ING. VLADIMÍR BENKO, PhD.: KORPUSY ARANEA

Datum konání: 11. 12. 2020

Místo konání: online – Microsoft Teams

Název přednášky: Korpusy Aranea

Přednášející: Ing. Vladimír Benko, PhD.

Počet účastníků: cca 20

Zpracovali: Trnovec Ondřej

Ing. Vladimír Benko, PhD., se mezi jinými zabývá webovými korpusy, morfologickou anotací a lexikografií. V současnosti působí na Jazykovednom ústave Ľudovíta Štúra Slovenskej akadémie vied a jeho stránky jsou k nahlédnutí zde https://juls.savba.sk/~vladob/.

Na úvod přednášky pan doktor Benko stručně prošel definici a historii korpusů, na to navázal výhodami webových korpusů nad tradičními, mezi které patří menší problémy s autorskými právy, větší výběr textů či zachycení dynamických tendencí (neologismů v jazyku). Poté jsme se dozvěděli, jak funguje proces tvorby webových korpusů.

Po úvodu se přešlo k samotnému projektu Aranea, který vznikl v roce 2013 a jehož cílem je tvorba rodiny porovnatelných webových korpusů v jazycích, kterými se mluví nebo které se vyučují na Slovensku a v sousedních zemích, momentálně to je 22 jazyků. Samotný název projektu znamená v latině „pavučiny“ a latina se používá i pro názvy a velikosti korpusů. S korpusy lze pracovat přes NoSketchEngine http://unesco.uniba.sk/ nebo http://aranea.juls.savba.sk/. Lze zde nalézt například jeden z největších ruských korpusů a u několika jazyků se rozlišují u teritoriální varianty.

Po představení následovaly ukázky použití korpusů. Atribut *ztag* (momentálně pouze pro české korpusy) lze použít k vyhledání archaických slov či třeba zeměpisných názvů (vizte třetí „tahák“ na konci zprávy). Velmi užitečný může být i unikátní atribut *hlemma*

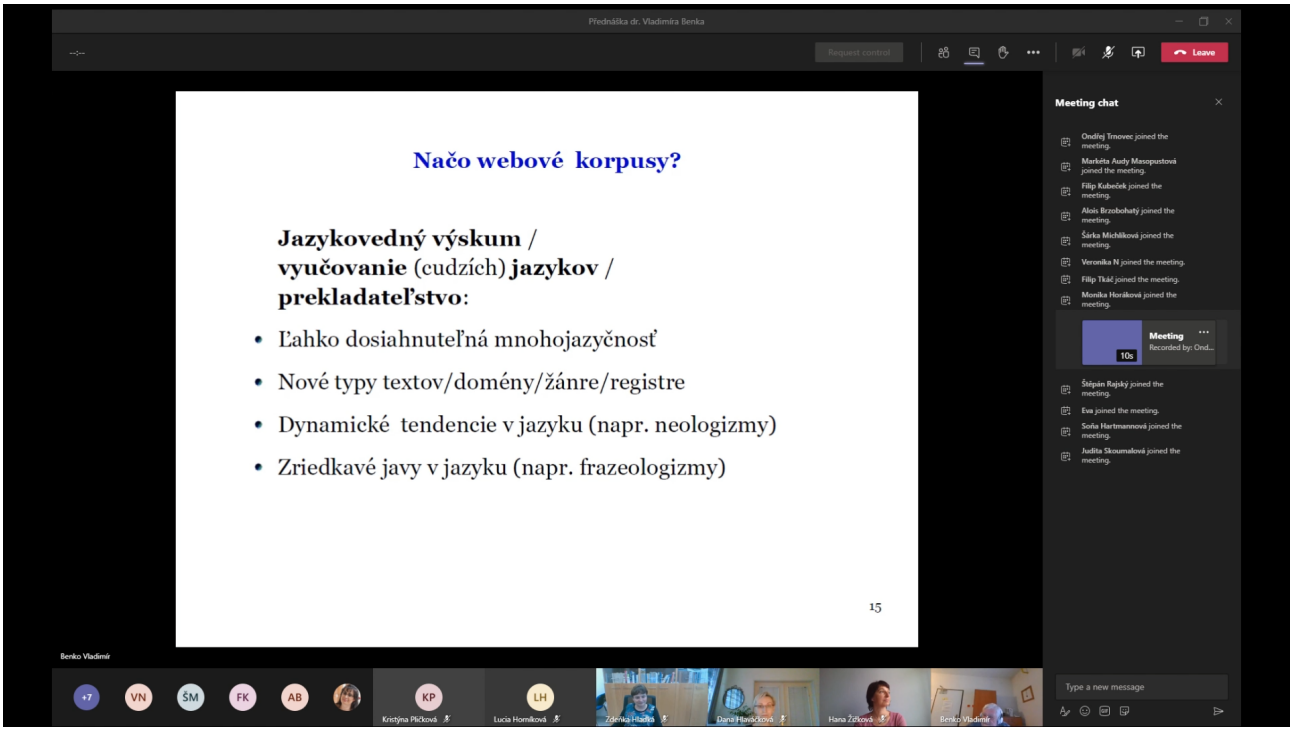
(momentálně pro češtinu), který pro slovesa zobrazuje nejen slovesné formy, ale i jmenné, což je umožněno díky morfologickému slovníku MorfFlex. Také jsme se dozvěděli, jak jednoduše vytvářet uživatelské subkorpusy a jak s nimi následně pracovat.

Přednáška nás zajímavou formou nalákala k používání korpusů Aranea, mezi jejichž výhody patří několik speciálních funkcí a to, že jsou porovnatelné. Určitě je nyní budeme používat více. Děkujeme panu doktorovi za přednášku online formou a příště se snad již sejdeme opět osobně na fakultě.

Aranea

Veľkosti

- **Maius** (väčší) ... základná verzia, 1,25 miliardy tokenov, t. j. približne. 1 miliard slov
- **Minus** (menší) ... 10 % vzorka korpusu Maius (používaný pri výučbe)
- Minimum (minimálny) ... 1 % vzorka (určená na testovanie nástrojov)
- **Maximum** (maximálny) ... koľko sa podarí



K přednášce jsme dostali k dispozici tyto „taháky“:

Rozloženie českej klávesnice MS Windows:

https://milo.juls.savba.sk/~vladob/resources/20191105_keyb_cs_bis.pdf

Súbor morfologických značiek (tagset) HM:

https://milo.juls.savba.sk/~vladob/resources/20191106_hm_tagset_cheatsheet.pdf

Slovnodruhové, sémantické a štylistické príznaky databázy Morfflex:

https://milo.juls.savba.sk/~vladob/resources/20200328_morfflex.pdf

Penn Tegset + AUT + Regex:

https://milo.juls.savba.sk/~vladob/resources/20180404_ukl_cheatsheet_1.pdf