KORPUSOVÁ LINGVISTIKA

VÝVOJ KORPUSOVÉ LINGVISTIKY

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

PRZA009

Vývoj korpusové lingvistiky

- **raná korpusová lingvistika**

(90. léta 19. st. – 50. léta 20. st.)

- **předěl** – generativní lingvistika (50. léta 20. st.)
- **počátky počítačové techniky** (50.–80. léta 20. st.)
- **rozvoj počítačové techniky**

(od. 80. let 20. st.)

Raná korpusová lingvistika

- metody založené na zkoumání souborů textů a na **empirii**
- shromažďování **jazykového materiálu**, nahrávky výpovědí
- archiv, kartotéka, 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

1. frekvence a lexikografie
2. akvizice jazyka
3. komparativní lingvistika
4. dialektologie a výzkum indiánských jazyků

Raná korpusová lingvistika

Raná korpusová lingvistika

frekvence a počátky moderní lexikografie – excerpční lístky – kartotéky, výpisky z beletrie, novin, zapojení slova v kontextu (**konkordance**)

frekvenční studie – **Friedrich Wilhelm Käding**, 1897–1898 (11 mil. slov), *Häufigkeitswörterbuch der deutschen Sprache,* na dlouhou dobu nejrozsáhlejší jazykový materiál v podobě frekvenčních seznamů a frekvenčního slovníka

výuka jazyka pro cizince – frekvenční seznamy slov, frekvenční slovníky, navazující slovníky a učebnice k výuce jazyka pro cizince

Raná korpusová lingvistika

akvizice jazyka – zápisy dětské mluvy, rodičovské deníky, později malý vzorek dětí a dlouhodobé sledování

William Thierry Preyer (1841–1897)

zakladatel dětské psychologie

- **empirické** pozorování a **experimenty**
- k výzkumu využívá **rodičovské deníky**
- významné dílo *Die Seele des Kindes* – vývojová psychologie

Raná korpusová lingvistika

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

dialektologie a zapisování indiánských jazyků

- **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, studie indiánských kmenů
  - vystudoval fyziku a geografii
  - při výpravě do severní Kanady ho okouzlil jazyk a kultura domorodých kmenů
  - emigroval do USA – profesorem antropologie na Columbia University

Korpusový přístup – kritika

Kritika zkoumání přirozeného jazyka

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

X

rozvoj počítačové techniky

po 2. sv. v.

Korpusová lingvistika a počátky výpočetní techniky

- vývoj i pod kritikou N. Chomského a jeho stoupenců
- využívání prvních **počítačů**
- konkordanční seznamy, strojově čitelné texty

počátky Digital Humanities

- **Roberto Busa** (1913–2011) – italský jezuitský kněz, studium spisů Tomáše Akvinského
- 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

Průkopníci korpusové lingvistiky a první počítačový korpus

Henry Kucera (Jindřich Kučera), 1925–2010

- studoval filozofii a lingvistiku na UK
- po r. 1948 emigrace do USA, doktorát na Harvardu, od r. 1955 profesor na Brown University (Slavic Department)

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)

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 (80 kategorií)
  - na delší dobu vzor pro další korpusy
- ***American Heritage Dictionary of the English Language***, 1969 – 1. slovník založený na korpusu (Brown Corpus, třířádkové citace, preskripce i deskripce), Boston

LOB

Lancaster-Oslo/Bergen Corpus (LOB), 1970–**1978**

Geoffrey Leech (1936–2014), **Stig Johansson**

- 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

SEU – pracoviště

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
- vzorky **psané** a **mluvené** britské angličtiny (půl na půl), **200** textů, každý **5000** slov, z let 1955 až 1985
- původně na papíře (lístky 6 x 4 palce) s podrobnou gramatickou anotací, později převeden do počítačově čitelné podoby (**Svartvik**)

Materiál byl použit pro jednu z nejdůležitějších korpusově založených gramatik – ***Comprehensive Grammar of the English Language*** (Quirk, Greenbaum, Leech, Svartvik, 1985)

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)

LLC

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** , 1964, Mannheim, Leibnitz-Institut für Deutsche Sprache
  - dnes 61,5 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), 270 mil. slov, metainformace o textech, Analyse et Traitement Informatique de la Langue Française (ATILF, Université de Lorraine, Nancy)

Korpusová lingvistika v ČR

- **Marie Těšitelová** – ***Korpus věcného stylu*** *(1971–1985)*
- Ústav pro jazyk český ČSAV – *Oddělení matematické a aplikované lingvistiky*
- **věcný styl** – odborná literatura, publicistika, administrativní texty
- **540 000 slov**, každý text **3 000 slov**
- **75 % texty psané**, **25 % mluvené projevy**
- ručně morfologicky a syntakticky značkovaný *Český akademický korpus*, ÚFAL MFF UK, 2007
  - Jaroslav Jelínek, Josef V. Bečka, Marie Těšitelová – Frekvence slov, slovních druhů a tvarů v českém jazyce, 1961

Korpusová lingvistika v ČR

- 1988 ***Iniciativní skupina pro přípravu počítačových korpusů, textů a slovníků*** (Pala, Čermák, Schmiedtová, Hajičová ad.)
- 1992 ***Počítačový fond češtiny***, *Skupina pro počítačový fond češtiny* – Čermák, Králík, Pala, Hajič, Hajičová, Sgall, Schmiedtová, Benko, Kučera
- 1993–95 ***Počítačový korpus českých psaných textů*** (GAČR)
- 1994 – založení **Ústavu Českého národního korpusu**
- první korpus **SYN2000**
  - **Akademický slovník současné češtiny**

Přístupy ke korpusům

- korpusový manažer **Kontext**
- https://korpus.cz/signup

- korpusový manažer **Sketch Engine**
- https://www.sketchengine.eu/
- Institutional Login
  - Masaryk University
  - UČO + sekundární heslo