

Korpusová lingvistika – 3

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

Typy jazykových korpusů

- znalost různých typů korpusů
- schopnost vybrat správný korpus pro svoji práci
- v průběhu času přibývají nové typy
- různá kritéria dělení
 - z hlediska obsahu, rozsahu, způsobu budování, ...
- stejné typy korpusů v celosvětovém měřítku
- národní korpusy a subkorpusy
- každý korpus je charakterizován více přívlastky

Jak vybrat správný korpus

- zařazení do typologie – obsah
- rozsah (frekvenční studie, statistická významnost)
- metodologie tvorby korpusu
- typ značkování a metadata
- přístupnost

Typy jazykových korpusů

- typ zachycené komunikace
 - **psané** korpusy (Written Corpora)
 - hraniční typ – soukromá korespondence, KSK
 - v ČNK – korpusy řady SYN, PUB
 - **mluvené** korpusy (Spoken Corpora)
 - včetně čtení předem připraveného psaného projevu
 - v ČNK – PMK, BMK, řada ORAL, ORTOFON
 - **multimodální** – DIALOG (ÚJČ), MALACH (MFF UK)

Typy jazykových korpusů

- časový záběr
 - **synchronní** – cca 2. pol. 20 st. – současnost
 - korpusy řady SYN, PUB
 - **diachronní** – 13. st. – 1945 (beletrie)/1989 (publicistika, odborné texty)
 - DIAKORP

Typy jazykových korpusů

- zachycený jazyk
 - **jednojazyčné** – např. SYN2015
 - **paralelní** – InterCorp, ORWELL (stejně texty v překladech do různých jazyků)
 - psané **srovnatelné** korpusy (Comparable Corpora) – Aranea (14 jazyků, V. Benko)
 - shodná metodika tvorby, srovnatelná velikost, čas pořízení textů, způsob budování
- účel
 - **všeobecné** – řada SYN, ORAL (psané, mluvené, synchronní, bez speciálního zaměření)
 - **specializované** – např. KSK, BMK, DIALEKT
 - **žákovské** korpusy (Learner Corpora) – psané texty češtiny produkované nerodilými mluvčími

Typy jazykových korpusů

- způsob budování
 - **tradiční** – poskytovatelé textů (autorský zákon), elektronické texty, OCR
 - SYN2013PUB
 - **mluvené** – nahrávka, přepis
 - **webové** – texty stahované z internetu
 - ukWaC, deWaC, CzTenTen12 a CzTenTen17 (CZPJ FI MU)
- možnost rozšíření
 - uzavřené – **referenční**, neměnný v čase, zpětně dostupný
 - jiný význam – při porovnávání frekvence jevů – standard, vzor obecného užívání jazyka
 - otevřené (monitorovací) – **nereferenční**

Typy jazykových korpusů

- značkování
 - **neoznačované** (morfologicky) – BMK, KSK, DIAKORP
 - značkové (morfologicky)
 - **lemmatizace** (přiřazení základních tvarů slov)
 - **tagging** (lemmata + morfologické značky), PoS tagging
 - značkové (foneticky, syntakticky)
 - **fonetická transkripce**
 - OMK (Katedra bohemistiky UP Olomouc, dr. Petr Pořízka), ORTOFON
 - **syntax** – SYN2015, PDT (ÚFAL MFF UK Praha)
- verzované korpusy – SYN v. 6, InterCorp v. 10, DIAKORP v. 6

British National Corpus (BNC)

- vedoucí role – Geoffrey Leech
- **jednojazyčný, synchronní, vyvážený, reprezentativní, široký průřez britskou angličtinou z konce 20 st., 1991–1994**
- BNC Consortium (Oxford University Press), British Library, British Academy
- **100 mil. slov, 90 % psaný jazyk (vzorky po 45 tis. slovech), 10 % mluvený**
- **publicistika, beletrie, odborná literatura, formální i neformální promluvy**
- **PoS Tagging**

Ústav Českého národního korpusu

- **ÚČNK FF UK**, Panská ul., Praha, www.korpus.cz
- založen 1994, ředitel prof. F. Čermák, od r. 2013 doc. V. Cvrček, od 2015 dr. M. Křen
- spolupráce s Ústavem teoretické a počítačové lingvistiky FF UK a Ústavem formální a aplikované lingvistiky MFF UK
- budování ČNK
- publikační činnost
- výuka pro magisterský stupeň
- výuka pro doktorský stupeň – obor Matematická lingvistika

Český národní korpus

- korpusy řady **SYN** (všeobecný korpus, psaný synchronní jazyk, referenční korpusy, celkem 4 mld. slov, lemmatizace, morf. značkování)
 - **SYN2000** – 1990–1999, 100 mil. slov
 - beletrie 15 %, odborná lit. 25 %, publicistika 60 %
 - **SYN2005** – 2000–2004, 100 mil. slov
 - beletrie 40 %, odborná lit. 27 %, publicistika 33 %
 - **SYN2010** – 2005–2010, 100 mil. slov
 - beletrie 40 %, odborná lit. 27 %, publicistika 33 %
 - **SYN2015** – 2011–2015, 100 mil. slov
 - beletrie 33 %, odborná lit. 33 %, publicistika 33 %
 - **SYN2006PUB** – 1989–2004, 300 mil. slov
 - **SYN2009PUB** – 1995–2007, 700 mil. slov
 - **SYN2013PUB** – 2005–2009, 935 mil. slov

Český národní korpus

- specializované korpusy, např.
- **KSK-DOPISY** – ručně psané dopisy z let 1990–2004, 800 tis. slov
- **ORWELL** – román G. Orwella 1984, ručně značkovaný, 80 tis. slov, 2003 (MULTEXT-East, 12 jazyků, paralelní korpus)
- **CzeSL-Plain** – žákovský korpus nerodilých mluvčích, 2 mil. slov, 2012 (eseje cizinců, odborné závěrečné práce, slohové práce romských žáků)
- **LINK** (Lingvistův Narozeninový Korpus k výročí prof. F. Čermáka), odborné lingvistické texty z let 1985–2010, 1,8 mil. slov

Český národní korpus

- diachronní korpus **DIAKORP** v. 6, 2005
- 3,4 mil. slov
- texty od konce 13. st. po hranice synchronní složky (48 % z 19. st.)
- různorodost textů – pravopisné systémy, tisky, rukopisy
- transkripce (rekonstrukce)
- vnětextové i vnitrotextové značkování, např. nezřetelné/nečitelné úseky, cizojazyčné citáty, poznámky pod čarou (bez morfologie)
- využití tzv. hyperlemmat (kůň – kuoň, kóň)

Český národní korpus

- paralelní korpus **InterCorp**, 0. verze 2008, 19 jazyků
- 1,46 mld. slov (verze 10, 2017), 39 jazyků
- párování (alignment), čeština = pivot
- jádro (core) korpusu – ručně zarovnané hl. beletristické texty
- kolekce (collection) – texty zarovnané automaticky
 - publicistika a zpravodajství z webu
 - právní texty Evropské unie
 - zápisy z jednání Evropského parlamentu 2007–2011 (Europarl)
 - filmové titulky z databáze Open Subtitles
 - překlady Bible

Český národní korpus

- srovnatelné webové korpusy **Aranea** (dr. Vladimír Benko, Bratislava)
- 2014, nereferenční, 14 jazyků, stejná velikost, stejná metodika a technologie tvorby
- open-source (volně dostupné) nástroje, PoS tagging
- základní velikost
 - 1,2 mld. slov = maius
 - 10% vzorek, 120 mil. = minus, určeny pro vyučování
 - 10 mil. = minimum, pro testování nástrojů
 - maximum – „koliko sa podarí“, 5,5 mld. jen pro češtinu

Korpusy v CZPJ FI MU

- **CzTenTen12** – webový korpus z r. 2012,
 - značkováný (morfologie), 4,2 mld. slov
- **czTenTen17** – 2017, untagged
- **Czes** – novinové články z let 1995–1998 a 2002, 350 mil. slov
- **OPUS2 Czech** – česká část paralelního webového korpusu
 - Jörg Tiedemann, Uppsala University
- **CNPK** – česko-německý paralelní korpus, Katedra německého jazyka a literatury PedF MU, 2005, celkem přes 7 mil. slov, snaha o vyváženost
- **chyby2** – texty studentů FI MU s vyznačenými chybami, 50 tis. slov
- **DESAM** – 1 mil. slov, ruční desambiguace, referenční korpus

Přístup ke korpusům

- Český národní korpus <http://wiki.korpus.cz/doku.php/cnk:uvod>
- Sketch Engine <http://ske.fi.muni.cz>
- Aranea <http://unesco.uniba.sk/aranea>