Korpusy neslovanských jazykov

História

- začiatok 17. storočia - Samuel Johnson - Dictionary of English Language
  - 40 tisíc hesiel
  - 150 000 citácií = 1 mil. slov
- 1828 - Noah Webster – Dictionary of the English Language
- do roku 1928 - Oxford English Dictionary
  - práce trvali 71 rokov - 2000 spolupracovníkov
  - skoro 0,5 mil. hesiel
  - 5 mil. citácií = 50 mil. Slov
  - pre školskú prax - vytvorené zoznamy najčastejších slov – pravopis
  - pre gramatiky
- 1962 - Chomsky – s dešpektom sa vyjadril o akejkoľvek možnosti prínosu korpusu k rozvoju jazykovedy a poznania jazyka

Survey of English Usage

- Randolph Quirk, Londýn
- nie elektronická forma
- písaná a hovorená angličtina (dnes 1953 – 1987)
- 1 milión slov
- výstup: A Comprehensive Grammar of the English Language
- http://www.ucl.ac.uk/english-usage/

Korpusy prvej generácie

- 1961 – 1964 Brown Corpus – BC
- 1970 – 1978 Lancaster-Oslo/Bergen - LOB Corpus
- 1975 Survey of Spoken English – SSE (Prehľad hovorenej angličtiny)
- London-Lund Corpus – LLC

Brown Corpus

- 1961 – 1964
- Henry Kučera, Nelson Francis
- Brown University
- americká angličtina z roku 1961
- neimaginatívna i imaginatívna – próza
- 1 milión slov
- anotovaný
- prístupný: CD-ROM
- http://khnt.hit.uib.no/icame/manuals/brown/index.htm

LOB Corpus

- 1970 – 1978
- G. Leech, S. Johansson, K. Hofland
- britská angličtina z roku 1961
- 1 milión slov
- anotovaný
- prístupný: CD-ROM
- http://khnt.hit.uib.no/icame/manuals/lob/index.htm

Survey of spoken English London Lund Corpus

- SSE
  - 1975
  - prepisovanie hovorenej časti **SEU**
- London – Lund Corpus
  - J. Svartvik, Švédsko – Lund
  - hovorená angličtina
  - 0,5 mil. slov
  - anotovaný
  - prístupný: CD-ROM
- http://khnt.hit.uib.no/icame/manuals/londlund/index.htm

Korpusy druhej generácie

- Longman/Lancaster English Language Corpus
- International Corpus of English
- Lancaster/IBM Spoken English Corpus – SEC
- Corpus of Spoken American English – CSAE
- Kolhapur Corpus of Indian English
- Wellington Corpus of Written New Zealand English
- Australian Corpus of English – ACE
- Macquarie Corpus of Written Australian English
- Corpus of English-Canadian Writing
- ...

Longman Lancaster English Language Corpus

- R. Quirk, G. Leech
- 30 miliónov slovných tvarov
- reprezentatívny
- využitie v lexikografii: slovníky vydavateľstva Longman
- http://www.longman.com/dictionaries/corpus/lclonlan.html

Korpusy tretej generácie

- British National Corpus (BNC)
- American National Corpus (ANC)
- COBUILD Corpus – Bank of English (BoE)
- Korpusy IDS v Mannheime
- 1998 CORIS/CODIS – CORpus di Italiano Scritto
- The Longman Spoken and Written English Corpus– LSWE Corpus
- národné korpusy: ...

British National Corpus

- Oxford University, Lancaster University, British Library, Oxford University Press, Longman, Chambers-Larousse
- synchrónny všeobecný vyvážený korpus
- britská angličtina z konca 20. storočia (od roku 1960)
- hovorené (22 %) i písané (78 %) texty
- 100 miliónov slov (10 miliónov z transkribovaných hovorených prejavov)
- vonkajšia anotácia - 60% kníh, 25% periodík
- POS – automaticky (Lancasterská univerzita – UCREL – Unit for Computer Research on the English Language)
- 2 milióny - **The BNC Sampler** – POS – ručne desambiguované

British National Corpus

- **SARA** – korpusový manažér cez internet
  - prístup: jednoduché vyhľadávanie na internete
  - online 60 libier
- **The BNC Sampler**- 30 libier
- **The BNC World edition**(2 CD-ROMy) – 50 libier pre jednotlivca, 250 libier pre licenciu na sieť - len na vedeckovýskumné účely
- **The BNC Consortium**– schvaľuje použitie na komerčné účely
- http://www.natcorp.ox.ac.uk/

Collins COBUILD/Bank of English

- 1980 **COBUILD Corpus – Collins Birmingham University International Language Database**
  - štandardná angličtina z roku 1960
  - 25% hovorené texty
- 1991 **Bank of English – BoE**
  - John Sinclair
  - vydavateľstvo HarperCollins Publishers a univerzita v Birminghame
  - monitorovací korpus hovorenej i písanej angličtiny po roku 1990
  - 524 miliónov
  - POS - automaticky anotovaný
  - Fred Karlsson - Helsinská univerzita - 200 miliónov syntakticky anotovaných

Collins COBUILD/Bank of English

- Prístup:**CobuildDirect**
  - **anglický:** Wordbanks***Online*** English corpus
    - 56 miliónov slov z **BoE**
    - písaný aj hovorený (americká angličtina)
    - anotovaný - automaticky
  - **francúzsky:** 78 miliónov slov – moderná francúzština
  - **španielsky:** 73 miliónov slov – európska aj latinskoamerická šp.
- jednoduché hľadanie, kolokácie, regulárne výrazy, tagy
- http://www.collins.cobuild.co.uk/direct_demo.html
- telnet titania.cobuild.collins.co.uk

American National Corpus

- ANC Consortium: univerzity – New York, Pennsylvania, Arizona, Colorado, California, Berkeley a komerčné firmy – IBM, Sony, Microsoft, vydavateľstvá
- americká angličtina – korpus porovnateľný s BNC
- hovorená podoba (telefóny, rozhovory) a písaná podoba (The New York Times, turistickí sprievodcovia, publikácie univerzít)
- 100 miliónov slov plánovaných - teraz 10 miliónov (testovací)
- automaticky anotované – POS, lemy
- Ciele: audio a video záznamy hovorených textov, dokončenie do 2005, sprístupnenie bude v 2008
- prístup: **ANC Consortium** – exkluzívny prístup, zadarmo pre univerzity - je tam možné sa prihlásiť za 40 000 $ (AKCIA!!)
- **Linguistic Data Consortium** - pre jednotlivcov 75 $
- http://americannationalcorpus.org/

International Corpus of English

- S. Greenbaum
- 10 národných skupín a variet angličtiny z celého sveta: India, Filipíny, Singapur, VýchodnáAfrika, Veľká Británia, Nový Zeland
- anotácia:
  - externá
  - lingvistická: morfologická, syntaktická
- http://www.ucl.ac.uk/english-usage/ice/

Korpusy IDS v Mannheime

- Das Institut für Deutsche Sprache (IDS), Mannheim
- písaná nemčina
- skoro 2 miliardy slov, 26 mil. automaticky anotovaných
- **Mannheimský korpus 1** (1856,- EUR)
  - 2,2 mil. slov z rokov 1950 – 1967
- **Mannheimský korpus 2** (169,- EUR)
  - 0,3 mil. slov z rokov 1949, 1952, 1960 – 1974
- **Bonnský korpus novinových textov** (2250,- EUR)
  - 3,1 mil. slov z rokov 1949, 1954, 1959, 1964, 1969, 1974
- http://www.ids-mannheim.de/kt/projekte/korpora/

CORIS/CODIS – CORpus di Italiano Scritto

- R. Rossini Favretti – CILTA – Centrum teoretickej a aplikovanej lingvistiky, Univerzita v Bologni
- písaná taliančina (80-te a 90-te roky)
- všeobecný, synchrónny
- 100 miliónov slov – každé tri roky obnovovaný monitorovací korpus
- od 2001 prístupný
- pre porovnávacie potreby – **CODIS** – dynamický, prispôsobiteľný
- 1997 The Bononia Legal Corpus – BOLC
- *Rema Rossini Favretti, Fabio Tamburini, Sinclair* ako poradca
- porovnávací – multilingválny: talianske a anglické právne texty
- 33,5 milióna talianskych slov + 21 miliónov angl. slov – prevažne z internetu

Corpas Náisiúnta na Gaeilge ITÉ National Corpus of Irish

- 15 miliónov slov
- 8 miliónov: externá anotácia – SGML
- ciele: morfosyntaktická anotácia; zatiaľ 150 000 slov (30 tis. ručne)
- prístup: CDROM, 50 euro
- http://www.ite.ie/corpus/

Maďarský národný korpus
Magyar Nemzeti Szövegtár

- Maďarská akadémia vied, Lingvistický výskumný ústav, Tamás Váradi
- 100 miliónov slov
- reprezentatívny všeobecný korpus súčasnej maďarčiny
- 5 subkorpusov (médiá, literatúra, odborná l., administr. a právo, neformálne texty)
- vyhľadávanie on-line po registrácii
- morfosyntaktická anotácia, lematizácia
- http://corpus.nytud.hu/mnsz/

Lietuvių kalbos tekstynas

- Vytautas Magnus University – Centre of Computational Linguistics
- Rūta Marcinkevičienė, Vytautas Zinkevičius, Artūras Gruodis, Andrius Utka
- súčasná písaná litovčina
- všeobecný monitorovací korpus
- texty po roku 1994
- ešte nelematizovaný a neanotovaný
- MorfoLema – nástroj na morfologickú analýzu a syntézu
- http://donelaitis.vdu.lt/tekstynas/

Špecializované korpusy

- hovorené korpusy
- historické korpusy
- syntakticky anotované korpusy: **treebanky**

Corpus of London Teenager Language

- 1994
- súčasť BNC
- 13 – 17-roční teenageri
- 0,5 milióna slov
- http://helmer.aksis.uib.no/colt/

Map Task Corpus

- Glasgow
- hovorená podoba jazyka škótskych študentov
- 147 tisíc slov
- prístupný: CD-ROM
- ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/info/MapTaskCorpus

Lancaster/IBM Spoken English Corpus

- 1984 – 1987
- hovorená angličtina
- 54 000 slov
- anotovaný – prozódia, syntax, pravopis
- zvukové nahrávky
- http://www.ling.lancs.ac.uk/staff/gerry/SEC.htm

Corpus of Spoken American English

- hovorené prejavy - dialekty
- 1 milión slov
- transkripcia
- zvukové nahrávky
- http://www.ldc.upenn.edu/Projects/SBCSAE/

Guangzhou Petroleum English Corpus

- koniec 80. rokov
- čínsky
- angličtina v ropnom priemysle
- 411 tisíc slov
- zoznam najfrekventovanejších slov

Helsinki Corpus of English Texts

- 1984 - 1991
- M. Rissanen, O. Ihalainen, M. Kytö
- diachrónny (r. 750 – 1700)
- 1,5 milióna slov
- prístupný: CD-ROM

SUSANNE Corpus

- Surface and underlying structural analysis of natural English
- Geoffrey Sampson, Hélène Knight, Robin Haigh, Tim Willis a Nancy Glaister
- Lancaster University, LOB Corpus – inšpirácia – bol potrebný ručne syntakticky označkovaný korpus
- podpora The Economic & Social Research Council
- písaná americká angličtina
- 130 tisíc slov z Brown Corpusu – syntakticky anotované
- v analýze uplatnená syntax bezprostredných zložiek
- voľne prístupný
- http://www.grsampson.net/RSue.html

The Penn Treebank

- LINC Laboratory of Computer and Information Science Department
- Univerzita v Pennsylvánii
- zdroje: Wall Street Journal, Brown Corpus, prepisy zvuk. Nahrávok
- POS anotácia, syntaktická anotácia
- prístupné: CD-ROM - 1 milión slov z roku 1989 (Wall Street Journal), anotovaný
- vzniká aj Proposition Bank (sémantika),Chinese Treebank, Korean Treebank
- http://www.cis.upenn.edu/~treebank/home.html

Negr@ Corpus

- Hans Uszkoreit, Thorsten Brants
- Universität des Saarlandes, Saarbrücken
- nemecké texty z periodika Frankfurter Rundschau
- 355 tisíc tokenov
- pri morfologickej anotácii použitý TnT tager a Stuttgart-Tübingen Tagset STTS
- syntakticky anotovaný – poloautomaticky
- http://www.coli.uni-sb.de/sfb378/negra-corpus/

Niektoré jazykovedné práce založené na korpuse

- Longman Dictionary of Contemporary English
- Collins Cobuild English Language Dictionary
- Collins Cobuild English Grammar
- Oxford English Dictionary
  - 2. vydanie
  - elektronická podoba – CD-ROM
- 1981 Dictionary of Old English
  - textový materiál k štúdiu angličtiny – citácie – Complete Corpus of English