PLIN041 Vývoj počítačové lingvistiky
Korpusová lingvistika a počítačová lexikografie

Mgr. Dana Hlaváčková, Ph.D.

od 60. let 20. st.

Henry Kucera, Winthrop Nelson Francis

- 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)
- průkopník korpusové lingvistiky, 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), průkopník korpusové lingvistiky
- v r. 1977 spoluzakladatel ICAME (International Computer Archive of Modern and Medieval English)

Brown Corpus

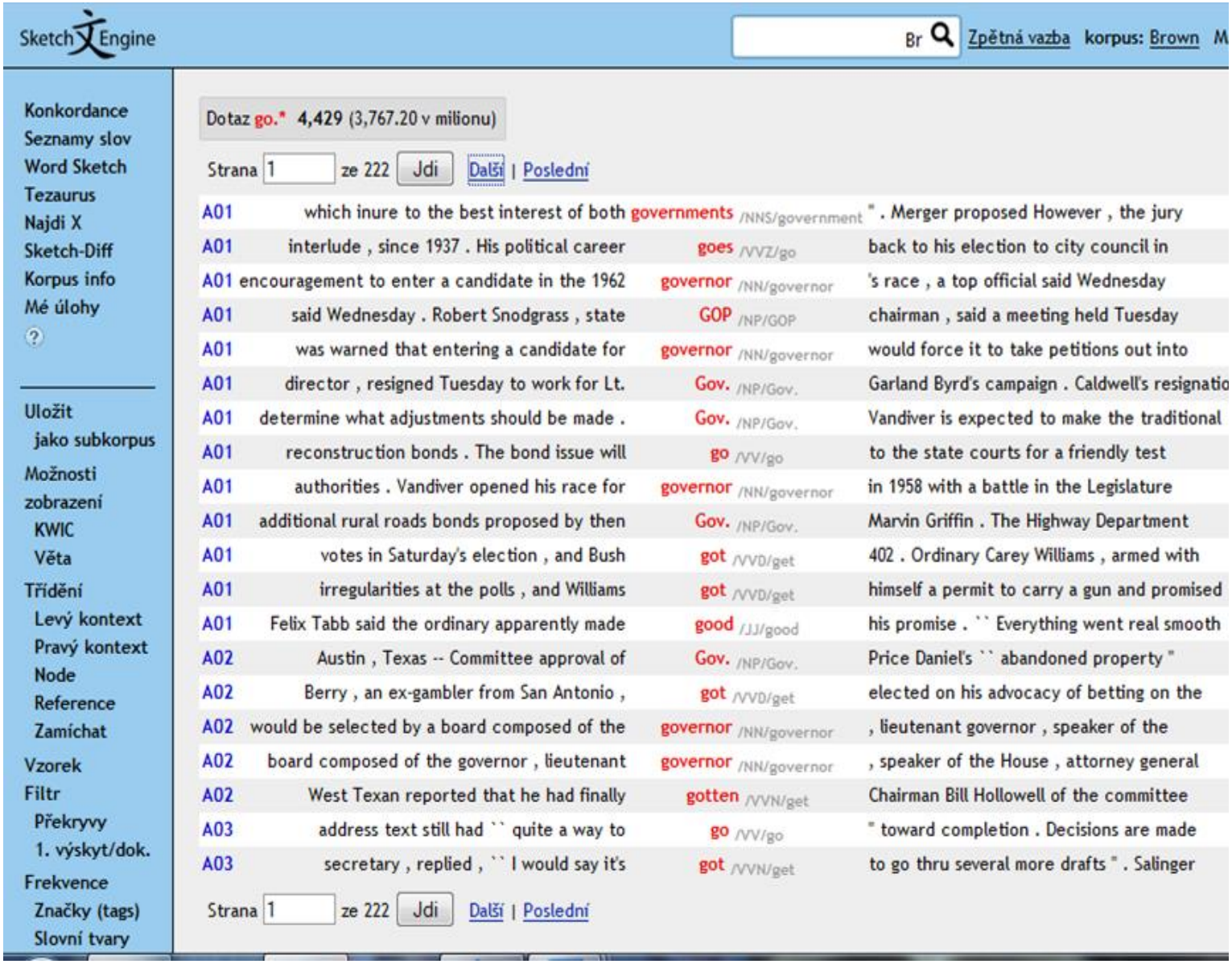
- ***Brown Corpus*** *(Brown Standard Corpus of Present-Day American English)*, 1963–**1964**, Brown University
- americká angličtina rodilých mluvčích
- **500** textových vzorků (vždy **2000** slov)
- **15** žánrových kategorií (časopisy, noviny, beletrie, odborná lit.), snaha o vyváženost
- **1 mil.** slov, vše z roku 1961
  - morfologicky označkován (PoS tagging – 80 kategorií)
  - na delší dobu vzor pro další korpusy
  - dostupný přes Sketch Engine
- **The Freiburg-Brown corpus of American English (Frown)**
- stejná struktura, Albert-Ludwigs-Universität Freiburg
- zveřejněn 1992, PoS tagging 2007

Publikace založené na Brown Corpus

- ***Computational Analysis of Present-Day American English,*** 1967, statistické studie (lingvistika, psychologie, statistika, sociologie)

- ***Frequency Analysis of English Usage***, 1967

- ***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

Geoffrey Leech, Stig Johansson

- **Geoffrey Leech** (1936–2014)
- **Stig Johansson** (1939–2010)
- narodil se ve Švédsku, později se stal norským občanem
- studoval na Lund University (Švédsko) a Indiana University (USA)
- profesor moderní angličtiny na Oslo University (Norsko)
- společně autoři korpusu

Lancaster-Oslo/Bergen Corpus

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 (Knut Hofland, programátor)
- originální verze – 1976
- značkovaná verze (PoS tagging) 1981–1986
- **The Freiburg–LOB Corpus of British English (F-LOB)**
- stejná struktura, z roku 1991, zveřejněn 1999, PoS tagging 2007)
- Albert-Ludwigs-Universität Freiburg

Randolph Quirk, SEU

Randolph Quirk (1920)

- korpus a pracoviště ***The Survey of English Usage (SEU)***, 1959
  - University College London, první korpusové pracoviště
  - 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)
- SEU byl použit pro jednu z nejdůležitějších korpusově založených gramatik – *Comprehensive Grammar of the English Language* (Quirk, Greenbaum, Leech, Svartvik, 1985)