SWE-CLARIN

RAPPORTER FRÅN SWE-CLARIN
SWE-CLARIN REPORT SERIES
sweclarin.se/scrs
[SCR-01-2024]

Named entity recognition in 19th century Swedish texts

Annotation guidelines

Eva Pettersson[1], Erik Lenas[2], Lars Borin[3], and Catharina Dahlgren[4]
[1]Uppsala University • eva.pettersson@lingfil.uu.se
[2]Riksarkivet • erik.lenas@riksarkivet.se
[3]Språkbanken Text, University of Gothenburg • lars.borin@svenska.gu.se
[4]Riksarkivet • catharina.dahlgren@riksarkivet.se

Contents

1 Introduction

1.1 Aims and motivation

Named entity recognition is the task of searching texts for names of persons, places, organisations and other name-like entities of interest. For this task, annotated corpora and guidelines 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 this project, we aim at manually annotating a corpus of historical Swedish texts, initially covering the time period 1730–1900, with named entity types identified as significant for this time period. This corpus will serve as a gold standard that can be used for search purposes, or for training and evaluation of automatic named entity recognition systems adapted to historical Swedish text.

1.2 Data

Our aim is to create a corpus covering as much as possible of the targeted time period (1730–1900), while at the same time including texts from a variety of genres, to make the corpus useful for as many research interests as possible. For the first version of the corpus, we have included police reports from the Swedish National Archives,[1] petitions from the project *Speaking to one's superiors*,[2] newspaper articles from the Kubhist2 Corpus[3] and literary texts from the Swedish Diachronic Corpus.[4]

[1] https://riksarkivet.se/startpage
[2] https://gaw.hist.uu.se/suppliker
[3] https://spraakbanken.gu.se/resurser
[4] Pettersson and Borin (2022): https://cl.lingfil.uu.se/svediakorp/

1.3 Entry types

The types of name-like phrases to be annotated in our corpus are listed in Table 1. In addition to named entities, annotators are also asked to mark the sentence boundaries in the text. This is because sentence segmentation is crucial for many corpus-based investigations. However, the texts that are used as input in the named entity annotation process are initially not segmented into sentences, due to state-of-the-art sentence segmentation tools not being adapted to handle historical text.

Category	**Annotation Abbr.**	**Example**
Person	PER	*Alfred Andersson*
Location	LOC	*Stockholm*
Organisation		
Company	ORG-COMP	*Handelsfirman J. O. Grén & Co.*
Institution	ORG-INST	*Rådhusrätten*
Other	ORG-OTH	
Measurement		
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		
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	

Table 1: Named entity categories covered in our project, listed together with the abbreviations used during annotation, and examples for each category.

2 Related work

Our annotation guidelines are based on the guidelines defined for contemporary Swedish by Ahrenberg, Frid and Olsson (2020), with some modifications and additions to better suit historical text. We are very grateful to Ahrenberg, Frid and Olsson (2020) for giving us the permission to reuse substantial parts of their text in our guidelines.

3 Annotation guidelines

3.1 General guidelines

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.

We also list the same general notes as Ahrenberg, Frid and Olsson (2020):

- 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 includes 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', *hon* 'she', and deictic adverbs such as *då* 'then', *här* 'here', should as a rule not be marked as named entities.
- Verbs are generally not marked as named entities. Participles (and in some rare cases relative clauses) may be part of a naming phrase, however.
- Tokenisation (word segmentation) may sometimes be 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'
    'vid besök å Augusta Sandslånekontor'

We want to annotate *Petter Wortiain* and *Augusta Sands* as persons, but *mågPetter* and *Sandslånekontor* have been segmented as one word, and may not be split.

– ‘härigenom kungöres, af Gefle den 30 Januarii 1837.
Auctions=Kammarens Deputerade’
The date splits the ORG entity *Gefle Auctions=Kammarens*.

In these cases, annotate the faulty tokenisation segments (*mågPetter*, *Sandslånekontor* and *den 20 Januarii 1837.*) with the TOK label, so that we can easily find and correct these instances at a later stage. Also, annotate the named entities as accurate as possible, ignoring the tokenisation errors (PER: *mågPetter Wortiain*, PER: *Augusta Sandslånekontor* and ORG: *Gefle den 30 Januarii 1837. Auctions=Kammarens*).

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 (PER), while the subsequence *trädgårdsmästaren* ‘the gardener’ should additionally be marked as an occupation (OCC).

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.

Words where a name is part of a longer word, which is typically a **compound** or a morphological derivation, should generally not be marked as named entities. Thus, *göteborgare* ‘Gothenburger’ and *Danmarksresa* ‘trip to Denmark’ are not marked as named entities. This rule applies also in the case where a compound has been split erroneously as in *kramp kännedom* ‘seizure awareness’, where *kramp* ‘seizure’ otherwise would have been tagged as a symptom.

Phrases where the head has undergone **ellipsis** should also be marked as named enitities. Consider for example the phrase *inte denna vecka men nästa* ‘not this week but the next’. In this context, both *denna vecka* ‘this week’ and *nästa* ‘the next’ should be marked as temporal expressions.

Phrases with **misspelled words** should also be marked, for example *nästa vekca*, for *nästa vecka* ‘next week’.

Phrases in **a foreign language** that occur naturally within a Swedish text should also be marked, for example: *the Corvette*.

Note 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).

3.2 Guidelines for specific entity types

In the following chapters, we list each entity type targeted by our annotation, together with a short description and a set of positive and negative examples (related expressions that should not be marked by this category). We also list cases of potential conflict with other named entity types, and how they should be resolved. The descriptions of conflicts are duplicated in the different chapters, so that all specific information that pertains to a given type can be found in one chapter.

It could also be noted that some of the examples given throughout this document are taken from the corpus to be annotated, while others are borrowed from Ahrenberg, Frid and Olsson (2020). In the latter case, the actual names in the example might not be relevant to historical text, but still serve the purpose of illustration.

4 Sentence segmentation (SENT)

In historical text, sentence boundaries are not always marked with the punctuation used by convention in present-day text. It is quite common that sentence boundaries in older texts are marked by other punctuation than full stop, such as semicolon (;) or slash (/). Sometimes, sentence boundaries are only marked by an initial upper-case letter, without a preceding punctuation sign. It also happens that sentence boundaries are not marked at all. This makes it harder for both humans and computers to detect sentence boundaries in the text. At the same time, sentence segmentation is a vital part of many natural language processing tools, meaning that it is important to get the sentences correct when building a corpus. Therefore, sentence segmentation is included as an important annotation category in our project, even though our main focus is on named entity recognition.

In the sentence segmentation step, we aim to end up with sequences that would be regarded as syntactically complete sentences by present-day standards (regardless of punctuation). One approach to decide on the sentence boundaries in tricky cases, could be to read the text out loud. This may be of guidance for finding the natural way of segmenting the text into sentences.

A special case in the sentence segmentation process regards listings. As a rule of thumb, regard full stop, exclamation mark, question mark, colon and semicolon as sentence delimiters.

Example sequence: *Enligt denna åſigt har Utſkottet funnit: att någon aflöning till Förſamlingarnes Skollärare af Statens medel icke bör tillſtyrkas; att Förſamlingarne icke böra åläggas, hwarken något wißt ſätt för Skollärares aflöning, eller art owilkorligen inrätta ſockenſkolor; att Comminiſtrarne icke kunne tillförbindas beſtrida Lärarebefattningen, och att icke någon må till Klockare antagas, ſom icke ådagalagt, att han för Barna=Underwisningen eger nödiga kunſkaper.*

Segmentation:

1. Enligt denna åſigt har Utſkottet funnit:
2. att någon aflöning till Förſamlingarnes Skollärare af Statens medel icke bör tillſtyrkas;
3. att Förſamlingarne icke böra åläggas, hwarken något wißt ſätt för Skollärares aflöning, eller art owilkorligen inrätta ſockenſkolor;
4. att Comminiſtrarne icke kunne tillförbindas beſtrida Lärarebefattningen, och att icke någon må till Klockare antagas, ſom icke ådagalagt, att han för Barna=Underwisningen eger nödiga kunſkaper.

In come cases, a text that has been provided for annotation ends with an incomplete sentence, meaning that the text has been erroneously cropped. In these cases, you should annotate any named entities included in the cropped sentence, but you should not annotate the segment as a sentence.

Headings are to be annotated as SENT, even when they are not syntactically complete sentences.

5 Persons (PER)

This category includes people of any kind, whether real or fictional. Gods and mythical characters are included, but not animals or other creatures.

The following entities should be marked:

1. Proper names referring to a person, either by itself or as part of a longer sequence. Examples: *Johan*, *Lovisa Mathilda Larm*.

2. Plural references should also be marked: *Svenssons*, *familjen Lundgren* 'the Lundgren family'.

3. When a proper name reference is preceded by a title or epithet, or any other attribute that classifies or restricts the referent, they should all be included. Examples: *apotekare Lundin* 'pharmacist Lundin', *ogifta Mathilda Larsson* 'unmarried Mathilda Larsson', *ynglingen Adolf Eliasson* 'the youngster Adolf Eliasson', *Poliskonstapeln No 77 Ekberg* 'policeman No 77 Ekberg'.

4. Also epithets succeeding the proper noun, for example as a prepositional phrase, should be included in the entity, as long as the phrase is classifying or restricting the referent in any way, as in *Johan i Hult* 'Johan in Hult'.

5. Initials and prepositions should be included as part of a name: *C. A. Strömbom*, *Carl von Linné*. Initials that appear on their own should be annotated when they abbreviate a name, as in: *L kom hit* 'L came here'.

6. Due to faulty tokenisation, a full stop belonging to an initial may appear on its own. If so, it should be marked as part of the person name, as in: *A . Andersson*.

7. Nicknames are treated as proper names. They may be marked as part of a longer phrase, *Olle ”Bagarn” Larsson*, or as a separate name if occurring on its own, or in apposition to a proper name: *Olle Larsson, kallad ”Bagarn”* ‘Olle Larsson, called “Bagarn”’, in the second case as two different name expressions (*Olle Larsson* and *”Bagarn”*). The quotation marks should be annotated as part of the proper noun, even if they have been separated as single tokens.

8. When names are coordinated, as in *handlanden A. Andersson och arbetskarlen Victor Adolf Strömbäck* ‘the trader A. Andersson and the worker Victor Adolf Strömbäck’, the different names that are part of the coordination should be annotated as separate entities, meaning that *handlanden A. Andersson* is annotated as one item and *arbetskarlen Victor Adolf Strömbäck* as another item. In addition, for coordinations preceded by an attribute that classifies or restricts the referents in the coordination as a group, as in *ynglingarne Adolf Eliasson och Johan Peter Bergström* ‘the youngsters Adolf Eliasson and Johan Peter Bergström’ or *detektivkonstaplarne Andrén, Rönnbäck och Bruun* ‘the police detectives Andrén, Rönnbäck and Bruun’, the whole group should also be annotated as a person name, including the preceding attribute.

The following entities should not be marked as persons:

1. References based on a family role: *mamma* ‘mum’, *brodern* ‘the brother’, *hans hustru* ‘his wife’.

2. Common references based on an attribute of a person such as *lillan* ‘little one’, unless it is clearly established as a nickname.

3. Prepositions preceding a name reference of a person should not be marked, meaning that in the expression *till Anders* ‘to Anders’, only *Anders* will be marked as a person name.

4. Words referring to monarchs, such as *kungen* ‘the king’ or *prinsen* ‘the prince’ should not be marked as person names.

5. Titles should not be annotated as person names, e.g., *Herr Polititidirektören* ‘Mr Police director’, *H. Maj:t* ‘His Majesty’, *Ers Exc* ‘Your Excellency’.

6 Locations (LOC)

This category includes geographical locations of any kind, real or fictional, big or small: continents, countries, regions, cities, villages, areas, parks, streets, mountains, rivers, and so on.

The following entities should be marked:

1. Proper nouns referring to an entity of these kinds should be marked. Examples: *Lund*, *Kungstorget*, *Bangatan*, *Europa*.

2. Addresses in historical text may be referred to in a somewhat different manner than today, in which case the whole address should be marked as a location, as in: *huset No 61 i Ms 1ste rote* 'the house No 61 in Ms 1ste district'.

3. In cases where a proper name is preceded by an article or possessive pronoun, that should also be marked: *mitt Stockholm* 'my Stockholm'.

4. Common nouns referring to locations can be marked when they have developed the character of a standard, namelike reference, as in *Gamla stan* 'Old Town', *Östergötlands län* 'Östergötland county', or *Lundby socken* 'the Lundby parish'.

5. Sometimes it might be hard to determine the boundaries of the location entity. Consider the following two examples:

   (a) *arbetareföreningarne från södra* ***Sverige*** *och Danmark*
   'the workers' associations from southern **Sweden** and Denmark'

   (b) *för att tillbringa sin tjenstledighet i* ***Södra Sverige***
   'to spend his leave of absence in **Southern Sweden**'

   In both examples, the writer is referring to the southern part of Sweden. In example (a), however, *södra* is written with an initial lower-case letter,

signalling that the writer refers to *Sverige* 'Sweden' as a location name, and uses the adjective *södra* 'southern' to further specify the part of the location. In this case we therefore annotate only *Sweden* as a location. In example (b), on the other hand, the writer uses a word-initial upper-case letter for *Södra*, signalling that s/he is referring to the whole segment *Södra Sverige* 'Southern Sweden' as a location. Therefore, in example (a) we will annotate *Sverige* as LOC, while in example (b) we will annotate the longer span *Södra Sverige* as LOC.

The following entities should **not** be marked as locations:

1. Location names that are a part of a brand name. Example: *Norrlands* in *Norrlands Guld* (a beer brand) is not marked as a location but as a work of art.
2. Indexical references using adverbs or common nouns, such as *hemma* 'at home', *i utlandet* 'abroad', should not be marked.
3. Prepositions preceding a location name should not be marked, meaning that in the expression *till Malmö* 'to Malmö', only *Malmö* will be marked as a location name.
4. Common names of rooms such as *köket* 'the kitchen' should not be marked.

Conflicts of Location with other categories:

- **LOC :: ORG** Organisations usually have offices which may serve as landmarks. Similarly, location names are commonly used as metonyms for organisations. Thus, there may be a conflict for Location with Organisation. Solve the conflict by considering the referent. If the phrase is part of a sentence, you can test whether the sentence answers a question introduced by *var* or *vart* 'where', in which case the entity should be marked as a location. If, instead, the sentence answers a question introduced by *vem/vilka* 'who' or *vems/vilkas* 'whose', the entity should be marked as an organisation.
  - *förfäljes på* ***Gefle Stads Auctions=Kammare***
    'sold at **Gävle auction house**'
    LOC: Answers the question *var säljs den?* 'Where is it sold?'
  - *uttagit polett å* ***riddarhufet***
    'given a token at **the knight's house**'

LOC: Answers the question *var mottogs poletten?*
'Where was the token given?'

– ***Papiſterne** hafwa fått ſin benämning af deras tro på Påfwens ofelbarhet*
'**The Papists** got their name from their belief in the infallibility of the Pope'
ORG: Answers the question *vem/vilka fick sitt namn?*
'who got their name?'

– *i så måtto att den höll **Englands** spira*
'to the extent that it held the sceptre of **England**'
ORG: Answers the question *vems spira?* 'whose sceptre?'

- **LOC :: EVN** An event is often referenced together with a location reference. The annotation will then depend on how established the location reference is as part of the name of the event. Example: *kalabaliken i Bender* 'the Skirmish at Bender' may be considered to be an established name, so the whole sequence is labelled EVN. In parallel, the subsequence *Bender* is to be annotated as LOC.

- **LOC :: WRK** Statues, buildings and other may also be used as metonymic for their locations. If the phrase is part of a sentence, you can test whether the sentence answers a question introduced by *var* or *vart* 'where', in which case the entity should be marked as a location. In other cases, it should be marked as a work of art.

  Example:

  WRK: *Östra Kungsgatubron försvann*
  'Östra Kungsgatubron disappeared'
  'vad/*var försvann?' 'what/*where disappeared'

  When both views are possible, use location as default.

7 Organisations (ORG)

This category includes companies, organisations, governments, political parties and NGOs, public bodies, sports clubs, schools, hospitals and generally anything with a legal status in a society. For a more fine-grained annotation, this category is further divided into three subcategories: company (COMP), institution (INST) and other (OTH).

The following entities should be marked:

1. Proper nouns and acronyms referring to entities of this category, could be either COMP (*handelsfirman J. O. Grén & Co* 'trading company J. O. Grén & Co') or INST (*Swenska Läkaresällskapet* 'Swedish Society of Medicine').

2. Common nouns that have established themselves as names, should typically be annotated as INST: *Nationalförsamlingen* 'the National Assembly', *Socialdemokraterna* 'the Social Democrats', *Poliskammaren* 'the Police Chamber', *Svenska Akademien* 'the Swedish Academy'. Also, occasionally nominalised adjectives: *de vita* 'the Whites'.

3. Common nouns or abbreviations that pick out a societal institution, such as *Pkm* 'the Police Chamber', *Riksdagen* 'the Parliament', *Rådhusrätten* 'the District Court' are annotated as INST.

4. Political committees should be marked as INST, e.g., *Allmänna Beſwärs= och Ekonomi=Utſkottet*, *Comiteen till öfwerſeende af Rikets allmänna underwisningswerk*.

5. Church communities should be marked as INST, e.g., *Grekiſka eller Ryſka Kyrkan är gammal, och liknar i det närmaſte den Romerſka* 'The Greek or Russian Church is old, and closely resembles the Roman one'. In this example, also note that *den Romerſka* 'the Roman one' should be

annotated, even though *kyrkan* 'Church' is omitted, see further information about ellipsis in Section 3.1.

6. When a name and an abbreviation occur together they should be marked both as separate references and as a whole. This means that in the example *Socialstyrelsen (SoS)*, the whole entity *Socialstyrelsen (SoS)* should be marked as an Institution, but also *Socialstyrelsen* and *SoS* as separate items.

7. In cases where a proper noun referring to an organisation is preceded or succeeded by an attribute that classifies or restricts the referent, the whole sequence should be included in the annotation: *Apoteket Uttern* 'the pharmacy Uttern'.

8. If an entity seems to refer to an organisation, but it is hard to distinguish as a company or an institution, choose the OTH category ('Other').

The following entities should not be marked:

1. Prepositions preceding an organisation name should not be marked.

2. Collective descriptive references to organisations or their memberships should not be marked. Examples: *NATO-medlemmar* 'NATO members', *de röda trupperna* 'the red troops'.

Conflicts of ORG with other categories:

- **ORG :: LOC** Organisations usually have offices which may serve as landmarks. Similarly, location names are commonly used as metonyms for organisations. Thus, there may be a conflict for Location with Organisation. Solve the conflict by considering the referent. If the phrase is part of a sentence, you can test whether the sentence answers a question introduced by *var* or *vart* 'where', in which case the entity should be marked as a location. If, instead, the sentence answers a question introduced by *vem/vilka* 'who' or *vems* 'whose', the entity should be marked as an organisation.

  – *förfäljes på* ***Gefle Stads Auctions=Kammare***
    'sold at **Gävle auction house**'
    LOC: Answers the question *var säljs den?* 'Where is it sold?'
  – *uttagit polett å* ***riddarhuſet***
    'given a token at **the knight's house**'

LOC: Answers the question *var mottogs poletten?*
'Where was the token given?'

- ***Papiſterne*** *hafwa fått ſin benämning af deras tro på Påfwens ofelbarhet*
  '**The Papists** got their name from their belief in the infallibility of the Pope'
  ORG: Answers the question *vem/vilka fick sitt namn?*
  'who got their name?'
- *i så måtto att den höll* ***Englands*** *spira*
  'to the extent that it held the sceptre of **England**'
  ORG: Answers the question *vems spira?* 'whose sceptre?'

- **ORG :: WRK** Product names often include the name of the company that makes the product. Use the context to decide whether the product or the company is involved. Consider the two examples with 'Netflix' below:

  - Work of Art
    Jag tittar på Netflix
    'I watch Netflix'
  - Organisation
    Nu har Netflix sagt att serien ska få en varningstext
    'Now, Netflix has said that the series will get a warning text'

8 Measurement entities

The measurement category includes units for measuring and is further divided into seven subcategories: monetary (MON), weight (WEI), length (LEN), distance (DIST), area (AREA), volume (VOL) and other (OTH).

The following entities should be annotated:

1. Expressions describing the monetary value of things should be annotated as MON: *en winteröfwerrock värd* **20 kr** 'a winter coat worth **20 SEK**'.

2. Expressions describing the weight of things should be annotated as WEI: ***5 ℔*** *kaffe* '**5 ℔** coffee'.

3. Expressions describing the length of things should be annotated as LEN: *100 meter* '100 meters'.

4. Expressions describing the distance to something should be annotated as DIST, e.g.: ***1 1/2 engelska mil*** *norr om S:t Mary Mission Kansas* '**1 1/2 English miles** north of St. Mary Mission Kansas' and *få fot från land* 'a few feet from land'

5. Expressions describing the area should be annotated as AREA: *7 hektar* '7 hectares'.

6. Expressions describing volume should be annotated as VOL: *30 läster* '30 loads', ***3 buteljer*** *öl* '**3 bottles of** beer', ***tvenne kannor*** *mjölk* 'two jugs of milk', ***åtta koppar*** *säd* '**eight cups** of grain'.

7. If an entity refers to a measurement subcategory not listed above, and you think it would be useful to add this particular subcategory to the annotation task, you could choose the OTH category ('Other') and add a comment to the annotation, stating what the new subcategory would be.

The following entities should not be marked:

1. References to measurements that are not related to money, weight, length, distance, area or volume, such as *4 st knifvar* '4 [pieces of] knives', *1 mejsel* '1 chisel'.

9 Temporal entities (TME)

For temporal entities, we follow the guidelines in Ahrenberg, Frid and Olsson (2020), covering time points and continuous intervals on a presumed timeline from the beginning of time to the present and including the future.

Note that many types of references that are temporal in some other sense are not included. These include durations (answering the question *hur länge?* 'for how long?'), frequencies (answering the question *hur ofta?* 'how often?'), and age references (answering the question *hur gammal?* 'how old?'). This is compatible with Kokkinakis (2004), but is more restrictive than the Sparv SweNER web service provided by Språkbanken Text.

For a more fine-grained annotation, we further divide this category into four subcategories: date expressions (DATE), time expressions (TIME), intervals (INTRV) and other temporal expressions (OTH).

The following entities should be marked:

1. Standard references to dates, weeks, months, years, seasons, decades and centuries are annotated as DATE: *den 12 December*, *2 Januari 1880*, *1800-talet* 'the 19th century', *november 1827*.

2. If the dating is defining an interval, the expression should instead be annotated as INTRV: *1817–19*.

3. Standard references to times of the day are annotated as TIME: *kl. 4 e. m.* '4 pm', *kvart över sex* 'a quarter past six'. This also includes more vague references, such as *vid middagstiden* 'at noon' and *på aftonen/morgonen* 'in the evening/morning'.

4. Special names for holidays such as *Påskdagen* 'Easter Day', *Nyårsafton* 'New Year's Eve' are annotated as DATE, when the temporal reference

is prominent (make sure to distinguish between holidays used as time references and holidays referring to events, se further below).

5. Deictic references related to the current speech-time with a nominal head word are annotated as DATE, such as *i morgon* 'tomorrow', *i sommar* 'this summer', *nästa vecka* 'next week', *förra månaden* 'last month', *för ett år sen* 'a year ago', *på tisdag* 'on Tuesday', *om tre år* 'in three years', *i morse* 'this morning', and even items such as *igår* 'yesterday', *idag* 'today', *i fjol* 'last year', *i natt* 'tonight' that are adverbial-like but still could be seen as prepositional phrases with a nominal head. The time reference could also be a bit more vague, as in *nästa vinterting* 'the next winter court session' and *hösten samma år* 'the autumn of the same year'.

6. Vague references are included if they have a nominal head word of a temporal entity: *om ett par timmar* 'in a couple of hours', *för några veckor sedan* 'a few weeks ago'.

7. If the temporal phrase includes a preposition, determiner or adjective, these should normally also be marked: *på torsdag* 'on Thursday'. However, some prepositions may change the interpretation from a specific interval to a duration and then we only mark the words that name the interval. So in the phrase *från och med i morgon* 'from tomorrow', only the sequence *i morgon* 'tomorrow' is marked as a date. Likewise, in the expression *sedan 1885* 'since 1885', only *1885* is marked as a date.

The following entities should not be marked:

1. Deictic adverbs such as *nu* 'now', *då* 'then', *senare* 'later', *tidigare* 'earlier', *samtidigt* 'at the same time', *nyss* 'just recently'.

2. Phrases of any kind expressing duration (unless the duration is expressed as a fixed interval, as described above): *länge* 'for a long time', *i två timmar* 'for two hours', *under de senare 12 åren* 'during the last twelve years'.

3. Phrases of any kind expressing age: *tre år gammal* 'three years old'.

4. Phrases of any kind expressing frequency: *ofta* 'often', *varje dag* 'every day', *på torsdagar* 'on Thursdays', *på kvällarna* 'in the evenings', *två gånger om året* 'twice a year', *varje vecka* 'every week'.

5. Phrases implying a reference point other than the current speech-time: *tre år senare* 'three years later', *efteråt* 'afterwards'.

6. If a temporal reference is written as part of a word sequence with quite a different meaning, it should not be marked: *född -58* 'born in -58'.

7. Phrases with vague indeterminate nouns such as *om ett tag* 'in a while', *om en stund* 'in a moment', *för länge sedan* 'a long time ago'.

Conflicts of TME with other categories:

- **TME :: EVN** An event is located in time, so it may be hard to judge whether a time reference is part of the event reference or a separate reference. Sometimes, the sequence as a whole could be annotated as an event, whereas part of it is also annotated as a date. For example, the phrase *Danmarks fälttåg 1848* 'Denmark's military campaign 1848' is annotated as EVN, while at the same time *1848* is annotated as DATE. Conversely, a temporal reference may use an event reference as a part: *innan jul* 'before Christmas', *under Andra världskriget* 'during the Second World War'. In those contexts, the sequence as a whole is to be annotated as a temporal expression, while at the same time parts of the expressions are also annotated as EVN (*jul* 'Christmas' and *Andra världskriget* 'the Second World War', in these examples).

  Quite often both interpretations seem to be present. A question test can sometimes be used. If the phrase answers the question *när?* 'when', it should be marked TME, if not it is likely to be EVN. Examples:

  – EVN: **Julen** *närmar sig*
    '**Christmas** is approaching'
    Answers the question: *Vad närmar sig?* 'What is approaching?'
  – DATE: *Vi måste vara färdiga* **innan jul**
    'We have to be finished **before Christmas**'
    Answers the question: *När måste vi vara färdiga?*
    'When do we have to be finished?'

  If both options are possible, use EVN as default for holidays that imply some kind of celebration.

- **TME :: SYMP** A symptom may be particularly serious if it is recurring. Thus, just as we would annotate *frekventa kräkningar* 'frequent vomiting' as SYMP, we also annotate *två timmar mellan kräkningar* 'two hours between vomiting' as SYMP.

10 Events (EVN)

Events cover all types of events listed in Kokkinakis (2004), namely historical and political events, weather phenomena and natural disasters, cultural events such as festivals, sports competitions and events of a religious nature and holidays. However, we do not provide labels for subevents, so EVN is used for all of them.

The following entities should be marked:

1. Historical or political events, such as battles, wars, scandals, campaigns and crimes. Example: *Danmarks fälttåg 1848* 'Denmark's military campaign 1848'.
2. Weather phenomena and natural disasters such as hurricanes and storms: *stormen Gudrun* 'the storm Gudrun'.
3. Cultural events, like festivals and fairs: *Lucie Marknad i Wimmerby* 'Lucia market in Wimmerby'.
4. Religious events, like holiday celebrations: *Påsk* 'Easter', *Julafton* 'Christmas Eve' (make sure to distinguish between holidays used as time references and holidays referring to events, se further below).
5. Sports events: *Olympiska spelen* 'the Olympic Games'.
6. Also smaller events are annotated, such as *bolagsstämma* 'annual general meeting' and *vinterting* 'winter court session'.

Conflicts of EVN with other categories:

- **EVN :: LOC** An event is often referenced together with a location reference. The annotation will then depend on how established the location

reference is as part of the name of the event. Example: *kalabaliken i Bender* 'the Skirmish at Bender' may be considered to be an established name, so the whole sequence is labelled EVN. In parallel, the subsequence *Bender* is to be annotated as LOC.

- **EVN :: TME** An event is located in time, so it may be hard to judge whether a time reference is part of the event reference or a separate reference. Sometimes, the sequence as a whole could be annotated as an event, whereas part of it is also annotated as a date. For example, the phrase *Danmarks fälttåg 1848* 'Denmark's military campaign 1848' is annotated as EVN, while at the same time *1848* is annotated as DATE. Conversely, a temporal reference may use an event reference as a part: *innan jul* 'before Christmas', *under Andra världskriget* 'during the Second World War'. In those contexts, the sequence as a whole is to be annotated as a temporal expression, while at the same time parts of the expressions are also annotated as EVN (*jul* 'Christmas' and *Andra världskriget* 'the Second World War', in these examples).

  Quite often both interpretations seem to be present. A question test can sometimes be used. If the phrase answers the question *när?* 'when', it should be marked TME, if not it is likely to be EVN. Examples:

  – EVN: **Julen** *närmar sig*
    '**Christmas** is approaching'
    Answers the question: *Vad närmar sig?* 'What is approaching?'
  – DATE: *Vi måste vara färdiga* **innan jul**
    'We have to be finished **before Christmas**'
    Answers the question: *När måste vi vara färdiga?*
    'When do we have to be finished?'

  If both options are possible, use EVN as default for holidays that imply some kind of celebration.

11 Works of art and other artefacts (WRK)

This category includes name or title references to works of art, such as books, plays, brand names of commercial products, newspapers and journals.

The following entities should be marked:

1. Proper nouns referring to a product: *Pepsodent*, *Honda Civic*, *Windows NT*.

2. In cases where a proper noun referring to a work of art is preceded or succeeded by an attribute that classifies or restricts the referent, the whole sequence should be included in the annotation: *ångfartyget ”Konung Oskar”* ‘the steamship “Konung Oskar”’.

3. Noun phrases used as product names or titles: *Dagens Nyheter*, *Macbeth*.

4. Phrases of other kinds, including complete clauses when used as the title of work of art. Note that all words should then be marked, including function words: *Till Damaskus* (the play).

5. A product or work of art using the name of another category, such as person or location: *Hamlet* (the play), *Jerusalem* (the book).

The following entities should not be marked:

1. Names of natural kinds such as *potatis* ‘potato’, *ros* ‘rose’, *lax* ‘salmon’ should not be marked. In contrast, *Bintje* is a developed product and should be marked.

Conflicts of WRK with other categories:

- **WRK :: LOC** Statues, buildings and other may also be used as metonymic for their locations. If the phrase is part of a sentence, you can test whether the sentence answers a question introduced by *var* or *vart* 'where', in which case the entity should be marked as a location. In other cases, it should be marked as a work of art.

  Example:

  WRK: *Östra Kungsgatubron försvann*
  'Östra Kungsgatubron disappeared'
  *vad/*var försvann?* 'what/*where disappeared'

  When both views are possible, use location as default.

- **WRK :: TREAT** A medicine is a product but also a kind of treatment. Let the context decide. If the medicine is referred to as a treatment, the sequence should be marked as TREAT. In other cases as WRK.

- **WRK :: ORG** Product names often include the name of the company that makes the product. Use the context to decide whether the product or the company is involved. Consider the two examples with 'Netflix' below:

  – Work of Art
    Jag tittar på Netflix
    'I watch Netflix'
  – Organisation
    Nu har Netflix sagt att serien ska få en varningstext
    'Now, Netflix has said that the series will get a warning text'

12 Symptoms (SYMP)

The guidelines for the Symptom category are modelled on the guidelines by Ahrenberg, Frid and Olsson (2020), which are in turn based on the annotation guidelines for the 2010 i2b2/VA challenge (i2b2 tranSMART Foundation 2010), and the concept 'medical problems' as defined there. A symptom phrase is a phrase that contains observations made by patients, clinicians or others about the patient's body or mind, that are thought to be abnormal or caused by a disease. It is important that the state reported is deviant and that it can be treated as a disease or illness. As our data go beyond patient records we do not restrict occurrences of such phrases to clinical data, but mark the phrases also when symptoms are discussed more generally or related to causes. It should also be noted that since we work with historical text, names of diseases and medical conditions may be quite different from today.

The following entities should be marked:

1. Noun phrases that name a disease (*kolera* 'cholera'), syndromes or abnormal states (*hosta* 'coughing', *bruten arm* 'broken arm'), specific viruses or bacteria, or indicative test results (*lågt blodtryck* 'low blood pressure').

2. Adjectival phrases, including participles, that do the same, such as *blek* 'pale', *förvirrad* 'confused', *mycket ont* 'in much pain'. Note that a preceding verb is not marked, so in the phrase *har ont* 'is in pain', only *ont* 'in pain' is marked as a symptom.

The following entities should not be marked:

1. **NB!** Verbs should never be marked, even if they indicate a problem. This means that a phrase like *blöder mycket* 'bleeds a lot' is not marked as a symptom.

2. General words such as *sjukdom* ‘disease’, *sjuk* ‘ill’, *virus*.
3. Words such as *bra* ‘good’ or *normalt* ‘normal’ should not be marked, even if referring to bodily phenomena such as blood pressure.
4. Even though negations of normality may indicate a problem, it should not be marked either: *inte bra* ‘not good’.
5. States that are the result of normal, everyday activities should not be marked: *blev trött* ‘became tired’.
6. Measurements, even if they can be inferred to be outside normal range: *blodtryck 160/100* ‘blood pressure 160/100’.
7. Naturally occurring states or phases that are not to be regarded as diseases or illnesses: *pubertet* ‘puberty’, *gravid* ‘pregnant’.
8. Words or phrases that could be taken as symptoms when they are related to a person, should not be marked if they don’t refer to a person or a body: *virus och kräklukt städades bort* ‘virus and vomit odor were cleaned away’.
9. Symptoms should only be annotated when related to illness, meaning that more literary references to emotional states should not be marked, e.g. *där trifves tungsinthet och ångest* ‘there sadness and anxiety thrive’ and *skakade honom i en paroxysm af obeslutsamt raseri* ‘shook him in a paroxysm of indecisive fury’.

Conflicts of SYMP with other categories:

- **SYMP :: TME** A symptom may be particularly serious if it is recurring. Thus, just as we would annotate *frekventa kräkningar* ‘frequent vomiting’ as SYMP, we also annotate *två timmar mellan kräkningar* ‘two hours between vomiting’ as SYMP.

13 Treatments (TREAT)

The guidelines for treatments are also modelled on the annotation guidelines for the 2010 i2b2/VA challenge (i2b2 tranSMART Foundation 2010), which employs treatment as a concept. Treatment phrases are phrases that describe procedures, interventions, and substances given to a patient in an effort to resolve a medical problem. They include both therapeutic and preventive measures, pharmacological substances, clinical drugs and drug delivery devices.

It should be noted that since we work with historical text, treatments pointed out in the text may be quite unconventional as compared to present-day standards, and names of medicines may also be different from today.

The following entities should be marked:

1. Noun phrases that refer to medications (*morfin* 'morphine'), biological substances (*blodtransfusion* 'blood transfusions'), hardware (*kateter* 'catheter') and general terms used for treatments (*terapin* 'the therapy', *läkemedel* 'medication').

2. Preventive measures, if prescribed by doctors or an organisation: *renlighet* 'cleanliness', *sund föda* 'healthy food', *vaccinering* 'vaccination'.

3. General terms referring to a patient's treatments: *hennes medicinering* 'her medication'.

4. Noun phrases that refer to substances that are not usually used as medications, if they are clearly part of a treatment, such as *filmjölk* 'sour milk' in *fick filmjölk på recept* 'got sour milk on prescription'.

5. Adjective phrases that do the same (though they seem to be rare).

The following entities should not be marked:

1. Precautionary measures that are not prescribed by a doctor, but occur regularly in everyday life: *måste få vila nu* ‘need to rest now’.

2. **NB!** Verbs should never be marked. This means that in the phrase *ge morfin var fjärde timme* ‘give morphine every four hours’, only *morfin var fjärde timme* ‘morphine every four hours’ is marked as TREAT, excluding the verb.

3. Phrases referring to tests used in order to diagnose a patient: *blodprov* ‘blood sample’.

Conflicts of TREAT with other categories:

- **TREAT :: WRK** A medicine is a product but also a kind of treatment. Let the context decide. If the medicine is referred to as a treatment, the sequence should be marked as TREAT. In other cases as WRK.

14 Occupations (OCC)

Common nouns describing what people do for a living should be marked as occupations: *läkare* ‘doctor’, *arbetskarl* ‘worker’, *piga* ‘maid’. This also includes more unorthodox ways of making a living, such as *tjuv* ‘thief’.

Occupational titles are often used as attributes to person names, as in *tullvaktmästare B. J. Lundgren* ‘customs officer B. J. Lundgren’. In these cases, the sequence as a whole (*tullvaktmästare B. J. Lundgren*) should be annotated as a person, while at the same time the occupational title (*tullvaktmästare* ‘customs officer’) should also be annotated as an occupation.

Words describing people in power refer to occupation-like titles and should thus be annotated as occupations, e.g., *kung* ‘king’, *kejsare* ‘emperor’, *furste* ‘ruler’ and *drottning* ‘queen’. Words such as *prins* ‘prince’, *greve* ‘earl’ or *enkedrottning* ‘widowed queen’ should however not be marked as occupations, since these are pure titles rather than a way of earning money or making a living.

The following entities should not be marked as occupations:

1. *tjänare* ‘servant’ in a submissive meaning: *Eders Höga Nådes ödmiukaste tienare* ‘Your Most Gracious Highness’s humble servant’.
2. Titles of persons: *Ers Exc.* ‘Your Excellency’, *H. Maj:t* ‘His Majesty’.
3. Titles of plays and other works of art: *Tillfället gör Tjufwen* ‘Opportunity makes the thief’.
4. Activities that could rather be seen as commissions or assignments than a way of making a living, e.g.: *ledamot* ‘commissioner’, *borgenär* ‘creditor’ and *nämndeman* ‘lay judge’.

15 Miscellaneous (MISC)

If name-like references are found that are hard to categorise into any of the above entity types, it could be marked as MISC. In those cases, the annotator is encouraged to add a comment to the annotation, describing why s/he finds this sequence suitable to annotate as a named entity, and why it is hard to categorise this particular entity into the present scheme.

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.

i2b2 tranSMART Foundation. 2010. 2010 i2b2 / VA Challenge Evaluation: Concept Annotation Guidelines.

Kokkinakis, Dimitrios. 2004. Reducing the effect of name explosion. *Proceedings of the LREC workshop: Beyond named entity recognition, semantic labelling for NLP tasks. Fourth Language Resources and Evaluation Conference (LREC)*. Lisbon: ELRA.

Pettersson, Eva and Lars Borin. 2022. Swedish diachronic corpus. Darja Fišer and CLARIN Andreas Witt (eds), *The infrastructure for language resources*. Berlin: De Gruyter Mouton.