

University of West Bohemia

Faculty of Applied Sciences

Czech Historical Named Entity Corpus v 1.0

Annotation Manual

Helena Hubková, Pavel Král

(c) Copyright 2019 Department of Computer Science & Engineering and New Technologies for the Information Society of the University of West Bohemia in Pilsen, Czech Republic.

These resources are licensed under the Attribution-NonCommercial-ShareAlike 3.0 Unported License. Commercial use in any form is excluded.

For more information, please see

<http://creativecommons.org/licenses/by-nc-sa/3.0/>

Pilsen 2019

Návod pro anotaci pojmenovaných entit v historických textech

(English version is following)

V historických článcích z novin *Posel od Čerchova* z roku 1872 budeme anotovat pojmenované entity (NE – named-entity). Texty byly nejprve naskenovány a poté metodou OCR (Optical Character Recognition) převedeny do počítačové textu.

Pokyny pro anotování:

1. Rozlišujeme 6 základních tagů:

p: personal names
i: institutions
g: geographical names
t: time expressions
o: artifact names „objekty“
a: ambiguous „nevím, kam zařadit“.

Co pod tyto skupiny patří, můžete najít v přehledu níže.

2. Na co si dát pozor:

a. entita může být napsána i s malým písmenem,
b. pokud je součástí entity interpunkce – jedná se o zkratku, pak označujeme vše včetně interpunkce (*14 . t . m .* znamená „toho měsíce“)
c. označujeme jak entitu *Chodské náměstí*, tak *náměstí Chodské*
d. pokud si nejsem jistý, kam entita patří, nebo zda se jedná o entitu, značím **a** : ambiguous - nejednoznačné

Rozlišujeme **6 základních** NE (named-entity) skupin:

1. PERSONAL NAMES (jména osob)

značíme malým písmenem p

- křestní jména a/nebo příjmení (*Antonína, Kostlivýho, Antonína Kostlivýho, Julie M. Prushaková*)
- umělecká jména, přídomky a přezdívky (*Havlíček Borovský, Karel Hájek z Libočan, Mnich sázavský, Rudolf ze Stadionu, Jestřáb*, atd.)
- (akademické) tituly (*Med et Chir., Mistr, Dr., MUDr., Ing., ...*, tedy celé: *Dr. Antonín Marčan, Sv. i svatý Vojtěch*)
- jména panovníků a historických postav (*Jan Lucemburský, Karel IV.* (označujeme včetně tečky), *Sámo*, atd.)
- jména rodů a rodin (*Lucemburkové, Habsburkové, Novákovi*, atd.)
- jména mýtických a literárních postav (*Přemysl Oráč, Švejk*, atd.)

2. INSTITUTIONS, (názvy institucí a organizací)

značíme malým písmenem i

- názvy institucí (*Československá obchodní akademie v Praze, Jenerální zastupitelství rakouského ústředního stavitelského spolku ve Vídni, C. k. vlastenecko-hospodářská společnost, občanská škola domažlická, Universita Odesská, Finanční výbor ve*

Vídni, archiv Domažlický, pozor: zkratka C. k. = "císařsko-královský" často na začátku instituce)

- názvy spolků a klubů a politických shromáždění (*Sbor ostrořelecký, Stavitelský spolek ve Vídni, Sokol, Sokol domažlický, hasičský sbor Domažlice, haličský sněm, atd.*)
- názvy podniků, obchodů a pohostinství (*Cukrovar Domažlice, Umělecký parní mlýn v Havlovicích, hostinec u zlaté koruny, knihkupectví Jiř. Prunara atd.*)
- označení kolektivů (*benediktini, husité, republikané, atd.*)

3. GEOGRAPHICAL NAMES, (geografické názvy)

značíme malým písmenem g

- názvy kontinentů a států včetně historických (*říše rakouská, Evropa, Čechy, habsburská monarchie, Rakousko-Uhersko, atd.*)
- názvy územně-správních jednotek včetně historických (*panství Koutského a Trhanovského, Plzeňský kraj, okres podbořanský, Domažlicko, Bavorsko, atd.*)
- názvy měst, obcí a jejich částí (*Pešť, Varšava, Plzeň, Horšův Týn, Nové Kdyně, atd.*)
- názvy ulic a veřejných prostranství (*poštovská ulice, dolejší předměstí v Domažlicích, Chodské náměstí, hradskou ulicí, Tomanova ulice, atd.*)
- pomístní názvy (*Svaté Dobrotivé, Na Hrázi, Pod Starým hradem, atd.*)
- názvy přírodních útvarů (*vrch sv. Anny, dolehy strouberských, Šumava, Mže, Úhlava, kopec Pohorí, atd.*)

4. TIME EXPRESSIONS (časové údaje)

značíme malým písmenem t

- čas (*12:00, v půl jedné odpůldne, atd.*)
- denní data (*6. 2. 2019, 6. února 1872, středa 5. srpna, 18. t. m. (znamená tohoto měsíce), 12. t. r. (znamená tohoto roku), 22. - 23. srpna, atd.*)
- letopočty (*1872, MCCCLXXI, 1654, 2019, včetně 1037-1055, atd.*)
- století (*6. století př. n. l., 18. století, osmnácté století, 650 po Kristu, atd.*)
- časová období (*novověk, středověk, raný novověk, moderní doba, atd.*)
- období dle uměleckých směrů (*gotika, baroko, barokní, funkcionalismus, atd.*)
- svátky a významné dny (*Boží hod vánoční, Velikonoce, svátek Všech svatých, den sv. Josefa, svatojánská pouť, atd.*)
- názvy dějinných událostí (*bitvě na Bílé hoře, Pražská defenestrace, bitva u Slavkova, atd.*)
- názvy oficiálních opakujících se událostí (*Mezinárodní filmový festival Karlovy Vary, Vsesokolský slet, atd.*)

5. ARTIFACT NAMES, (označení objektů, produktů, dokumentů a staveb)

značíme malým písmenem o

- významné dokumenty (*Zlatá bula sicilská, Kutnohorský dekret, Charta 77, atd.*)
- umělecká díla (*Hej Slované, opera Drahomíra, Vyšebrodský oltář, Krajinka v zimním hávu, Malá noční hudba, atd.*)
- názvy produktů (*Turecké železniční losy, Uherské prémiové losy, Kladrubské pivo, atd.*)
- knihy, časopisy, edice ad. tiskoviny (*Posel od Čerchova, Osvěta, Svoboda, Životy posledních Rožmberků, Ženská bibliothéka, Slovanský kalendář Ottův, Monumenta Egrana, Minulostí západočeského kraje, Pilsner Tagblatt, atd.*)

- stavební objekty konkrétní (věž u svatých, *kostel sv. Bartoloměje, zámek Kozel, klášter benediktinský u Davle, hrad Domažlický*, atd.)
 - názvy měn (*kr, zl, zl. r. č.* = zlatých rakouského čísla, *kr. r. m.* = krejcarů rakouské měny, *zlatých, tolar, krejcarů* atd.)
6. AMBIGUOUS, NEJEDNOZNAČNÉ
značíme malým písmenem a
- neumím rozlišit
 - cokoliv, o čem jsem přesvědčená/y, že je pojmenovaná entita, ale nejsem schopen/schopna určit, do které kategorie výše patří

Annotation manual for NER in Czech historical texts

In the historical articles from the newspaper called *Posel od Čerchova* (1872) newspaper, we will annotate named entities (NE). The texts were first scanned and then converted to digitized text using OCR method (Optical Character Recognition).

We distinguish 6 basic tags:

- p**: personal names
- g**: geographical names
- i**: institutions
- t**: time expressions
- o**: artifact names „objects“
- a**: ambiguous (I believe it is a NE but I am not sure which one)

1. PERSONAL NAMES [p] include:

- first names and surnames (*Antonína, Kostlivýho, Antonína Kostlivýho, Julie M. Prushaková*, etc.),
- artistic names and nicknames (*Havlíček Borovský, Karel Hájek z Libočan, Mnich sázavský*, etc.),
- (academic) titles (*Med et Chir., Mistr, Dr., MUDr., Ing.*, etc.),
- royal (family) names, family names and names of historical persons (*Jan Lucemburský, Karel IV., Sámó, Lucemburkové, Novákoví*, etc.),
- names of mythical and literary characters (*Přemysl Oráč, Švejk*, etc.).

2. INSTITUTIONS [i] include:

- names of institutions (*Československá obchodní akademie v Praze, Jenerální zastupitelství rakouského ústředního stavitelského spolku ve Vídni, C. k. vlastenecko-hospodářská společnost, občanská škola domažlická, Universita Odesská, Finanční výbor ve Vídni*, etc.),
- names of organizations and clubs (*Sbor ostrožřelecký, Stavitelský spolek ve Vídni, Sokol, Sokol domažlický, hasičský sbor Domažlice*, etc.),
- names of companies and shops (*Cukrovar Domažlice, Tatra, knihkupectví Jiř. Prunara* etc.),

- names of historical collectives, e.g. religious orders, political parties (*benediktini, husité, republikané*, etc.).

3. GEOGRAPHICAL NAMES [g] include:

- names of continents and (historical) states (*říše rakouská, Evropa, Čechy, habsburská monarchie, Rakousko-Uhersko*, etc.),
- names of (historical) territorial-administrative units (*panství Koutského a Trhanovského, okres podbořanský, Domažlicko, Bavory*, etc.),
- names of towns and their parts (*Pešť, Varšava, Plzeň, H. Týn, Nové Kdyně*, etc.),
- streets and public places (*poštovská ulice, dolejší předměstí v Domažlicích, Chodské náměstí, hradskou ulicí, ulice Tomanova*, etc.),
- names of natural monuments including local names (*vrch sv. Anny, dolech strousberských, Šumava, Mže, kopec Pohoří*, etc.),
- local names (*Svaté Dobrotivé, Na Hrázi, Pod Starým hradem*, etc.).

4. TIME EXPRESSIONS [t] include:

- names of date (*6. 2. 2019, 6. února 2019, 18. t. m.*, etc.),
- names of hours (*12:00, v půl jedné*, etc.),
- names of years (*MCCCLXXI, 1654, 1872*, etc.),
- names of centuries (*6. století př. n. l., 18. století, osmnácté století, 650 po Kristu*, etc.),
- names of epochs (*novověk, středověk, raný novověk, moderní doba, gotika, baroko*, etc.),
- holidays and important days (*Boží hod vánoční, Velikonoce, svátek Všech svatých, den sv. Josefa*, etc.),
- historic events (*bitvě na Bílé hoře, Pražská defenestrace, bitva u Slavkova*, etc.).

5. ARTIFACT NAMES ("objects") [o] include:

- names of documents (*Zlatá bula sicilská, Kutnohorský dekret*, etc.),
- names of artworks (*Hej Slované, opera Drahomíra, Vyšebrodský oltář, Krajinka v zimním hávu, Malá noční hudba*, etc.),
- names of products (*Turecké železniční losy, Uherské prémiové losy*, etc.),
- names of books and newspapers (*Posel od Čerchova, Svoboda, Osvěta, Životy posledních Rožmberků, Minulostí západočeského kraje, Pilsner Tagblatt*, etc.),
- names of buildings (*věž u svatých, kostel sv. Bartoloměje, zámek Kozel, klášter benediktinský u Davle*, etc.),
- names of currency (*kr, kr., zl, zl., zlatých, tolar*, etc.).

6. AMBIGUOUS [a]:

I can't distinguish. Anything I believe is an entity, but I am unable to determine which category it is.