Návod pro anotaci pojmenovaných entit z historických textů

(doplňující materiály k zadání SP z předmětu KIV/UIR 2018/19)

V historických článcích z novin *Posel od Čerchova* z roku 1872 budeme anotovat pojmenované entity (NE – named-entity) [1]. Texty byly nejprve naskenovány a poté metodou OCR (Optical Character Recognition) převedeny do počítačové textu.

Pokyny pro anotování:

1. Nejprve přečíst zadání až do konce (především, co všechno se označuje, jak se daná entita značí, atd.)

2. Pro anotování využijeme online aplikaci: https://demo.lighttag.io/individual/#/, přeskočíme reklamu, poté zvolíme možnost „Write/paste Text“. Vložíme text, který budeme anotovat, potvrdíme tlačítkem „Submit“. Poté si v levém sloupci pomocí „Add a New Tag“ vložíme 6 základních NE + 1 speciální: pouze malá písmena, kterými dané skupiny značíme, tzn. **p**, **i**, **g**, **t**, **o**, **a, e** (reprezentují
   - **p**: personal names
   - **i**: institutions
   - **g**: geographical names
   - **t**: time expressions
   - **o**: artifact names „objekty“
   - **a**: ambiguous „nevím, kam zařadit”.
   - **e**: chyba v OCR - není NE, pomůže při “čištění” korpusu

   Co pod tyto skupiny patří, můžete najít v přehledu níže. **NUTNÉ** předem prostudovat a při anotování do přehledu nahlížet.

3. Co ignorovat:
   a. interpunkci (,.?!: atd.),
   b. to, že kontext nenavazuje,
   c. části textu, které jsou psány latinsky nebo německy.

4. Na co si dát pozor:
   a. entita může být napsána i s malým písmenem

[1] *„Za pojmenované entity jsou považována slova a slovní spojení, která v textu vystupují jako jména osob, geografické názvy, jména produktů, názvy organizací, ale také jako časové údaje apod. - tedy výrazy, které nemají apelativní (pozn. rozuměj obecný) význam, ale odkazují ke konkrétní osobě (např. příjmení Brousek), časovému úseku (např. číslo ve spojení rok 1959 nemá běžný kvantifikační význam, ale odkazuje ke konkrétnímu časovému úseku v trvání jednoho roku) apod.*“ (Ševčíková et al., *Zpracování pojmenovaných entit v českých textech*, 2007.)

b. pokud je součástí entity interpunkce – jedná se o zkratku, pak označujeme vše (*14 . t . m .* nebo *t . r .* znamená „toho měsíce“ nebo „toho roku“)
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

5. Pokud jsme s anotováním hotovi, stáhneme pomocí tlačítka v pravém rohu a uložíme pod názvem, který bude definován následujícím způsobem. (Původní název byl X.txt, kde X je Vaše přidělené číslo. Název anotovaného souboru bude X.lab)

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

1. PERSONAL NAMES (jména osob) [p]
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*)
- jména panovníků a historických postav (*Jan Lucemburský, Karel IV., Sámo, atd.*)
- jména rodů a rodin (*Lucemburkové, Habsburkové, Novákovi, Langobardi, atd.*)
- jména mytických a literárních postav (*Přemysl Oráč, Švejk, atd.*)

2. INSTITUTIONS, (názvy institucí a organizací) **[i]**
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, Městská rada, Universita Odesská, Finanční výbor ve Vídni*)
- názvy spolků a klubů (*Sbor ostrostřelecký, Stavitelský spolek ve Vídni*, *Sokol*, *Sokol domažlický, hasičský sbor Domažlice, atd.*)
- názvy podniků (*Cukrovar Domažlice*,*Tatra, atd.*)
- označení kolektivů (*benediktini, husité, republikané, atd.*)

3. GEOGRAPHICAL NAMES, (geografické názvy) **[g]**
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, Bavory, atd.*)
- názvy měst, obcí a jejich částí (*Pešť, Varšava, Plzeň, Horšův Týn, Nové Kdyně, Plzeň-Bory, Jižní Předměstí, Litice, 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, dolech strousberských, Šumava, Mže, Úhlava, kopec Pohoří, atd.*)

4. TIME EXPRESSIONS (časové údaje) **[t]**
značíme malým písmenem t

- čas (*12:00*, *v půl jedné, atd.)*
- denní data (*6. 2. 2019, 6. února 2019, 18. t . m. (znamená tohoto měsíce), atd.*)
- dny v týdnu (*středa, neděle, atd.*)
- měsíce (*únor, listopad, atd.*)
- letopočty (*MCCCLXXI, 1654, 2019, 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, atd.*)
- názvy dějinných událostí (*bitvě na Bílé hoře*, *Pražská defenestrace, bitva u Slavkova, atd.*)
- názvy oficiálních opakujících se událostí (*Mezinárodní filmový festival Karlovy Vary, Všesokolský slet, atd.*)

5. ARTIFACT NAMES, (označení **o**bjektů, produktů, dokumentů a staveb) **[o]**

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*)
- knihy, časopisy ad. tiskoviny (*Posel od Čerchova, Svoboda, Životy posledních Rožmberků, Monumenta Egrana, Minulostí západočeského kraje, Pilsner Tagblatt, atd.*)
- stavební objekty konkrétní (věž u svatých, *kostel sv. Bartoloměje, zámek Kozel, klášter benediktinský u Davle, atd.*)
- názvy měn (*Kr, zl, zlatých, tolar, atd.*)

6. AMBIGUOUS, NEJEDNOZNAČNÉ **[a]**

značíme malým písmenem a

- neumím rozlišit
- cokoliv, o čem jsem přesvědčená/ý, že je entita, ale nejsem schopen/schopna určit, do které kategorie výše patří