

# PLIN041 Vývoj počítačové lingvistiky

## Korpusová lingvistika od 70. let 20. st.

Mgr. Dana Hlaváčková, Ph.D.

PLIN041 Vývoj počítačové lingvistiky  
Korpusová lingvistika a počítačová lexikografie

Mgr. Dana Hlaváčková, Ph.D.

*od 60. let 20. st.*

# Henry Kucera, W. 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

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 Corpus

- **Geoffrey Leech, 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 (Knut Hofland, programátor)
  - značkováná verze (PoS tagging) 1981–1986

# LOB Corpus

- **The Freiburg–LOB Corpus of British English (F-LOB)**
- stejná struktura, z roku 1991, zveřejněn 1999, PoS tagging 2007)
- Albert-Ludwigs-Universität Freiburg



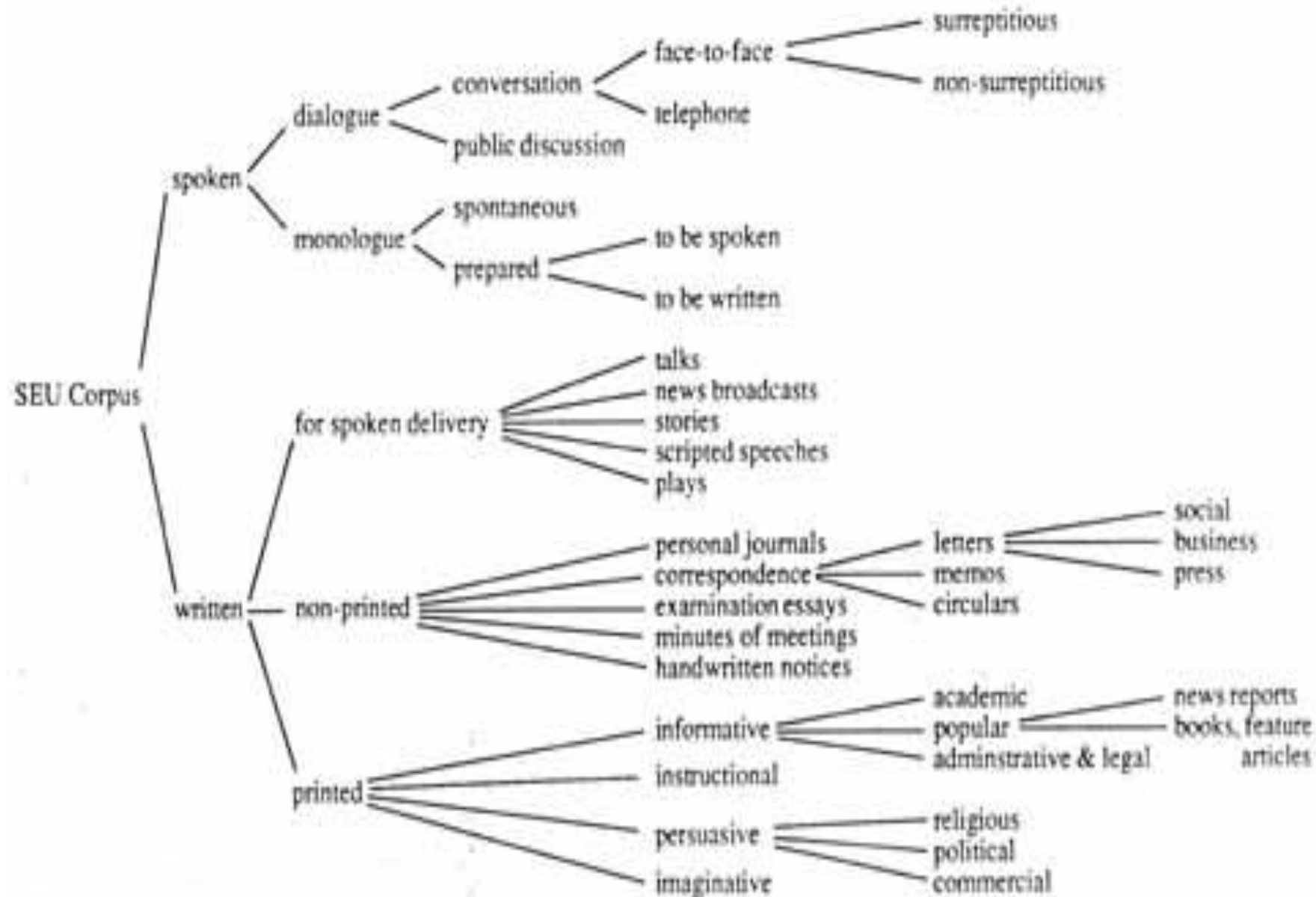
*Stig Johansson*

- **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)



# SEU

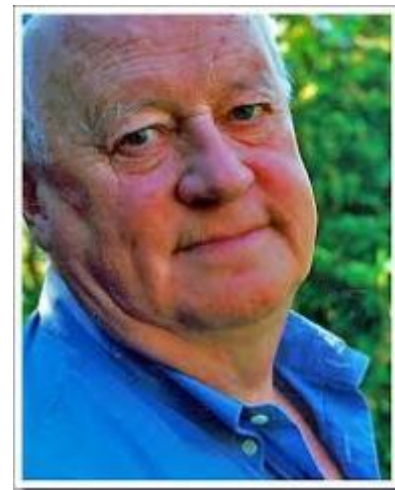
- **Randolph Quirk** (1920–2017)
- korpus a pracoviště *The Survey of English Usage (SEU)*, zal. 1959
  - University College London, první korpusové pracoviště
  - korpus – vzorky psané a mluvené britské angličtiny (půl na půl), z let 1955 až 1985
  - 200 textů, každý 5000 slov, mluvené – monology i 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 – Gang of Four, 1985)



# LLC

- **Jan Svartvik, 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)

# Jan Svartvik (1931)



- švédský lingvista (anglistika)
- studium na univerzitě v Uppsale, stáže na Brown University, UCL a dalších univerzitách
- 1959–60 na University of Durham u R. Quirka
- profesor angličtiny na univerzitě v **Lundu** (1970–1995)
- přední představitel korpusové lingvistiky, člen **Gang of Four (1961–64 SEU na UCL)**
- ve švédské lingvistice měl text ústřední postavení, jako nerodilý mluvčí nemohl využívat introspekce (důvod, proč se věnoval korpusové lingvistice)
- členství a předsednictví v mnoha vědeckých organizacích ve Švédsku i ve světě
- čestný doktorát z univerzity v Bergenu a z MU (1998, 1. Švéd v historii MU)
- spolupráce s českým lingvistou Janem Firbasem
- spolupráce s katedrami anglistiky FF a PedF MU

# Jan Svartvik

- pojmenoval nový směr – **forezní lingvistika** (forensic linguistics)
  - *The Evans Statements: A Case for Forensic Linguistics*, 1968
  - rozbor tvrzení T. J. Evanse falešně obviněného a odsouzeného za vraždu manželky a dcery (1950)
  - na základě kvalitativní a kvantitativní analýzy předkládá pochybnosti o autorství a pravosti tvrzení
- *A Grammar of Contemporary English* (1972) – v té době Svartvik v Lundu, komunikace na dálku
- *A Comprehensive Grammar of the English Language* (1985)
- *A Life in Linguistics* (2005) – vzpomínky
- *English – One Tongue, Many Voices*, s G. Leechem (2006)

# Propojení lexikografie s korpusovou lingvistikou

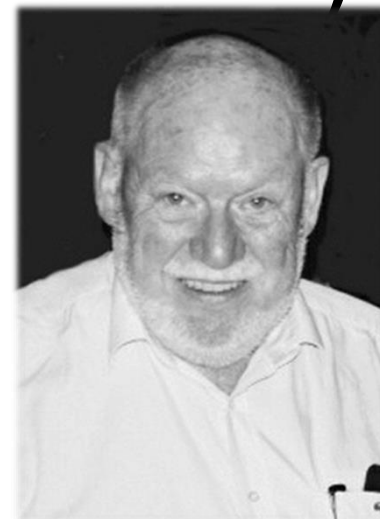
- **COBUILD** – Collins Birmingham University International Language Database, britské výzkumné centrum na University of Birmingham, od r. 1980 založeno vydavatelstvím Collins (dnes HarperCollins Publishers), na počátku vedl profesor **John Sinclair** (1933–2007)
- cílem vydání slovníku pro výuku angličtiny
- korpus **Birmingham Collection of English Text** (BCE), 1980, jako první využil OCR
  - Centre for Corpus Research University of Birmingham (od 70. let)
  - 20 mil. slov, hlavně psaná britská angličtina
  - jiná struktura než první korpusy (noviny, brožury, letáky, knihy, časopisy, korespondence), oproti LOB vyloučena poezie a drama

# Propojení lexikografie s korpusovou lingvistikou

- ***Collins COBUILD English Language Dictionary*, 1987, J. Sinclair**
  - pro výuku angličtiny jako cizího jazyka
  - první slovník založený na současné, běžně užívané angličtině
- **Bank of English (BoE)**, 650 mil. slov (na poč. 90. let 200 tis.), mluvená i psaná angličtina, pro výuku angličtiny, označován
  - dnes součástí ***Collins Corpus*** – 4,5 mld., stále se rozrůstá

# John McHardy Sinclair (1933–2007)

- britský lingvista (původem Skot)
- studoval na univerzitě v Edinburghu
- profesor moderní angličtiny –  
**Birmingham University** (1965–2000)
- jako emeritní profesor stále publikoval
- oblasti zájmu:
  - průkopník **analýzy diskurzu**
  - ***idiom principle*** (v komunikaci používáme předpřipravené konstrukce frází)
  - *Towards the Analysis of Discourse*, 1975
  - **lexikografie** – kolokace, **korpusová lingvistika**, výuka angličtiny





# John Sinclair

- *Corpus, Concordance, Collocation* (Oxford University Press, 1991)
- *Reading Concordances* (2003)
- *Trust the Text* (2004)
  
- 5 dětí, druhá žena *Elena Tognini-Bonelli*
  - italská anglistka, korpusová lingvistka, působí na univerzitě v Sieně (Toskánsko)
- založil asociaci *Tuscan Word Centre*, Certosa di Pontignano, výzkumné centrum korpusové lingvistiky

# Korpusy – Německo, Francie

- **Deutsches Referenzkorpus** (DeReKo), 1964, (*Mannheim corpora, IDS corpora, COSMAS corpora*), Institut für Deutsche Sprache
  - dnes 49,6 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, metainformace o textech (laboratoř ATILF, Nancy)

# 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)
- *BNC World Edition*, 2001 (metainformace)
- poslední *BNC XML Edition* z r. 2007, jinak se korpus nemění (původně v SGML)
- 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)

# Lidé kolem BNC

- **Geoffrey Leech** – *100 Million Word of English* (English Today, 1993)
- **Tony McEnery**, Andrew Wilson – *Corpus Linguistics: An Introduction* (Oxford University Press, 2001)
- Paul Baker, Lancaster University
- **Sue Atkins**, Jeremy Clear, Nicholas Ostler – *Corpus Design Criteria* (Literary and Linguistic Computing, 1992)
- **Michael Rundell** (Lexicography MasterClass)

# První korpusy

- *Brown Corpus*, 1963–1964 (Kučera, Francis)
- *Lancaster-Oslo/Bergen Corpus (LOB)*, 1970–1978 (Leech, Johansson)
- korpus a pracoviště *The Survey of English Usage (SEU)*, 1959, 1985 (Quirk, Gang of Four)
- *The London-Lund Corpus of Spoken English (LLC)*, 1980 (Svartvik, Greenbaum, Quirk, Hofland)
- *Birmingham Collection of English Text*, 1980 (John Sinclair)
- *International Corpus of English*, 1990 (Greenbaum)
- *Bank of English (BoE)*, poč. 90. let (John Sinclair)
- *British National Corpus*, 1994