HUMBOLDT-UNIVERSITÄT · ZU BERLIN ·

Guidelines for Full Text Annotations in the SoNAR (IDH) Corpus

Sina Menzel, Josefine Zinck & Vivien Petras

•

Berlin School of Library and Information Science
Humboldt-Universität zu Berlin

Abstract:

This document presents guidelines for the manual annotations made on a representative sample of the full text corpus in the SoNAR (IDH) project (www.sonar.fh-potsdam.de).

Contents

1. Introduction ... 3
2. Data set ... 3
2 Annotation environment ... 4
3 Quality management ... 4
4 Annotation of named entities ... 4
5.1 General annotation rules ... 4
5.2 Person (PER) ... 6
5.3 Organization (ORG) ... 7
5.4 Location (LOC) ... 8
5.5 Conference (CONF) ... 8
5.6 Event (EVT) ... 9
5.7 Works and expressions (WORK) ... 9
6. Excluded full text sections ... 9
7. Full text correction ... 9
8. References ... 12
Appendix A: Example tag-set ... 13
Appendix B: List of exceptions ... 16

1. Introduction

The following guidelines present the detailed ruleset and specifications for the manual annotation of named entities (NE) within a representative sample of the full text corpus in the SoNAR (IDH) project. "Annotation" means the manual tagging of appearances of predetermined semantic units within a text and is therefore an enrichment of full texts with metadata. Section 2 introduces the corresponding full text corpus, section 3 and 4 specify the environment and quality management for the annotation process.

The semantic units of interest for the SoNAR (IDH) full text corpus annotation are named entities of the following classes:

Persons (PER),
Organizations (ORG),
Locations (LOC),
Conferences (CONF),
Events (EVT),
Works and expressions (WORK).

Section 5 defines these classes and the corresponding annotation rules in detail, and section 6 lists exeptions from these rules.

The ruleset of the present guidelines is being developed iteratively along with the ongoing annotation (please see version no. above). The purpose of the guidelines is to secure consistency and coherence in the annotation process, in order to achieve optimal quality of the annotation's outcome: The gold standard that supports the evaluation of an automated process of named entity recognition (NER) in the realm of the project. The guidelines build upon former work (Fort et al. 2009; Rosset et al. 2011; Reznicek 2013; Reiter 2017) as well as the German Integrated Authority File (GND) hosted by the German National Library (DNB). The latter will be the main knowledge base used for named entity linking (NEL) and therefore serves as orientation for ambiguous cases.

In a broader sense, annotation includes adjustments of the original text, such as character correction as described in section 7.

2. Data set

The complete dataset of full texts in the SoNAR (IDH) annotation process consists of 2,123,393 historical German text documents derived from the *Zeitungsinformationssystem* repository (ZEFYS), hosted by the Berlin State Library. The documents are newspaper pages from the following periodicals (late 19th and early 20th century, see table 1).

Title	Time span	# of documents	Shares in %
Berliner Börsenzeitung	1872-1931	642,480	30.26
Berliner Tageblatt	1877-1939	489,983	23.08
Berliner Volkszeitung	1890-1930	142,403	6.71
Deutsches Nachrichtenbüro	1936-1940	7,429	0.35
Neueste Mittheilungen	1882-1894	1322	0.06
Norddeutsche Allgemeine Zeitung	1878-1930	165,622	7.80
Provinzial Correspondenz	1863-1884	1090	0.05
Teltower Kreisblat	1856-1896	25822	1.22
Vossische Zeitung	1857-1917	647,242	30.48

Table 1: Newspapers in the data set for SoNAR (IDH) full text annotations.

From this data set, a representative subset is derived, which is manually annotated over the course of the project.

3. Annotation environment

The annotations are made browser based by a single human annotator with the project's in-house-tool named *neath*[1] (named entity annotation tool in html). *Neath* is adjusted iteratively along the annotation process to any necessities that might occur due to specifics of the textual content.

4. Quality management

The quality of the annotation is secured by the present guidelines as well as sample checks of the annotated texts by the co-annotation of sample documents, which allows to take agreement measures. The latter is expected to bring forward disagreement cases that show loopholes in the guidelines. Additionally, we introduced the "TODO"-tag in *neath*, which may be used for ambiguous or uncertain tokens in order to support discussion and clarification on the guidelines in regular meetings of the annotation team. After each completion of an annotated text document, a revision session by the annotator is required.

5. Annotation of named entities

> *"For […] efficient NE annotation […], it is important to focus, not on* how *to annotate, but rather on* what *to annotate […]."* – Fort et al. 2009, p. 147.

The following section defines characteristics of named entities as well as the different semantic entity classes considered in the annotation process. More examples, as well as exceptions and special cases can be found in appendix A.

5.1 General annotation rules

The following rules are partly extracted from Reznicek 2013, p. 2ff.

1. The value of precision is favored over recall in the annotation process. For this reason, ambiguous cases are not marked as named entities, but with the label "TODO" for discussion in the annotation team. Should a suspected NE not be decodeable by the annotation team (e.g. due to the historical origin of the corpus), it is not to be annotated.

[1] https://github.com/qurator-spk/neath

Example: A suspected organization that is not known to the annotation team and not included in the GND nor on Wikipedia.

2. Named entitities occur as proper nouns, full nominal phrases, as well as derivations and abbreviations of the former.
   Example: Die [erfurter]LOC Innenstadt.

3. Pronouns are not to be marked as named entities.

4. Determiner (e.g. articles) are not part of named entities. See appendix for exceptions.
   Example: Die [Parteimitglieder]PER

5. Named entities may include at least one and up to x tokens.

6. If named entities occur in the plural, they are to be treated the same way as in the singular.

7. Named entities might occur as part of a token, e.g genitive case. In these cases, the entire token is to be labeled with the corresponding type of entity. Compounds are to be separated into several tokens, see rule no. 8.
   Example: [Frankreichs]LOC Käsevielfalt
   [Kreuzberger]LOC Nächte sind lang
   [Lisa's]PER Geburtstag

> Note: An NE as part of a token is NOT the same as an embedded NE!
> In the first case, the other parts of the token are no entities. Nevertheless, the entire token is to be annotated in cases of embedded entities, too.

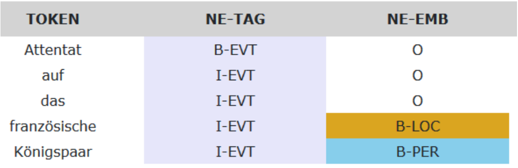
8. If named entities occur in compounds, they are to be split up into several tokens.
   Example: Die Verleihung des [Humboldt]PER-Preises

9. Named entities may be embedded in other named entities (second level NE). This might also occur in compounds, if more than one component is a separate entity.
   Example: Die [[Heinrich Böll]PER-Stiftung]ORG
   Die [[SPD]ORG-[Bundestags]ORG abgeordnete]PER

10. If one entity marks the entire (group of) token(s) while the other entity marks only parts of it/them or derives from it, the latter is the second level entity. If the order of levels is not clear, the annotator may choose, which class to mark on first and which on second level.
    Example: [Stonehenge]WORK/LOC
    (annotator decides, which is embedded)
    Die [deutsch[französische]LOC]LOC Freundschaft
    (annotator decides, which is embedded)
    [[Köln]LOC-Minden]LOC Bahnlinie
    (annotator decides, which is embedded)

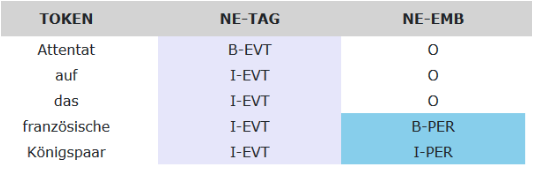
11. If more than one named entity is embedded in another named entity, the annotator chooses which entity is to be marked on second level by evaluating the nesting levels: Subject/object of the sentence is the first level entity, while its direct attribute is the second level entity. The third level component is to be left out.
    Example: Das [Attentat auf das [französische Königspaar]PERemb]EVT
    ([französische]LOC is to be left out in this case, because it refers to the second level entity PERemb)

incorrect



TOKEN	NE-TAG	NE-EMB
Attentat	B-EVT	O
auf	I-EVT	O
das	I-EVT	O
französische	I-EVT	B-LOC
Königspaar	I-EVT	B-PER

correct



TOKEN	NE-TAG	NE-EMB
Attentat	B-EVT	O
auf	I-EVT	O
das	I-EVT	O
französische	I-EVT	B-PER
Königspaar	I-EVT	I-PER

12. Enumerations of related entities are to be annotated separately. This also applies if one token does not represent the entire entity.
    Example: [Ost-]LOC und [Westdeutschland]LOC

13. Named entities may occur within metonymic references. In these cases, the referenced entity is to be annotated on the first level. If the referring token(s) may also mark a named entity, in which case they are annotated on the second level (embedded).
    Example: Der [[Kreml]ORG]LOCemb hat entschieden.

14. Any exceptions to the aforementioned rules must be agreed upon by the annotation team and will be listed in appendix A or B.

15. After the completion of the annotation of a document, there is a mandatory revision session on the same document by the annotator in order to secure the best possible gold standard.

5.2 Person (PER)

The following rules build upon Rosset et al. 2011, p. 21 and Reznicek 2013, p. 6. For orientation in ambiguous cases: Guidelines of the GND and RDA-Toolkit (Section 9, 10). See appendix A for a complete list of subclasses.

1. Named entities referring to definite individuals may be classified as "person" with the label "PER".

2. The label may also be given to tokens referring to families.
   Example: Die Intrigen der [Borgia]PER

Note: Bands are considered organizations, see rule no. 3 in section 5.3.

3. Descriptors referring to unambiguous, exclusive family connections with information on the person they are referring to are to be annotated as person entities. This includes the temporal context of the source, e.g. the point in time a newspaper was published.
   Examples: [Max Mustermann]PER verhielt sich genau wie [Max Mustermanns Vater]PER.
   [Max Mustermann]PER verhielt sich genau wie sein Bruder.
   (in this case, the connection is not exclusive)
   [Max Mustermann]PER verhielt sich genau wie [Max Mustermanns Frau]PER.
   [Max Mustermann]PER verhielt sich genau wie seine Ex-Frau.
   (in this case, the connection is not exclusive)

4. Populations are not to be marked as persons.
   Example: Die Amerikaner
   Sowjets

5. The PER-class may refer to first names, middle names, family names, nicknames, fictional characters, pseudonyms. Nicknames do not have to be unique, but unambiguous.

6. Titles (academic titles, titles as Mrs. or Mr., as well as titles of nobility as Sir, Madame, Duke, Dutchess, military titles as General or Lieutenant and the like) are not part of named entities. See appendix B for exceptions.

7. Job titles and functions are not to be annotated unless they describe unambiguious, exclusive positions of a definite individual in the context of the text section they appear. Descriptions of jobs (e.g. “Her Royal Highness”, “His Excellence”, “Her Majesty”) are to be treated the same way.
   Example: Die [Bundeskanzlerin]PER
   Der [englische Botschafter]PER
   (in the latter case, the article referred to the country of Turkey, which makes the position unambiguous)

   If the job title and name of the person appear together, tokens in between are to be included in the entity.
   Example: Die [Bundeskanzlerin Frau Dr. Angela Merkel]PER
   (underlined tokens are not considered entities if they appear separately or at the margin of other entities, see rule 6 in this section)
   Der [[russische]LOCemb Ministerpräsident Stolyvin]PER

8. Definite descriptions may be marked as persons in cases of referential unicity. Unclear cases are to be marked with the tag “TODO” for discussion in the annotation team.
   Exception: Unambiguous cases with referential uniqueness are to be marked as “PER”.

   Example: [The Iron Lady]PER
   [The Rock]PER

5.3 Organization (ORG)

The following rules build upon Reznicek 2013, p. 6-7 and Rosset et al. 2011, p. 29ff. For orientation in ambiguous cases: Guidelines of the GND and RDA Toolkit (Section 11). See appendix A for a complete list of subclasses of this class.

1. Named entities that refer to bodies, companies and the like are to be classified as "organization" with the corresponding label "ORG". See appendix A for a complete list of subclasses of this label.

2. ORG-entities may occur as acronyms or nominalizations.
   Example: Die [NATO]ORG
   Der [ADAC]ORG

3. Bands and similar professional collectives are not considered persons, but organizations.
   Example: Das Konzert der [Queens of the Stoneage]ORG

4. Generic descriptions in front of an ORG-entity are not to be annotated.
   Example: Firma [[G.H. Friedländer]ORG]PERemb
   Fleischereifachgeschäft [Wurstbasar]ORG

5.4 Location (LOC)

The following rules build upon Reznicek 2013, p. 7. For orientation in ambiguous cases: Guidelines of the GND and RDA-Toolkit (Section 16). See appendix A for a complete list of subclasses of this class.

1. Named entities referring to "politically or geographically defined location (cities, provinces, countries, international regions, bodies of water, mountains, etc.)" (MUC-6 task definition 1995) are to be labeled with "LOC". Locations might be fictional.
   Example: Die Tür nach [Narnia]LOC

2. If more than one location is described in one token, one of them is to be marked as embedded entity. See 5.1 no. 9 for first and second level disambiguation.
   Example: Die [[spanisch]LOCdeutsche]LOCemb Frau

3. According to rule no.6 in 5.2, definite descriptions of locations may not be labeled "LOC". Unclear cases are to be marked with the tag "TODO" for discussion in the annotation team.
   Exception: Unambiguous cases with referential uniqueness are to be marked as "LOC".
   Example: [The Big Apple]LOC

4. Locations embedded in descriptors of populations are to be marked on first level for populations are not considered named entities (see also rule no. 4 under 5.2).
   Example: Die [Amerikaner]LOC
   [Sowjets]LOC

5.5 Conference (CONF)

For orientation in ambiguous cases: Guidelines of the GND. See appendix A for a complete list of subclasses and examples.

1. Named entities referring to uniquely named gatherings of individuals on a certain pre-defined scientific topic, goal or shared purpose as well as a pre-defined ending point are to be classified as "conferences" with the label "CONF".

2. CONF-entities may occur as acronyms or nominalizations.
   Example: Der diesjährige [CLEF-Task]CONF
   (see rule no. 6 in section 5.1)
   Die [DHd]CONF

3. If a CONF-entity holds a time tag, the latter is to be marked as part of the entity.
   Example: Der [Bibliothekartag 2018]CONF

5.6 Event (EVT)

1. Named entities referring to uniquely identifiable events apart from conferences are to be tagged with the label "EVT". This class is annotated for experimental purposes and therefore does not follow a strict definition.

2. In contrast to conferences, events may be of spontaneous nature.

3. Topics of events may vary (e.g. military, political, cultural…).

5.7 Works and expressions (WORK)

The following rules build upon Rosset et al. 2011, p. 39ff. For orientation in ambiguous cases: Guidelines of the GND and RDA-Toolkit (Section 6). Please make sure to check the token in question in the current version of the GND catalogue. See appendix A for a complete list of subclasses.

1. Named entities referring to titled human creations are to be classified as works or expressions. The corresponding label is "WORK".

2. Separate parts of the works, such as acts in plays are not to be annotated.
   Example: Der zweite Akt von [Romeo und Julia]WORK.

6. Excluded full text sections

There is no exclusion of any sections in the full texts, the documents are to be completely annotated.

7. Full text correction

Since the annotation sample is based on print originals, the digitization process required the automated recognition of optical characters within the scanned documents (OCR). Under current software solutions, this process still comes with an inevitable error rate (Kugler 2018, p. 42) which might affect the recognition of named entities by the human annotator and certainly affects the recognition of named entities by current learning algorithms for automated NER (Kettunen/ Ruokolainen 2017).

The following types of errors might occur (based on Zumstein/Baierer 2016, p. 74-75):

I. Character errors
   This is the most frequent and most relevant type of error in the annotation process. It includes mistakes in the recognition of characters.

II. Segmentation errors
These errors are a special type of character error, where spaces between tokens are not recognized correctly. This leads to the incorrect splitting or merging of tokens.

III. Word errors
Word errors are character errors of full words. This frequently occurs in correlation with shifting fonts or if automated post-OCR-normalizations apply. The latter are usually based on wordlists that might disimprove individual tokens.

IV. Sectional errors
This type refers to formatting errors regarding the layout or other textual sections, e.g. sentence boundaries.

Corrections on the SoNAR annotation sample concentrate on error types I, II, and IV. They exclusively concern errors occurring in named entities. There is no OCR-correction of the entire full text! For this purpose, *neath* supports changes in the character strings as well as merging and splitting of tokens.

7.1 Marking of sentences

Since the data format in *neath* is based on the format used in the GermEval2014 Named Entity Recognition Shared Task, sentence boundaries are indicated by an empty line (position 0, see User Guide). For this reason, error type IV. is being corrected in the annotation process only if it concerns sentence boundaries.

1. Colons do not mark the beginnings of sentences.

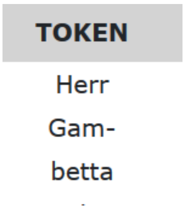
7.2 OCR-correction

1. If a token is predicted to have an error, but the corresponding word is not recognizable neither by OCR results nor by the original scan, the token is not to be corrected, but to be annotated if the type of entity is clear from the context of the sentence.

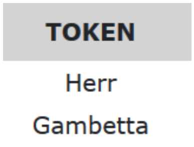
2. There is no correction of orthography due to the historical context of the sample. The adjustment of a token's characters therefore has to follow the printed original on the scan, even if the spelling does not align with current orthography. This also applies to suspected spelling and printing mistakes within the original (ger.: Aufnahme nach Vorlageform). Ambiguous cases (spelling vs. OCR) are to be discussed by the annotation team, possible exceptions will be captured in the guidelines. This also applies to punctuation characters (e.g. "=" instead of "-" to mark compounds).

Exception: Hyphenations of named entities over two lines in the original are to be counted as sectional errors. This also applies to composita that are divided into two lines in the original scan.

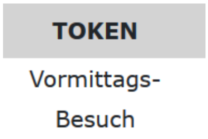
Examples: incorrect correct incorrect correct



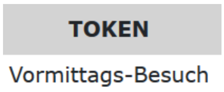
TOKEN
Herr
Gam-
betta



TOKEN
Herr
Gambetta



TOKEN
Vormittags-
Besuch

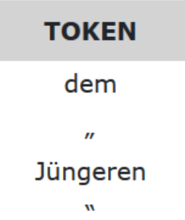


TOKEN
Vormittags-Besuch

3. Some newspapers in the corpus in gothic type do not distinguish between capital I and capital J. In these cases, the OCR is interpretation considered correct, since verification though the snippet is impossible.

4. Completely missing words due to OCR errors to be manually refilled, if the missing word(s) is/are recognized to be a named entity by the original scan.

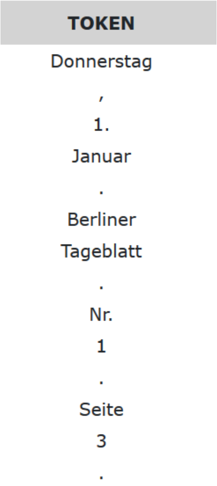
5. Punctuation characters are to be counted as separate tokens each.
Example:



TOKEN
dem
„
Jüngeren
“

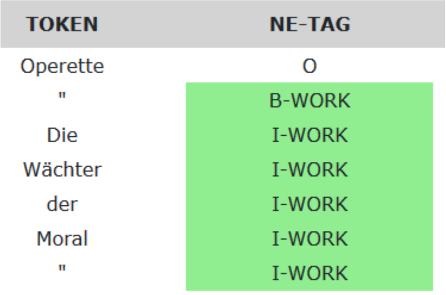
Exception: Punctuation characters as parts of abbreviations (e.g. “St.”) and numberings (e.g. “4.” for “fourth”) are part of the token and therefore not to be counted separately.

Example:



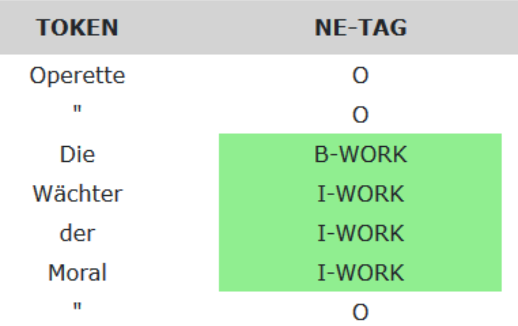
TOKEN
Donnerstag
,
1.
Januar
.
Berliner
Tageblatt
.
Nr.
1
.
Seite
3
.

6. Should an entity be surrounded by punctuation characters, the latter are not to be included in the annotation of the entity.
Example: incorrect



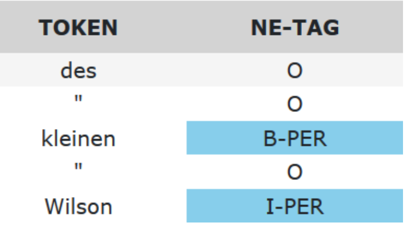
TOKEN	NE-TAG
Operette	O
"	B-WORK
Die	I-WORK
Wächter	I-WORK
der	I-WORK
Moral	I-WORK
"	I-WORK

correct

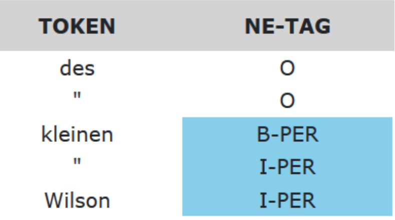


TOKEN	NE-TAG
Operette	O
"	O
Die	B-WORK
Wächter	I-WORK
der	I-WORK
Moral	I-WORK
"	O

7. Should one or more punctuation characters be embedded between two or more tokens that mark a single entity, they are to be included in the annotation of the entity.
   Example: incorrect correct

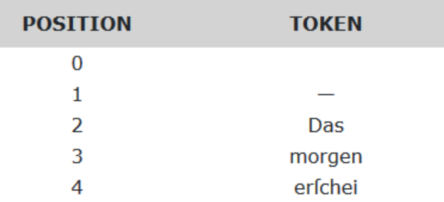


TOKEN	NE-TAG
des	O
"	O
kleinen	B-PER
"	O
Wilson	I-PER

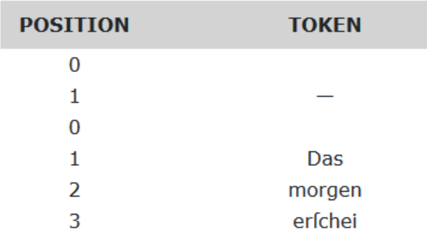


TOKEN	NE-TAG
des	O
"	O
kleinen	B-PER
"	I-PER
Wilson	I-PER

8. Extended dashes are to be corrected to a single dash in case of entity compounds ("–" → "-").
9. If a sentence starts with a punctuation character, the first is to be considered a separate sentence.
   Exception: Quotation marks ("")
   Example: incorrect correct



POSITION	TOKEN
0	
1	—
2	Das
3	morgen
4	erſchei



POSITION	TOKEN
0	
1	—
0	
1	Das
2	morgen
3	erſchei

10. Special characters are to be taken into account, if they are part of the basic Latin or extended German alphabet ("Ü","ü","Ö","ö","Ä","ä","ß"). Additionally, the following accents are to be taken into account: aigu (é), grave (è), circonflex (ê), as well as historical characters: ſ

8. References

Zumstein, Philipp; Baierer, Konstantin (2016): Verbesserung der OCR in digitalen Sammlungen von Bibliotheken. In: 027.7 Zeitschrift für Bibliothekskultur 4 (2). DOI: 10.12685/027.7-4-2-155

Fort, Karën; Ehrmann, Maud; Nazarenko, Adeline (2009): Towards a Methodology for Named Entities Annotation. In: Proceedings of the Third Linguistic Annotation Workshop (LAW III), p. 142-145. Available at: https://www.aclweb.org/anthology/W09-3025.pdf

Kettunen, K.; Ruokolainen, T. (2017): Names, Right or Wrong: Named Entities in an OCRed Historical Finnish Newspaper Collection. In: Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage - DATeCH2017. Göttingen, Germany: ACM Press, S. 181–186. Available at: http://dl.acm.org/citation.cfm?doid=3078081.3078084

Kugler, Anna (2018): Automatisierte Volltexterschließung von Retrodigitalisaten am Beispiel historischer Zeitungen. 33-54 Seiten / Perspektive Bibliothek, Bd. 7, Nr. 1 (2018). DOI: 10.11588/PB.2018.1.48394.

Rosset, Sophie; Grouin, Cyril; Zweigenbaum, Pierre (2011): Entités Nommées Structurées : guide d'annotation Quaero (Sophie Rosset, Cyril Grouin, Pierre Zweigenbaum), Technical report. Available at: http://www.quaero.org/media/files/bibliographie/quaero-guide-annotation-2011.pdf

Rosset, Sophie; Grouin, Cyril; Fort, Karën; Galibert, Oliver; Kahn, Juliette; Zweigenbaum, Pierre (2012): Structured Named Entities in two distinct press corpora: Contemporary Broadcast News and Old Newspapers. In: Proceedings of the Sixth Linguistic Annotation Workshop, p. 40-48. Available at: https://www.aclweb.org/anthology/W12-3606.pdf

Reiter, Nils (2017): How to Develop Annotation Guidelines. Blog post. Available at: https://sharedtasksinthedh.github.io/2017/10/01/howto-annotation/

Reznicek, Marc (2013): Linguistische Annotation von Nichtstandardvarietäten —Guidelines und „Best Practices". Guidelines NER. Version 1.5. Available at: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-ner-1.5

Appendix A: Example tag-set

This is an extended and modified list based on the NoSta-D-TagSet (Rezincek et al. 2013, p. 6ff.).

Entity class	Subclass	Example	Exceptions
Person (PER)	Vorname/Mittelname	Nina Hannelore Winfried Emil	
	Familienname	Feuerstein Winkler-Eversberg	
	Dynastie, Geschlecht	Borgia Habsburger	
	Künstlername/Pseudonym	P!nk Felix Brummer Marilyn Monroe	
	Fiktiver oder religiöser Charakter	Harry Potter Miss Piggy Heiliger Antonius Buddha	
	Spitzname/Nickname	honeylove86 chatbotchatter123 Naddel Müller “Der Jüngere”	Der „Kleine“ Schatz
	Berufliche Funktionen (exklusiv)	Der Finanzminister Seine königliche Hoheit Die Kaiserin	
Organization (ORG)			~~Delegationen~~ ~~Expertengruppe~~ ~~Die Großmächte~~ ~~Die kleinen Völker~~ ~~Der Feind~~ ~~Bundesgenossen~~ ~~Die Verbündeten~~ ~~Die Christen~~ ~~Bildungseinrichtung nicht-explizit: z.B. „die Schule“, „das Waisenhaus“~~

	Öffentliche o. politische Organisation/Körperschaft	NATO EU Deutscher Bundestag Parlament Regierung Ministerien Kabinett Kommission Expertengruppe Die Pforte Feuerwehr Polizei Eisenbahn Schweizerische Westbahnen Zoll Börse Aufsichtsräte (WENN Firma ersichtlich!)	~~MinisterORGemb~~ ~~ParlamentarierORGemb~~
	Kaufhäuser (unique)	Kaufhaus des Westens	Keine LOC!
	Unternehmen	Microsoft VW	
	Institut	Dt. Inst. f. Menschenrechte	
	Bildungseinrichtung (explizit)	Freie Universität Berlin	
	Kultureinrichtung (nicht explizit)	Cinemaxx SPK	Explizite Einrichtungen = LOC (Bsp. Pergamonmuseum)
	Presse	Berliner Zeitung Tagesspiegel Die Pforte	~~Zeitungsname im Titel der aktuellen Ausgabe~~
	Vereine, Clubs	Füchse Berlin VfB Stuttgart Lions Club	Mannschaften = PER
	Sender, Rundfunkanstalten	ZDF Arte Radio Bremen	
	Restaurants und Hotels	Adlon Zur Linde Sassella	Keine LOC!
	Bands, Musikgruppen, Orchester	The Beatles	
	Militäreinheiten	Blauhelme, Armeen, Heere, Sondereinsatzkommando	
	Modelabel	Chanel	
	Politische Parteien	FDP Die Grünen	
			~~Schiffe~~ ~~Sozialisten~~ ~~Kommunisten~~
Location (LOC)			~~Himmelsrichtungen~~ ~~Ausland~~

			~~Inland~~ ~~Feindesland~~ ~~International~~ ~~Die ganze Welt~~ ~~Unsere Welt~~ ~~Deine Welt~~ ~~Die Front~~ ~~Ostfront~~ ~~Inselreich~~
	Städte	(Hansestadt) Hamburg New York City Kapstadt	
	Länder, Nationen, Staaten	Südafrika	
	Stadtteile, Bezirke, Kieze	Schöneberg Köln Deutz	
	Sehenswürdigkeiten	The Bean Brandenburger Tor	
	Planeten, Galaxien	Erde Milchstraße	
	Landschaften	Lüneburger Heide	
	Straßen, Plätze	Alexanderplatz Bernauer Strasse	
	Gewässer, Flüsse, Seen, Meere etc.	Viktoriasee Spree	
	Kontinente	Südamerika	
	Geografische Räume (kulturell)	Der Orient Das Abendland	
	Geografische Räume (juristisch oder politisch)	Frz. Hoheitsgebiet Deutsches Zollgebiet Französicher Kollonialraum	
	Gebäude	Pentagon Kreml Bundestag	
Conference (CONF)	Kongresse, Tagungen	CLEF	
Event (EVT)	Demontrationen	Freitagsdemo von Fridays for Future	
	Paraden	Christopher Street Day	
	Parteitage	SPD Parteitag	
	Demonstrationen	Freitagsdemo von Fridays for Future	
	Festivals	Lollapalooza 2015	
	Kriege	Der Zweite Weltkrieg	
	Firmenspezifische Treffen	Generalversammlung Aufsichtsratsitzung	
Works and expressions (WORK)			~~Technische Serientypen, z.B. Automobiltypen (VW Käfer)~~ ~~Patente~~ ~~Software~~

	Gemälde	Mona Lisa Guernica	
	Plastiken und Skulpturen	Venus von Willendorf	
	Literarische Werke	Die Räuber Frankenstein Der Abrogans	
	Religiöse Werke	Die Bibel Das Alte Testament 95 Thesen	
	Aufführungen und Inszenierungen	Schwanensee	
	Filme	Some Like It Hot	
	Musikstücke, Songs, Alben	Abbey Road Jingle Bells 9. Sinfonie	
	Gesetze, Verträge	Handelsvertrag zwischen Deutschland und Oesterreich vom [Datum (DD)-MM-YY] Meistbegünstigungsvertrag Hartz IV	

Appendix B: List of exceptions

Rule no.	Exception	Explanation
5.1 no. 4: Determiner	The Big Apple	This definite descriptions of a location includes articles.
5.2 no. 5: Titles	Kaiser/in Minister/in des Inneren	