

# Korpusová lingvistika – 2

## Vývoj korpusové lingvistiky

Mgr. Dana Hlaváčková, Ph.D.  
CJBB105

# Vývoj korpusové lingvistiky

- **raná korpusová lingvistika**, počítačová lexikografie, frekvenční studie (90. léta 19. st. – 50. léta 20. st.)
- **předěl** – generativní lingvistika (50. léta 20. st.)
- **rozvoj počítačové techniky** (od. 60. let 20. st.)

# 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, **rozsah** je důležitým parametrem
- nemluvíme o korpusech ani o korpusové lingvistice – archiv, kartotéka, deníky, seznamy, slovníky
- žánrová **vyváženost** souboru textů
- zkoumání významů slov a homonymie
- problematika slovní jednotky a lemmatizace
- morfologické, syntaktické i sémantické analýzy jazyka na základě textového materiálu

# Raná korpusová lingvistika

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

# Raná korpusová lingvistika

- **zápisy dětské mluvy** – rodičovské deníky, studium akvizice jazyka (1876–1926), od 1927 analýzy jazyka, 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

- **komparativní lingvistika** – srovnávání významů slov z různých jazyků, studium jazyka Bible a dalších kanonických textů (užívání konkordancí)
- **zapisování indiánských jazyků**
- 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
  - byl proti tzv. vědeckému rasismu – např. stavba lebky se řídí rasou
  - stavba kostry je ovlivněna okolním prostředím a výživou
  - chování lidí není výsledkem biologické predispozice, ale ovlivněno sociálním prostředím a výchovou

# Korpusový přístup – kritika

## Kritika

- kolem 1950 – **Noam Chomsky** – generativní lingvistika, odpor ke korpusovému přístupu k jazyku, korpusy nejsou v lingvistice potřebné, poskytují *pokřivená data*
- předpočítačové období – ruční hledání v rozsáhlých datech je příliš pracné
- **X** rozvoj počítačové techniky

# 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

Konkordance  
Seznamy slov  
Word Sketch  
Tezaurus  
Najdi X  
Sketch-Diff  
Korpus info  
Mé úlohy



Uložit  
jako subkorpus  
Možnosti  
zobrazení  
KWIC  
Věta  
Třídění  
Levý kontext  
Pravý kontext  
Node  
Reference  
Zamíchat  
Vzorek  
Filtr  
Překryvy  
1. výskyt/dok.  
Frekvence  
Značky (tags)  
Slovní tvary

Dotaz **go.\*** 4,429 (3,767.20 v milionu)Strana  ze 222 [Jdi](#) [Další](#) | [Poslední](#)

A01	which inure to the best interest of both	<b>governments</b> /NNS/government "	. Merger proposed However , the jury
A01	interlude , since 1937 . His political career	<b>goes</b> /VVZ/go	back to his election to city council in
A01	encouragement to enter a candidate in the 1962	<b>governor</b> /NN/governor	's race , a top official said Wednesday
A01	said Wednesday . Robert Snodgrass , state	<b>GOP</b> /NP/GOP	chairman , said a meeting held Tuesday
A01	was warned that entering a candidate for	<b>governor</b> /NN/governor	would force it to take petitions out into
A01	director , resigned Tuesday to work for Lt.	<b>Gov.</b> /NP/Gov.	Garland Byrd's campaign . Caldwell's resignatio
A01	determine what adjustments should be made .	<b>Gov.</b> /NP/Gov.	Vandiver is expected to make the traditional
A01	reconstruction bonds . The bond issue will	<b>go</b> /VV/go	to the state courts for a friendly test
A01	authorities . Vandiver opened his race for	<b>governor</b> /NN/governor	in 1958 with a battle in the Legislature
A01	additional rural roads bonds proposed by then	<b>Gov.</b> /NP/Gov.	Marvin Griffin . The Highway Department
A01	votes in Saturday's election , and Bush	<b>got</b> /VVD/get	402 . Ordinary Carey Williams , armed with
A01	irregularities at the polls , and Williams	<b>got</b> /VVD/get	himself a permit to carry a gun and promised
A01	Felix Tabb said the ordinary apparently made	<b>good</b> /JJ/good	his promise . `` Everything went real smooth
A02	Austin , Texas -- Committee approval of	<b>Gov.</b> /NP/Gov.	Price Daniel's `` abandoned property "
A02	Berry , an ex-gambler from San Antonio ,	<b>got</b> /VVD/get	elected on his advocacy of betting on the
A02	would be selected by a board composed of the	<b>governor</b> /NN/governor	, lieutenant governor , speaker of the
A02	board composed of the governor , lieutenant	<b>governor</b> /NN/governor	, speaker of the House , attorney general
A02	West Texan reported that he had finally	<b>gotten</b> /VVN/get	Chairman Bill Hollowell of the committee
A03	address text still had `` quite a way to	<b>go</b> /VV/go	" toward completion . Decisions are made
A03	secretary , replied , `` I would say it's	<b>got</b> /VVN/get	to go thru several more drafts " . Salinger

Strana  ze 222 [Jdi](#) [Další](#) | [Poslední](#)

# 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
- originální verze – 1976
- značkováná verze (PoS tagging) – 1981–1986

# SEU

- **Randolph Quirk** (1920) – *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
  - původně na papíře (lístky 6 x 4 palce), 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 corpora, IDS corpora, COSMAS corpora*), Institut für Deutsche Sprache
  - dnes 31,6 mld. slov (největší na světě)
  - texty cca od r. 1950
  - 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), 500 děl, metainformace o textech



# 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**