4. Korpusová lingvistika

Pokus o definici:

Korpusová lingvistika je spíš metodologickým přístupem než samostatnou lingvistickou disciplínou (přístup = doložené jazykové jevy, metoda = pozorovat tyto doložené jevy ve vzorku, kterým je korpus; nástroje tohoto pozorování mají blízko ke statistice, a jsou tedy spíš kvantitativního charakteru)

Program - rozděleno do 3 přednášek

(Podrobnější úvod do korpusové lingvistiky bude mít v LS dr. Nádvorníková – bude vypsán jako Úvod do románské korpusové lingvistiky - ASZRS0024)

1) Definice a typologie *korpusů*. Prezentace významných korpusů (ČNK, BNC, *La Repubblica*). Projekt InterCorp - budování paralelních korpusů na FF UK. Ukázka konkrétní studie na paralelním korpusu: kauzativní konstrukce ve španělštině a italštině vs. čeština + *zamyšlení nad „překladatelštinou“*

2) Základní pojmy korpusové lingvistiky. Zpracování textů (značkování, lematizace). Statistické pojmy a funkce (*type frequency / token frequency*, frekvence - absolutní/relativní, MI-score, T-score)

3) Speciální případy: Základní pojmy lexikální statistiky (frekvenční seznam, frekvenční spektrum, křivka nárůstu slovníku, tempo nárůstu slovníku, teoretické modely frekvenční distribuce → zachycení morfologické produktivity)

Ad 1) Základní literatura /velmi bohatá.../

- Lüdeling, A., Kytö, M. (eds.) (2008). *Corpus linguistics: An international handbook, volume 1*. Berlin: Mouton de Gruyter.
- Lüdeling, A., Kytö, M. (eds.) (2009). *Corpus linguistics: An international handbook, volume 2*. Berlin: Mouton de Gruyter

DEFINICE.

- Korpus = (dnes již) rozsáhlý soubor jazykových dat, který je zpracován tak, aby sloužil jako *reprezentativní vzorek* daného jazyka; obecně jakákoli „kolekce“ textů (např. korpus textů jednoho autora, žánru apod.).
- Korpus = **vzorek** (*sample*) reprezentující **populaci** (*population*) (populaci bude odpovídat jazyk) (popř. korpus = výběrový soubor; populace = základní soubor; k terminologii srov. Volín, J. *Statistické metody ve fonetickém výzkumu*, Praha 2007, s. 18-22)
- **Reprezentativní** vzorek zachycuje stejnoměrně kvalitativní i kvantitativní rysy jazyka (kvalitativní = textová typologie, např. poezie, publicistika, beletrie atd.; kvantitativní = kolik poezie? vs. kolik publicistiky? atd.)
- Příklad: složení SYN korpusů (

TYPOLOGIE

- **Kvantitativní - z hlediska velikosti korpusu (dnes nejspíš jinak :)**
  - Chiari, 2007, s. 45:

nedostatečný (nereprezentativní) korpus	< 15 tis. slovních tvarů/textových slov
malý k.	15 tis. - 100 tis.
střední k.	100 tis. - 1 mil.
středně velký k.	1 mil. - 50 mil.
standardní k.	50 mil - 100 mil.
velký k.	Přes 100 mil.
Webové korpusy	Přes 1 miliardu; celý SYN 2,2

  - ČNK (SYN2000 = přes 120 mil., SYN2006PUB = přes 300 mil., PMK = přes 600 tis., DIAKORP = teď už přes 2 mil.)
  - BNC (http://www.natcorp.ox.ac.uk/) = přes 110 mil.
  - *La Repubblica* (http://sslmitdev-online.sslmit.unibo.it/) = 380 mil. (330 mil bez interpunkce)

- **Kvalitativní - z hlediska obsahu (viz struktura ČNK http://ucnk.ff.cuni.cz/struktura.html)**
- synchronní vs. diachronní
- psaný vs. mluvený
- smíšený vs. žánrově vymezený (např. KSK - Korpus soukromé korespondence)
- jeden vs. více jazyků (paralelní korpusy)

VÝZNAMNÉ KORPUSY (přehled mj. na http://ucnk.ff.cuni.cz/jine_korpusy.html; nezmiňuji ty, s kterými jako romanisté už dávno pracujeme, např. Frantext (placený...), Corpus del Español, CREA aj.)

ČNK, BNC - přístupné přes KonText, NoSke, ...
La Repubblica - přístupné přes webové rozhraní (http://sslmitdev-online.sslmit.unibo.it/) (příklady vyhledávání příště)

PROJEKT INTERCORP (http://www.korpus.cz/intercorp/)
- mnozí spolupracují, takže spíš než ilustraci postupu ukázka konkrétní studie

- vychází ze společného příspěvku s Petrem Čermákem:
- původně předneseno na konferenci InterCorp 2009, září 2009, Praha FF UK (publikováno jako ČERMÁK Petr - ŠTICHAUER Pavel. Španělské a italské kauzativní konstrukce hacer / fare + sloveso a jejich české ekvivalenty. In ČERMÁK František and KOCEK Jan (et alii). Mnohojazyčný korpus Intercorp: Možnosti studia. Nakladatelství Lidové Noviny, 2010, s. 70-90.)
+ nejnověji rozpracováno a rozšířeno na fr. + port. in Čermák, P. – Nádvorníková, O. *Románské jazyky ve světle paralelních korpusů*, Praha: Karolinum, 2015, kap. 2.