SWENER-1800: A Corpus for Named Entity Recognition in 19th Century Swedish

Eva Pettersson[1,*,†], Lars Borin[2,†] and Erik Lenas[3,†]

[1]*Uppsala University*
[2]*University of Gothenburg*
[3]*Swedish National Archives*

Abstract

Named entity recognition (NER) is the process of automatically identifying persons, places, organisations and other name-like entities in text, in order to perform natural language processing tasks such as automatic extraction of metadata from text, anonymisation/pseudonymisation of sensitive personal data, or as a preprocessing step for linking different terms describing the same entity to a single reference. While NER is a mature language technology, it is generally lacking for historical language varieties. We describe our work on compiling SWENER-1800, a large (half a million words) reference corpus of historical Swedish texts, covering the time period from the first half of the 18th century until about 1900, and manually annotating it with named entity types identified as significant for this time period, as well as with sentence boundaries, notoriously difficult to recognise automatically in historical text. This corpus can then be used to train and evaluate NER systems and sentence segmenters for historical Swedish text. An additional concrete contribution from this work is a manual for annotation of named entities in historical Swedish.

Keywords

NLP for historical text, named entity recognition, NER, corpus linguistics, Swedish

1. Introduction

Named entity recognition (NER) is the process of automatically identifying persons, places, organisations and other name-like entities in text (Nadeau and Sekine 2007). NER is included in many natural language processing applications, to perform tasks such as automatic extraction of metadata from running text, information extraction and retrieval (see for example Brandsen et al. 2022), anonymisation/pseudonymisation of sensitive personal data (e.g., Bridal 2021 or Papadopoulou et al. 2022), or as a preprocessing step for linking different terms describing the same entity to a single reference, e.g., *New York*, *NY*, and *Big Apple*.

Digital Humanities in the Nordic and Baltic Countries, May 27–31, 2024, Reykjavík, Iceland
*Corresponding author.
†These authors contributed equally.
eva.pettersson@lingfil.uu.se (E. Pettersson); lars.borin@svenska.gu.se (L. Borin); erik.lenas@riksarkivet.se (E. Lenas)
https://www2.lingfil.uu.se/person/pettersson/ (E. Pettersson); https://spraakbanken.gu.se/om/personal/lars (L. Borin)
0000-0002-1447-4501 (E. Pettersson); 0000-0001-5434-9329 (L. Borin)

Digital Humanities in the Nordic and Baltic Countries Publications – ISSN: 2704-1441

Annotated corpora and guidelines for NER have been developed for contemporary texts in Swedish and in other languages, but these resources are less useful for historical texts, due to different spellings and name conventions in historical settings.

In the project described here, our aim has been to manually annotate a corpus of historical Swedish texts, covering the time period from the first half of the 18th century until about 1900, with named entity types identified as significant for this time period. This corpus will serve as a gold standard that can be used for training and evaluation of automatic named entity recognition systems adapted to historical Swedish text. It could also be a useful dataset for historical research per se, and for evaluation of historical language models.

The rest of the paper is structured as follows: In Section 2, we give an overview of related work in the field of NER for both contemporary and historical Swedish text. Section 3 presents the workflow that was implemented for creating the corpus, including focus group interviews, development of annotation guidelines, text collection and annotation rounds. In Sections 4 and 5, we describe the annotation scheme and the annotation guidelines, respectively. In Section 6, we present and discuss the results achieved. Finally, our main contributions are summarised in Section 7.

2. Related work

As indicated above, there are both NER systems and corpora annotated with named-entity information for contemporary Swedish. The first high-performing NER system for Swedish was a rule-based system developed by Kokkinakis (2004) in the framework of the Nordic *Nomen Nescio* collaboration, where several NER systems (both rule-based and machine-learning based) and small evaluation corpora were developed for the Continental Nordic languages (i.e., Danish, Norwegian, and Swedish; Johannessen et al. 2005).

More recently, the Swedish National Library's digital humanities lab KBLab have released a NER system achieving good performance based on a large language model (KB-BERT; Malmsten, Börjeson, and Haffenden 2020) trained on the library's extensive text holdings (Kurtz and Öhman 2022).

All kinds of automatic linguistic annotation by computers are crucially dependent on *evaluation*, i.e. a "sanity check" of the extent to which the automatic annotations correspond to ground truth. For good methodological reasons, evaluation is most often carried out using a – typically manually prepared – *gold standard* dataset. NER forms no exception in this regard, and the one-million word Stockholm-Umeå Corpus (SUC; Gustafsson-Capková and Hartmann 2006) has long served as the main such dataset for general Swedish. However, due to both the restrictive license of SUC and its age,[1] an initiative was taken some years back by the Swedish CLARIN consortium to compile a new NER gold-standard corpus – Swe-NERC[2] – consisting of more modern texts[3] without licensing restrictions. This corpus and its annotation are described by Ahrenberg, Frid, and Olsson (2020).

[1] SUC comes with an individual license restricting its use to research purposes only, and it is compiled from formally published texts from the early 1990s. Hence, no internet texts appear in SUC.

[2] https://repo.spraakbanken.gu.se/xmlui/handle/10794/121

[3] The corpus contains approximately 140,000 words, and most of its 70-plus text samples are from around 2010.

The Swedish NER system of Kokkinakis (2004) has been refined over the years and deployed mainly as a component of systems in medical informatics (e.g. Kokkinakis and Thurin 2007). It has been reimplemented using a state-of-the-art finite-state framework (Kokkinakis et al. 2014). Relevant to the present initiative, it has also seen some use in research in digital humanities, applied to older literary texts (Borin, Kokkinakis, and Olsson 2007; Borin and Kokkinakis 2010; Oelke, Kokkinakis, and Malm 2012; Borin, Dannélls, and Olsson 2014). Karsvall and Borin (2018) apply it to the Swedish National Archives' medieval charters, although the NER is actually performed on modern summaries of the medieval texts (which are often originally written in Latin rather than in Old Swedish, to boot).

3. Method

The workflow for creating the corpus is illustrated in Figure 1.

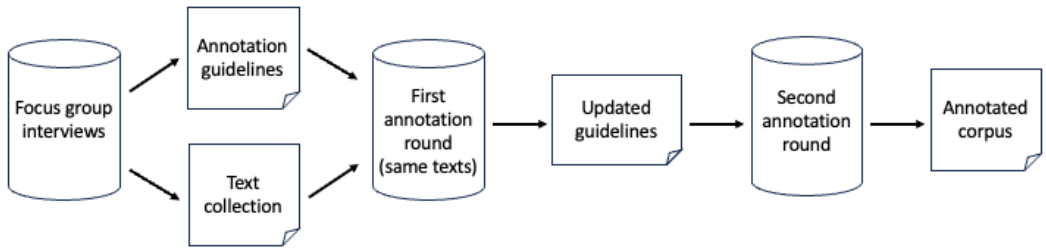

Figure 1: Workflow for creating the NER corpus

3.1. Focus group interviews

The first step in the creation of the NER corpus was to put together a focus group of researchers in history, historical linguistics, (Swedish) language history, and history of science and ideas, as potential users of the resulting corpus. In our discussions with the focus group, we brought up issues concerning text collection, how to make the corpus balanced in different ways, and what entities to aim for in the annotation process. These discussions led to many useful insights. For example, historians – most famously perhaps the British historian Eric Hobsbawm (see Evans 2019) – talk about the "long nineteenth century" (1789–1914)[4] as a distinct period in Europe's past, and our initial thought was to restrict our corpus to texts from this period. However, the language historians and historical linguists in our focus group pointed out that setting the starting year for text collection to 1730 would better reflect the Late Modern Swedish period[5] in the Swedish language development. We also got suggestions for entities of special interest for

[4]The long nineteenth century is bracketed by the French Revolution and the outbreak of the First World War.

[5]The Late Modern Swedish period is conventionally defined as extending between 1732 and the spelling reform of 1906. Following the suggestions of our focus group, we may wish to take into account that the old spelling remained in written sources for about two decades after 1906 (Callin 2014). Hence, the ideal periodisation for our corpus would be 1730–1925, but we have not (yet) included any texts from the 20th century.

historical research, such as occupational titles and a distinction under the *Organisation* top-level category between *Company* and *Institution*.

3.2. Annotation guidelines

The second step was to develop a first set of annotation guidelines. We based these guidelines mainly on the guidelines defined for contemporary Swedish by Ahrenberg, Frid, and Olsson (2020), with modifications and additions to better suit historical text, in accordance with the insights gained from the focus group.

3.3. Text collection

In parallel to developing annotation guidelines, we started collecting the texts to be included in the corpus. In this process, we have aimed for a corpus covering as much as possible of the targeted time period, while at the same time including texts from a variety of genres, to make the corpus useful for as many research interests as possible. The texts should also be judged to include a sufficient number of named entities, meaning that for example legal texts are not very suitable for our purpose. Furthermore, the texts should be digitally accessible in a reasonable format.

Following these criteria, the texts included in the corpus are:

- police reports from the Swedish National Archives[6]
- petitions from the project *Speaking to one's superiors*[7]
- newspaper articles from the *Kubhist 2 Corpus*[8]
- court records from the *Swedish Diachronic Corpus* (Pettersson and Borin 2022)[9]
- literary texts from the *Swedish Diachronic Corpus* (Pettersson and Borin 2022)[10]

3.4. Annotation rounds and updated guidelines

For the annotation phase, we decided to use Label Studio, an open-source data labelling platform that we adapted to our annotation task.[11] Label studio is a highly customisable and flexible annotation environment. It offers a user friendly interface and lowers the barrier to entry for the annotators. The platform also has robust data management and a strong community support. Figure 2 shows a screenshot from working with our annotation task in Label Studio.

For the actual annotation, we recruited three people: two students and one of the authors of this paper. In the first annotation round, all annotators annotated the same texts, and afterwards we extracted the cases where the annotators disagreed, and discussed these to reach a consensus. Following these discussions, we also updated the annotation guidelines. Towards the end of the project we did another round of double annotations to check how inter-annotator agreement

[6] https://sok.riksarkivet.se/nad?postid=Arkis+d3267232-77a0-11d5-a6f2-0002440207bb&s=Balder
[7] https://gaw.hist.uu.se/suppliker
[8] https://spraakbanken.gu.se/en/resources/kubhist2
[9] https://www2.lingfil.uu.se/person/pettersson/svediakorp
[10] https://www2.lingfil.uu.se/person/pettersson/svediakorp
[11] https://labelstud.io/

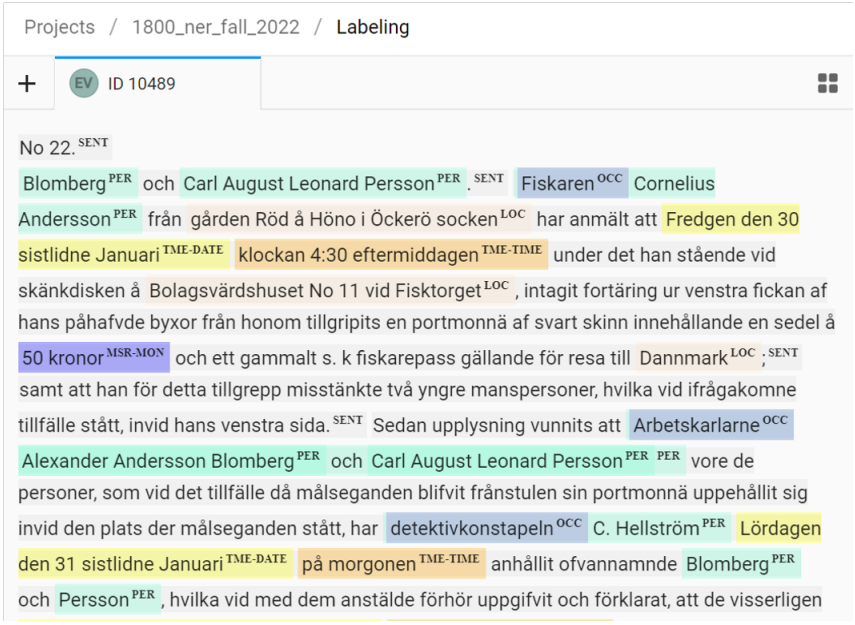

Figure 2: Annotation of named entities in Label Studio. PER = person name. OCC = occupation. TME-DATE = date. TME-TIME = time. MSR-MON = monetary

had progressed from the initial round of double annotations. The main part of the corpus was annotated by one annotator only. All in all, we ended up with a corpus of 573,605 tokens.

4. Annotation scheme

The named entity types annotated in our corpus are listed in Table 1.

There are some features in our annotation scheme that are quite unique to our specific task. One such thing is the outcome of the discussions with the focus group, where some entities are regarded as important for annotation in historical text, even though they would usually not be found in traditional NER systems, e.g. annotation of occupational titles.

We also see a relatively large number of subcategories for some entities. Some of these subcategories were suggested by the focus group (e.g. the previously mentioned distinction between *Organisation: Company* and *Organisation: Institution*), whereas others emerged during the annotation process. One example of the latter is the main category *Measurement*, for which we ended up distinguishing between monetary, weight, length, distance, area and volume.

It is also worth noting that some entities occur more seldom than others. Consequently, in the process of training a NER system based on this corpus resource, it will be possible (and probably also desirable) to merge (some of the) subcategories.

Another distinguishing feature is that we allow for nested entities, which is quite uncommon in existing NER systems. An example would be the phrase *bagaren Johan Nilsson* 'the baker Johan Nilsson', where the whole phrase is annotated as a personal name (since 'the baker' is

Table 1
Named entity categories covered in our project, listed together with the abbreviations used during annotation, and examples for each category. The MISC category may be used for adding a unit that the annotator would like to categorize as a named entity, but of a type that is not covered by the current scheme.

Category	Annotation Abbr.	Example
Person	PER	*Alfred Andersson*
Location	LOC	*Stockholm*
Organisation	(ORG)	
Company	ORG-COMP	*Handelsfirman J. O. Grén & Co.*
Institution	ORG-INST	*Rådhusrätten*
Other	ORG-OTH	
Measurement	(MSR)	
Monetary	MSR-MON	*20 kr*
Weight	MSR-WEI	*5 ℔*
Length	MSR-LEN	*100 meter*
Distance	MSR-DIST	*1 1/2 engelska mil*
Area	MSR-AREA	*7 hektar*
Volume	MSR-VOL	*5 tunnor*
Other	MSR-OTH	
Temporal	(TME)	
Date	TME-DATE	*den 2e Januari 1880*
Time	TME-TIME	*kl. 4 e. m.*
Interval	TME-INTRV	*1817-19*
Other	TME-OTH	
Event	EVN	*påsk*
Work of Art	WRK	*ångfartyget "Konung Oskar"*
Symptom	SYMP	*kolera*
Treatment	TREAT	*läkemedel*
Occupation	OCC	*trädgårdsmästare*
Miscellaneous	MISC	

used to point out which Johan Nilsson we are referring to), while at the same time the subpart *bagaren* 'the baker' is also annotated as an occupational title, as is illustrated in Figure 3.

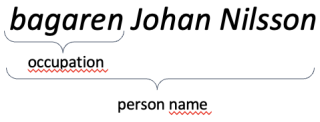

Figure 3: Nested named entities

In addition to named entities, the annotators were also asked to mark sentence boundaries in the text. Automatic sentence segmentation is crucial for many studies using text corpora, in

language technology, linguistics, and other disciplines, and almost taken for granted. However, sentence boundaries are often marked by different, not obviously form-based principles in historical text compared to modern sentence-marking conventions, and are hard for NLP tools to detect automatically. There were also other abbreviations containing full stop than those used today, meaning that sentence segmentation systems trained on present-day Swedish will not recognise these from the training data. We therefore hope that, as a side-effect of our project, we will be able to also train a tool for better sentence segmentation of historical (Swedish) text.

5. Annotation guidelines

The full annotation guidelines are specified in Pettersson et al. (2024). Some of the more general instructions are also described here.

In the annotation process, a sequence of words should be marked as a named entity if it is part of a name-like phrase that refers to any of the categories listed in Table 1. We adopt the approach described by Ahrenberg, Frid, and Olsson (2020), where the decision on which named entity category a certain sequence of words belongs to is primarily based on semantics, i.e., what kind of entity it is referring to in the specific context where it occurs.

It is also worth mentioning that the notion of 'name-like phrase' can be different for different entity types. However, it should in general be a syntactic phrase of some sort, that is an established standard reference for an entity, or include such a standard reference as its main part. A name-like phrase may thus include words that are not proper nouns but are rather referring to attributes of the referent.

Pronouns, such as *han* 'he' and *hon* 'she', deictic adverbs such as *då* 'then' and *här* 'here', and verbs in general should as a rule not be marked as named entities.

Genitive forms are marked in the same way as nominative forms. Thus, in a phrase such as *Olssons handelsbod* 'Olsson's general store', *Olssons* is marked as a person.

As a rule, each unique sequence in a text should not be assigned more than one named entity type. We do, however, allow for **nested annotations**, where a shorter sequence of words may be annotated as one named entity type, while at the same time being part of a longer sequence of words with another named entity label. For example, the whole sequence *trädgårdsmästaren Alfred Andersson* 'the gardener Alfred Andersson' should be marked as a person, while the subsequence *trädgårdsmästaren* 'the gardener' should additionally be marked as an occupation.

In the corpus at hand, we have noticed that tokenisation (word segmentation) is sometimes unorthodox, due to faulty automatic (or manual) segmentation. This may affect the possibility to annotate named entities, as in the following examples:

- *min mågPetter Wortiain* 'my brother-in-lawPetter Wortiain'
- *besök å Augusta Sandslånekontor* 'visit at Augusta Sand'sloanoffice'

In the above examples, we would have wanted to annotate *Petter Wortiain* and *Augusta Sands* as persons, but the two-word sequences *mågPetter* and *Sandslånekontor* have been tokenised as single "words", and may thus not be split by the annotator. In such cases, the faulty tokenisation segments (*mågPetter* and *Sandslånekontor*) are annotated with a tokenisation error label, so that we can easily find and correct these instances at a later stage. In parallel, the named entities are

annotated as accurately as possible, ignoring the tokenisation errors, meaning that *mågPetter Wortiain* and *Augusta Sandslånekontor* are annotated as person names.

Phrases with **misspelled words** should be annotated, for example *nästa vekca* (misspelled variant of *nästa vecka* 'next week'). In historical text, it is also not always clear whether a non-standard word form is a misspelling or just reflects spelling variation, quite common in historical texts.

It is also worth mentioning that in historical Swedish text, capitalisation is not an infallible indication of proper nounhood. During some time periods and in some genres, capitalisation of nouns has been used as a means of emphasis (in addition to indicating proper nouns).

6. Results and discussion

6.1. Corpus composition and annotation statistics

The final corpus comprises a total of 573,605 tokens and 27,640 sentences. The distribution of texts across the different genres is shown in Table 2. Short fiction and newspapers make up the lion's share of the corpus, ensuring varying styles and content. Most of the texts have their origin in the 19th century. This reflects, not so much our intention, as it does the scarcity of digitised Swedish text from the 18th century on the one hand, and copyright restrictions concerning even early 20th century texts on the other.

Table 2
Distribution of texts in the corpus.

	Number of tokens	Year of origin
Short fiction	290,447	1849–1899
Newspapers	172,794	1819–1895
Police records	64,624	1860–1885
Court records	27,763	1809–1818
Petitions	17,977	1709–1782
Total	573,605	1709–1899

The number of entities annotated within the main categories is shown in Table 3. An even distribution of different entities across the corpus was not something we aimed for since modern named entity recognition systems based on transfer learning require fewer examples to learn from (Devlin et al. 2018). However, a large enough number of instances of an entity in the corpus does not in itself guarantee that a named entity recognition system can learn to recognise the entity. It also has to be consistently annotated, as discussed in Section 6.2.

Tables 4, 5, and 6 show breakdowns of the measurement, organisation and time expression supercategories into subcategories. Even though some of these categories probably contain too few examples for a traditional NER system to learn from, a system like the one suggested by Ashok and Lipton (2023), that utilises a combination of heuristics and a large language model few-shot setting, might accomplish it. Therefore, even the subcategories represented by only a few examples throughout the corpus were kept separated.

Table 3
Number of instances annotated for each entity type.

	PER	LOC	OCC	TME	MSR	ORG	WRK	EVN	SYMP	TREAT
Instances	12,412	6,315	5,001	4,732	2,432	2,242	945	360	207	15

Table 4
Number of instances of the measurements (MSR) subcategories.

(MSR-)	MON	VOL	WEI	LEN	DIST	AREA	OTH
Instances	1,890	220	102	75	42	37	66

Table 5
Number of instances of the organisation (ORG) subcategories.

(ORG-)	INST	COMP	OTH
Instances	1,751	321	170

Table 6
Number of instances of the time-expressions (TME) subcategories.

(TME-)	DATE	TIME	INTRV	OTH
Instances	3,520	1,173	37	2

6.2. Inter-annotator agreement

We used Krippendorff's α (alpha) as a measure for inter-annotator agreement during the annotation process. This measure is particularly appropriate in scenarios with nominal data, multiple raters and uneven distribution of categories (Zapf, Castell, Morawietz, et al. 2016). The values of Krippendorff's α range from 0 (no agreement) to 1 (perfect agreement) where a value between 0.41 and 0.61 is considered moderate agreement, between 0.61 and 0.81 substantial agreement, and between 0.82 and 1.0 almost perfect agreement (Landis and Koch 1977). Table 7 compares the Krippendorff's α score for the main categories from the first control round to the second control round at the end of the annotation process. The numbers merit a few comments.

First of all, we did not start out from scratch, but for many of the entities we used the guidelines provided by Ahrenberg, Frid, and Olsson (2020) as a starting point. We also used the main recommendation from that work, that in the case of ambiguity, the annotators should look at what the span in question refers to, as determined by its context, and decide on an entity type based on that. This solid point of departure that we had in place already before the annotation process began most likely explains why the measures from the first and second

control rounds do not differ that much. The most marked improvement is shown in the organisation category, and this was a clear focus for discussions after the first control round since annotating organisations can be difficult in historical text, as was shown by the questions and comments of the annotators. Persons and works of art also showed clear improvements from the first to the second round. Locations, occupations and time-expressions, on the other hand, showed slight disimprovement. The event category, a very difficult one judging by the scant data collected from the control rounds, had too few instances in the second round to say anything conclusive about improvement or disimprovement.

Table 7
Krippendorff's α score for the different entities, where control round 1 at the beginning of the annotation process is compared to control round 2 at the end of the annotation process.

	PER	LOC	OCC	TME	ORG	MSR	WRK	EVN
Round 1 – No. of inst.	755	457	424	278	147	89	75	13
Krippendorff's α	0.84	0.83	0.77	0.88	0.59	0.86	0.63	0.37
Round 2 – No. of inst.	165	208	131	151	69	68	40	3
Krippendorff's α	0.89	0.79	0.74	0.86	0.69	0.86	0.69	0.15

Tables 8, 9, and 10 show a breakdown into subcategories of the Krippendorff's α scores for the supercategories measurement, organisation and time expression, respectively. The most marked improvement is seen in the case of institution, a subcategory of organisation. Clearly the discussions and updates to the annotations guidelines between the rounds had an effect there. The measurement subcategories, on the other hand, show a disimprovement. However, for all but the monetary subcategory the measured instances are too few to draw any conclusions. Time expressions show a slight improvement in dates but a slight disimprovement regarding time expressions proper, but these numbers are within the margin of error.

Table 8
Krippendorff's α score for measurement (MSR) subcategories.

(MSR-)	MON	AREA	LEN	WEI	VOL	OTH
Round 1 – No. of inst.	63	6	4	3	0	13
Krippendorff's α	0.87	0.88	0.68	0.79	-	0.22
Round 2 – No. of inst.	37	3	6	1	9	9
Krippendorff's α	0.81	0.16	0.94	0.74	0.43	0.29

In conclusion, from what we have gathered from the annotators' questions and comments, and from our work with the annotation guidelines, annotating historical text comes with extra challenges, difficulties and ambiguities. But putting work into the guidelines and allotting time for discussing difficulties with the annotators ultimately make for a more coherent dataset.

Table 9
Krippendorff's α score for organisation (ORG) subcategories.

(ORG-)	INST	COMP	OTH
Round 1 – No. of inst.	116	17	14
Krippendorff's α	0.55	0.43	0.0
Round 2 – No. of inst.	45	10	14
Krippendorff's α	0.72	0.32	0.0

Table 10
Krippendorff's α score for time-expression (TME) subcategories.

(TME-)	DATE	TIME	INTRV
Round 1 – No. of inst.	212	64	2
Krippendorff's α	0.83	0.82	0.45
Round 2 – No. of inst.	123	26	2
Krippendorff's α	0.85	0.80	0.22

7. Conclusions

Summing up, the work presented above has resulted in the following contributions:

First, there is a manual for annotation of named entities in historical Swedish (Pettersson et al. 2024), based on, and largely compatible with, the corresponding recent manual for present-day Swedish (Ahrenberg, Frid, and Olsson 2020).

Second, we have produced SWENER-1800, a large dataset of historical Swedish text annotated with named entities, and released under an open license.

Third, as added value SWENER-1800 also contains manual sentence segmentation, which could form a basis for a system for sentence segmentation of historical (Swedish) text.

A separate publication is under preparation where we will describe our currently ongoing work on developing a NER system trained on historical Swedish text, and on systematic comparison based on evaluations of systems trained on historical and modern text, respectively.

Acknowledgments

The work presented here was conducted as part of the activities of the Swedish national research infrastructure Språkbanken and Swe-Clarin, funded jointly by the Swedish Research Council (2018–2024, contract 2017-00626) and the 10 participating partner institutions.

We are also very grateful to Ahrenberg, Frid, and Olsson (2020) for giving us the permission to reuse substantial parts of their text in our annotation guidelines.

References

Ahrenberg, Lars, Johan Frid, and Leif-Jöran Olsson. 2020. *A new gold standard for Swedish named entity recognition: Version 1 contents.* SWE-CLARIN Report Series SCR-01-2020. https://sweclarin.se/swe/v%C3%A5ra-resurser/rapportserie.

Ashok, Dhananjay, and Zachary C. Lipton. 2023. "PromptNER: Prompting For Named Entity Recognition." *arXiv preprint arXiv:2305.15444.*

Borin, Lars, Dana Dannélls, and Leif-Jöran Olsson. 2014. "Geographic Visualization of Place Names in Swedish Literary Texts." *Literary and Linguistic Computing* 29 (3): 400–404.

Borin, Lars, and Dimitrios Kokkinakis. 2010. "Literary onomastics and language technology." In *Literary education and digital learning,* edited by Willie van Peer, Sonja Zyngier, and Vander Viana, 53–78. Information Science Reference. https://doi.org/10.4018/978-1-60566-932-8.ch003.

Borin, Lars, Dimitrios Kokkinakis, and Leif-Jöran Olsson. 2007. "Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature." In *Proceedings of LaTeCH 2007).* 1–8. Prague: ACL, June. https://aclanthology.org/W07-0901.

Brandsen, Alex, Suzan Verberne, Karsten Lambers, and Milco Wansleeben. 2022. "Can BERT Dig It? Named Entity Recognition for Information Retrieval in the Archaeology Domain." *J. Comput. Cult. Herit.* (New York, NY, USA) 15, no. 3 (September). ISSN: 1556-4673. https://doi.org/10.1145/3497842. https://doi.org/10.1145/3497842.

Bridal, Olle. 2021. *Named-entity recognition with BERT for anonymization of medical records.* Linköping University, Department of Computer and Information Science.

Callin, Markella. 2014. *Gammalstafning eller nystavning? En komparativ analys av tidningsspråket efter 1906 års stavningsreform.* Bachelor's thesis, Uppsala University, Dept. of Nordic Languages. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-230365.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." *arXiv preprint arXiv:1810.04805.*

Evans, Richard J. 2019. *Eric Hobsbawm: A Life in History.* London: Little, Brown.

Gustafsson-Capková and Britt Hartmann. 2006. *Manual of the Stockholm Umeå Corpus version 2.0.* Https://spraakbanken.gu.se/parole/Docs/SUC2.0-manual.pdf.

Johannessen, Janne Bondi, Kristin Hagen, Åsne Haaland, Andra Björk Jónsdottir, Anders Nøklestad, Dimitris 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. https://doi.org/10.1093/llc/fqh045. eprint: https://academic.oup.com/dsh/article-pdf/20/1/91/2815218/fqh045.pdf. https://doi.org/10.1093/llc/fqh045.

Karsvall, Olof, and Lars Borin. 2018. "SDHK meets NER: Linking Place Names with Medieval Charters and Historical Maps." In *Proceedings of DHN 2018,* 38–50. Aachen: CEUR-WS.org.

Kokkinakis, Dimitrios. 2004. "Reducing the Effect of Name Explosion." In *Proceedings of the LREC Workshop: Beyond Named Entity Recognition, Semantic labelling for NLP tasks. Fourth Language Resources and Evaluation Conference (LREC)*. Lisbon: ELRA.

Kokkinakis, Dimitrios, Jyrki Niemi, Sam Hardwick, Krister Lindén, and Lars Borin. 2014. "HFST-SweNER — A New NER Resource for Swedish." In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14),* 2537–2543. Reykjavik, Iceland: European Language Resources Association (ELRA), May. http://www.lrec-conf.org/proceedings/lrec2014/pdf/391_Paper.pdf.

Kokkinakis, Dimitrios, and Anders Thurin. 2007. "Identification of Entity References in Hospital Discharge Letters." In *Proceedings of the 16th Nordic Conference of Computational Linguistics (NODALIDA),* 329–332. Tartu, Estonia.

Kurtz, Robin, and Joey Öhman. 2022. *The KBLab Blog: SUCX 3.0 – NER.* https://kb-labb.github.io/posts/2022-02-07-sucx3_ner/.

Landis, J Richard, and Gary G Koch. 1977. "The measurement of observer agreement for categorical data." *Biometrics* 33 (1): 159–174.

Malmsten, Martin, Love Börjeson, and Chris Haffenden. 2020. *Playing with Words at the National Library of Sweden – Making a Swedish BERT.* arXiv.org. eprint: 2007.01658 (cs.CL). https://arxiv.org/abs/2007.01658.

Nadeau, David, and Satoshi Sekine. 2007. "A Survey of Named Entity Recognition and Classification." *Lingvisticae Investigationes* 30 (August). https://doi.org/10.1075/li.30.1.03nad.

Oelke, Daniela, Dimitrios Kokkinakis, and Mats Malm. 2012. "Advanced Visual Analytics Methods for Literature Analysis." In *Proceedings of LaTeCH 2012,* 35–44. Avignon: ACL. https://aclanthology.org/W12-1007.

Papadopoulou, Anthi, Yunhao Yu, Pierre Lison, and Lilja Øvrelid. 2022. "Neural Text Sanitization with Explicit Measures of Privacy Risk." In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers),* edited by Yulan He, Heng Ji, Sujian Li, Yang Liu, and Chua-Hui Chang, 217–229. Online only: Association for Computational Linguistics, November. https://aclanthology.org/2022.aacl-main.18.

Pettersson, Eva, and Lars Borin. 2022. "Swedish Diachronic Corpus." In *The Infrastructure for Language Resources,* edited by Darja Fišer and Andreas Witt. Berlin: De Gruyter Mouton. https://doi.org/10.1515/9783110767377-022.

Pettersson, Eva, Erik Lenas, Lars Borin, and Catharina Dahlgren. 2024. *Named entity recognition in 19th century Swedish texts: Annotation guidelines.* SWE-CLARIN Report Series SCR-01-2024. https://sweclarin.se/sites/sweclarin.se/files/SCR-01-2024.pdf.

Zapf, A., S. Castell, L. Morawietz, et al. 2016. "Measuring inter-rater reliability for nominal data – which coefficients and confidence intervals are appropriate?" *BMC Medical Research Methodology* 16 (1): 93. https://doi.org/10.1186/s12874-016-0200-9.

A. Online Resources

- *SWENER-1800: Annotated corpus*: TBA
- *Named entity recognition in 19th century Swedish texts: Annotation guidelines*: https://sweclarin.se/sites/sweclarin.se/files/SCR-01-2024.pdf
- *Swedish Diachronic Corpus*: https://www2.lingfil.uu.se/person/pettersson/svediakorp
- *Swe-NERC*: https://spraakbanken.gu.se/en/resources/swe-nerc