08 Named Entities

Named entities (NEs) are nouns and noun phrases that specifically designate a person, location, organisation or other distinct object existing in real space and time, In a broader sense, they can also include (possessive) adjectives derived from a person's name, such as DERIV-PER[Obamova] izvolitev). In Slovene, named entities are typically indicated orthographically by capitalization (e.g., "Slovenska tiskovna agencija") or abbreviations (e.g., "STA"). It's important to note, however, that a capital letter or an abbreviation doesn't always signify a named entity (e.g. the Slovene acronym BDP, translated to 'GDP' in English, represents a common noun phrase). The ability to accurately identify named entities in text plays a crucial role in numerous natural language processing tasks, including information extraction, coreference resolution, sentiment analysis, and more.

Introduction to Labels

This chapter summarises labels for named entities (NEs). A more detailed presentation can be found in the guidelines in the Annotation Guidelines chapter.

Category	Subcategory	Examples	Doesn't belong in the category
PER	Person (name and/or surname)	Janez Novak, da Vinci, Ludvik XIV.	dr., gospa, sv.
	Pet name	Fifi	
	Artistic name, pseudonym	Madonna, mati Terez(ij)a, Banksy	
	Fictional characters (from books, films etc.)	Ana Karenina, Rdeča kapica	
	Nicknames	(Boštjan Gorenc -) Pižama, Zvezdica89	
	Named group of people (placerelated or family name)	Angleži, Nemec, Ljubljančan; Novakovi	
	Twitter mentions	@pizama, @Nike	
DERIV-PER	Personal possessive adjectives	Novakov (pes)	Alzheimerjeva (bolezen)
ORG	Organizations	EU, Nato, Rimskokatoliška cerkev	parlament, vlada
	Companies	Microsoft, Pasadena d.o.o.	
	Airport operators	Aerodrom Ljubljana	Letališče Jožeta Pučnika
	Educational institutions	Filozofska fakulteta	
	Institutes	Institut “Jožef Stefan”	
	Museums, libraries	Prirodoslovni muzej	
	Theatres, cinemas etc.	MGL, Kinodvor	
	Media (TV, radio, newspaper etc.)	Dnevnik, Delo, Radio Center	
	Restaurants, hotels, bars, pubs etc.	Kavarna Zvezda, [hH]otel Lev	
	Healthcare facilities	[zZ]dravstveni dom Ribnica	
	Music bands and other art-related groups	U2, Beatli, [aA]nsambel Avsenik	

Category	Subcategory	Examples	Doesn't belong in the category
	Other public and private institutions	[oO]bčina Piran, NPK	
	Political parties, civic societies, NGOs	DeSUS, Zveza potrošnikov Slovenije	
	Sports clubs, associations	(HDD SIJ) Acroni Jesenice, (FC) Barcelona	
	Cultural organizations (also amateur)	[mM]ešani pevski zbor Divača	
LOC	Celestial bodies (planets, comets etc.)	Mars, Andromeda, Halleyjev komet	
	Continents	Južna Amerika	
	Countries, provinces, lands (historic and modern)	Slovenija, Združene države (Amerike)	EU
	Regions	Primorska, Valonija, Nova Anglija	
	Cities and settlements (including parts)	Ljubljana, Šiška, Vrhnika, Na klancu	
	Streets, squares	Jamova cesta 39	A2, gorenjska AC
	Shopping centres	Citypark, Supernova	
	Airports	Letališče Jožeta Pučnika	
	Churches (named building)	[cC]erkev sv. Nikolaja	Rimskokatoliška cerkev
	Local sights (cultural, natural)	Tromostovje, Triglavski narodni park	
	Other named buildings (without org. structure)	[kK]ulturni dom Ljubno, WTC 2	Cankarjev dom (ima org. strukturo, npr. direktorja)
	Mountains, lakes, rivers and other named geographical objects	Triglav, Blejsko jezero, Sava, Logarska dolina	
MISC	Computer systems, programs, apps	Windows 10, Word, Android 5.1 Lollipop	.docx, pdf, OCR
	Titles of books, films, paintings and other works of art; titles of documents	Vojna in mir, Ko jagenjčki obmolknejo, Sopranovi, Guernica; Uradni list RS	
	Registered names or models of products (cars, mobile phones, computers, games etc.) and other commercial products (brands)	Galaxy Note 7, Nokia Lumia 950, Toyota RAV4, Minecraft, Človek ne jezi se	

Category	Subcategory	Examples	Doesn't belong in the category
	Titles of events	Oskarji, Zlata lisica, 10. mednarodna konferenca Jezikovne tehnologije	shod nacifašistov
	Project names	Obzorje 2020	
	Stock market indices	SBI20, Dow Jones, Nasdaq	Bonitetne ocene (AAA)

Annotation Guidelines

This chapter summarizes the annotation guidelines for named entity recognition (NER) as applied to Slovene texts. The guidelines are arranged from the latest, up-to-date version to the oldest version.

Version 1.1

Project Development of Slovene in a Digital Environment
ZUPAN, Katja; LJUBEŠIĆ, Nikola in ERJAVEC, Tomaž, 2023: *Annotation guidelines for Slovenian named entities Janes-NER*: Version 1.1. Clean copy for the Development of Slovene in a Digital Environment project. [PDF]

References and Links

This chapter compiles relevant references and provides links to projects where named entity recognition (NER) has been developed and applied to Slovene texts.

Projects, in which the system has been developed

MUC-6 Named Entity Task Definition

CONLL 2003

BSNLP 2017 shared task

Janes - Resources, Tools and Methods for the Research of Nonstandard Internet Slovene

References

Marc Reznicek: *Linguistische Annotation von Nichtstandardvarietäten / Guidelines und „Best Practices*" *Guidelines NER* (version 1.5).

https://www.linguistik.huberlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-ner-1.5

LDC - Linguistic Data Consortium: ACE (Automatic Content Extraction) English Annotation Guidelines for Entities, Version 6.6 2008.06.13, http://projects.ldc.upenn.edu/ace (Accessed on 2 November 2020).