# Workshop: Korpus českého znakového jazyka (1)

Úvodní informace:

* moodle: VS Workshop korpus ČZJ ACN300505 – heslo: „korpus“
* atestace:
  + max 2 absence (setkávání jednou za dva týdny)
  + práce na skupinovém projektu
  + krátký report (cca na půl strany)
* myšlenka semináře:
  + hlavně něco nahrát, práce s vlastními krátkými nahrávkami
  + nahrávky budou sice pilotní, ale měly by představovat materiál, který lze již nějakým způsobem využít
  + práce v semináři bude kopírovat jednu studii
  + měli bychom si vyzkoušet sběr a zpracování nahrávek, nebude následovat psaní odborných článků (ale ani to není vyloučené)
  + každý vytvoří několik nahrávek (hned na příští hodině), pro jednotlivé zpracování budou v celkovém rozsahu cca 15 minut.
  + nahrávat se budou jak rodilí, tak nerodilí mluvčí, nahrávky budou elicitovány na základě otázek
  + zaměření semináře hodně záleží na zájmu a preferencích studentů (čemu by se chtěli věnovat, jaký jazykový jev sledovat)
  + budou vytvořeny pracovní skupiny po dvou, nebo se bude pracovat jednotlivě (podle počtu přihlášených studentů)
    - každá skupina bude pracovat na své vybrané oblasti
  + cílem není vytvoření plnohodnotného korpusu, ale spíše praktické nahlédnutí do dané problematiky
  + zkušenosti ze semináře se mohou stát podkladem pro téma diplomové práce

Co je to korpus?

* Soubor vzorků jazykové produkce
  + Pozn.: tou může být také např. lístkový katalog
* Definiční rysy:
  + Musí být digitalizovaný formát (elektronická podoba), softwarové zázemí
  + Alespoň základní prvky prohledávání
  + Metalingvistická anotace
  + Ideálně by měl být korpus veřejně přístupný
* Může se google označit za korpus? (obsahuje texty, je prohledatelný apod.)
  + V korpusech je snaha o vyvážený materiál (aby byly do určité míry reprezentativní)
  + Výzkumy by se měly provádět na tzv. uzavřených korpusech (do kterých se již nepřikládají texty) – to u googlu není
  + Google nemusí pokaždé vyhledat stejný výsledek, některé stránky se mohou ztratit apod.
  + Dalším problém je kopírování textů (ten stejný text se vyskytuje na různých stránkách s většími nebo menšími úpravami) – výsledek by neodpovídal úzu.

Dimenze nahlížení na korpusy:

* autoři: McEnery & Hardie, 2011

1) modalita (jaký jazykový materiál korpus obsahuje)

* a) unimodální (1 modalita)
  + buď korpusy psaných textů, nebo transkripce mluvených záznamů
    - př.: korpusy politických debat – neexistuje nahrávka, ale pouze zápis
  + počátky: 1. mluvený korpus → britský korpus (ze 60. let, obsahoval pouze přepisy, na jeho vzniku se podíleli Leech a Quirk)
  + naprostá většina korpusů je unimodální
* b) multimodální
  + mluvené: obsahující transkripci a zvuk
    - může být také připojen obraz
  + znakované: obraz + transkripce

2) metoda sběru materiálu

* u unimodálních korpusů je k dispozici větší materiál (např. noviny, knižní produkce, korespondence, školní práce)
* u multimodálních obvykle převažují dvě metody: elicitace a záznam spontánní reakce
  + rozlišuje se, jestli reakci předcházel nějaký zásah výzkumníka
  + pozn.: problém se záznamem spontánní reakce je ten, že se v ní při natáčení nemusí objevit zkoumaný jev
  + používá se např. elicitace narativu (zajímají mě slovesa pohybu – respondentům se pustí vhodný animovaný film a pak jsou požádáni o jeho převyprávění)

3) účelové × obecné korpusy

* účelové: jsou vytvářeny s nějakým výzkumným cílem, ale mohou se stát zdrojem pro obecný korpus
* obecné: všeobecné zaměření, spíše větší projekty a využitelnost

4) sampling × reprezentativnost

* někdo může korpus označit jako nereprezentativní, protože v něm něco chybí (např. u publicistických textů nebudou zařazeny časopisy o zahradnictví)
  + → žádný korpus nemůže být 100% reprezentativní
  + musí se dbát jak na diverzitu textů, tak na diverzitu mluvčích
    - pozn.: korpus může být zdrojem pro sociolingvistický výzkum
  + každé použití filtru (určený výběr) omezuje použitelnost korpusu
* sampling: pomáhá tomu, aby byl korpus co možná nejreprezentativnější (snaha o co nejvíce ratifikovaný vzorek)
  + diverzita: jak podle věku, tak podle sociálního statusu, stupeň dosaženého vzdělání, regionální diverzita, pohlaví atd.

5) počet jazyků

* a) vícejazyčné
  + paralelní subkorpusy (např. korpusy filmových titulků, překladů bible apod.)
    - stejný text v různých jazycích
  + vícejazyčný korpus (obsahuje míšení kódů)
    - např. mluva bilingvální osoby, co bude střídat kódy
* b) jednojazyčné
  + produkce v L1 (v mateřském jazyce)
  + produkce v L2 (v druhém a dalším jazyce)
    - např. psaná čeština českých neslyšících (pro ty, jejichž mateřským jazykem je ČZJ)
  + narušená produkce: např. afatické korpusy
* vedle toho:
  + korpusy interakcí člověk-počítač
  + korpusy primátů, včel apod.

6) technika výstavby

* rozděluje se na základě přístupu v korpusové analýze, metody práce s daty
* a) corpus-based
  + deduktivní (máme předpoklad z pozorování, nebo z introspekce -> korpusové doklady → potvrzení/vyvrácení hypotézy)
* b) corpus-driven
  + induktivní (na základě pozorování v korpusu se vytváří hypotéza/tvrzení → ta se ověřuje dalšími procesy (např. jiným vzorkem) → zobecnění výsledků → teorie)

Multimodální korpusy (také tedy korpusy zn. j.)

* poměrně mladé období lingvistiky (zásadnější rozvoj od 90. let)
* znakové jazyky poměrně opomíjená oblast → při vytváření korpusů zn. j. překážky:
* A) čas
  + zpracování zvukové nahrávky asi desetkrát náročnější, než zpracování textu
  + u projevů ve zn. j. ještě náročnější
* B) technika
  + nutné vybavení
  + zajištění koordinace kamer
  + pouze jeden pokus (opakování již nepředstavuje spontánní reakci)
  + poruchovost
* C) nereliabilita anotace
  + záleží na lidském faktoru (nízká mezianotátorská shoda)
  + Další otázky: Co je znak, co gesto? Jak dlouhý je vlastně znak? Jaké jsou hranice znaku? Apod.
* D) lidské zdroje
  + potřeba velký počet lidí, hl. anotátorů
* E) peníze
  + nejen na pokrytí platu pracovníků
  + ideálně by také měli něco dostat respondenti
  + problém – výstavba korpusu trvá déle, než je délka standardního grantu
* F) etické problémy
  + zveřejňování tváří lidí (nutný souhlas)
  + ve velkých korpusech se tolik neanonymizuje, ale např. v korpusu DEAF (psaná čj českých neslyšících), nebo v korpusu ROMI (texty romských mluvčí češtiny) k anonymizaci dochází
    - → pracuje se s jazykovou komunitou (dalo by se určit, kdo je autorem textu)