ÚVOD DO KORPUSOVÉ LINGVISTIKY

JAZYKOVÝ KORPUS

- Rozsáhlý soubor elektronicky uložených jazykových dat, obvykle označkovaný, organizovaný se zřetelem k využití pro určitý cíl, vůči němuž je také považován za reprezentativní.

Čermák, F. Jazykový korpus: Prostředek a zdroj poznání. In *Studie z korpusové lingvistiky*. Praha: Karolinum, 2000, s. 15-38.

KORPUSOVÁ LINGVISTIKA

- Studium jazyka založené na jeho přirozeném kontextovém užívání
- Studuje a analyzuje data získaná z korpusů
- Metodologie
- Podstatná část počítačové lingvistiky

ZÁKLADNÍ POJMY

- **textové slovo** – řetězec znaků oddělený z obou stran mezerami
- **pozice** (token, tokenizace)
- **lemma** (základní slovní tvar)
- **konkordance**, konkordanční řádek, konkordanční seznam
- **KWIC** – key word in context
- **korpusový prohlížeč**, korpusový manažer
- **regulární výraz**
- **kolokace** – ustálená slovní spojení

PŘEDNOSTI JAZYKOVÝCH KORPUSŮ

- velký rozsah s možností dalšího rozšiřování
- jazyková data v přirozené kontextové podobě
- převaha typických jazykových jevů nad okrajovými
- reprezentativní korpus je schopen zachytit variabilitu jazyka
- zrychlení a usnadnění lingvistické práce

DVA TYPY PŘÍSTUPŮ K JAZYKU

- *Raná „korpusová“ lingvistika* - „korpusový“ přístup k jazykovému materiálu, dostatečně velký soubor přirozeně se vyskytujících jazykových dat (konec 19. st. – 50. léta 20. st.)
- Předěl (50. léta 20. st.) – N. Chomsky a generativní lingvistika
- Empirický přístup, observace **X** intuice a introspekce

  Corpus linguists study real language, other linguists just sit at their coffee table and think of wild and impossible sentences.
- *Ch. Fillmore:*

  „I don‘t think there can be any corpora, however large, that contain information about all of the areas of English lexicon and grammar that I want to explore … [but] every corpus I have had the chance to examine, however small, has taught me facts I couldn‘t imagine finding out any other way. My conclusion is that the two types of linguists need one another.“

KORPUSOVÁ LINGVISTIKA V ČR

- Lexikální archiv ÚJČ, od r. 1911, 12 mil. ručně psaných lístků
- 1988 Iniciativní skupina pro přípravu počítačových korpusů, textů a slovníků (sdružení lingvistů, matematiků a programátorů)
- 1991 Počítačový fond češtiny - projekt lexikografického počítačového korpusu a tezauru češtiny (Čermák, Sgall, Pala, Hajič, Hajičová, Králík, Schmiedtová, Kučera, Benko)
- 1994 založení ÚČNK

TYPY KORPUSŮ

- druh zachycené komunikace – psané (written corpora)
  – mluvené (spoken corpora)
- časový záběr – diachronní
  – synchronní
- účel – všeobecné
  – specializované
- jazyk – jednojazyčné
  – paralelní
- možnost rozšíření – uzavřené
  – otevřené
- značkování – tagging (POS tagging, morfologie)
  – parsing (syntax, treebank)
  – alignment (párování)

REPREZENTATIVNOST

- Relativní

- v závislosti na účelu korpusu (kvantita x kvalita)
- malý vzorek vzhledem k celku jazyka
- nezobrazuje reálné užití jazyka

- Snaha zachytit variabilitu textů

- texty informativní
- texty imaginativní

Atkins, S., Clear, J., Ostler, N. Corpus Design Criteria. In *Literary and Linguistic Computing* 7/1 1992, s. 1-6.
(přel. Čermák, Schmiedtová – Studie z korpusové lingvistiky)

SYN2000
denní tisk / 60 %
naučná literatura / 25 %
krásná literatura / 15 %

SYN2005
publicistika / 33 %
odborná literatura / 27 %
beletrie / 40 %

LINGVISTICKÉ VYUŽITÍ KORPUSŮ

- závislost na typu a velikosti korpusu
- objevování nových jazykových jevů x ověřování hypotéz
- fonetika, fonologie, morfologie, slovotvorba, syntax, stylistika
- lexikografie, lexikologie
- sémantika, dialektologie, terminologie
- diachronní studie
- frekvenční studie
- konfrontace s cizími jazyky, překlad