Korpusová lingvistika a počítačová lexikografie

Od 60. let 20. st.

Raná korpusová lingvistika (konec 19. st – 50. léta 20. st., Early corpus linguistics)

- 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

- počátky moderní lexikografie – excerpční lístky (ručně, na stroji), výpisky z beletrie, novin, zapojení slova v kontextu (konkordance)

  - **zápisy dětské mluvy** – rodičovské deníky, akvizice jazyka (1876–1926), od 1927 analýzy jazyka, později malý vzorek dětí a dlouhodobé sledování, W. T. Preyer – zakladatel dětské psychologie
  - **frekvenční studie** – Käding (11 mil. slov), na dlouhou dobu nejrozsáhlejší jazykový materiál
  - **výuka jazyka pro cizince** – frekvenční seznamy slov, frekvenční slovníky, např. E. Thorndike – *The Teacher's Word Book*, 1921

Raná korpusová lingvistika

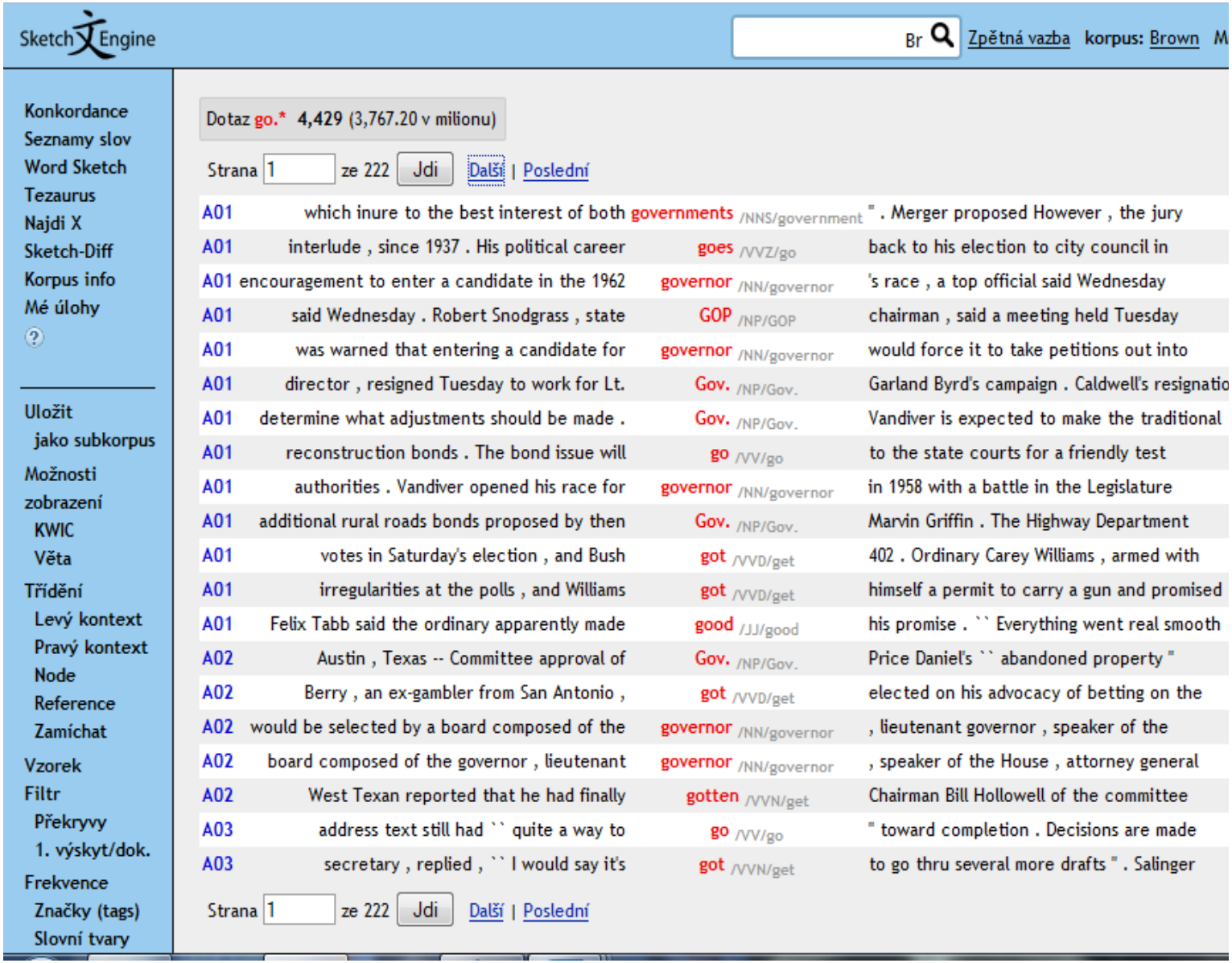
- **komparativní lingvistika** – srovnávání významů slov z různých jazyků
- **zapisování indiánských jazyků** – Franz Boas, 1940, zakladatel moderní americké antropologie, studie indiánských kmenů

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

- **Henry Kučera, W. Nelson Francis** – ***Brown Corpus***, 1960–**1964**, Brown University
  - 500 textových vzorků (vždy 2000 slov), 15 žánrových kategorií, 1 mil. slov
  - *Computational Analysis of Present-Day American English,* 1967 (lingvistika, psychologie, statistika, sociologie)
  - později v 70. letech označkován (PoS tagging)
  - vzor pro další korpusy
  - 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


Sketch Engine
Br
Zpětná vazba
korpus: Brown
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 1 ze 222 Jdi Další | Poslední
A01 which inure to the best interest of both governments /NNS/government " . Merger proposed However , the jury
A01 interlude , since 1937 . His political career goes /VVZ/go back to his election to city council in
A01 encouragement to enter a candidate in the 1962 governor /NN/governor 's race , a top official said Wednesday
A01 said Wednesday . Robert Snodgrass , state GOP /NP/GOP chairman , said a meeting held Tuesday
A01 was warned that entering a candidate for governor /NN/governor would force it to take petitions out into
A01 director , resigned Tuesday to work for Lt. Gov. /NP/Gov. Garland Byrd's campaign . Caldwell's resignatio
A01 determine what adjustments should be made . Gov. /NP/Gov. Vandiver is expected to make the traditional
A01 reconstruction bonds . The bond issue will go /VV/go to the state courts for a friendly test
A01 authorities . Vandiver opened his race for governor /NN/governor in 1958 with a battle in the Legislature
A01 additional rural roads bonds proposed by then Gov. /NP/Gov. Marvin Griffin . The Highway Department
A01 votes in Saturday's election , and Bush got /VVD/get 402 . Ordinary Carey Williams , armed with
A01 irregularities at the polls , and Williams got /VVD/get himself a permit to carry a gun and promised
A01 Felix Tabb said the ordinary apparently made good /JJ/good his promise . `` Everything went real smooth
A02 Austin , Texas -- Committee approval of Gov. /NP/Gov. Price Daniel's `` abandoned property "
A02 Berry , an ex-gambler from San Antonio , got /VVD/get elected on his advocacy of betting on the
A02 would be selected by a board composed of the governor /NN/governor , lieutenant governor , speaker of the
A02 board composed of the governor , lieutenant governor /NN/governor , speaker of the House , attorney general
A02 West Texan reported that he had finally gotten /VVN/get Chairman Bill Hollowell of the committee
A03 address text still had `` quite a way to go /VV/go " toward completion . Decisions are made
A03 secretary , replied , `` I would say it's got /VVN/get to go thru several more drafts " . Salinger
Strana 1 ze 222 Jdi Další | Poslední

Korpusová lingvistika a počítačová lexikografie

- **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
- značkovaná verze (PoS tagging) – 1986

Korpusová lingvistika a počítačová lexikografie

- **Randolph Quirk** (1920) – korpus ***The Survey of English Usage (SEU)***, 1959
  - University College London, první korpusové pracoviště
  - v týmu také Jan Firbas (český jazykovědec, anglista)
  - vzorky psané a mluvené britské angličtiny (půl na půl), z let 1955 až 1985
  - 200 textů, každý 5000 slov, mluvené – monology i dialogy (shromažďováno 30 let)
  - původně na papíře (lístky 6 x 4 palce), později převeden do počítačově čitelné podoby (Svartvik)
- R. Quirk – *Towards a description of English Usage*, 1960, publikace o SEU

Korpusová lingvistika a počítačová lexikografie

- SEU 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 – ***The London-Lund Corpus of Spoken English***, Lund University, Sweden (100 přepisů, 500 tis. slov, zveřejněn až 1980)
  - 1. počítačový korpus mluveného jazyka
  - **SEU** – 13 textů mluvené angličtiny
  - **Survey of Spoken English (SSE)**, Jan Svartvik, Lund University, 1975 jako sesterský projekt London Survey
    - 87 textů mluvené angličtiny

Korpusová lingvistika a počítačová lexikografie

- **COBUILD** – Collins Birmingham University International Language Database, britské výzkumné centrum na University of Birmingham, od r. 1980 založeno vydavatelstvím Collins, na počátku vedl profesor

John Sinclair (1933–2007)

- **Birmingham Collection of English Text** (Collins Corpus), 1980, jako první využil OCR
- ***Collins COBUILD English Language Dictionary***, 1987, Sinclaire (pro výuku angličtiny jako cizího jazyka), první slovník založený na současné, běžně užívané angličtině

Corpus, Concordance, Collocation, (Oxford University Press, 1991)

Korpusová lingvistika a počítačová lexikografie

- **Deutsches Referenzkorpus** (DeReKo), 1964, Mannheim Korpus, Institut für Deutsche Sprache

- **LIMAS** (Linguistik und Maschinelle Sprachbearbeitung), 1970, Universität Bonn
  - německá varianta Brown Corpus – 500 textů, 15 kategorií, 1 mil. slov, texty z let 1969–70