



TECHNISCHE
UNIVERSITÄT
DARMSTADT

RUHR
UNIVERSITÄT
BOCHUM

RUB



F-AG 7: Angewandte Sprachwissenschaft, Computerlinguistik Kurationsprojekt 2

Linguistische Annotation von Nichtstandardvarietäten — Guidelines und „Best Practices“

Guidelines NER

Version 1.6

Stand: 20.10.13

Marc Reznicek

Burkhard Dieterle

Darina Benikova



Kafka@WebAnno

Guidelines für die Named Entity Recognition. Sie bauen auf den Guidelines in den [STTS-Guidelines](#) (Schiller et al. 1999), dem [Stylebook for Tübingen Treebank](#) (Telljohann et al. 2012) und der [MUC-6 Named Entity Task Definition](#) auf.

Inhalt

Inhalt.....	2
Einführung: Named Entity Recognition.....	2
Wie finde ich eine NE?.....	3
Zu welcher semantischen Klasse gehört ein Eigennamen?.....	5
Wie finde ich Ableitungen von NEs?.....	5
Zu welcher semantischen Klasse gehört eine Ableitung?.....	6
Faustregel zur Unterscheidung einer Klassenbezeichnung und eines Namens.....	6
NoSta-D-TagSet.....	6
Bibliografie.....	17
Beispiele.....	18

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 (Gallmann 2009:149).

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.

*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: marc.reznicek@staff.hu-berlin.de. So können wir die Guidelines sukzessive verbessern.

Wie finde ich Ableitungen von NEs?

Schritt 1:

- 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

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

Semantische Klasse	Semantische Subklasse	Beispiele	Ausnahmen
PER - Person	Person	<i>Hans Winkler</i>	
	Zuname	<i>(Familie) Feuerstein</i>	
	Tiernamen	<i>(Schweinchen) Babe</i>	
	Künstlernamen	<i>Madonna</i>	
	Charaktere	<i>Schneewitchen, Miss Piggy</i>	
	Nicknames	<i>Sternchen333</i>	
	Superhelden	<i>Batman</i>	
ORG - Organisation	Organisationen	<i>Nato, EU, Landgericht Darmstadt, Bundesverwaltungsgericht, Weimarer Republik</i>	Parlament
	Unternehmen	<i>Microsoft, Bertelsmann</i>	
	Flughäfen	<i>Fraport</i>	Frankfurter Flughafen
	Betreiber	<i>Lotto 6 aus 49</i>	
	Institute	<i>Institut für chinesische Medizin</i>	
	Museen	<i>Pergamonmuseum</i>	
	Zeitungen	<i>Süddeutsche Zeitung, Der Spiegel</i>	
	Clubs	<i>VfB Stuttgart</i>	
	Theater, Kinos	<i>Metropol-Theater, CinemaxX</i>	
	Festivals	<i>Eurovision Song Contest</i>	
	Ausstellungen	<i>Körperwelten</i>	
	Universitäten	<i>Technische Universität Darmstadt</i>	
	Rundfunksender	<i>Arte, Radio Bremen</i>	
	Restaurants und Hotels	<i>Sassella, Adlon</i>	

	Militäreinheiten	<i>Blauhelme</i>	
	Krankenhäuser, Pflegeheime	<i>Charité, Klinikum Ingolstadt</i>	
	Modelabels	<i>Chanel</i>	
	Sportereignisse	<i>Olympische Spiele, Wimbledon</i>	
	Festspiele	<i>Berlinale</i>	
	Bands	<i>Beatles, Die Fantastischen Vier</i>	
	Institution	<i>Bundestag</i>	
	Bibliotheken	<i>Amerika Gedenkbibliothek</i>	
	Parteien	<i>SPD, CDU</i>	
LOC - Ort	Bezirke	<i>Schöneberg</i>	
	Sehenswürdigkeiten, Kirchen	<i>Brandenburger Tor, Johanniskirche</i>	
	Planeten	<i>Mars</i>	
	Landschafts-bezeichnungen	<i>Königsheide</i>	
	Straßen, Plätze	<i>Söogestraße, Alexanderplatz, A 5</i>	
	Einkaufszentren	<i>Luisencenter, Allee-Center</i>	
	Berge, Seen, Flüsse	<i>Alpen, Viktoriasee, Spree</i>	
	Kontinente	<i>Europa, Asien</i>	
	Länder, Staaten	<i>Frankreich, Hessen, Assyrien, USA</i>	
	Städte	<i>Berlin, Babylon</i>	
	Regionen	<i>Gazastreifen</i>	
LOCderiv - Ortsableitungen	Wettbewerbe	<i>Deutsche (Meisterschaft)</i>	Wenn spezifisch, dann ORG.
OTH - Andere	Betriebssysteme	<i>DOS</i>	
	Buch-, Filmtitel etc.	<i>Faust, Schlaflos in Seattle</i>	
	Kriege	<i>Zweiter Weltkrieg</i>	
	Politische Aktionen	<i>7. Bremer Protesttag gegen Diskriminierung</i>	
	Projektnamen	<i>Agenda 21</i>	
	Währungen	<i>Euro</i>	
	Marktindex	<i>Dow Jones, Dax</i>	
	Reihennummerierungen	<i>SZ-Magazin 41/07</i>	
	Sprachen	<i>Deutsch, Latein</i>	
	Buchtitel mittels Autor	<i>Helbig et al. ([Helbig]_{PER} et al.)_{OTH}</i>	
	Spiele	<i>Mensch-ärgere-dich-nicht, Halo</i>	
	Kunstwerke	<i>Mona-Lisa</i>	
	Epochen	<i>Barock, Romantik (auch Neubildungen: „Neuzeit“)</i>	
	Webseiten	<i>www.ebay.de, google, www</i>	
	Sprachen	<i>Hochdeutsch, Englisch</i>	
OTHderiv – Ableitung von OTH	modifizierte Sprachadjektive	<i>hochdeutsche Verben</i>	
	Chatraum	<i>Katzenkörbchen, Erdbeerworld</i>	

Keine NE	Datums- und Zeitangaben	<i>Montag, April, Feiertage</i>	
	Maßeinheiten	<i>Meter, Liter</i>	PERderiv: Celsius
	Religionen	<i>Christentum, muslimisch</i>	Götternamen
	Tiernamen	<i>Gepard, Schlange</i>	

	Bezeichner/Fachwörter	<i>Phosphat, Geodäten, Ikonen</i>	Produkt-namen: Aspirin
	Himmelsrichtungen	<i>südlich, Norden</i>	Bestimmte Götter: Vishnu
	Mottos	<i>Carpe diem!</i>	
	Titel/Anrede	<i>Frau, König</i>	
		<i>Gott</i>	
	Dynastien und Geschlechter	<i>Habsburger, Wittelsbacher</i>	
	Politische Strömungen	<i>Kommunismus, Sozialismus</i>	Stämme: Moari

Regeln

Regel	Beispiele	NE
Klassen werden unabhängig von der semantischen Rolle im Kontext vergeben. ABER: Grammatische Hinweise entscheiden.	Nils Petersen geht ... zu [Bremen] ORG 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	✓
	<i>die anglikanische Kirche</i>	✗

Formen

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

Einzelfälle

Begriff	Semantische Klasse	Semantische Subklasse	Kommentar
<i>Bundesliga</i>	ORG	Organisationen	kontextabhängig
<i>Creditreform-Mittelstandsindex</i>	ORGpart	Unternehmen	
<i>Darmstadtium</i>	ORG/LOC	Veranstalter / Veranstaltungsort	
<i>Bibel</i>	OTH	Buchtitel	wie Planeten
<i>Hotel Bellevue</i>	ORG	Hotels	
<i>Milchstraße</i>	LOC	Himmelskörper	
<i>Evangelium</i>	keine NE	Bezeichnung	
<i>Gott</i>		Bezeichnung	
<i>Polizei & Feuerwehr</i>		Gruppen	

<i>Indianer</i>	Bezeichnung	einzelne Stämme : ORG
<i>Bundesregierung</i>	Bezeichnung	
<i>Weltmeisterschaft</i>	Bezeichnung	Bestimmte WM, zB. Fußball-WM: ORG
<i>ISBN</i>	Bezeichnung	

Bibliografie

- **Schiller, Anne; Teufel, Simone; Stöckert, Christine; Thielen, Christine (1999):** Guidelines für das Tagging deutscher Textkorpora mit STTS. Technical Report. University of Stuttgart; University of Tübingen. [BibTex](#)
- **Telljohann, Heike; Hinrichs, Erhard W.; Kübler, Sandra; Zinsmeister, Heike; Beck (2012):** Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Seminar für Sprachwissenschaft. Universität Tübingen, Germany. <http://www.sfs.uni-tuebingen.de/resources/tuebadz-stylebook-1201.pdf>. [BibTex](#)
- **MUC-6 Named Entity Task Definition:**
http://cs.nyu.edu/faculty/grishman/NEtask20.book_1.html

Beispiele

Im Musikbereich vertritt das **Bureau International des Sociétés gérant les Droits d'Enregistrement et de Reproduction Mécanique (BIEM)** die Interessen der Urheber im mechanischen Recht.

Der Deutsche wird weiterhin in einem Bergversteck im Süden der Provinz Wardak, südwestlich von Kabul, von einer Gruppe um den Anführer Nissam Udin festgehalten.

Während des 17-minütigen Testfluges des **RAF-Testpiloten Gerry Sayer** erreichte die Maschine eine Höchstgeschwindigkeit von 545 km/h.

BIBERACH (cc) Nach dem Sieg im Derby gegen Bad Wurzach soll die Leistungskurve der Landesliga-Volleyballer der SG Biberach-Ochsenhausen weiter nach oben zeigen.

Eine weitere Hauptrolle erhielt Lexa Doig im Fantasy-Drama **Second Sight** (2007) und in der zweiten und dritten Staffel von **EUREKA – Die geheime Stadt** einen Gastauftritt.

PKN Orlen gaben 1,6 % nach und **BZ WBK** 1,4 %.

Johannesburg - 14. August - Die Satzung der Afrikanischen Union (AU) sieht vor, dass durch einen Militärputsch an die Macht gekommene Staatschefs nicht anerkannt werden.

Dass Rathold ursprünglich die Gründung eines Klosters beabsichtigte, ist weder be- noch widerlegbar: Möglicherweise wollte er zunächst bloß vermeiden, dem bayrischen Heerbann gegen die Ungarn Folge leisten zu müssen.

Den Nürnbergern, die defensiv zudem sehr gut agierten, kamen die Schwächen des Meisters gelegen, besonders **Simon Fischhaber** (30.), **Greg Leeb** (43.), **Adrian Grygiel** (45.) und **André Savage** (60.).

(**Gerhard Johann David von Scharnhorst**: Nutzen der militärischen Geschichte).

Nürnberg (dpa) Torjäger **Robert Vittek** ist nach fünfwöchiger Verletzungspause ins Mannschaftstraining des Fußball-Bundesligisten **1. FC Nürnberg** zurückgekehrt.

Schwarz: Ich bin eine Leseratte, habe "Das Parfum" mit 13 regelrecht verschlungen und sofort an Freunde weitergereicht.

Das sagte telegate-Vorstandschef **Andreas Albath** der "Financial Times Deutschland" (Mittwochausgabe).

Sprintwertung Frauen Sprintwertung Männer Spitzenreitertrikots Der aktuelle Gesamtführende des Skilanglauf-Weltcup-Finals trägt das Gesamt-Spitzenreiter-Trikot.

