Co se peče v Českém národním korpusu

Datum konání: 1. 3. 2019

Místo konání: Ústav Českého národního korpusu, Praha

Počet účastníků: 17

Zpracovali: Vacíková Michala

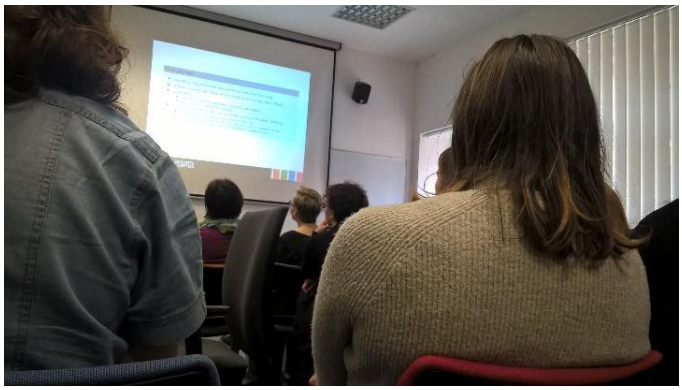
Úvod

V pátek 1. 3. 2019 se početná skupina bohemistů vydala z nádraží Židenice směrem ku Praze, kde je čekala exkurze do Ústavu Českého národního korpusu (ÚČNK), který je pracovištěm Filozofické fakulty Univerzity Karlovy. O korpusech, novinkách a projektech ústavu se rozpovídal sám ředitel ÚČNK doktor Michal Křen.

Korpusy a nástroje

Nejprve krátce představil jednotlivé typy korpusů (synchronní, diachronní, mluvené, paralelní a další). Mimo jiné zmínil, že velký synchronní korpus SYN, který zahrnuje všechny jednotlivé korpusy řady SYN, umožňuje využívat jednotlivé subkorpusy (například SYN2005). Ty jsou ovšem na rozdíl od samostatných korpusů stojících mimo velký SYN přeznačkovány nejnovějšími nástroji, a nabízejí tak kvalitnější výsledky.

Zmínil také korpus mluveného jazyka ORTOFON, který byl zveřejněn roku 2017. Uvedl, že při sběru dat byl kladen zvláštní důraz na vyváženost v oblasti pohlaví, věku, vzdělání i regionu mluvčích. Zároveň je tento korpus výjimečný tím, že je v něm použita dvouúrovňová transkripce, tj. fonetická a ortografická. Je lemmatizován a morfologicky označkován.



Upozornil rovněž na nářeční korpus DIALEKT, který byl zveřejněn v minulém roce, ovšem zahrnuje data již od šedesátých let. Tento korpus byl tvořen pomocí řízených rozhovorů a obsahuje ortografické i dialektologické přepisy, lemmatizaci i morfologické značky.

Zmínil se také o paralelním korpusu InterCorp, kde se nyní používají národní taggery, jejichž způsob značkování je nejednotný, ale v budoucnu by mělo dojít ke sjednocení prostřednictvím Universal Dependencies.

Upozornil také na další korpusové nástroje, jako je SyD (konkurence variant), Morfio (složení slov, derivační morfologie) nebo KWords (analýza klíčových slov).

Plány do budoucna

Mezi plány do budoucna uvedl doktor Křen snahu propojit diachronní korpusy s korpusem SYN. V hledáčku je také internetový jazyk jakožto prostředek zkoumání poloformálního jazyka blížícího se mluvenému (příspěvky na blozích, diskuze pod články apod.). Za tímto účelem by měl být vytvořen samostatný, morfologicky označkovaný korpus. Měl by také vzniknout nástroj, který by umožnil získat nejrůznější informace o hledaném výrazu na jednom místě, ukazoval by tedy synonymické varianty, vývojové trendy, kolokace aj.

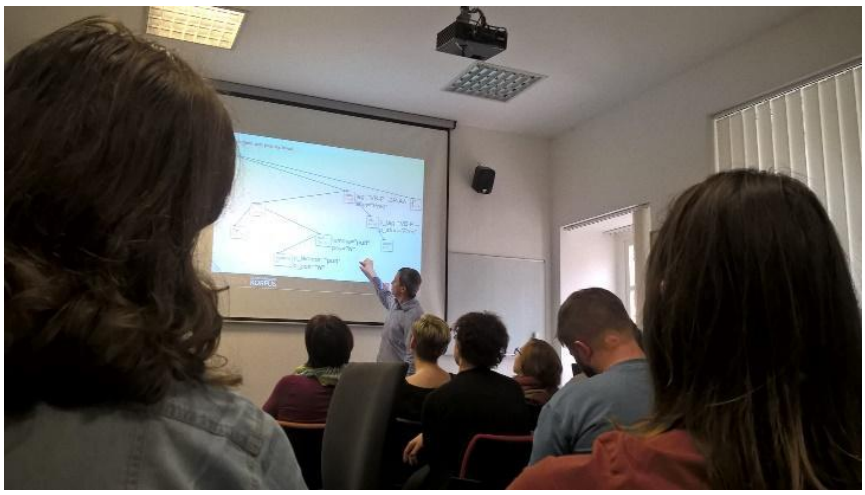
Aktuální projekty

Jako nejvýznamnější aktuální projekt uvedl multidimenzionální analýzu, která se zabývá vnětextovými i vnitrotextovými vztahy a snaží se rozpoznat a kvantifikovat různé rysy, například morfologické nebo hláskoslovné.

K tomu byl využit specializovaný korpus Koditex, který si kladl za cíl pokrýt co nejširší spektrum textů. Tento korpus také zahrnuje anotaci frazémů či pojmenovaných entit.

Z více než tří tisíc vzorků bylo získáno 122 rysů, přičemž byla posuzována relativní frekvence daného rysu ve vzorku. Rysy byly poté pomocí modelu seskupeny do dimenzí, přičemž dvěma hlavními dimenzemi se ukázaly být statičnost vs. dynamičnost projevu a spontánnost vs. připravenost.

Nakonec upozornil také na syntaktické značkování, které se poprvé objevilo v korpusu SYN2015 a na jehož vylepšování se pracuje. Uvedl, že tento způsob značkování přináší řadu dalších atributů, pomocí kterých je možné v korpusu vyhledávat.



Závěr

Exkurze poskytla poměrně komplexní vhled do práce členů Ústavu Českého národního korpusu. Výklad byl přehledný, srozumitelný a zajímavý i pro člověka, který už základní informace o korpusech znal. Nejednalo se totiž pouze o přehled korpusů, ale také upozornění na nové či méně známé funkce, představení aktuálních projektů a nastínění dalších plánů. Celou exkurzi hodnotím pozitivně a bude-li se podobná akce opakovat, studentům účast na ní vřele doporučuji.