NoSta-D Named Entity Annotation for German: Guidelines and Dataset

Darina Benikova[1], Chris Biemann[1], Marc Reznicek[2]

(1) FG Language Technology, Comp. Sci. Dept., TU Darmstadt, Germany
(2) Facultad de Filologa, Universidad Complutense de Madrid, Spain
darina.benikova@stud.tu-darmstadt.de, biem@cs.tu-darmstadt.de, mreznice@ucm.es

Abstract

We describe the annotation of a new dataset for German Named Entity Recognition (NER). The need for this dataset is motivated by licensing issues and consistency issues of existing datasets. We describe our approach to creating annotation guidelines based on linguistic and semantic considerations, and how we iteratively refined and tested them in the early stages of annotation in order to arrive at the largest publicly available dataset for German NER, consisting of over 31,000 manually annotated sentences (over 591,000 tokens) from German Wikipedia and German online news. We provide a number of statistics on the dataset, which indicate its high quality, and discuss legal aspects of distributing the data as a compilation of citations. The data is released under the permissive CC-BY license, and will be fully available for download in September 2014 after it has been used for the GermEval 2014 shared task on NER. We further provide the full annotation guidelines and links to the annotation tool used for the creation of this resource.

Keywords: German Named Entity Recognition and Classification, NERC, annotated data, nested span annotation, annotation practices, guidelines

1. Introduction

1.1. Motivation

Named Entity Recognition (NER, a.k.a NERC) is the detection and classification of proper name spans in text. Automatic NER is an important pre-processing step in tasks such as information extraction, question answering, automatic translation, data mining, speech processing and biomedical science. Also, it serves as a pre-processing step for deeper linguistic processing such as syntactic or semantic parsing, and co-reference resolution.

NER for German is especially challenging, as not only proper names, but all nouns are capitalized, which renders the capitalization feature less useful than in other Western languages such as English or Spanish. Furthermore, adjectives derived from Named Entities (NEs) such as "*englisch*" are not capitalized. A baseline established on capitalized words therefore fails to show even moderate accuracy levels for German.

Despite German being a wide-spread and comparatively well-resourced language, German NER has not received a lot of attention, and has so far been only trained on the CoNLL-data (Tjong Kim Sang and De Meulder, 2003). Since this data was annotated by non-native speakers and it is known to be somewhat inconsistent, system performance is typically in the 70%-75% range, as opposed to a recognition rate of close to 90% for a comparable task on English (Tjong Kim Sang and De Meulder, 2003). More recently, Faruqui and Padó (2010) have extended this data for evaluation purposes, and made available a German NER module for the Stanford NER tagger. However, the entire training data is cumbersome to obtain due to copyright issues, and its use is only permitted for academic purposes. In contrast to this, our data set is freely available for download under a permissive license.

In this project, we have annotated nested NE annotations, which is more complex than the prevalent BIO tagging scheme (Tjong Kim Sang and De Meulder, 2003), consider e.g. "*Real Madrid*" referring to an organization, with a nested location "*Madrid*".

Moreover, NER has previously been regarded as a rather syntactic task. Phrases only partly containing names ("*Germany-wide*" - "*deutschlandweit*") or adjectives referring to NEs ("*Euclidean*" - "*euklidisch*"), were ignored by most named entity projects, but cannot be ignored for semantic tasks, e.g. identifying the locations in a news article. From an information extraction perspective, in an example like "*the president traveled from Hungary to his Danish allies*", it is more interesting that the destination of the described travel event was Denmark than the fact that this was expressed using an adjective.

The dataset presented in this paper was annotated by native speakers, according to semantic-based guidelines containing derivations along with phrases partly containing NEs. All data was annotated by at least two annotators, and adjudicated by a third.

1.2. Related Work

To our knowledge, German NER data has only been released as part of the CoNLL-2003 challenge, on which the Stanford NER tagger (Faruqui and Padó, 2010) and similar projects on German NER (Chrupała and Klakow, 2010; Rössler, 2004) have been trained. Apart from an extension by Faruqui and Padó (2010) for out-of-domain evaluation purposes, we are not aware of any other German NER data set.

Nested span annotation for NER (e.g. (Byrne, 2007), and see below) is encountered rarely, which has mainly been attributed to technological reasons (Finkel and Manning, 2009).

Rosset et al., (2012) describe what they call multi-level annotation for French, where parts of spans are subclassified (e.g. first and last names) in a rather fine-grained ontology of name types. In contrast to this, we restrict our-

selves to the four main NER types, but annotate them in nested form when a long span would contain a shorter one, usually of different type. The GENIA corpus (Kim et al., 2003) labels biomedical entities for bio-text mining, 17% of which are embedded in other entities (Finkel and Manning, 2009). Another dataset including nested entities is the Spanish and Catalan newspaper text corpus AnCora (Taulé et al., 2008), containing 50% nested entities (Finkel and Manning, 2009). These numbers illustrate that by ignoring nested named entities, a large part of the information in the data is lost. Although training on a nested named entity set is not trivial, there exist well-performing classifiers for this task on the AnCora or the GENIA corpora (Finkel and Manning, 2009; Màrquez et al., 2007) that should also be applicable for nested NER.

2. Source Data Selection and Distribution

2.1. Source Data

The German source text was sampled sentence-wise from Wikipedia articles and online newspapers, using corpora from LCC[1] (Richter et al., 2006) as a basis. The sampling procedure allows distributing the data as a compilation of citations without violating copyrights, for details see Section 2.2.

While in the general case, a considerable amount of context information is lost when only processing sentences in random order rather than full documents, the sentence context is sufficient to decide for NER spans and their classes in the overwhelming number of cases. With this stratified sampling of sentences across documents, we also avoid burstiness effects (Church, 2000) that lead to the overrepresentation of certain low-frequency names.

2.2. Legal Aspects

The re-distibution of sentences from Wikipedia is allowed due to its CC-BY license. For sentences randomly sampled from news articles, the following legal situation applies. A sentence, as compared to a full document, is a comparably short snippet of text and not regarded as a complete work. It can be literally cited, if the source of the sentence is clearly stated and the sentence was, intentionally and for an extended period of time, available to the public. Since LCC corpora are collected from online newspapers and subsequently down-sampled and randomized in order, the original articles are neither complete nor reconstructable, thus the copyright of the articles as a whole is not violated.

According to American copyright law, there is no copyright on a common phrase, which is taken out of its context[2].

According to German copyright law, even short phrases may have copyright, if individual features of the originator exist. As no stylistic devices or other features of individual thoughts of style are to be expected in randomly chosen sentences from newspaper articles, the set of sentences in our data set can be regarded as literal quotes with citations without individual copyright in at least German and American legislation. The source of every sentence is contained as a comment above every sentence in the data set, as exemplified in Table 3. The distribution of the sentences cited in the set are justified by the 17 U.S.C Art. 106(3) - Right to distribute copies of the copyrighted work to the public. As these randomly chosen sentences are unprotected by German copyright (see Schricker and Loewenheim (2010), Art. 62 Rn. 1), they may be distributed to the public. For further details, see (Lucke, 2010), pp. 231ff.

3. Annotation Guidelines

In this section, we give a short introduction to the guidelines and define our named entity classes. Then, we discuss the genesis of the final guidelines, which were iteratively improved by a first set of annotators and subsequently tested and refined on a second set of annotators. The full guidelines (in German language) are given in the Appendix below, and are also available for download along with the data in their original form.

3.1. Named Entity Classes and Nesting

The guidelines we used in this study have been developed as part of a larger project dealing with the extension of given annotation guidelines to non-standard varieties (Dipper et al., 2013). Taking a mainly qualitative perspective on out-of-domain linguistic phenomena, we had to deal with the long-lasting theoretical discussion on the distinction between proper nouns and common nouns, which has revealed its intrinsically gradual nature (Ernst, 2002; Koß, 1995). Since we had no resources to develop a linguistically satisfying procedure for gradual named entity annotation, we based our guidelines on the most commonly used guidelines for part-of-speech annotation for German (Schiller et al., 1999) and on the Stylebook for the Tübingen Treebank (Telljohann et al., 2012), which has been employed in the NE annotation of one of the largest manually annotated newspaper corpora of German (TüBa-D/Z). We conflated categories to raise inter-annotator agreement (location & geopolitical entity) and added new categories (e.g. virtual locations for chat logs, not relevant to the dataset described here). The guidelines distinguish two partial tasks where the first is to locate named entities and the second to classify them. In the first task we had to account for tokenization difficulties like the in example 1) where "Heinrich Böll" describes a person and "Heinrich Böll-Stiftung" is an organization. In this case the person NE does include only part of the second token. We solved the problem by introducing categories the "part"-class marking tokens that partly include named entities. Our solution would resolve in 2).

1) Daß er das aber als Vorstand der [[Heinrich Böll]1-Stiftung]2 tut ...
2) Daß er das aber als Vorstand der [[Heinrich]1 [Böll-Stiftung]2]3 tut ...

Figure 1 shows the annotation as visualized in the WebAnno web-based annotation tool[3] (Yimam et al., 2013)

[1] http://corpora.uni-leipzig.de/

[2] "It is well established that copyright or literary rights do not extend to words or phrases isolated from their context, nor do they extend to abstract ideas or situations." (O'BRIEN v. CHAPPEL & CO. 159 F.Supp. 58 (1958))

[3] https://code.google.com/p/webanno/

we used for this annotation project. Most of the classes are illustrated. We have annotated four main classes (PERson, ORGanization, LOCation and OTHer). Each class can appear in its pure form, as a part of a token (e.g. ORGpart in Figure 1 to express that "*EU"* is an organization, which forms part of the token "*EU-Verwaltung*"), or as a derivation, such as e.g. in "*österreichischen*" in the figure, which is an adjective derivation of Austria.

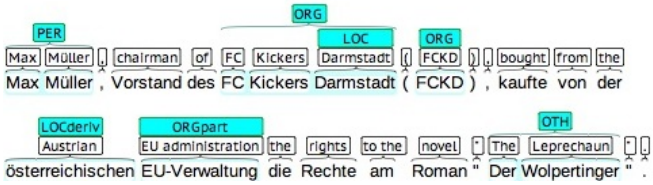

Figure 1: Sample annotation made with WebAnno to illustrate the tagset. English gloss provided for information only, and is not displayed during annotation.

3.2. Iterative Improvement of Guidelines

To improve the guidelines and consequently the quality of the dataset, meetings with the annotation group for clarifying the guidelines were held. As a result of these group meetings, more detailed examples were added and rules were clarified in order to prevent misunderstandings. After convergence, the guidelines were tested by a new team of annotators, which received no other instructions than the written guidelines. With these improved guidelines, the new group was able to work measurably quicker and more consistently: whereas the four members of the first group annotated 120 sentences per hour on average, the three members of the second group annotated 180 sentences per hour. Moreover, the pairwise kappa agreement amongst annotators of the first group was between 0.55 - 0.69 with an average of 0.63. The second group reached a kappa between 0.73 - 0.75 with an average of 0.74. As improvements in this comparatively high kappa range (Carletta, 1996) indicate much less disagreements on the span level, the speed of curation doubled for the second group.
Only minor adjustments to the guidelines were requested by the second group. All of these indicators show that the guidelines are consistent, comprehensive, understandable and thus practical for the source data. Hence, the dataset annotated with these guidelines should provide a suitable training set for further processing.

3.3. Scaling up

A group of native German linguistics students annotated and categorized NEs in the sentences using WebAnno. All data was annotated independently by at least two annotators, and subsequently curated by a third annotator, using the convenient functionalities of WebAnno regarding user management, data import and export, and visually supported curation. The curator can choose correct annotations in case of conflict, and also add missing or change/delete wrong annotations. Figure 2 shows the curation interface of WebAnno: sentences with disagreement are marked in red on the left side. A curator can quickly resolve disagreement between annotators (as shown on the right side). Curators were asked to check all sentences, not only sentence with disagreement.

4. Characteristics and Dataset

4.1. Size and Characteristics

The dataset is publicly available for download[4] under the permissive CC-BY license. It consists of a total of 31,300 sentences respectively 591,006 tokens, marked up with 41,005 span annotations, about 7% of them being nested and about 15% being either derivations or partly NEs. Table 1 displays the distribution of the annotations per class for both nested and simple span annotations for the entire dataset.

Class	All annotations	Nested
Location	12,165	1,452
Organization	7,175	281
Person	10,500	488
Other	4,041	59
Location_deriv	4,403	790
Location_part	707	36
Organization_deriv	55	4
Organization_part	1,073	9
Person_deriv	95	19
Person_part	275	29
Other_deriv	282	3
Other_part	234	3
Total	41,005	3,173

Table 1: Distribution of classifications in our dataset of 31,300 sentences.

The overall dataset contains 41,005 annotations. For comparison, the entire CoNLL data and the extension by Faruqui and Padó (2010) sum up to 421,682 tokens in 23,388 sentences, and 19,434 span annotations. The distribution of classification of the extended CoNLL data may be viewed below:

Class	Number of occurrences
LOC	7,154
ORG	4,323
PER	5,840
OTH/MISC	2,117
Total	19,434

Table 2: Distribution of classifications in the extended CoNLL data (2003 challenge and extension by Faruqui and Padó (2010).

The table shows that our dataset contains 33% more sentences and 40% more tokens than the previous German NER datasets combined. Further, due to the choice of the source data and the more inclusive annotation guidelines, our dataset contains more than twice as many overall annotations than the previous datasets.

[4] http://www.lt.informatik.tu-darmstadt.de/de/data/german-named-entity-recognition/

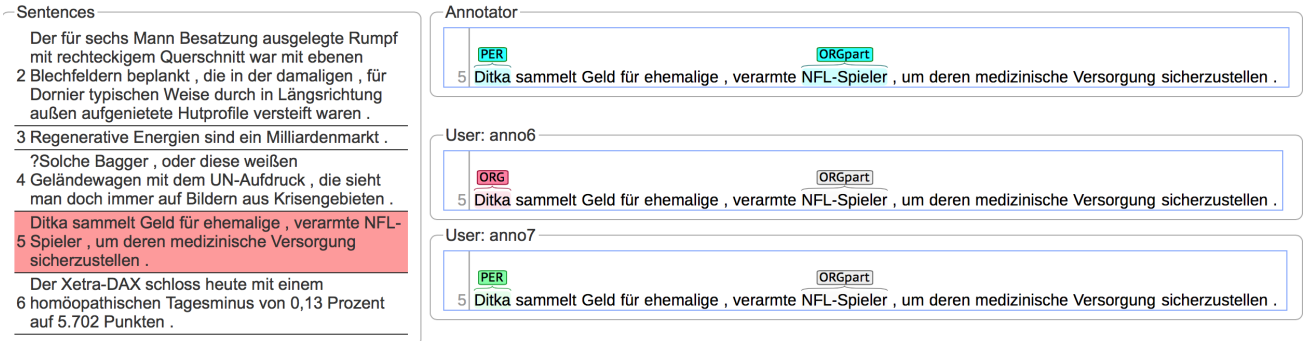

Figure 2: Curation interface of WebAnno. A disagreement on the class of *Ditka* in Sentence 5 between annotators anno6 and anno7 has been resolved by the curation annotator.

4.2. File Format

We distribute the dataset in a tab-separated format similar to the CoNLL-Format. In contrast to the original CoNLL-NER-Format, we have added token numbers per sentence in the first column, and a comment line before each sentence that indicates source and date. We use the BIO-scheme to encode named entity spans, and use two columns for this: The first NER column encodes the outer spans, and the second column contains nested/embedded spans. Despite having observed a few (about one every 1000 sentences) cases where these two levels of nesting do not suffice, we have decided to only provide these two levels for the sake of simplicity. Table 3 shows an example of the data format for one sentence from Wikipedia.

#	http://de.wikipedia.org/wiki/Manfred_Korfmann [2009-10-17]		
1	Aufgrund	O	O
2	seiner	O	O
3	Initiative	O	O
4	fand	O	O
5	2001/2002	O	O
6	in	O	O
7	Stuttgart	B-LOC	O
8	,	O	O
9	Braunschweig	B-LOC	O
10	und	O	O
11	Bonn	B-LOC	O
12	eine	O	O
13	große	O	O
14	und	O	O
15	publizistisch	O	O
16	vielbeachtete	O	O
17	Troia-Ausstellung	B-LOCpart	O
18	statt	O	O
19	,	O	O
20	„	O	O
21	Troia	B-OTH	B-LOC
22	-	I-OTH	O
23	Traum	I-OTH	O
24	und	I-OTH	O
25	Wirklichkeit	I-OTH	O
26	”	O	O
27	.	O	O

Table 3: Data format illustration. The example sentence contains five named entities: the locations *Stuttgart*, *Braunschweig* and *Bonn*, the noun including a location part *Troia-Ausstellung*, and the title of the event *Troia - Traum und Wirklichkeit*, which contains the embedded location *Troia*.

5. Conclusion and Further Work

The iterative improvement approach, the double annotation and the tool-supported curation step ensure a high consistency of this dataset. Advanced features such as nested annotation and classification categories, namely of derivations and NE parts, but also the annotation by native speakers make this dataset the most comprehensive freely available German dataset for NER.

We will use this data for a shared task in the GermEval 2014 competition on Named Entity Recognition[5]: While 26,200 sentences are already available, the remaining 5,100 sentences will be made available in September 2014 after the shared task submission deadline, since they will used as blind test data.

As a future step, we plan to train and test a NER tagger for German to assess learnability of the new categories (derivations and parts) and the nested representation. Further, we will investigate whether it is beneficial to combine our dataset with the extended CoNLL dataset for training, despite its known issues with consistency and annotation quality.

6. Acknowledgements

This work was supported by a German BMBF grant to the CLARIN-D project, the Hessian LOEWE research excellence program as part of the research center “Digital Humanities” and a travel grant by “Vereinigung von Freunden der Technischen Universität zu Darmstadt e.V.”. Thanks goes to Uwe Quasthoff for pointing us to the legal regulations regarding the distribution of short snippets of textual data. We thank Burkhard Dietterle for his contributions to the guidelines, and Eva Jahnson, Jascha Jung, Kinga Milan, Franz-Xaver Ott, Rebekka Raab and Isabel Steinmetz for annotating.

[5] https://sites.google.com/site/germeval2014ner/

7. References

Armin Burkhardt. 2004. Nomen est omen? : zur Semantik der Eigennamen. In *Landesheimatbund Sachsen-Anhalt e. V. (Hrsg.): "Magdeburger Namenlandschaft" : Orts- und Personennamen der Stadt und Region Magdeburg*, pages 7–22, Halle, Germany. Druck-Zuck.

Kate Byrne. 2007. Nested named entity recognition in historical archive text. In *International Conference on Semantic Computing (ICSC)*, pages 589–596. IEEE.

Jean Carletta. 1996. Assessing agreement on classification tasks: the kappa statistic. *Computational linguistics*, 22(2):249–254.

Nancy Chinchor. 1995. MUC-6 Named Entity Task Definition (Version 2.1). In *6th Message Understanding Conference*, Columbia, Maryland, USA.

Grzegorz Chrupała and Dietrich Klakow. 2010. A Named Entity Labeler for German: exploiting Wikipedia and distributional clusters. In *Proceedings of the Conference on International Language Resources and Evaluation (LREC)*, pages 552–556, Malta, Valetta.

Kenneth W. Church. 2000. Empirical estimates of adaptation: the chance of two Noriegas is closer to p/2 than p 2. In *Proceedings of the 18th conference on Computational linguistics (COLING) -Volume 1*, pages 180–186, Hong Kong, China.

Stefanie Dipper, Anke Lüdeling, and Marc Reznicek. 2013. NoSta-D: A corpus of German non-standard varieties. In Marcos Zampieri and Sascha Diwersy, editors, *Non-Standard Data Sources in Corpus-Based Research*, pages 69–76. Shaker.

Peter Ernst. 2002. *Pragmalinguistik: Grundlagen. Anwendungen. Probleme*. Walter de Gruyter.

Manaal Faruqui and Sebastian Padó. 2010. Training and evaluating a German named entity recognizer with semantic generalization. In *Proceedings of Konferenz zur Verarbeitung natürlicher Sprache (KONVENS)*, pages 129–133, Saarbrücken, Germany.

Jenny R. Finkel and Christopher D. Manning. 2009. Nested named entity recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP): Volume 1*, pages 141–150, Singapore.

Jin-Dong Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl 1):180–182.

Gerhard Koß. 1995. Die Bedeutung der Eigennamen: Wortbedeutung/Namenbedeutung. *Eichler, Ernst/Hilty, Gerold/Löffler, Heinrich/Steger, Hugo/Zgusta, Ladislav (Hrsg.)*, pages 458–463.

Bettina Lucke. 2010. *Die Google Buchsuche nach deutschem Urheberrecht und US-amerikanischem Copyright Law*. Verlag Peter Lang, Frankfurt a.M.

Lluís Màrquez, Luis Villarejo, Maria A. Martí, and Mariona Taulé. 2007. Semeval-2007 task 09: Multilevel semantic annotation of Catalan and Spanish. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 42–47, Prague, Czech Republic.

Matthias Richter, Uwe Quasthoff, Erla Hallsteinsdóttir, and Chris Biemann. 2006. Exploiting the Leipzig corpora collection. In *Proceesings of the IS-LTC*, Ljubljana, Slovenia.

Sophie Rosset, Cyril Grouin, Karën Fort, Olivier Galibert, Juliette Kahn, and Pierre Zweigenbaum. 2012. Structured named entities in two distinct press corpora: Contemporary broadcast news and old newspapers. In *Proceedings of the Sixth Linguistic Annotation Workshop (LAW)*, pages 40–48, Jeju, Republic of Korea.

Marc Rössler. 2004. Corpus-based learning of lexical resources for German named entity recognition. In *Proceedings of Conference on International Language Resources and Evaluation (LREC)*, Lisbon, Portugal.

Anne Schiller, Simone Teufel, and Christine Stöckert. 1999. Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical report, IMS, University of Stuttgart and SfS, University of Tuebingen, Germany.

Gerhard Schricker and Ulrich Loewenheim. 2010. *Urheberrecht – Kommentar, 4. Auflage*. C.H. Beck.

Mariona Taulé, Maria A. Martí, and Marta Recasens. 2008. Ancora: Multilevel annotated corpora for Catalan and Spanish. In *Conference on International Language Resources and Evaluation (LREC)*, Marrakech, Morocco.

Heike Telljohann, Erhard W. Hinrichs, Sandra Kübler, Heike Zinsmeister, and Kathrin Beck. 2012. Stylebook for the Tübingen treebank of written German (TüBa-D/Z). Technical report, Universität Tübingen, Seminar für Sprachwissenschaft.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147, Edmonton, Canada.

Seid M. Yimam, Iryna Gurevych, Richard E. de Castilho, and Chris Biemann. 2013. WebAnno: A flexible, web-based and visually supported system for distributed annotations. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–6, Sofia, Bulgaria.

Appendix: Original Guidelines

This appendix contains the final version of the German guidelines as used to annotate the dataset described in this work. While the layout was modified, the content is complete, except for some examples, which are redundant as the whole corpus of over 31,000 sentences was annotated according to the guidelines.

Guidelines für die Named Entity Recognition. Sie bauen auf den Guidelines in den STTS-Guidelines (Schiller et al., 1999), (Telljohann et al., 2012) und (Chinchor, 1995) auf.

Einführung: Named Entity Recognition

Unter der Named Entity Recognition (NER) versteht man die Aufgabe, Eigennamen (named entities) in Texten zu erkennen. Technisch gesehen sind hierzu zwei Schritte notwendig. Zuerst müssen in einem laufenden Text die Token gefunden werden, die zu einem Eigennamen gehören (Named Entity Detection: NED), danach können diese Eigennamen semantischen Kategorien zugeordnet werden (Named Entity Classification). Prototypisch ist dabei der Unterschied zwischen Eigennamen und Appellativa der, dass letztere eine Gattung oder eine Klasse beschreiben, während erstere einzelne Individuen oder Sammlungen von Individuen unabhängig von gemeinsamen Eigenschaften bezeichnen (Burkhardt, 2004). Die vorliegenden Guidelines sollen es Annotatoren ermöglichen, Eigennamen in Texte aus Standard und Nichtstandard-Varietäten konsistent zu annotieren. In diesen Guidelines werden die beiden Aufgaben der NED und NEC nicht unterschieden, da die Konzentration auf Beispiele in diesem Dokument, die Trennung künstlich erzeugen müsste und nicht zu erwarten ist, dass die Resultate sich dadurch verbessern würden. In Anlehnung an die oben genannten Guidelines für Zeitungssprache werden in NoSta-D vier semantische Hauptklassen unterschieden (Personen, Organisationen, Orte und Andere). Diese werden teilweise um spezifische Verwendungen erweitert (virtuelle Orte), Eigennamen, die Teile größerer Einheiten sind, werden als solche markiert (NEpart), oder Appellativa, die von Eigennamen abgeleitet sind, werden gesondert behandelt (NEderiv).

Wie finde ich eine NE?

Schritt 1:
Nur volle Nominalphrasen können NEs sein. Pronomen und alle anderen Phrasen können ignoriert werden.
Schritt 2:
Namen sind im Prinzip Bezeichnungen für einzigartige Einheiten, die nicht über gemeinsame Eigenschaften beschrieben werden.
Beispiel:

[Der Struppi] folgt [seinem Herrchen].

Hier gibt es zwei Nominalphrasen als Kandidaten für einen Eigennamen (NE). "*Der Struppi*" bezeichnet eine einzige Einheit. Es kann auch mehrere Struppis geben, aber diese haben an sich keine gemeinsamen Eigenschaften, bis auf den gemeinsamen Namen, daher handelt es sich um einen Eigennamen.
"*seinem Herrchen*" bezeichnet zwar (typischerweise) auch nur eine einzige Person allerdings können wir diese nur über die Eigenschaft identifizieren, dass sie ein Herrchen ist und dass dies für Struppi zutrifft. Struppi könnte auch mehrere Herrchen haben, die alle die Eigenschaften teilen, die ein Struppi-Herrchen beinhaltet (z.B. darf Struppi streicheln, muss ihn ausführen und füttern etc.)
Schritt 3:
Determinierer sind keine Teile des Namens.
Beispiel:
Der [Struppi]NE folgt seinem Herrchen.
Schritt 4:
Eigennamen können mehr als ein Token beinhalten.
Beispiel:
Viele Personennamen (PER für person):
[Peter Müller]PER
Filmtitel (OTH für other):
[Knockin' on Heavens Door]OTH
Schritt 5:
Eigennamen können auch ineinander verschachtelt sein.
Beispiel:
Personennamen in Filmtiteln:
[[Shakespeare]PER in Love]OTH
Orte (LOC für location) in Vereinsnamen (ORG für organisation):
[SV [Werder [Bremen]LOC]ORG]ORG
Schritt 6:
Titel, Anreden und Besitzer gehören NICHT zu einem komplexen Eigennamen. Besitzer können natürlich selber Eigennamen sein. Beispiel:
Referenz auf Musiktitel:
[Vivaldis]PER [Vier Jahreszeiten]OTH
Referenz auf Personen:
Landesvorsitzende Frau [Ute Wedemeier]PER
Schritt 7:
Eigennamen treten auch als Teil eines komplexen Tokens auf. Hier wird für das gesamte Token annotiert, dass es einen Eigennamen enthält. Beispiel:
mit Firmen Assoziiertes:
[DAEWOO-Millionen]ORGpart
mit bestimmten Personen verbundene Erfindungen/Arbeiten:
[Hartz-Reformen]PERpart
[Ottomotor]PERpart
ABER: Wenn auch das Gesamttoken einen Eigennamen darstellt, dann wird nur dieser annotiert. Beispiel:
Stiftungen: *[Böll-Stiftung]ORG*
Schritt 8:
Kann in einem Kontext nicht entschieden werden, ob eine NP sich als Eigennamen oder Appellativ verhält, wird es nicht als NE markiert. Beispiel:
Ortsnamen vs. -beschreibungen:
...und zogen mit ihren großen Transparenten gestern vom [Steintor] über den [Ostertorsteinweg]LOC zum [Marktplatz].
Schritt 9:
Wenn ein Name als Bezeichnung für bestimmte Gegenstände in die Sprache übergegangen ist und in seiner Nutzung nicht als NE fungiert, so wird dieser nicht annotiert. Beispiel:

[Teddybär] (NICHT PERderiv)
[Colt] (NICHT PERderiv)
Schritt 10:
Bei Aufzählungen mit Hilfe von Bindestrichen oder Vertragen eines Teils der NE auf spätere Wörter, wird die NE so annotiert, als sei sie voll ausgeschrieben.
Beispiel:
[Erster]OTH und [Zweiter Weltkrieg]OTH
[Süd-]LOC und [Nordkorea]LOC

Zu welcher semantischen Klasse gehört ein Eigenname?

Wenn der Namenskandidat in der Liste unter der Klasse "keine NE" aufgeführt wird, dann handelt es sich nicht um eine NE im Sinne dieser Guidelines.
Wenn der Eigenname in eine der Klassen in der Liste Faustregel zur Unterscheidung einer Klassenbezeichnung und eines Namens gehört, dann annotiere die zugehörige Klasse.
Sollte die gefundene NE Rechtschreibfehler enthalten, wird sie dennoch annotiert.
In Zweifelsfällen hilft auch die Tabelle NoSta-D-TagSet und alle Untertabellen, insbesondere die Beispiele mit dem weiter.
Wenn nicht klar ersichtlich ist, ob eine NE NEderiv oder NEpart ist, weil sie beiden Klassen zugeordnet werden könnte, gilt folgende Regel: Steht eine vollständige NE in der Phrase, so ist wird es NEpart zugeordnet, ansonsten NEderiv.
Beispiel:
[deutschlandweit]LOCpart
[norddeutsche]LOCderiv Stämme
Wenn eine Kombination aus NEderiv und NEpart auftritt, so wird die Klassifizierung nach der ersten NE gemacht.
Beispiel:
[Linke-Europaabgeordnete]ORGpart
Jahreszahlen in ORGanisationen werden nicht markiert.
Beispiel:
[Fußball-WM]ORG 2006
[Eurovision Song Contest] 2013
Wenn der Eigennamen in KEINE der vorhandenen Klassen passt, dann markieren ihn mit ***UNCLEAR*** und notiere Dir bitte das Beispiel und schicke uns eine E-Mail an: xx@y.z. So können wir die Guidelines sukzessiv verbessern.

Wie finde ich Ableitungen von NEs?
Eigennamen, die durch morphologische Derivation in andere Wortarten überführt wurden, werden als solche markiert. NEderivs müssen keine vollen Nominalphrasen sein. Deklination in diesen Guidelines nicht als Derivation angesehen und deshalb als direkte NE annotiert.
Beispiel:
Ortsadjektive: *die [Bremer]LOCderiv Staatsanwaltschaft*
Personenadjektive: *die [Merkelsche]PERderiv Begeisterung für Europa*
ABER: Genitive: *[Deutschlands] LOC beste Fußballspieler*

Zu welcher semantischen Klasse gehört eine Ableitung?
Die Klasse setzt sich aus dem Tag der Klasse zusammen, in die der ursprüngliche Eigennamen gehört und dem Marker für die Ableitung "deriv".
Beispiel:
Ortsadjektive:
[Bremen]LOC
die [Bremer]LOCderiv Staatsanwaltschaft
Faustregel zur Unterscheidung einer Klassenbezeichnung und eines Namens:

- Elemente der fraglichen Einheit verbinden die gleichen Eigenschaften → Klasse → keine NE
- Christen glauben an Christus → Christ glaubt an Christus → keine NE
- Die Elemente der fraglichen Einheit verbindet nur der Name oder Element ist Einheit bezeichnet ein spezifisches Individuum → Name → NE
- Barock bezeichnet spezifische Epoche

ABER: die [Deutschen]LOCderiv

NoSta-D-TagSet

Sem. Subklasse	**Beispiele**
Person	*Hans Winkler*
Zuname	*(Familie) Feuerstein*
Tiernamen	*(Schweinchen) Babe*
Künstlernamen	*Madonna*
Charaktere	*Schneewitchen, Miss Piggy*
Nicknames	*Sternchen333*
Superhelden	*Batman*

Table 4: Semantische Klasse 'PER-Person'

Sem. Subklasse	**Beispiele**
Bezirke	*Schöneberg*
Sehenswürdigkeiten, Kirchen	*Brandenburger Tor,Johanniskirche*
Planeten	*Mars*
Landschafts-bezeichnungen	*Königsheide*
Straßen, Plätze	*Söogestraße, Alexanderplatz, A 5*
Einkaufszentren	Luisencenter, Allee-Center
Berge, Seen, Flüsse	*Alpen, Viktoriasee, Spree*
Kontinente	*Europa, Asien*
Länder, Staaten	*Frankreich, Hessen, Assyrien, USA*
Städte	*Berlin, Babylon*
Regionen	*Gazastreifen*

Table 5: Semantische Klasse 'LOC-Ort'[6]

[6]Die semantische Klasse LOCderiv enthält Ortsableitungen wie sie der semantischen Subklasse von Wettbewerben genutzt werden, wie beispielsweise das *[Deutsche]LOCderiv Meisterschaften*. Vorkommen spezifischer Wettbewerbe gehören zur Klasse ORG
[7]Ausnahme: Parlament
[8]Ausnahme: Frankfurter Flughafen
[9]Die semantische Klasse OTHderiv enthält die Subklasse modifizierter Sprachadjektive, wie zum Beispiel die Phrase *[hochdeutsche]OTHderiv Verben*
[10]Ausnahme: Götternamen

Sem. Subklasse	Beispiele
Organisationen	*Nato, EU, Landgericht Darmstadt, Bundesverwaltungsgericht, Weimarer Republik*[7]
Unternehmen	*Microsoft, Bertelsmann*
Flughäfen	*Fraport*[8]
Betreiber	*Lotto 6 aus 49*
Institute	*Institut für chinesische Medizin*
Museen	*Pergamonmuseum*
Zeitungen	*Süddeutsche Zeitung, Der Spiegel*
Clubs	*VfB Stuttgart*
Theater, Kinos	*Metropol-Theater, CinemaxX*
Festivals	*Eurovision Song Contest*
Austellungen	*Körperwelten*
Universitäten	*Technische Universität Darmstadt*
Rundfunksender	*Arte, Radio Bremen*
Restaurants und Hotels	*Sassella, Adlon*
Militäreinheiten	*Blauhelme*
Krankenhäuser, Pflegeheime	Charit, Klinikum Ingolstadt
Modelabels	*Chanel*
Sportereignisse	*Olympische Spiele, Wimbledon*
Festspiele	*Berlinale*
Bands	*Beatles, Die Fantastischen Vier*
Institution	*Bundestag*
Bibliotheken	Amerika Gedenkbibliothek
Parteien	*SPD, CDU*

Table 6: Semantische Klasse 'ORG-Organisation'

Semantische Subklasse	Beispiele
Betriebssysteme	*DOS*
Buch-, Filmtitel etc.	*Faust, Schlaflos in Seattle*
Kriege	*Zweiter Weltkrieg*
Politische Aktionen	*7. Bremer Protesttag gegen Diskriminierung*
Projektnamen	*Agenda 21*
Währungen	*Euro*
Marktindex	*Dow Jones, Dax*
Reihennummerierungen	*SZ-Magazin 41/07*
Sprachen	*Deutsch, Latein*
Buchtitel mittels Autor	Helbig et al. ([Helbig]PER et al.)OTH
Spiele	*Mensch-ärgere-dich-nicht, Halo*
Kunstwerke	*Mona-Lisa*
Epochen	*Barock, Romantik (auch Neubildungen: 'Neuzeit'))*
Webseiten	*www.ebay.de, google, www*
Sprachen	*Hochdeutsch, Englisch*

Table 7: Semantische Klasse 'OTH-Andere'[9]

Semantische Subklasse	Beispiele
Maßeinheiten	*Meter, Liter*
Religionen	*Christentum, muslimisch*[10]
Tiernamen	*Gepard, Schlange*
Bezeichner/Fachwörter	*Phosphat, Geodäten, Ikonen*[11]
Himmelsrichtungen	*südlich, Norden*
Mottos	*Carpe diem!*
Titel/Anrede	*Frau, König*
	Gott[12]
Dynastien und Geschlechter	*Habsburger, Wittelsbacher*[13]
Politische Strömungen	*Kommunismus, Sozialismus*

Table 8: Semantische Subklassen, die keine NEs sind

Regel	Beispiele	NE?
Klassen werden unabhängig von der semantischen Rolle im Kontext vergeben. ABER: Grammatische Hinweise entscheiden.	Nils Petersen geht a) zu [Bremen]ORG b) nach [Bremen]LOC Die [Wolfsburger]LOCderiv entwickeln Spitzentechnik. (eigentlich VW in Wolfsburg)	✓
Marken- oder Erfindernamen die als Universalbegriffe genutzt werden werden nicht als NEderiv markiert	Pampers, Tempo, Teddybär, Celsius, Watt, olympische	×
Klassen	Gepard-Klasse, A-Klasse	×
Ableitungen NEderiv werden nur dann annotiert, wenn sie den Stamm mit einer NE teilen.	die [decartessche]PERderiv Philosophie	✓
	die anglikanische Kirche	×

Table 9: Regeln zu NEs

Regel	Beispiele	NE
abgetrennt Kompositionsglieder	in [West-]LOC und ganz besonders in [Ost-Berlin]LOC [Adenauer-]ORG und [Böll-Stiftung]ORG	✓
Ortsteile	[West-Afrika]LOC [Nord-Berlin]LOC	✓
Adelstitel	Herr [von [Hohenzollern]LOC]PER Queen [Elisabeth]PER	✓
Gesetze	Art . 2 Nr. 18	×
Abkürzungen	*Amis, Sowjets*	×

Table 10: Formeln von NEs

Begriff	Sem. Klasse	Sem. Subklasse
Bundesliga	**ORG**	Organisationen
Creditreform-Mittelstandsindex	**ORGpart**	Unternehmen
Darmstadtium	**ORG/LOC**	Veranstalter / Veranstaltungsort[14]
Bibel	**OTH**	Buchtitel
Hotel Bellevue	**ORG**	Hotels
Milchstraße	**LOC**	Himmelskörper[15]
Evangelium	**keine NE**	Bezeichnung
Gott	**keine NE**	Bezeichnung
Polizei & Feuerwehr	**keine NE**	Gruppen
Indianer[16]	**keine NE**	Bezeichnung
Bundesregierung	**keine NE**	Bezeichnung
Weltmeisterschaft[17]	**keine NE**	Bezeichnung
ISBN	**keine NE**	Bezeichnung

Table 11: Einzelfälle

[11] Ausnahmen sind Produktbezeichnungen wie Aspirin
[12] Ausnahmen sind bestimmte Götternamen wie Vishnu
[13] Ausnahmen sind Stämme wie Maori
[14] kontextabhängig
[15] wie Planeten
[16] einzelne Stämme: ORG
[17] Bestimmte WM, zB. Fußball-WM: ORG