Schůzka Hičkok 20. 3. 2024

přítomni: Dan Zeman, Jiří Pergler, Anna Michalcová, Olga Navrátilová, Jana Zdeňková, Ondřej Svoboda, Klára Pivoňková, Václav Cvrček, Martin Stluka

- jádro a doplněk u staročeských textů domluven; KlarGlosA do doplňku, BiblOl Mt, Mc, L BiblDrážď jádro, J BiblOl jádro; MartKronT jádro

- první pokus anotace UD: 2 texty po dvojicích; cca. 350 rozdílů mezi anotátory; bude použit script, který ještě nebyl spuštěn (DZ) > možné další rozdíly a změny

- jmenná adjektiva (participia) budou mít pády

- u před retnicemi a jinde > bude mít lemma „u“ a patřičný tag specifikující pád

- problematika novočeského lemmatu u staré češtiny – bude (pro projekt chápáno lemma jen jako „zastřešující kód“); novočeské lemma dostanou jevy typu: a) souhláskové skupiny (čso > co, stkvít > skvět; pravedlnost > spravedlnost); b) haplologie (inhed > ihned); kvantita (pokáranie > pokárání); d) asimilace (spósob > způsob) atd.

- nelze víceré lemma, vždy vybrat jen jedno lemma (novočeské); v případě pochybností, jaké lemma vybrat, je rozhodující jeho frekvence v nové češtině s přihlédnutím k tomu, jak je to v UD

- komparativ a superlativ adj. a adv.: v případě přidělení stupně je lemma vždy v pozitivu; jmenné tvary mají složené lemma; rozlišujeme typy *dál* (došel jsem dál > komparativ a lemma daleko x dál o tom nemluv > pozitiv a lemma dál, pod. např. *více, dříve...*)

- zájmeno *veš* včetně všech svých slovotvorných odvozenin má lemma VŠECHEN

- příklonné částice typu *takež, takož* neznačíme, bude lemma *tak*; tyto částice zpracováváme pouze u substantiv > *přítelť* – bude rozděleno na 2 tokeny

- obecně supletivní tvary (člověk – lidé; rok - let) > využije se lemmatizace v UD (PDT)

- problém substantivizace adjektiv > bude adj., pokud není jasné, že to existuje jen jako substantivum (tzn. jinými slovy, pokud se to v různých kontextech může vyskytnout jako adjektivum i jako substantivum, např. *cestující matka nastoupila* a *cestující nastoupil* > bude to vždy adjektivum)

- *sám* a *samý* bude mít lemma *sám*; *onen* a *oný* bude lemma jen *onen* (vše jsou zájmena)

- trpná příčestí budou mít adjektivní lemma a tag (v něm bude specifikováno, že se jedná o příčestí)

- Dan zjistí, jak vypadá obecně systém UD anotace; např. jak vypadá lemmatizace osobních a osobních přivlastňovacích zájmen (časem zapíše na projektovou wiki)

- sklonná a nesklonná (absolutní) zájmena – řešeno na zájmenech *jeho, její, ješto, ježto, jenžto* – obecně asi existují ještě jiné případy, nicméně mustr k řešení bude i u nich obdobný;

a) zájmena *jeho* a *její* nesklonná – pokud to bude podle kontextu jistě nesklonné, tak v tagu bez kategorií pádu, rodu..., pokud sklonné, tak tam budou dané morfologické kategorie

b) vztažné *ješto* a *ježto* bude mít vždy lemma *jenžto*, pokud to bude na místech, kde je to součástí paradigmatu zájmena *jenžto* (v těchto případech s uváděním náležitých morf. kategorií > pád, rod...); vztažné *ježto* (*ješto*) pokud podle kontextu jistě nesklonné, bude mít lemma *ježto* a bez uvedení daných morf. kategorií; pokud to bude spojka, tak bude lemma *ježto* < *ježto* i *ješto*

Jaký je následný postup prací:

- ještě jednou dvojí kontrola anotace, vyřeší se ModlKunh, ŽaltU

- pak po sjetí scriptem (DZ) přijdou na řadu další texty pro srovnání

- kdy budou hotové desambiguace stč. textů: po ustálení na anotaci – letos nejméně 10 textů rozhodnutých (celkem jich je 80); zároveň natrénování něčeho, co bude mít lepší výsledky

- v srpnu bude navržena schůzka na září (v případě potřeby i jindy dříve) > domluva o a) ustálenosti a výsledcích stč. anotace; b) o možnosti postupného překlápění etalonu 19. stol. do UD