

Korpusová lingvistika – 2

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

konec 19. st – 50. léta 20. st.

- 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

Raná korpusová lingvistika

- frekvence a lexikografie
- akvizice jazyka
- komparativní lingvistika
- dialektologie a výzkum indiánských jazyků

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 (11 mil. slov), *Häufigkeitwörterbuch 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 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 okouznil jazyk a kultura domorodých kmenů
 - emigroval do USA – profesorem antropologie na Columbia University

Korpusový přístup – kritika

Kritika

- 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 (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
 - 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

Brown Corpus, lemma „student“, 351 výskytů

KONKORDANCE Získat více místa ? ! 👤

lemma **student** • 351
298,55 na milion tokenů • 0.03%

Detaily Levý kontext KWIC Pravý kontext

1	<input type="checkbox"/>	<input type="checkbox"/>	various st counties .</s><s>The bill is designed to provide special schooling for more deaf students in the scholastic age at a reduced cost to the state .</s><s>There was no debate a	<input type="checkbox"/>
2	<input type="checkbox"/>	<input type="checkbox"/>	various aid the nation<g/>'s 92 medical and 47 dental schools can<g/>not now handle the student load needed to meet the rising need for health care .</s><s>Moreover , he said , m	<input type="checkbox"/>
3	<input type="checkbox"/>	<input type="checkbox"/>	various ical and dental school equal to \$<g/>1,500 a year for one-fourth of the first year students .</s><s>The schools could use the money to pay 4<g/>-year scholarships , based c	<input type="checkbox"/>
4	<input type="checkbox"/>	<input type="checkbox"/>	various ey to pay 4<g/>-year scholarships , based on need , of up to \$<g/>2,000 a year per student .</s><s>In addition , the government would pay a \$<g/>1,000 `` cost of education "	<input type="checkbox"/>
5	<input type="checkbox"/>	<input type="checkbox"/>	various ont and kept the lead through two thirds of the race .</s><s>Then Kerr , a graduate student from Illinois , moved past him on a straightaway and held off Mills<g/>'s challenge	<input type="checkbox"/>
6	<input type="checkbox"/>	<input type="checkbox"/>	various re for the Christmas holidays and ca<g/>n't return .</s><s>Young Peter Wendell , a student at the Westminster school , has measles , and his sister , Mrs. Andrew Thomas , ar	<input type="checkbox"/>
7	<input type="checkbox"/>	<input type="checkbox"/>	various s><s>They will live in Corpus Christi .</s><s>Miss Shirley Joan Meredith , a former student of North Texas State University , was married Saturday to Larry W. Mills , who has i	<input type="checkbox"/>
8	<input type="checkbox"/>	<input type="checkbox"/>	various Mass. .</s><s>Her mother is the former Miss Stella Hayward .</s><s>Mr. Wall is a student at Tulane university , where he is a member of Delta Kappa Epsilon fraternity .</s><	<input type="checkbox"/>
9	<input type="checkbox"/>	<input type="checkbox"/>	various uties , fund-raising events were being planned by a homeowners association and a student council for the hard-hit Henry Kowalski family , 34220 Viceroy .</s><s>Mrs. Eleano	<input type="checkbox"/>
10	<input type="checkbox"/>	<input type="checkbox"/>	various ir fund drive throughout their subdivision on behalf of the Kowalski family .</s><s> Students help out</s><s>Houghton said 6 p.m. Friday had been set for a canvass of all 480	<input type="checkbox"/>
11	<input type="checkbox"/>	<input type="checkbox"/>	various illage Station .</s><s>Vicky Kowalski meanwhile learned that several of her fellow students had collected almost \$<g/>25 for her family during the lunch hour yesterday at Fu	<input type="checkbox"/>
12	<input type="checkbox"/>	<input type="checkbox"/>	various vate hospital nurse Monday for Mrs. Kowalski .</s><s>Funds from dances</s><s> Student Council officers announced today the Kowalski family would be given the combine	<input type="checkbox"/>
13	<input type="checkbox"/>	<input type="checkbox"/>	various n a school dance held two weeks ago , and another dance for Fuhrmann<g/>'s 770 students this Friday night .</s><s>`` Fuhrmann<g/>'s faculty is proud that this has been a sp	<input type="checkbox"/>
14	<input type="checkbox"/>	<input type="checkbox"/>	various aculty is proud that this has been a spontaneous effort , started largely among the students themselves , because of fondness for Vicky and sympathy for her entire family , Pc	<input type="checkbox"/>
15	<input type="checkbox"/>	<input type="checkbox"/>	various niversity<g/>'s Board of Trustees announced Friday that it was prepared to accept students of any race as soon as the state<g/>'s tax laws made such a step possible .</s><s>	<input type="checkbox"/>
16	<input type="checkbox"/>	<input type="checkbox"/>	various niversity<g/>'s charter and by-laws have never required admission or rejection of students on the basis of race " , board chairman Henry L. Bowden stated .</s><s>But an offi	<input type="checkbox"/>
17	<input type="checkbox"/>	<input type="checkbox"/>	various past two private colleges in the Atlanta area now or in the past have had integrated student bodies , but their tax-exempt status never has been challenged by the state .</s><	<input type="checkbox"/>

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
- značkováná verze (PoS tagging) – 1981–1986

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 prepisů 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 57,6 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), 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 značkový (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