Schůzka Hičkok 23.1.2024

Přítomni: Václav Cvrček, Martin Stluka, Klára Pivoňková, Daniel Zeman, Anna Michalcová, Olga Navrátilová, Jiří Pergler, Ondřej Svoboda

1) **staročeské etalony**

Ondřej Svoboda: současný stav - řeší se ještě nějaké chyby v textech, hlavně počítačové zpracování – téměř vyřešeno

**Ondřej Svoboda: do 2 týdnů** dodá Danu Zemanovi pročištěná data pro etalony: 2x 100 tisíc tokenů k novočeské anotaci (100 tisíc ze 14.-15. stol. a 100 tisíc ze 16.-18. stol.)

Předpokládaný průběh anotace

Dan Zeman: současný stav - po prosincové přípravné fázi jsou scripty upravené pro převedení do formátu potřebného pro UD;

Předpokládaný průběh:

1. první anotace bude chybová (parser a tagger u neznámých tokenů se někdy pokusí vytvořit lemma, jindy dá do lemmatu samotný tvar)
2. po ruční korekci části oanotovaného materiálu (více zaměřený na nejstarší materiál, ale také alespoň menší část středněčeského) se natrénuje model a využije se pro novou verzi anotace

Pro následnou ruční práci s oanotovanými daty je dobré se pokusit, aby výsledná data prošla alespoň jednou za celý proces přes dva anotátory – ovšem v závislosti na časových možnostech

**Dan Zeman: konec února** – první anotace k dispozici k opravám

**Ondřej Svoboda: polovina března** – analýza problematických jevů, které se ukázaly v první anotaci: problematické jevy při přičleňování novočeského lemmatu diachronním podobám lexémů, ale dnes existujícím, seznam nalezených vymřelých slov a další. Analyzovaný materiál poslouží také jako podklad pro dořešení otázek typu kongruence přechodníků, lemmatizace proprií atd.

Nástroj pro práci s oanotovanými etalony

ÚJČ se spolu s Danem Zemanem rozhodne pro některý dostupný nástroj, který se případně dále přizpůsobí na míru – nabízí se několik možností:

- tabulkový procesor EXCEL plus/ openrefine / FEAT (využívaný v ÚČNK na desambiguaci etalonu pro 19.st.)

**Dan Zeman:**  napíše Jiřímu Hanovi, zda má a mohl by poskytnout zdrojová data pro FEAT;

**Jiří Pergler/Ondřej Svoboda:** po zohlednění všech pro a proti vyberou koncem února před opravami anotace anotační nástroj (po dohodě s Danem Zemanem)

2) LEMMATIZACE

Česká verze UD má momentálně už nastavena některá řešení – např. u agregátů není povrchový segment anotován, anotovány jsou jeho jednotlivé části, které UD rozsegmentuje, pokud to slovo pozná. Trpná příčestí (n-ová a t-ová participia) jsou značena jako adjektiva. Tokeny s pomlčkou jsou rozdělovány. Přechodníky anotovány podle současné kodifikace bez ohledu na shodu apod.

Václav Cvrček: Je žádoucí neodklánět se v projektu od všeobecné verze českého UD. Lze ale uvažovat o možnostech, které se zároveň s naším projektem do všeobecné verze UD následně aplikují.

Strategie lemmatizace pro Hičkok:

Obecně budou přidělována novočeská lemmata a lemmatizace bude jednoúrovňová, token bude mít jednu jednoznačnou interpretaci

**Václav Cvrček:** Pro tyto účely poskytne z Intercorpu (značkovaném v UD) seznam novočeských lemmat

Problematická otázka formy lemmatu u dnes neexistujících slov – buď lemma adekvátní nejnovější doložené podobě (míra náročnosti souvisí s počtem takových slov v korpusu a s ohraničením materiálu, ve kterém by se hledala nejnovější doložená podoba); nebo lemma rekonstruované do novočeského stavu (nevýhoda v potenciální nedohledatelnosti lemmatu pro situace, kdy uživatel buď slovo zná, pak ho zná jako staré a tak ho vyhledává)

Rozhodnutí na další schůzce na základě podkladů z analýzy první anotace - příklady dnes neexistujících slov

**Martin Stluka: na konci února** upřesní termín další schůzky

**Václav Cvrček:** vytvoří pracovní prostor na wiki ÚČNK přístupný jen po přihlášení členům týmu

**Jiří Pergler:** pro pracovní účely zprostředkuje možnost sdílet poznámky k dřívějšímu značkování staročeského Matoušova evangelia

V době zápisu je už hotovo:

1. Jiří Pergler: zprostředkovány poznámky k dřívějšímu značkování stč. Matoušova evengelia
2. Václav Cvrček: a) vytvořen pracovní prostor na Wiki; b) zpřístupněn seznam novočeských lemmat z Intercorpu (značkovaném v UD)