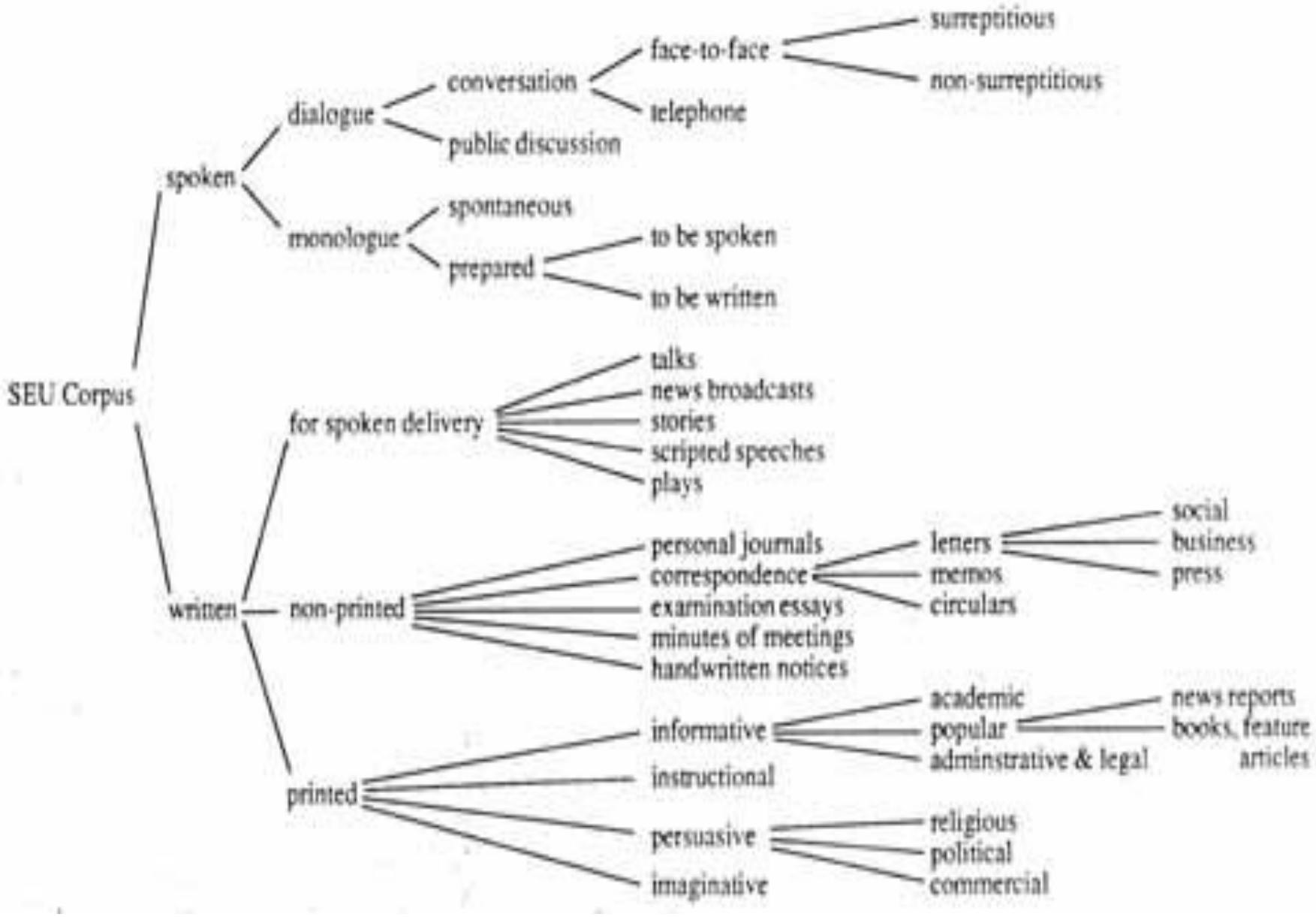
PLIN041 Vývoj počítačové lingvistiky

Korpusová lingvistika
od 70. let 20. st.

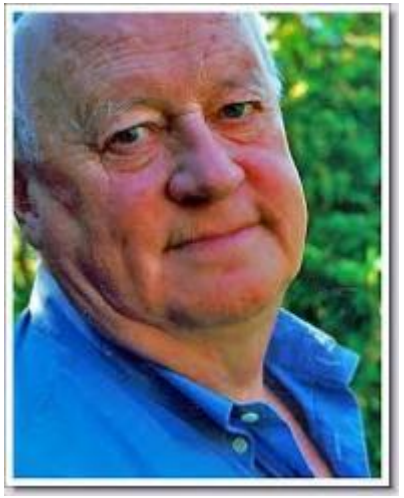
Mgr. Dana Hlaváčková, Ph.D.

Korpusová lingvistika

- **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 přepisů nahrávek, 500 tisíc slov, zveřejněn až 1980
  - fonetická transkripce, značeny prozodické vlastnosti
  - někteří mluvčí o nahrávání nevěděli (spontánní projev)


SEU Corpus
spoken
written
dialogue
monologue
conversation
public discussion
spontaneous
prepared
face-to-face
telephone
surreptitious
non-surreptitious
to be spoken
to be written
for spoken delivery
non-printed
printed
talks
news broadcasts
stories
scripted speeches
plays
personal journals
correspondence
examination essays
minutes of meetings
handwritten notices
letters
memos
circulars
social
business
press
informative
instructional
persuasive
imaginative
academic
popular
adminstrative & legal
news reports
books, feature articles
religious
political
commercial

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 – **forenzní 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. Sinclaire
  - 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, označkován
  - 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 29 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

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*