RAPPORTER FRÅN SWE-CLARIN
SWE-CLARIN REPORT SERIES
sweclarin.se/scrs
[SCR-01-2020]

A New Gold Standard for Swedish Named Entity Recognition

Annotation Guidelines

Lars Ahrenberg[1], Johan Frid[2], and Leif-Jöran Olsson[3]

[1]Linköping University • lars.ahrenberg@liu.se
[2]Lund University Humanities Lab, Lund University • johan.frid@humlab.lu.se
[3]Språkbanken Text, University of Gothenburg • leif-joran.olsson@svenska.gu.se

Contents

1 Introduction

The CLARIN research infrastructure[1] (short for Common Language Resources and Technology Infrastructure) aims to make digital language resources available to researchers from all disciplines, with a special focus on the humanities and social sciences. Language resources do not only consist of language data, but also of tools and systems for data analysis. For this purpose it is beneficial to have access to manually annotated data, which enables both system development and system evaluation.

In this work we are concerned with the annotation of *named entities* in Swedish texts, i.e., words and word sequences that refer to an entity of a specific semantic type, such as a person or a place. Systems that can identify such sequences in texts are referred to as systems for Named Entity Recognition and Categorisation (henceforth NERC). Often the interest is focused on phrases where the head word is a proper name (hence the term ’named entity’), but for many types the name, or commonly accepted reference, is not built on proper nouns at all; think of events, temporal references or titles of books or films.

The work has been done as part of the activities in the Swedish CLARIN node, *Swe-Clarin*[2], by a special project group. The invitation to participate was open, and a couple of open meetings were held in the beginning of the project. After this initial phase the three authors of this report have been the main actors. The decisions reported here have been made by us in meetings and discussions, mostly over the Internet.

In this report we describe and motivate the guidelines as they have been developed for the first distribution of the gold standard resource. The main aim is to describe the criteria that have been used to delineate the categories and how they relate to taxonomies and criteria used in other NERC projects. We start,

[1] https://www.clarin.eu/

[2] https://sweclarin.se/

in the rest of this chapter, by stating the aims and scope of the project, and the work as it has developed this far. Chapter 2 gives a summary of related work, and Chapter 3 presents the annotation format and the general guidelines. The following eight chapters provide our guidelines and examples for each one of the selected entity types.

In a companion report (Ahrenberg, Frid and Olsson 2020) the actual contents of the resource and some benchmarking data are described in more detail.

1.1 Aims

Currently available gold standards for Swedish NERC either represent language from the 1990ies or a single genre. With this resource we wished to include more recent language, in particular as it is used in social media. As NERC has gained increased interest in medical applications over the years, we have also included two semantic types from the medical domain. Still, our aim is far from producing an all-embracing resource; rather, the scope is limited to eight different categories that provide different challenges for automatic named-entity recognition.

The aims have been:

- to provide a free resource for research and development
- to provide at least 1000 instances for each selected category
- to select categories that are relevant and at the same time provide challenges of different kinds for developers
- to develop detailed criteria and guidelines for the categories that can be distributed with the resource
- to base the resource on annotations from three different annotators with known inter-annotator agreement
- to provide benchmarks on the basis of state-of-the-art software

To guarantee that the resource can be distributed freely we selected data that are already available from Språkbanken Text[3]. This has meant that the sentences are often scrambled, but the resource also contains data from unscrambled documents. We have also aspired to collect data from the same time period; and most of the texts included were produced in or around 2010.

[3] However, SIC, the Stockholm Internet Corpus has been created by Robert Östling at Stockholm University and can be found at https://www.ling.su.se/english/nlp/corpora-and-resources/sic/stockholm-internet-corpus-sic-1.99019

1.2 Work process

The first decisions concerned the categories to be included and the genres and sub-corpora that were to supply the data. We settled for eight categories as listed in Table 1.

Category	**Abbreviation**
Persons	PRS
Location	LOC
Organisation	GRO
Event	EVN
Time point or interval	TME
WorkOfArt / Product	WRK
Symptom	SMP
Treatment	MNT

Table 1: The eight entity types of the resource with abbreviations.

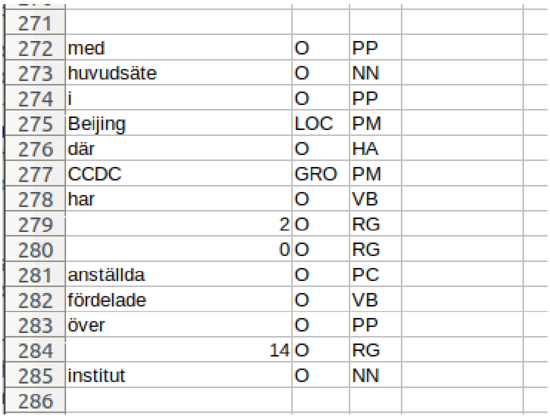
Secondly, we made decisions on which data to use. The current corpus has actually been developed incrementally, as we found that certain categories were not well represented in the first selection of data sources. This motivated the addition of the Smittskydd- and Wikipedia-krig corpora. An overview of the corpus is given in Table 2.

Source	**Genre**	**Subject**	**Mode**
bloggmix	blog texts	life of a youth	scrambled
familjeliv-barnhälsa	social forum	children’s health	scrambled
flashback-fordon	social forum	vehicles, esp. cars	scrambled
SIC	blog texts	everyday activities	unscrambled
Göteborgsposten	news text	varied	scrambled
Smittskydd	Medical journal	health protection	scrambled
Wikipedia-krig	Wikipedia text	war history	unscrambled

Table 2: Data sources of the project.

Documents have been sampled from the corpora with a size of 2000-2500 tokens and formatted in a spreadsheet. Tokenisation is automatic and not seldom at odds with standard Swedish orthography. The spreadsheets have three columns, one for tokens, one for an automatically generated named-entity tag, and one for a part-of-speech. An example is shown in Figure 1. Annotation is performed by changing the proposed named-entity tag, when it is found to be erroneous. Abbreviations for the categories were chosen so that it would normally be sufficient to press the key for the first letter to change the tag.

A fourth column is used for keeping track of instances (see below, chapter 3.1).



271					
272	med	O	PP		
273	huvudsäte	O	NN		
274	i	O	PP		
275	Beijing	LOC	PM		
276	där	O	HA		
277	CCDC	GRO	PM		
278	har	O	VB		
279	2	O	RG		
280	0	O	RG		
281	anställda	O	PC		
282	fördelade	O	VB		
283	över	O	PP		
284	14	O	RG		
285	institut	O	NN		
286					

Figure 1: Data as presented to annotators. Necessary changes are made in the second column.

Annotation and guidelines have been developed in tandem. The first version of the guidelines was based on guidelines produced in other projects within or related to the Swe-Clarin project. It was not very detailed and annotators relied to a large extent on intuitions derived from the characterization of the categories. After an initial round of annotating ten documents each, we collected problematic examples for discussion and decision. The decisions then resulted in changes and additions to the guidelines, and revisions of the annotations.

Entity type	**Iter 1**	**Iter 2**	**Iter 3**
Persons	0.898	0.946	0.930
Location	0.890	0.889	0.917
Organisation	0.759	0.789	0.846
Event	0.646	0.724	0.787
Time point or interval	0.463	0.699	0.836
WorkOfArt/Product	0.780	0.763	0.822
Symptom	0.537	0.664	0.750
Treatment	0.463	0.699	0.836
All	0.716	0.800	0.856

Table 3: Annotation progress. Fleiss' kappa at different stages in the annotation process

In the course of the project the annotation guidelines have been revised several times. This version is the outcome of the ninth revision.

Inter-rater agreements have been checked on several occasions. Initially, they were made with tight intervals including discussions of problematic examples in between. Progress was quite sharp in the initial phase, as shown in Table 3.

Before producing the final annotations to be included in the resource, inter-rater agreements for all annotators were computed again. One annotator was found to be deviating greatly from the others and we decided to discard those annotations in the final phase. Inter-rater agreements, using Fleiss' kappa ,for the other annotators are shown in Table 4.

The final annotations were produced using a spreadsheet where the available annotations, at least three for every token, were set side-by-side. One annotator was appointed for each sub-corpus to check disagreements against the guidelines one more time. In case a disagreement is a matter of interpretation, and not clearly specified in the guidelines, majority voting was applied. If that did not resolve the issue, the appointed annotator made the decision.

Annotators	**Kappa**
1, 2, 3	0.88
1, 2, 4	0.87
1, 2, 3, 4	0.78

Table 4: Inter-rater agreements, measured by Fleiss' kappa before final decisions were made.

1.2.1 Acknowledgements

We are indebted to Patricia Lindblad and Arne Jönsson for annotation work and discussion of the guidelines, to Elena Volodina for giving us the anonymisation guidelines of the SweLL project, and to Stian Rodven Eide for pointing out missing information in earlier versions of the guidelines.

2 Related work

Although the problem of name recognition was studied before, consorted efforts towards NERC started with the sixth Message Understanding Conference (MUC-6) (Grishman and Sundheim 1996). The named-entity task was defined there for the first time as a separate task, valuable in itself and seen as necessary towards the goal of information extraction.

2.1 Entity type selections

The MUC-6 conference introduced the three abbreviations ENAMEX, TIMEX, and NUMEX for 'entity-name expressions', 'time expressions' and 'numerical expressions', respectively. In the annotation these were used as SGML-tags and sub-typed; thus, 'person' and 'organisation' were values of the TYPE attribute for <ENAMEX>-tags, while the TYPE attribute for <NUMEX> had values such as 'money' and 'percentage' (Grishman 1995). Altogether, there were seven different types.

The splitting of the named-entity task into three sub-tasks was continued for the MUC-7 (Chinchor 1997). However, when NERC was defined for the Conferences on Natural Language Learning some years later, the terminology and the format of the annotation had changed (Sang 2002, 2003). While the task had been extended to more languages, Spanish and Dutch in 2002, English and German in 2003, the number of semantic types had been reduced to four: persons, locations, organisations, and miscellaneous, where the latter includes referents for numerical and time expressions.

The four-split datasets developed for the 2002 and 2003 CoNLL shared tasks are still in much use (Yadav and Bethard 2019). Some recent Nordic projects also follow this coarse-grained split (Ingólfsdóttir, Þorsteinsson and Loftsson 2019; Johansen 2019).

After these early initiatives the interest in NERC has been constant and has developed in many directions (Nadeau and Sekine 2007). One line of research focuses on special domains, where the biomedical domain is particularly noticeable with specific categories such as proteins and cell attributes (Kim et al. 2004) or general categories such as test, treatment and drug (Uzuner et al. 2011; Segura-Bedmar, Martínez and Herrero-Zazo 2013). Another trend is to attack the more difficult problems associated with noisy text such as that found in social media (e.g., Baldwin et al. 2015).

The three-way division of the MUC shared tasks into expression categories came with a simple hierarchy and a total of seven categories. Subsequent work sometimes worked with more elaborate divisions in what has been called fine-grained, or open-domain NERC. Conversely, categories that have been difficult to differentiate, such as names of cities and countries that may refer either to a geographical location or its government have sometimes been merged into categories such as geo-political entities. The type ’facility’ has been used to cover other entities such as banks that can appear in texts either as legal entities, i.e., organisations or as buildings in a landscape, i.e., as locations. There have also been attempts at developing grander schemes of entity types. These schemes try to cover everything that could be taken as a named entity and organise them neatly. For example, Sekine and Nobata (2004) defined a set of some 200 different types to support question answering and information extraction on newspaper text.

2.2 Swedish resources

2.2.1 SUC

The first larger gold standard for named entities in Swedish text was the Stockholm-Umeå corpus (SUC), which was supplied with named-entity annotation for its second version (Gustafsson-Capková and Hartmann 2006). In the most recent version, SUC3.0, the annotation has been checked further. Formally, named entities are marked using the start tag <name> and its corresponding end tag </name> with the first carrying an attribute, *type*, to indicate the entity type. The types used are: person, animal, myth (ological entity), place, inst (itutional entity), product, work (of art), event, and other.

In addition, numbers are identified as a separate kind of entity-referring expressions. The distribution is uneven over the categories with numbers having the most (18098), and events the fewest (245).

SUC2.0 was used by Salomonsson, Marinov and Nugues (2012) to build a four-split system, where the categories animal, myth, inst, product, event and other were merged into a miscellaneous category.

2.2.2 NomenNescio

A joint Nordic project developed a common framework for NERC on Scandinavian languages using six categories: PRS (Person), LOC (Location), ORG (Organization), EVT (Event), WRK (Work of Art), and OTH (Other) (Johannessen et al. 2005). The project compared and evaluated several methods, both manually and automatically on available gold standards. A conclusion of the project was the importance of gazetteers for achieving good performance.

2.2.3 SweNER

In the context of the NomenNescio project, Kokkinakis (2004) developed a NERC-system for a comprehensive taxonomy of types with eight top level types and altogether 47 subtypes. The top categories were: location, person, organisation, event, object, work and art, time, and measure. The object category covers products of various kinds but also prizes and, along with medical products, also names of diseases and genes. It is kept separate from the work and art category which, apart from works of creation also covers such products as newspapers.

The system was evaluated on a dataset of edited texts from different genres including newspaper texts of various kinds and excerpts from literature. It is not said whether these texts were different from the texts that the system was developed on. However, the evaluation set was large with more than 2000 tokens being parts of names. The evaluation was performed on a token basis with an average precision of 0.9422 on all types. Surprisingly, including the subtypes in the evaluation decreased the results with only 0.7%.

The SweNER system of 2004 were largely based on rules and large lists of relevant names and multiword phrases. It has later been developed and reimplemented for different tasks (e.g., Borin and Kokkinakis 2010). A major reimplementation is the HFST-SweNER which used the same eight categories as the previous system, but an enlarged set of subtypes (Kokkinakis et al. 2014). This time the system was evaluated on the SUC3.0 gold standard. However, due to the fact that the categories are not always one-to-one, some measures of harmonisation and re-mapping were needed. Although it could be shown that the output from HFST-SweNER overlapped with that of SweNER with only minor differences (1-2% of tokens), the performance this time was much poorer with an average precision of 79.02% and average recall at 70.56%.

The web service for named-entity recognition (Sparv) provided by Språkbanken Text is based on SweNER providing the eight top level entity types and several subtypes.

2.2.4 Special genres

Ek et al. (2011) developed a NERC system for Short Text Messages that ran on a mobile platform. They used the following entities: locations, persons, dates, times, and telephone numbers. As part of the project a corpus was built consisting of some 4,500 text messages and about 60,000 tokens. They used the IOB2 format from Sang (2002) for annotation yielding 11 tags to choose from, two for each entity type and 'O' for all tokens not belonging to any of the five types.

Three systems were evaluated: one based on regular expressions, a second using feature-based classification, and an ensemble system from the two. Evaluation was performed, as in the CoNLL shared tasks, by comparing tags proposed by a system with gold standard tags in two modes, strict and partial F-score. They showed that the classifier was slightly better, in particular in finding the start tags of locations and persons. They also showed the usefulness of gazetteers and other kinds of lists.

There have also been studies on Swedish for the medical domain, using annotated patient health records. These corpora are not public, though, for obvious reasons. The Stockholm EPR corpus has been annotated in several iterations with the number of categories ranging from 28 to eight (Velupillai et al. 2009; Henriksson, Dalianis and Kowalski 2014). Skeppstedt et al. (2014) used the same corpus for a project using four types: disorder, finding, drug and body structure.

Another (closed) medical corpus has been generated from the system of electronic health records at Sahlgrenska University Hospital in Göteborg (Kokkinakis and Thurin 2007). A small part of this corpus was used to evaluate the prospects of adapting the generic SweNER system for the purpose of de-identifying hospital discharge letters. For this purpose seven categories were tested: persons, locations, organisations, drugs, diseases, time and measure expressions.

2.2.5 Relation to existing Swedish resources

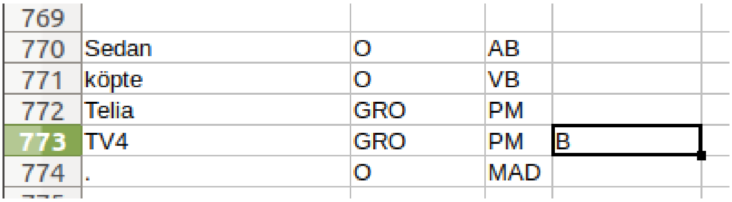
The choice of entity types in our project is most similar to that of SweNER. The main differences are that we include two types from the clinical domain and don't employ sub-typing. Sub-types may be added in future versions, however. We have also collapsed the two categories 'objects' and 'work and art' into one, at the same time excluding some marginal subtypes of these categories – the details can be found below in the guidelines for the category WRK. Finally, we have not included measures, a type for which existing systems usually show high performance.

3 General Guidelines

3.1 Annotation format

Each token of a text is given a tag. In case a token is not itself a name nor part of a phrase naming an entity it will carry the tag ’O’. For the categories we use short abbreviations (those in Table 1). To support fast annotation they all start with a distinct first letter.

A tag sequence of the same tag may occasionally cover more than one instance of an entity, for example when a direct object follows a subject as in *Sedan köpte Telia TV4.* (’Then Telia bought TV4’). Both tokens *Telia* and *TV4* will then be marked GRO, but as they refer to different entities they need to be kept apart. This is done by adding an extra B, for Beginning, in a separate column. See figure 2.



769				
770	Sedan	O	AB	
771	köpte	O	VB	
772	Telia	GRO	PM	
773	TV4	GRO	PM	B
774	.	O	MAD	

Figure 2: Annotation when two instances of the same type are referred to by adjacent tokens.

3.2 General guidelines

The basic principle is that a token should be marked with a tag other than ’O’ if and only if it is part of a name-like phrase that refers to an entity of the eight selected types. The description of a name-like phrase varies with the type. The definition of a type is primarily semantic: what kind of entity it is referring to in the context where it occurs.

Note the following:

- 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*, *hon*, and deictic adverbs such as *då*, *här*, should as a rule be marked ’O’. Possible exceptions can be found with WRKs and TMEs.
- Verbs are generally marked ’O’. Participles, however, may be part of a naming phrase.
- Tokenization must not be changed by annotators, even if unorthodox. This is because NERC systems are often applied to text which has been tokenised automatically.

To keep things simple we don’t allow a token to carry more than one tag (cf. Chinchor, 1999). If an annotator is uncertain about the choice of tag, he or she can add a comment and bring the matter up for discussion. Discussions cannot solve all problematic cases, however, so ultimately hard cases of conflict are solved by defaults.

Genitive forms are marked in the same way as nominative forms. Thus, in a phrase such as *Olles hund*, *Olles* is marked PRS.

If a name of a certain category is **part of a longer expression**, which can be analysed as a named entity of a different kind, the label appropriate for the longer phrase should be used. Example: *Uppsala universitet* should be marked ’GRO GRO’, not ’LOC GRO’. However, if this category is not one of the eight categories that are part of this work, the name should not be marked at all. Example: *Halleys komet* (as as reference to a celestial body).

An exception to the above rule are **Genitive attributes**. They are treated on a par with prepositional attributes as ’external’ to the naming expression. Thus a phrase such as ’USA:s förre president Bill Clinton’ is analysed as a sequence of two entity-referring phrases: ’USA:s/GRO förre/PRS president/PRS Bill/PRS Clinton/PRS’.

Also, we generally don’t mark tokens where a name is part of a longer token, such as **compounds**. Thus, *Stockholms-syndromet*, *göteborgare*, or *Danmark-*

sresa are tagged ’O’. This rule applies also in the case where a compound has been split erroneously as in *kramp kännedom*, where *kramp* otherwise had been tagged SMP. We make exceptions for compounds where the last part is a classifier for the name, as in *Zelda-spel*, *True Blood-box*. These compounds act as a kind of synonyms for the name.

When names are **coordinated**, there are two alternatives. Consider *Anna och Erik Eriksson* and *AMI-Hammarby och Södertälje*. If the coordination is taken to refer to two different entities, the conjunction should not be included; however if it is taken to refer to a single entity, the conjunction should be included as part of the name. In these examples the annotation is likely to be *Anna/PRS och/O Erik/PRS Eriksson/PRS*, *AMI-Hammarby/GRO, och/GRO, Södertälje/GRO*. Coordination with other conjunctions such as *eller* or the slash, /, are treated similarly.

Phrases where the head has undergone **ellipsis** should be marked: *inte/O denna/TME vecka/TME men/O nästa/TME*.

Phrases with **misspelled words** should also be marked: *nästa/TME vekca/TME*.

Phrases in **English or other foreign languages** that occur naturally within a Swedish text should be marked: *Take/O the/WRK Corvette/WRK*.

3.3 Guidelines for specific entity types

It is generally agreed that the demands of a given task are what should define the relevant entity types and their standard referring expressions. In this project, we don’t have such a task to base our definitions on. Nevertheless, we need to define the types and the expressions somehow. We then appeal to two dimensions: semantics for defining the entity type, and pragmatics for separating standard referring expressions for an entity from descriptions of an entity. Encyclopedias, including Wikipedia, may be consulted to decide whether a referring expression is standard enough to be regarded as such.

Each entity type is provided with (a) a short description, (b) an enumeration of different positive examples of the type, (c) an enumeration of negative examples: related expressions that should not be marked by this category, and, (d) a listing of cases of potential conflict with other types, and how they should be resolved. In the following chapters information on conflicts is duplicated; this is done so that all specific information that pertains to a given type can be found in one chapter.

4 Persons (PRS)

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 tokens should be marked:

1. Proper names referring to a person, either by itself or as part of a longer sequence. Examples: *Peter*, *Maria Eriksson*
2. Plural references should be marked: *Svenssons*, *familjen Lundgren*.
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. Example: *apotekare Lundin*, *morbror Ernst*, *den norske pianisten Leif Ove Andsnes*, *medborgare Vreeswijk*.
4. Initials and prepositions should be included as part of a name: *John A Ericson*, *Björn af Kleen*. Initials that appear on their own should be annotated when they abbreviate a name: *L/PRS kom/O hit/O*.
5. Due to faulty tokenisation a full stop belonging to an initial may appear on its own. If so, it should be marked PRS: *Ulysses/PRS S/PRS ./PRS Grant/PRS*.
6. Nick-names 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"*, in the second case as two different name expressions. The citation marks should be annotated even if they have been separated as single tokens.
7. When names are listed as part of a group that have accomplished something together, they should be marked as one instance, including any in-

tervening commas and conjunctions: *Hansson/PRS &/PRS Karlsson/PRS ,/O Monument/WRK*.

The following tokens should not be marked:

1. References based on a family role: *mamma*, *pappa*, *brorsan*, *hennes pojkvän*
2. Common references based on an attribute of a person such as *lillan* or *lillfian*, unless it is clearly established as a nick-name.
3. Prepositions preceding a name reference of a person should not be marked. *till/O Maja/PRS*

The following are common conflicts of PRS with other categories:

- **PRS :: GRO** The Person category may sometimes be hard to distinguish from the category Organisation when the reference is to a pop group, an orchestra, or a theater company. When it is possible to view the phrase as referring to people that are performing, the category should be PRS. When the reference is to an institution or company behind the group, the category should be GRO.

  - *Jag lyssnar på ABBA/PRS.*
  - *ABBA/PRS skapades 1971.*
  - *Dramaten/GRO sätter upp pjäs i Motala*

  If several labels seem appropriate, use PRS as default.

- **PRS :: WRK** It may also touch on the category WRK as when an artist performs under a special label such as *Prince of Assyria*, or if a person is represented in a statue. A statue is WRK, but a performing person is PRS. As a general rule people are PRS by default.

5 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 tokens should be marked:

1. Proper nouns referring to an entity of these kinds should be marked. Examples: *Stockholm*, *Vasaparken*, *Kungsgatan 24*, *Europa*
2. In cases when a proper name is preceded by an article or possessive pronoun, that also should be marked: *mitt Stockholm*
3. Common nouns referring to locations can be marked when they have developed the character of a standard, namelike reference, as in *Gamla stan*, *Östergötlands län*, or *Norrlands inland*.

The following tokens should **not** be marked LOC:

1. Location names that are a part of a name expression for another category. Example: *Norrlands* in *Norrlands Guld*, a beer brand, is not marked as LOC but WRK.
2. Location names are common in postal addresses. If the expression is clearly that of a postal address, a category not recognized in this project, use O as the annotation. Example: *P.O./O Box/O 1323/O X-stad/O*.
3. Indexical references using adverbs or common nouns such as *hemma*, *i utlandet*, should not be marked.
4. Prepositions preceding a location name should not be marked.

5. Common names of rooms such as *köket*, *vardagsrummet* should not be marked.

6. Web sites such as Blocket, Facebook are often linguistically treated as locations ("Jag var på fb hela förmiddagen") but should be analysed as organisations (GRO).

7. URLs should never be marked LOC. However, names of companies may sometimes look like URLs.

Conflicts of LOC with other categories:

- **LOC :: GRO** case 1. Organisations usually have offices or headquarters which may serve as landmarks. Thus, there may be a conflict for LOC with GRO. 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*.

  Vi träffades utanför Åhlens. (LOC, answers the question 'Var träffades vi')
  Åhlens drar ner på personal. (GRO, 'vilka/*vart drar ner på personal?').

- **LOC :: GRO** case 2. Location names are commonly used as metonyms for organisations, as in *Sverige spelade oavgjort mot Slovakien* or *Moskva avvisar alla anklagelser*. Johannessen et al. (2005) discuss two principles to handle this kind of metonymy in systems: *Form over function*, or *Function over form*. The first means that the intuitively most basic category of a proper noun, findable in a gazetteer, is used, while the other principle prefers to use a label appropriate for the intended referent. We follow the latter principle as we generally appeal to semantics for demarcations. In particular, if the referent is performing some kind of action such as playing, supporting, deciding, denying etc. it should be analysed as GRO. In the example sentences above, the proper nouns should all be tagged GRO, not LOC. However, if both views are possible, LOC is the default for place names.

- **LOC :: WRK** Statues, buildings and other may also be used as metonymic for their locations. A question may help also in this case. Example:

  Östra Kungsgatubron försvann (WRK, 'vad/*var försvann?').

  When both views are possible, use LOC as default.

6 Organisations (GRO)

This category includes companies, governments, political parties and NGOs, public bodies, sports clubs, schools, hospitals and generally anything with a legal status in a society.

The following tokens should be marked:

1. Proper nouns and acronyms referring to entities of this category: *Ericsson*, *ABB*, *St Görans*
2. Common nouns that have established themselves as names: *Företagarna*, *Socialdemokraterna*, *Änglarna* (as a football team), *Svenska akademin*. Also, occasionally nominalized adjectives: *de vita*.
3. Common nouns or abbreviations that pick out a societal institution such as *BVC*, *vårdcentralen*, *Riksdagen*.
4. When a name and an abbreviation occur together they should both be marked, but as separate references. Example: *Socialstyrelsen/GRO (/O SoS/GRO)/O*.
5. A proper noun referring to an organisation is sometimes preceded or followed by a common noun. These cases are treated as we do with persons: any attribute preceding the name should be as part of the name: *Apoteket Uttern*, *den svenska telekomjätten Ericsson*, *St Görans sjukhus*.
6. If a company is named as a web address, like *flygresor.se*, it should be marked GRO.

The following tokens should not be marked:

1. References to organisations are often restricted by its location. Unless the location is clearly part of the name for the organisation it should be marked separately: *Maxi/GRO i/O Södertälje/LOC*.

2. Prepositions preceding an organisation name should not be marked.
3. Collective descriptive references to organisations or their memberships should not be marked. Examples: *NATO-medlemmar*, *de röda trupperna*.
4. URLs should generally not be marked as GRO, only in case 6 above if it is used as the name of a company.
5. Projects are normally considered WRK, not GRO.

Conflicts of GRO with other categories:

- **GRO :: LOC** case 1. Organisations usually have offices or headquarters which may serve as landmarks and the same holds for structures such as bridges or statues. Thus, there may be a conflict for LOC with GRO. 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*.

  Vi träffades utanför Åhlens. (LOC, answers the question 'Var träffades vi')
  Åhlens drar ner på personal. (GRO, 'vilka/*vart drar ner på personal?').

- **GRO :: LOC** case 2. Location names are sometimes used as metonyms for organisations, as in Moskva avvisar alla anklagelser, or Sverige spelade oavgjort mot Slovakien. (Johannessen et al. 2005) discuss two principles to handle this kind of metonymy in systems: Form over function, or Function over form. The first means that the intuitively most basic category of a proper noun, findable in a gazetteer, is used, while the other principle prefers to use a label appropriate for the intended referent. We propose to follow the latter principle; thus, rely on the semantics. In particular, if the referent is performing some kind of action such as playing, supporting, deciding, denying etc. it should be analysed as GRO. In the example sentences above, the proper nouns should all be tagged GRO, not LOC. However, if both views are possible, LOC is the default for place names.

- **GRO :: PRS** GRO may sometimes be hard to distinguish from the category Person when the reference is to a pop group, an orchestra, or a theater company. When it is possible to view the phrase as referring to people that are performing, the category should be PRS. When the reference is to an institution or company behind the group, the category should be GRO.

Jag lyssnar på ABBA/PRS.

ABBA/PRS skapades 1971.

Dramaten/GRO sätter upp pjäs i Motala.

If several labels seem appropriate, use PRS as default.

- **GRO :: 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. When the context does not make it clear, use the category which seems most appropriate for the name. Examples:

  Jag föredrar Volvo. If the preference concerns cars use WRK, but if it concerns shares, use GRO. If the context does not make it clear use what is the standard referent for the name in isolation. Here probably GRO.

  Om Netflix/WRK används i diverse olika Mediacenters som t.ex Xbox 360.
  But: *Nu har Netflix/GRO sagt att serien ska få en varningstext.*

7 WORKS OF ART AND OTHER ARTEFACTS (WRK)

This category includes name or title references to works of art, such as books, films, plays, brand names of commercial products such as cars or toothpaste, newspapers and journals, names of software programs and cooperative undertakings such as projects. It covers the two categories WRK and OBJ in the taxonomy used by Kokkinakis (2004) with the exception of common names of plants and flowers, which are regarded as natural kinds and annotated O.

The following tokens should be marked:

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

2. Noun phrases used as product names or titles: *Dagens Nyheter*, *The Exorcist*, *Mitt liv som hund*

3. 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: *Härifrån/WRK till/WRK evigheten/WRK*

4. Scientific papers and books belong in this class. Tokens belonging to the title should be marked WRK, whereas authors, if mentioned, should be marked PRS, and publishers, if mentioned, should be marked GRO. Example: *Joakim/PRS Nivre/PRS ,/O Inductive/WRK Dependency/WRK Parsing/WRK ,/O Springer/GRO 2006/TME*

5. If a journal is identified with references to volume and issue no, these should be annotated as well, including any intervening parentheses or commas.

6. A product or work of art using the name of another category such as person or location is to be marked WRK: *Chicago* (the musical), *Manhattan* (the film)

7. A type reference accompanying a brand or model name should be marked as part of the reference. *Mercedes/WRK A40/WRK automat/WRK*

8. If the name heads a noun phrase, include all words of the noun phrase up to a following prepositional phrase: *en/WRK ny/WRK Mazda/WRK*. However, don't make the phrase longer than necessary, if the name occurs in an apposition: *min/O bil/O, en/WRK volvo/WRK från/O 2009/TME*.

The following tokens should not be marked:

1. Names of natural kinds such as *potatis, gullviva, abborre* should not be marked. In contrast, *Bintje* is a developed product and should be marked.

2. Types of products should not be marked. Thus if a car is described solely in terms of a type reference such as *en automat*, *en diesel*, the annotation should be O.

3. People should not be marked WRK but PRS.

4. A property expressed in a prepositional phrase should generally not be marked. Example: *volvon/WRK i/O artikeln/O*.

Conflicts of WRK with other categories

- **WRK :: LOC** Statues, buildings and other may also be used as metonymic for their locations. A question may help also in this case. Example:

  Östra Kungsgatubron försvann (WRK, 'vad/*var försvann?').
  When both views are possible, use LOC as default.

- **WRK :: MNT** A medicine is a product but also a kind of treatment. If so, the reference should be annotated as MNT.

- **WRK :: PRS** WRK may touch on the category PRS when an artist performs under a special label such as *Prince of Assyria*, or if a person is represented in a statue. A statue is WRK, but a performing person is PRS. As a general rule people should not be marked WRK.

- **WRK :: GRO** 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. When the context does not make it clear, use the category which seems most appropriate for the name. Examples:

Jag föredrar Volvo. If the preference concerns cars use WRK, but if it concerns shares, use GRO. If the context does not make it clear use what is the standard referent for the name in isolation. Here probably GRO.

Om Netflix används i diverse olika Mediacenters som t.ex Xbox 360. (WRK) But: *Nu har Netflix sagt att serien ska få en varningstext.* (GRO)

- **WRK :: EVN** A show or performance may be seen as an event, in particular if it is popular. But it may also be seen as a work of art. Example: *Idol* (a TV show). If, as a TV-viewer you describe yourself as taking part of the show, the label EVN is appropriate. Otherwise, use WRK as a default label.

8 Temporal entities (TME)

Temporal entities in this project cover time points and continuous intervals on a presumed timeline from the beginning of time to the present and including the future. We mark phrases that specify a temporal entity of this kind, whether absolutely or deictically, provided it is specific enough.

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

The following tokens should be marked:

1. Standard references to times of the day, dates, weeks, months, years, seasons, decades, centuries. These usually employ numerals or nouns: *2 april 1991*, *tisdagen den 30 januari 2018*, *1960-talet*, *1800-talet*, *2018-01-30*, *november 1989*, *1970-71*, *kl. 19.30*, *kvart över sex*

2. Special names for holidays such as *Påskdagen*, *Nyårsafton*, when the temporal reference is prominent.

3. Deictic references related to the current speech-time with a nominal head word such as *i morgon*, *i sommar*, *nästa vecka*, *förra månaden*, *denna vecka*, *för ett år sen*, *på tisdag*, *om tre år*, *i morse*, *i förrgår kväll*, and even items such as *igår, idag, i fjol, i natt* that are adverbial-like but still could be seen as PPs with a nominal head.

4. Vague references are included if they have a nominal head word of a temporal entity: *om ett par timmar*, *för några veckor sedan*.

5. If the temporal phrase includes a preposition, determiner or adjective,

these should normally also be marked: *på/TME torsdag/TME*. 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: *från/O och/O med/O i/TME morgon/TME*, or *sedan/O 1995/TME*.

The following tokens should not be marked:

1. Deictic adverbs such as *nu, då, senare, tidigare, samtidigt, nyss*
2. Phrases of any kind expressing duration: *länge, i två timmar*
3. Phrases of any kind expressing age: *tre år gammal*
4. Phrases of any kind expressing frequency: *ofta, varje dag, på torsdagar, på kvällarna, två gånger om året, varje vecka*. **Note** though that frequency may be included in a symptom or treatment, if it is an important aspect of it. But then it should be marked SMP, not TME. Example: *två/SMP timmar/SMP mellan/SMP kräkningar/SMP*.
5. Phrases implying a reference point other than the current speech-time: *tre år senare, efteråt*
6. If a temporal reference is written as part of a token with quite a different meaning it should not be marked: *född-58*
7. Phrases with vague indeterminate nouns such as *om ett tag, om en stund, för länge sedan*

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. Often they can be separated: *OS 2014* should be marked *OS/EVN 2014/TME*. Conversely, a temporal reference may use an event reference as a part: *innan jul*, *under Andra världskriget*. Quite often both interpretations seem to be present A question test can sometimes be used. If the phrase answers the question 'När?', it should be marked TME, if not it is likely to be EVN. Examples: *Julen/EVN närmar sig.* (Answers the question: Vad närmar sig?) *Vi måste vara färdiga innan/TME jul/TME.* (Answers the question: När måste vi vara färdiga?)

  If both options are possible, use EVN as default for holidays that imply some kind of celebration: *På/O julen/EVN ska vi vara glada.*

9 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 and conferences, sports competitions and events of a religious nature and holidays. However, we do not provide labels for the sub events, so EVN is used for all of them.

The following tokens should be marked:

1. Historical or political events, such as battles, wars, scandals, campaigns and crimes. Example: *Andra* världskriget
2. Weather phenomena and natural disasters such as hurricanes and storms: *stormen Gudrun*
3. Cultural events, like festivals and fairs: *Peace and Love*, *mello*
4. Religious events, like holiday celebrations: *Påsk*, *Julafton*
5. Sports events such as conferences and world championships: *Olympiska spelen*

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. Examples: *kalabaliken i Bender* may be considered to be an established name, so that all three tokens are labelled EVN. The Olympic Games are recurring events and so *OS i Sotji* may preferably be labelled *OS/EVN i/O Sotji/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. Often they can be separated: *OS 2014* should be marked *OS/EVN 2014/TME*. Conversely, a temporal reference may use an event reference as a part: *innan jul, under Andra världskriget*. Quite often both interpretations seem to be present. A question test can sometimes be used. If the phrase answers the question 'När?', it should be marked TME, if not it is likely to be EVN. Examples:

  Julen/EVN närmar sig. (Answers the question: Vad närmar sig?)
  Vi måste vara färdiga innan/TME jul/TME. (Answers the question: När måste vi vara färdiga?)

  If both options are possible, use EVN as default for holidays that imply some kind of celebration: *På/O julen/EVN ska vi vara glada.*

- **EVN :: WRK** A show or performance may be seen as an event, in particular if it is popular. But it may also be seen as a work of art. Example: *Idol* (a TV show). If, as a TV-viewer you describe yourself as taking part of the show, the label EVN is appropriate. Otherwise, use WRK as a default label.

10 Symptoms (SMP)

The guidelines for this category are modelled 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, as in an article of a medical journal.

The following tokens should be marked:

1. Noun phrases that name a disease (*lunginflammation*), syndromes or abnormal states (*ångestattacker*, *hosta*, *bruten arm*), specific virus or bacteria (*HI-virus*), or indicative test results (*lågt blodtryck*).

2. Adjectival phrases, including participles, that do the same such as *förvirrad*, *nedstämd*, *mycket ont*. Note that a preceding verb is not marked SMP: *har/O ont/SMP*.

The following tokens should not be marked:

1. **NB!** Verbs should never be marked even though they indicate a problem: *blöder/O mycket/O*. Also: *har/O svåra/SMP smärtor/SMP*.

2. General words such as *sjukdom, sjuk, krasslig, virus*.

3. Words such as *bra* or *normalt* should not be marked even if referring to bodily phenomena such as blood pressure.

4. While negations of normality may indicate a problem, it should not be marked: *inte/O bra/O*.
5. States that are the result of normal, everyday activities should not be marked: *blev/O trött/O*.
6. Measurements, even if they can be inferred to be outside normal range: *blodtryck/O 160/100/O*
7. Naturally occurring states or phases that are not to be regarded as diseases or illnesses: *pubertet/O, trotsålder/O, gravid/O*
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: *Jag städade undan virus/O och kräklukt/O*.

Conflicts of SMP with other categories:

- **SMP :: TME** A symptom may be particularly serious if it is recurring. Thus just as we would annotate *frekventa/SMP kräkningar/SMP*, we also annotate *två/SMP timmar/SMP mellan/SMP kräkningar/SMP*.

11 Treatments (MNT)

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.

The following tokens should be marked:

1. Noun phrases that refer to medications, (*Simvastatin, 40 mg*), biological substances (*blodtransfusion*), hardware (*respiratorbehandling, kateter*) and general terms (*terapin, astmamedicinen*) used for treatments.
2. Preventive measures if prescribed by doctors or an organisation: *vaccinering, trippelvaccin*
3. General terms referring to a patient's treatments: *hennes medicinering*
4. Noun phrases that refer to substances that are not usually used as medications, if they are clearly part of a treatment, t.ex. *fick/O filmjölk/MNT på/O recept/O*.
5. Adjective phrases that do the same (though they seem to be rare)

The following tokens should not be marked:

1. Precautionary measures that are not prescribed by a doctor but occur regularly in everyday life: *använder regelbundet solskyddskräm/O*, *måste få vila/O nu*.
2. **NB** Verbs should never be marked. Example: *ge/O morfin/MNT var/MNT fjärde/MNT timme/MNT*.

3. Phrases referring to tests used in order to diagnose a patient: *ultraljudsundersökning, hjärtröntgen, urinprovet*

4. Treatments used as metonyms for locations: *patienten/O hänvisades/O till/O reumatologin/O*

Conflicts of MNT with other categories:

- **MNT :: WRK** A medicine is a product but also a kind of treatment. If so, the reference should be annotated as MNT.

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.

Baldwin, Timothy, Marie Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter and Wei Xu. 2015. Shared tasks of the 2015 workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. *Proceedings of the workshop on noisy user-generated text*, 126–135. Beijing, China: Association for Computational Linguistics.

Borin, Lars and Dimitrios Kokkinakis. 2010. Literary onomastics and language technology. *Literary education and digital learning*, 53–78. Information Science Reference.

Chinchor, Nancy. 1997. MUC-7 Named Entity Task Definition.

Ek, Tobias, Camilla Kirkegaard, Håkan Jonsson and Pierre Nugues. 2011. Named entity recognition for short text messages. *Procedia - Social and Behavioral Sciences* 27: 178–187.

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

Grishman, Ralph. 1995. Named Entity Task Definition. Version 2.0 (31 May 1995).

Grishman, Ralph and Beth Sundheim. 1996. Design of the MUC-6 Evaluation.

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.

Henriksson, Aron, Hercules Dalianis and Stewart Kowalski. 2014. Generating features for named entity recognition by learning prototypes in semantic space: The case of de-identifying health records. *Ieee international conference on bioinformatics and biomedicine (bibm)*.

Ingólfsdóttir, Svanhvít Lilja, Sigurjón Þorsteinsson and Hrafn Loftsson. 2019.

Towards high accuracy named entity recognition for Icelandic. *Proceedings of the 22nd nordic conference on computational linguistics*, 363–369. Turku, Finland: Linköping University Electronic Press.

Johannessen, Janni Bondi, Kristin Hagen, Åsne Haaland, Andra Björk Jónsdottir, Anders Noklestad, 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.

Johansen, Bjarte. 2019. Named-entity recognition for Norwegian. *Proceedings of the 22nd nordic conference on computational linguistics*, 222–231. Turku, Finland: Linköping University Electronic Press.

Kim, Jin-Dong, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi and Nigel Collier. 2004. Introduction to the bio-entity recognition task at jnlpba. *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications*, JNLPBA '04, 70–75. Stroudsburg, PA, USA: Association for Computational Linguistics.

Kokkinakis, Dimitrios. 2004. Reducing the effect of name explosion. *Proceedings of the lrec workshop: Beyond named entity recognition, semantic labelling for nlp tasks. ourth language resources and evaluation conference (lrec).*

Kokkinakis, Dimitrios, Jyrki Niemi, Sam Hardwick, Krister Lindén and Lars Borin. 2014. HFST-SweNER — a new NER resource for Swedish. *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, 2537–2543. Reykjavik, Iceland: European Language Resources Association (ELRA).

Kokkinakis, Dimitrios and Anders Thurin. 2007. Identification of entity references in hospital discharge letters. *Proceedings of the 16th nordic conference of computational linguistics (nodalida)*, 329–332. Tartu, Estonia.

Nadeau, David and Satoshi Sekine. 2007. A survey of named entity recognition and classification. *Lingvisticæ Investigationes*, vol. 30:1.

Salomonsson, Andreas, Svetoslav Marinov and Pierre Nugues. 2012. Identification of entities in swedish. *Sltc 2012 : The fourth swedish language technology conference*, 63–64. SLTC.

Sang, Erik, F. Tjong Kim. 2002. Introduction to the conll-2002 shared task: Language-independent named entity recognition. *The 6th conference on natural language learning 2002 (conll-2002)*, COLING-02.

Sang, Erik, F. Tjong Kim. 2003. Introduction to the conll-2003 shared task:

Language-independent named entity recognition. *Proceedings of the 7th conference on natural language learning at hlt-naacl 2003*, 141–47.

Segura-Bedmar, Isabel, Paloma Martínez and María Herrero-Zazo. 2013. SemEval-2013 task 9 : Extraction of drug-drug interactions from biomedical texts (DDIExtraction 2013). *Second joint conference on lexical and computational semantics (*SEM), volume 2: Proceedings of the seventh international workshop on semantic evaluation (SemEval 2013)*, 341–350. Atlanta, Georgia, USA: Association for Computational Linguistics.

Sekine, Satoshi and Chikashi Nobata. 2004. Definition, dictionaries and tagger for extended named entity hierarchy. *Proceedings of the fourth international conference on language resources and evaluation (LREC'04)*. Lisbon, Portugal: European Language Resources Association (ELRA).

Skeppstedt, Maria, Maria Kvist, Gunnar H. Nilsson and Hercules Dalianis. 2014. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: An annotation and machine learning study. *Journal of Biomedical Informatics* 49: 148–158.

Uzuner, Özlem, Brett R South, Shuying Shen and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *JAMIA* 18(5): 552–556.

Velupillai, Sumithra, Hercules Dalianis, Martin Hassel and Gunnar H. Nilsson. 2009. Developing a standard for de-identifying electronic patient records written in swedish: precision, recall and f-measure in a manual and computerized annotation trial. *International Journal of Medical Informatics* 78(12) (December): e19–e26.

Yadav, Vikas and Steven Bethard. 2019. A survey on recent advances in named entity recognition from deep learning models. *arXiv.org*, no. 1910.11470. https://arxiv.org/abs/1910.11470.