PLIN041 Vývoj počítačové lingvistiky

Korpusová lingvistika

Mgr. Dana Hlaváčková, Ph.D.

Vývoj korpusové lingvistiky

- **raná korpusová lingvistika** (90. léta 19. st. – 50. léta 20. st.)
- **předěl** – generativní lingvistika (50. léta 20. st.)
- **počátky výpočetní techniky** (50.–80. léta 20. st.)
  - první korpusy
- **rozvoj výpočetní techniky** (od. 80. let 20. st.)
  - začátky moderních korpusů

Raná korpusová lingvistika

konec 19. st – 50. léta 20. st.

- strukturalistická tradice, americký deskriptivismus, metody založené na zkoumání souborů textů a na empirii
- shromažďování jazykového materiálu, nahrávky výpovědí (analýza bottom-up)
- **nejde o korpusy** – archivy, kartotéky, deníky, seznamy, slovníky
- společné prvky s pozdější korpusovou lingvistikou:
  - rozsah je důležitým parametrem
  - žánrová vyváženost souboru textů
  - zkoumání významů slov a homonymie
  - problematika slovní jednotky a lemmatizace (**lemma** = základní tvar slova)
  - morfologické, syntaktické i sémantické analýzy jazyka na základě textového materiálu

Raná korpusová lingvistika

- frekvence a lexikografie
- akvizice jazyka (mateřského i cizího)
- komparativní lingvistika
- dialektologie a výzkum jazyků původních obyvatel Ameriky a Kanady

Raná korpusová lingvistika

- **1) frekvence a počátky moderní lexikografie** – excerpční lístky (ručně, na stroji) – kartotéky, výpisky z beletrie, novin, zapojení slova v kontextu (**konkordance**)
  - **frekvenční studie** – Friedrich Wilhelm Käding, 1897–1898 (vycházel z 11 mil. slov), *Häufigkeitswörterbuch der deutschen Sprache,* na dlouhou dobu nejrozsáhlejší jazykový materiál v podobě frekvenčních seznamů a frekvenčního slovníku
  - **výuka jazyka pro cizince** – frekvenční seznamy slov, frekvenční slovníky, navazující slovníky a učebnice k výuce jazyka pro cizince, např. Edward L. Thorndike (am. psycholog) – *The Teacher's Word Book*, 1921

Raná korpusová lingvistika

- **2) akvizice jazyka** – zápisy dětské mluvy, rodičovské deníky, později malý vzorek dětí a dlouhodobé sledování
- William Thierry Preyer (1841–1897)
- narodil se v Anglii, studoval a žil v Německu
- působil v Jeně jako ředitel fyziologického ústavu
- zakladatel dětské psychologie
  - založena na empirickém pozorování a experimentech
  - k výzkumu využívá rodičovské deníky
  - významné dílo *Die Seele des Kindes* – vývojová psychologie

Raná korpusová lingvistika

- **3) komparativní lingvistika** – srovnávání významů slov z různých jazyků, studium jazyka Bible a dalších kanonických textů (užívání **konkordancí**)

- **4) dialektologie a zapisování jazyků domorodých kmenů**
- dialektologie
  - pro češtinu v souvislosti s národním obrozením (pol. 19. st.)
  - historickosrovnávací a později strukturalistický přístup

- Franz Boas (1858–1942), pův. Němec, zakladatel moderní americké antropologie, vystudoval fyziku a geografii
  - při výpravě do severní Kanady ho okouzlil jazyk a kultura domorodých kmenů
  - studie severoamerických domorodých kmenů (Inuité)
  - emigroval do USA – profesorem antropologie na Columbia University (1899)
  - důraz na interdisciplinární přístup a empirický výzkum

Korpusový přístup – kritika

Kritika

- kolem 1950 – Noam Chomsky – generativní lingvistika
- racionalismus x empirie, kompetence x performance
- odpor ke korpusovému přístupu k jazyku, korpusy (textové materiály) nejsou v lingvistice potřebné, poskytují *pokřivená data*
- předpočítačové období – ruční hledání v rozsáhlých papírových datech je příliš pracné

- **X** rozvoj počítačové techniky po 2. sv. v.

Korpusová lingvistika a počátky výpočetní techniky (50.–80. l. 20. st.)

- vývoj i pod kritikou N. Chomského a jeho stoupenců
- využívání prvních počítačů
- konkordanční seznamy, strojově čitelné texty
- **počátky Digital Humanities**
  - výzkum starověkých jazyků
  - Roberto Busa (1913–2011) – italský jezuitský kněz, studium spisů Tomáše Akvinského, využití počítačů pro lingvistické a literární analýzy
  - spojení s **IBM**, konkordance, lemmatizace, 30 let práce, 56 tištěných svazků (70. léta 20. st.)
  - Index Thomisticus (webová verze 2005)
  - The Busa Price v oblasti DH

Korpusová lingvistika a počítačová lexikografie (od 60. let 20. st.)

- **BROWN CORPUS – průkopníci korpusové lingvistiky**

- Henry Kucera (Jindřich Kučera), 1925–2010
- studoval filozofii a lingvistiku na UK v Praze
- po r. 1948 emigrace do USA, doktorát na Harvardu, od r. 1955 profesor na Brown University (Slavic Department)
- autor jednoho z prvních automatických korektorů pravopisu

- W. Nelson Francis, 1910–2002, americký lingvista
- studoval na Harvardu a University of Pennsylvania, literatura, angličtina, řečtina, latina a francouzština
- profesor na Brown University (navštěvoval Kučerův kurz počítačové lingvistiky)

Brown Corpus

- ***Brown Corpus*** *(Brown Standard Corpus of Present-Day American English)*, 1963–**1964**, Brown University
- americká angličtina rodilých mluvčích
- **500** textových vzorků (vždy **2000** slov)
- **15** žánrových kategorií (časopisy, noviny, beletrie, odborná lit.), snaha o vyváženost
- **1 mil.** slov, vše z roku 1961
  - morfologicky označkován (PoS tagging – 80 kategorií)
  - na delší dobu vzor pro další korpusy
  - na MU dostupný přes Sketch Engine
- ***American Heritage Dictionary of the English Language***, 1969 – 1. slovník založený na korpusu (Brown Corpus, třířádkové citace, preskripce i deskripce), Boston

LOB

- **Geoffrey Leech** (1936–2014), Stig Johansson – ***Lancaster-Oslo/Bergen Corpus (LOB)***, 1970–**1978**
- britský protějšek k *Brown Corpus*, stejná struktura (**1 mil** slov, **500** textových vzorků po **2000** slovech, **15** žánrů)
- psaná britská angličtina z r. 1961
- University of Lancaster, University of Oslo, Norwegian Computing Centre for the Humanities, Bergen
- značkovaná verze (PoS tagging) – 1981–1986

- Frown – *The Freiburg-Brown corpus of American English*
- F-LOB – *The Freiburg-LOB corpus of British English*
  - *texty z r. 1992, zveřejnění 1999*
- Brown + LOB + Frown + F-LOB = **Brown Family**

SEU

- **Randolph Quirk** (1920–2017) – ***The Survey of English Usage (SEU)***, 1959, University College London, **první korpusové pracoviště**
  - v týmu také Jan Firbas (český jazykovědec, anglista)
  - cílem bylo popsat gramatický repertoár dospělých, vzdělaných rodilých mluvčích v Británii
  - **SEU** – vzorky **psané** a **mluvené** britské angličtiny (půl na půl), 200 textů, každý 5000 slov, mluvené – monology i dialogy, z let 1955 až 1985
  - původně na papíře (lístky 6 x 4 palce) s podrobnou gramatickou anotací, později převeden do počítačově čitelné podoby (Svartvik)
- SEU byl použit pro jednu z nejdůležitějších korpusově založených gramatik – ***Comprehensive Grammar of the English Language*** (Quirk, Greenbaum, Leech, Svartvik, 1985)

LLC

- Jan Svartvik (1931), Sidney Greenbaum, R. Quirk, K. Hofland
- ***The London-Lund Corpus of Spoken English (LLC)***
- **1. počítačový korpus mluveného jazyka** (magnetické pásky)
- spojení dvou projektů
  - **Survey of Spoken English (SSE)**, Jan Svartvik, Lund University, **1975** jako sesterský projekt SEU
    - 87 textů mluvené angličtiny (britská angličtina vzdělaných mluvčích)
  - **SEU** – 13 textů mluvené angličtiny
- celkem 100 přepisů nahrávek, 500 tisíc slov, zveřejněn až 1980
  - fonetická transkripce, značeny prozodické vlastnosti
  - někteří mluvčí o nahrávání nevěděli (spontánní projev)

Propojení lexikografie s korpusovou lingvistikou

- **COBUILD** – *Collins Birmingham University International Language Database*, britské výzkumné centrum na University of Birmingham, od r. 1980 založeno vydavatelstvím Collins (dnes HarperCollins Publishers), na počátku vedl profesor **John Sinclair** (1933–2007)
- cílem vydání slovníku pro výuku angličtiny
- korpus **Birmingham Collection of English Text** (BCE), 1980, jako první využil OCR
  - **20 mil. slov**, hlavně psaná britská angličtina
  - jiná struktura než první korpusy (noviny, brožury, letáky, knihy, časopisy, korespondence), oproti LOB vyloučena poezie a drama
- ***Collins COBUILD English Language Dictionary***, 1987
  - pro výuku angličtiny jako cizího jazyka
  - první slovník založený na současné, běžně užívané angličtině

British National Corpus (1991–1994)

- **100 mil. slov**, vyvážený korpus (široké spektrum textů)
- vzorky – 45 tis. slov od jednoho autora
- psaná (90 %) i mluvená (10 %) angličtina (ortografická transkripce)
- značkování (PoS) – Lancaster University (Geoffrey Leech, Roger Garside a Tony McEnery)
- zaštiťuje *BNC Consortium* (Oxford, Lancaster, nakladatelství, firmy, akademie, knihovna apod.)
- subkorpusy
  - **BNC Sampler** (1 mil. psaný, 1 mil. mluvený)
  - **BNC Baby** (4 milionové vzorky ze čtyř různých žánrů)

Německo, Francie

- **Deutsches Referenzkorpus** (DeReKo), 1964, Mannheim, Leibnitz-Institut für Deutsche Sprache
  - dnes 50,6 mld. slov (největší na světě)
  - texty cca od r. 1950
  - otevřený, monitorovací, nevyvážený
- **LIMAS** (Linguistik und Maschinelle Sprachbearbeitung), 1970, Universität Bonn
  - německá varianta Brown Corpus – 500 textů, 15 kategorií, 1 mil. slov, texty z let 1969–70
- **Frantext** – databáze literárních textů ve francouzštině, od 10. do 21. st., (word, lemma, phrase), 264 mil. slov, metainformace o textech, Analyse et Traitement Informatique de la Langue Française (ATILF, Université de Lorraine, Nancy)