Annotation guidelines for Slovenian named entities Janes-NER

V1.1, 2017-09-18

Katja Zupan, Nikola Ljubešić, Tomaž Erjavec

[Changes from V1.0: added – subcategory "Twitter mentions" in the typology, point 1.4 to Sec. 4 and "hashtags" to 5.2]

Five categories of Named Entities (NEs) are distinguished:

- person, PER
- person derivative, DERIV-PER
- location, LOC
- organization, ORG
- miscellaneous, MISC

These are all standard types of NEs, except for DERIV-PER, which we introduce to mark personal (possessive) adjectives in order to enable better anonymization of personal data.

WHAT TO ANNOTATE AS A NAMED ENTITY?

Basic principle: **nouns and noun phrases** that identify a certain person, location, organization or other real world object, or in cases of person derivatives, **personal (possessive) adjectives** derived from a person name that refer to the person in question, e.g. DERIV-PER[Obamova] izvolitev.

At the level of orthography, NEs are most often written with **capitalized initial letters** (Slovenska tiskovna agencija) or as an **acronym** (STA) in all caps. However, not all capitalised words or acronyms are NEs (e.g. BDP/GDP). Noun phrases that refer to the same entity but whose spelling varies are to be treated in the same way, regardless of capitalisation, e.g. ORG[Ministrstvo za kulturo Republike Slovenije] = ORG[ministrstvo za kulturo]. However, common noun phrases should not be marked as NE, even if similar in spelling, e.g. not ORG[kulturno ministrstvo].

1 BASIC RULES

1. Only one category should be used for each NE.
2. Longer units are preferred to shorter ones, e.g. LOC[Jamova 39, Ljubljana] and not LOC[Jamova 39], LOC[Ljubljana].
3. In case of nested NEs, only the top-most entity should be annotated, e.g. ORG[Ministrstvo za kulturo Republike Slovenije], and not ORG[Ministrstvo za kulturo LOC[Republike Slovenije]].
4. If the first element of a noun phrase is a common noun (e.g. letališče, mejni prehod, občina, osnovna šola, zavod), it is to be annotated as part of the NE *only* if it is used as part of the official name of the NE (could thus be written with a capitalized initial according to the standard orthography) and separates one NE from another NE, e.g., separate locations of transport infrastructure vs. location of a settlement as in Peljem se na LOC[letališče/Letališče Portorož] vs. Počitnikujem v LOC[Portorožu].

2 WHICH CATEGORY TO USE

1. The main guideline is the Typology of NEs given in Sec. 3, with subcategories serving as further explanation of which types of NEs the category contains.
2. In cases of ambiguous NEs, typically the “default” or “primary” category of a NE should be chosen, so countries are categorised as locations, even if they are used as organisations (e.g. LOC[Slovenija] je zaprosila za članstvo v Natu), and institutions are categorised as organizations even if they are used as locations, e.g. Državno tekmovanje je potekalo na ORG[osnovni šoli Franceta Prešerna]).
3. In cases where it is difficult to choose the default interpretation, the context of the NE should be considered, e.g. if the name of a literary character (PERS) is also the title of a literary work (MISC), as in PERS[Faust] išče smisel življenja vs. OTH[Faust] je dramska pesnitev.
4. If a NE does not match any of the listed subcategories, MISC should typically be used.
5. If there is good reason, the NE can also be annotated as PER, ORG or LOC.

3 TYPOLOGY OF NAMED ENTITIES

Category	Subcategory	Examples	Does not belong
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, Buda*	
	Nicknames	*(Boštjan Gorenc -) Pižama, Zvezdica89*	
	Named group of people (place-related 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*	

	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*	
	Shopping centres	*Citypark, Supernova*	
	Airports	*Letališče Jožeta Pučnika*	*Aerodrom Ljubljana*
	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 (has org. structure, e.g. a director)*
	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*	
	Titles of events	*Oskarji, Zlata lisica, 10. mednarodna konferenca Jezikovne tehnologije*	*shod nacifašistov*
	Project names	*Obzorje 2020*	
	Stock markets	*SBI20, Dow Jones, Nasdaq*	

4 ADDITIONAL NOTES FOR EACH CATEGORY

1. PER
   - 1.1. Person names should not include titles, honorifics, and functions/positions.
   - 1.2. Initials, pseudonyms and prepositions in surnames (e.g. van, da, von), however, are treated as NEs.
   - 1.3. Named references to place-related and family-related groups of people (inhabitants of a LOC (e.g. Slovenci, Korošci, Angleži, Londončani) and family names (e.g. Novakovi)) are also PER NEs. However, adjectives are part of the NE only when capitalised (e.g. koroški Slovenci, Beneški Slovenci).
   - 1.4. Mentions of Twitter user accounts (@XYZ) should be annotated as PER in all cases, even if they refer to organisations.
2. DERIV-PER
   - 2.1. This category marks personal possessive adjectives and should thus be seen as an umbrella term for all descriptive references to a person, whenever they “reveal” information about this person, even if the person is a fictional one, e.g. PER[Faustovo] mnenje. If, however, something (disease, award etc.) was merely named after a certain person (e.g. Alzheimerjeva bolezen) rather than being the name of the person suffering from this disease, it should not be annotated.
3. LOC
   - 3.1. Key identifiers of LOC are *entities that are naturally or artificially situated in geographic space and that are, as such, often marked on maps*.
   - 3.2. Countries are always treated as location to avoid difficulties in interpreting geopolitical contexts and metonymy.
   - 3.3. However, if a sports club, school or similar is named after a location, it should be annotated as ORG (e.g. ORG[Barcelona] je premagala ORG[Madrid] in ORG[Bayern]).
4. ORG
   - 4.1. The key identifier of ORG is the *presence of some organizational structure, i.e people who officially preside/manage an NE*.
   - 4.2. However, common noun phrases related to management bodies should not be annotated as part of NE unless they are capitalised, as they do not separate one organization from another organization (e.g. uprava ORG[Gorenja], občinski svet ORG [občine Piran]).
   - 4.3. If the spelling does not use capital initials (e.g. ORG[evropska komisija]) or part of the official name is omitted (e.g. ORG[ministrstvo za kulturo] rather than the complete ORG[ministrstvo za kulturo Republike Slovenije]), it should still be annotated as an NE provided the reference is to a specific entity. If the reference is to a general description of a type of entity (common NP) or ambiguous, which is typically manifested in adjectival shift (e.g. Piran > piranski; greater deviation from the official name) and no capital initial is used, it should not be annotated (e.g. kulturno ministrstvo, piranska občina).
5. MISC
   - 5.1. Latin names for plants and anatomical parts are not NEs.
   - 5.2. URLs, email addresses and hashtags should not be annotated.

APPENDIX

- What about countries that do not exist anymore (e.g. Jugoslavija). Are they still to be annotated as loc? - **Yes.**
- Are holidays (e.g. Christmas) named entities? If so, I assume it falls under the Misc. category, correct? - **Yes.**
- Are entities such as Twitter or Facebook (seen in the context “Facebook page”) annotated as org or misc (given that they’re companies, but also apps)? - **The primary/default meaning is ORG.**
- What about the scope of annotating phrases such as "zapadna Hrvatska” - is it [zapadna Hrvatska] or zapadna [Hrvatska]? - **Uness it is an official geographical entity, written as 'Zapadna Hrvatska', then annotate as 'zapadna [Hrvatska]'**
- In a similar vein, a type of phrase that often pops up is “riječko Sveučilište” - what is to be annotated here? [riječko Sveučilište], riječko [Sveučilište] or neither? - **Conflicting feedback, but let's also go with 'riječko [Sveučilište]'**
- Double-checking, we only annotate explicit/surface named entities, right? So if there is a common noun in the sentence (like 'company') that refers to a specific named entity (in the text or in the world), this is not to be annotated, correct? - **Yes, do not annotate this.**
- If a named entity is somehow broken up, (e.g. "Europska je Unija donijela odluku..."), given that there's no option for a partial annotation, which way is preferable? [Europska je Unija] or [Europska] je [Unija]? - **Difficult to say, but let's go with [Europska je Unija].**
- "kosovski Srbi" - is it [kosovski Srbi] or kosovski [Srbi]? - **We'll go with kosovski [Srbi].**

REFERENCES

- Marc Reznicek: **Linguistische Annotation von Nichtstandardvarietäten** — Guidelines und „Best Practices" Guidelines NER (version 1.5): https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/nosta-d/nosta-d-ner-1.5
- **MUC-6 Named Entity Task Definition**: http://cs.nyu.edu/faculty/grishman/NEtask20.book_1.html
- **CONLL 2003**: http://www.cnts.ua.ac.be/conll2003/ner/annotation.txt
- **BSNLP 2017 shared task:** http://bsnlp-2017.cs.helsinki.fi/shared_task.html