM">riksgrefvin-
nans far</ENAMEX>, <ENAMEX TYPE=
"PRM">öfveramiralen</ENAMEX> […]
hade till <ENAMEX TYPE="FAF">mor
</ENAMEX> <ENAMEX TYPE="UNF">gre-
fvinnan</ENAMEX> Beata Wrangel från
[…]
```

Table (3) in Section 6.1 presents the results for the evaluation of this type of normative information. Note also, that in order to make the annotations more practical we have included the person name designators (e.g. 'herr' – 'Mr') in the markup as in (4); here *PRS* stands for *PeRSon*:

(4)
```
<ENAMEX TYPE="UNM">Herr</ENAMEX>
<ENAMEX TYPE="PRS" SBT="HUM">Boman
</ENAMEX> becomes <ENAMEX TYPE=
"PRS-UNM" SBT="HUM">Herr Boman
</ENAMEX>
```

4.3 Name Similarity

We can safely assume that the various system resources will not be able to identify all possible entities in the texts, particularly personal and location names. Although there is a large overlap between the names in the texts and the gazetteer lists, there were cases that could be considered as entity candidates but were left unmarked. This is because exhaustive lists of names even for limited domains are hard to obtain, and, in some domains even difficult to manage. Therefore, we also calculated the orthographic similarity between such words and the gazetteer content, according to the following criteria: a potential entity starts with a capital letter; it is ≥ 5 characters long; it is not part of any other annotation and it does not stand in the beginning of a sentence. We have empirically observed that the length of 5 characters is a reliable threshold, unlikely to exclude many NEs. As a matter of fact, only two such cases could be found in the evaluation sample, namely *ätten Puff* 'the family Puff' and "Yen-" in the context "Yen- kenberg"

As measure of orthographic similarity (or rather, difference) we used the Levenshtein distance (LD; also known as *edit distance*) between two strings. The LD is the number of deletions, insertions or substitutions required to transform a string into another string. The greater the distance, the more different the strings are. We chose to regard 1 and 2 as trustworthy values and disregarded the rest. We chose these two values since empirical observations suggest that contemporary Swedish and

19th century Swedish entities usually differ in one or two characters. In case of more than one match, we choose the most frequent alternative, as in the case of *Wenern* below. Table 1 illustrates various cases and the obtained results.

text word	#	gazeteer	LD	ann.	??
Dalarne	6	Dalarna	1	loc	yes
Asptomten	1	---	---	---	-
Härnevi*	1	Arnevi	2	prs	no
Sabbathsberg	1	Sabbatsberg	1	loc	yes
Wenern*	7	Werner,Waern Vänern	2 2	prs loc	no
Kaknäs	1	Valnäs,Ramnäs	2	loc	yes
Kallmar	1	Kalmar	1	loc	yes

Table 1. LD between potential NEs and the gazeteers; '*': both are locations;'??': correct annot.?

5 The Document Centered Approach

There is a known tradeoff between rule-based and statistical systems. Handcrafted grammar-based systems typically obtain better results, but at the cost of considerable manual effort by domain experts. Statistical NER systems typically require a large amount of manually annotated training data, but can be ported to other domains or genres more rapidly and require less manual work. Although the Swedish system is mainly rule-based, using a handcrafted grammar for each entity group, it can also be considered a hybrid system in the sense that it applies a document-centered approach (DCA) to entity annotation, which is a different paradigm compared to the local context approach, called *external evidence* by McDonald (1996). With DCA, information for the disambiguation of a name is derived from the entire document.

DCA as a term originates from the work by Mikheev (2000: 138), who claims that:

> important words are typically used in a document more than once and in different contexts. Some of these contexts create very ambiguous situations but some don't. Furthermore, ambiguous words and phrases are usually unambiguously introduced at least once in the text unless they are part of common knowledge presupposed to be known by the readers.

This implies a form of online learning from the document being processed where unambiguous usages are used for assigning annotations to ambiguous words, and information for disambiguation is derived from the entire document.

Similarly, label consistency, the preference of the same annotation for the same word sequence everywhere in a particular discourse, is a comparable approach for achieving qualitatively higher recall rates with minimal resource overhead (*cf.* Krishnan and Manning, 2006). Such an approach has been used, e.g., by Aramaki et al. (2006), for the identification of personal health information (age, id, date, phone, location and doctor´s and patient´s names).

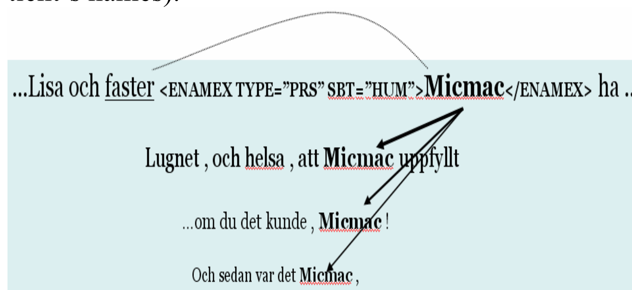

Figure 1. Example of label consistency

Figure 1 illustrates this approach with an example taken from *Almqvist's Collected Works, Vol. 30*. In this example, the first occurrence of the female person name *Micmac*, which is not in the gazetteer lists, is introduced by the author with the unambiguous designator *faster* 'aunt'. Many of the subsequent mentions of the same name are given without any reliable clue for appropriate labelling. However, as already discussed, there is strong evidence that subsequent mentions of the same name should be annotated with the same label, and since the same entity usually appears more than once in the same discourse, in our case a book, labelling consistency should guarantee better performance. There are exceptions for certain NE categories which may consist of words that are not proper nouns such as in the *Work&Art* category, and of course the temporal and measure groups which are blocked from this type of processing; *cf.* section 6.2.

6 Evaluation and Error Analysis

The system was evaluated twice, while no normalization or other preprocessing was applied to the original documents. Problems identified during the first evaluation round were taken under consideration and specific changes were suggested to the system by incorporating appropriate modifications.

During the first run, no adaptations or enhancements were made to the original NER system. After the first evaluation round, four major areas were identified in which the system either failed to produce an annotation or produced only partial or erroneous annotations. These failures were caused by:

- **Spelling variation**: particularly the use of <f/w/e/q> instead of <v/v/ä/k> as in modern Swedish. Most of the cases could be easily solved while other required different means such as calculating the LD between the name lists and possible name mentions in the texts (Section 4.3). One case that could be easily tackled was the addition of alternate spelling forms for a handful of keywords and designators, especially the preposition *av/af* common in temporal contexts, such as *i början af/av 1790-talet* 'in the beginning of the 1790s'; or words such *begge/bägge* 'both' and *qväll/kväll* 'evening';
- A number of **definite plural forms** of nouns, often designating a group of persons, with the suffix "erne" instead the "erna" as in modern Swedish, such as *Kineserne/Kineserna* 'the Chinese [people]' and *Svenskarne/Svenskarna* 'the Swedes';
- **Unknown names**: mentioned once with unreliable context;
- **Structure preservation**: the document structure of the texts in Litteraturbankens is designed to create a faithful rendering of the visual appearance of the original printed books. In extracting the texts from the XML format used in Litteraturbanken, we did not want to apply any kind of normalization or other processing. Such an approach would have altered the document structure. This implies that for a handful of the entities, for which the hyphenation in the original paper version has divided a name into two parts, as in (5), correct identification cannot be accomplished, while in some cases only a partial identification was possible, as in (6).

(5) `[…] Stock- holm`

(6) `<ENAMEX TYPE="PRS" SBT="HUM">Bertha von Lichten-</ENAMEX> ried`

6.1 Results

As a baseline for the evaluation we use the result of simple dictionary lookup in the single name gazetteer. This process is very accurate (w.r.t. precision). We could identify a number of cases with erroneous annotations, due to various circumstances: Names in the gazetteer lists may have multiple entity tags associated with them, and thus an entity may belong to more than one group that could not be disambiguated by the surrounding context, such as *Ekhammar* as a city and surname; many names are ambiguous with common nouns or verbs, such as *Stig* as a first name and as the verb 'step/walk'; the gazetteers contained a number of words that should not have been in the list in the first place, such as *Hvem* 'Who', *styrman* 'first mate' and *fänrik* 'lieutenant'. A probable cause of the latter problem is the fact that the name lists have been semi-automatically compiled from various sources including corpora and the Internet.

We performed two evaluations, based on two different random samples consisting of 500 segments (roughly 30,000 tokens) each. A segment consists of an integral number of sentences (up to 10–20). The overall results for all tests are shown in table 2. Results for individual entities using the whole system during both runs are found in table 3. The samples were evaluated according to precision, recall and f-score using the formulas:

Precision = (Total Correct + Partially Corrtect) / All Produced
Recall = (Total Correct + Partially Correct) / All Possible
*F-score =2*P*R/P+R.*

	1st run – no adaptations			**2nd run** – with adaptations		
	P	R	F	P	R	F
Baseline (gazeteer lookup)	88,8%	69,8%	78,1%	93,1%	86,2%	89,5%
Rule-Based System (no time&animacy)	91,8%	75,4%	82,7%	96,9%	86,9%	91,6%
Rule-Based System (all categ.)	93,8%	69,4%	79,7%	96%	87,9%	91,7%
Rule-Based System & DCA	95,4%	83,4%	89%%	96,1%	88,8%	92,3%
Rule-Based System&DCA+ED	96%	84%	89,6%	96,6%	89,4%	92,8%

Table 2. Overall performance of the NER

NE Categories	1st run				2nd run			
	# of NEs corr prod./ possible	P	R	F	# of NEs corr.prod./ possible	P	R	F
PERSON	424/441	92,3%	96,1%	94,1%	410/419	96%	97,8%	96,9%
LOCATION	83/123	100%	67,4%	80,5%	74/95	97,3%	77,8%	86,4%
ORGANIZATION	7/10	70%	70%	70%	4/4	40%	100%	57,1%
ARTIFACT	0/4	---	---		0/6	---	---	
WORK/ART	3/9	75%	33,3%	46,1%	4/10	100%	40%	57,1%
EVENT	0/2	---	---		0/2	---	---	
TEMPORAL	102/114	99%	89,4%	97%	106/118	100%	89,8%	94,6%
MEASURE	1/4	100%	25%	40%	4/16	66,6%	25%	36,3%
ANIMACY	207/277	98,1%	74,7%	84,8%	292/339	96,3%	86,1%	90,9%
TOTAL	***827/984***	***96%***	***84%***	***89,6%***	***894/999***	***96,6%***	***89,4%***	***92,8%***

Table 3. Performance of the NER on the individual named entities including animacy

Partially correct means that an annotation gets partial credit. For instance, if the system produces an annotation for the functional location *Nya Elementarskolan* as in (7) instead of the correct (8), then such annotations are given half a point, instead of a perfect score.

(7) `Nya <ENAMEX TYPE="LOC" SBT="FNC"> Elementarskolan</ENAMEX>`

(8) `<ENAMEX TYPE="LOC" SBT="FNC">Nya Elementarskolan</ENAMEX>`

If, on the other hand, the type is correct but the subtype is wrong, then the annotation is given a score of 0.75 points (e.g. a functional location instead of a geopolitical location).

6.2 Limitations of the Centering Approach

Labeling consistency and the DCA approach relies on the assumption that usage is consistent within the same document by the same author. However, we have observed that there are problems with entities composed of more than a single word, particularly within the group *Work&Art*, which can produce conflicting information, if we allow the individual words in such content (often nouns or adjectives) to be re-applied in the text.

For instance, the name of the novel *Syster och bror* occurred 32 times in one of the evaluation texts (Almqvist's Collected Works Volume 29). If we allow the individual words that constitute the title, *Syster*, *och* and *bror* to be re-applied in the text as individual words (2 common nouns and a conjunction), then we would have degraded the precision considerably since we would have allowed *Work&Art* annotations for irrelevant words. However, such cases can be resolved by simply letting the system ignore multiword *Work&Art* annotations during the DCA processing.

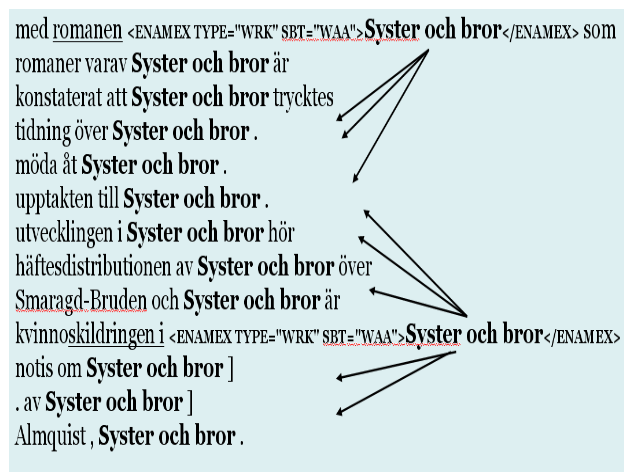

Figure 2. Occurrences of the multi-word entity *Syster och bror*; the rule-based system could reliably identify and annotate 2/32 occurrences.

Generally speaking, the experimental results have shown that any breaking of a multiword entity, except personal names, into its individual words often has a negative effect on performance. The best results are achieved when the DCA approach deals with single or bigram entities, particularly personal names.

7 Conclusions and Future Prospects

In this paper we have described the application of a generic Swedish named entity recognition system to a number of literary texts, novels from the 19th century, part of *Litteraturbanken*, the Swedish Literature Bank. We evaluated the results of the named entity recognition and identified a number of error sources which we tried to resolve and then introduce changes that would cover for such cases in the rule-based component of the system, in order to increase its performance (precision and recall) during a second evaluation round.

Entity annotations open up a whole new research spectrum for new kinds of qualitative and quantitative exploitations of literary and historical texts, allowing more semantically-oriented exploration of the textual content. In the near future, we will annotate and evaluate a larger sample and possibly integrate machine learning techniques in order to improve the results even more. We are also working to integrate the handling of named entity annotations into Litteraturbanken's search and browsing interfaces and hope to be able to conduct our first demonstrations and tests with users later this year.

References

Eiji Aramaki, Takeshi Imai, Kengo Miyo and Kazuhiko Ohe. 2006. Automatic Deidentification by using Sentence Features and Label Consistency. *Challenges in NLP for Clinical Data Workshop*. Washington DC.

Kalina Bontcheva, Diana Maynard, Hamish Cunningham and Horacio Saggion. 2002. Using Human Language Technology for Automatic Annotation and Indexing of Digital Library Content. *Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*.

Andrew Borthwick. 1999. *A Maximum Entropy Approach to Named Entity Recognition*. PhD Thesis. New York University.

Gregory Crane and Alison Jones. 2006. The Challenge of Virginia Banks: an Evaluation of Named Entity Analysis in a 19th-century Newspaper Collection. *ACM/IEEE Joint Conference on Digital Libraries*, JCDL. Chapel Hill, NC, USA. 31–40.

Karina van Dalen-Oskam and Joris van Zundert. 2004. Modelling Features of Characters: Some Digital Ways to Look at Names in Literary Texts. *Literary and Linguistic Computing* 19(3): 289–301.

Julia Flanders, Syd Bauman, Paul Caton and Mavis Cournane. 1998. Names Proper and Improper: Applying the TEI to the Classification of Proper Nouns. *Computers and the Humanities* 31(4): 285–300.

Michael Fleischman and Eduard Hovy. 2002. Fine Grained Classification of Named Entities. *Proceedings of the 19th International Conference on Computational Linguistics*. Taipei, Taiwan. 1–7.

Janne Bondi Johannessen, Kristin Hagen, Åsne Haaland, Andra Björk Jónsdottir, Anders Nøklestad, Dimitrios Kokkinakis, Paul Meurer, Eckhard Bick and Dorte Haltrup. 2005. Named Entity Recognition for the Mainland Scandinavian Languages. *Literary and Linguistic Computing*. 20(1): 91–102.

Dimitrios Kokkinakis. 2004. Reducing the Effect of Name Explosion. *Proceedings of the LREC-Workshop: Beyond Named Entity Recognition - Semantic Labeling for NLP*. Lisbon, Portugal.

Vijay Krishnan and Christopher D. Manning. 2006. An Efficient Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition. *Proceedings of COLING/ ACL 2006*. Sydney, Australia. 1121–1128.

David D. McDonald. 1996. Internal and External Evidence in the Identification and Semantic Categorisation of Proper Nouns. *Corpus-Processing for Lexical Acquisition*. James Pustejovsky and Bran Boguraev (eds). MIT Press. 21–39.

Andrei Mikheev. 2000. Document Centered Approach to Text Normalization. *Proceedings of the 23rd ACM SIGIR Conference on Research and Development in Information Retrieval*. Athens, Greece. 136–143.

Satoshi Sekine. 2004. Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. *Proceedings of the Language Resources and Evaluation Conference (LREC)*. Lisbon, Portugal.

Constantin Orasan and Roger Evans. 2001. Learning to Identify Animate References. *Proceedings of the Workshop on Computational Natural Language Learning (CoNLL-2001)*. ACL-2001. Toulouse, France.

Robert Shoemaker. 2005. Digital London. Creating a Searchable Web of Interlinked Sources on Eighteenth Century London. *Program: Electronic Library & Information Systems* 39(4): 297–311.