Korpusová lingvistika – 3

Typy korpusů

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

PRZA009

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 (např. psaný, synchronní, verzovaný)

Jak vybrat správný korpus

- zařazení do typologie – **obsah**
- **rozsah** (frekvenční studie, statistická významnost)
- metodologie **tvorby** korpusu (má často vliv na obsah)
- typ **značkování** a metadata
- **přístupnost** – volné verze, přihlašovací údaje, typy korpusových manažerů

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, ORATOR
  - **multimodální/multimediální** – DIALOG a MONOLOG (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 (14.–20. století)

Typy jazykových korpusů

- zachycený jazyk
  - **jednojazyčné** – národní, např. SYN2020
  - **paralelní** – stejné texty v překladech do různých jazyků
    - **InterCorp**
    - ORWELL (Multext East)
    - Česko-německý paralelní korpus (PedF MU)
    - mezinárodní OPUS (online texty), EuroParl
  - psané **srovnatelné** korpusy (Comparable Corpora) – **Aranea** (24 jazyků, V. Benko)
    - shodná metodika tvorby, srovnatelná velikost, čas pořízení textů, způsob budování

Typy jazykových korpusů

- účel
  - **všeobecné** – řada SYN, ORAL (psané, mluvené, synchronní, bez speciálního zaměření)
  - **specializované** – např. KSK, BMK, DIALEKT
    - **žákovské/akviziční** korpusy (Learner Corpora) – psané texty češtiny produkované nerodilými mluvčími

Typy jazykových korpusů

- **způsob budování**
  - **tradiční** – poskytovatelé textů (autorský zákon), elektronické texty, OCR
  - **mluvené** – nahrávka, přepis (souhlas s nahráváním)
  - **webové** – texty stahované z internetu
    - ukWaC, deWaC, CsTenTen12 a CzechWeb17
- **možnost rozšíření**
  - uzavřené – **referenční**, neměnný v čase, zpětně dostupný
  - otevřené (monitorovací) – **nereferenční**

Typy jazykových korpusů

- **značkování**
  - **neoznačkované** (morfologicky) – BMK, KSK, DIAKORP
  - značkované (morfologicky)
    - **lemmatizace** (přiřazení základních tvarů slov)
    - **tagging** (lemmata + morfologické značky), PoS tagging
  - značkované (foneticky, syntakticky)
    - **fonetická transkripce**

    – OMK (Katedra bohemistiky UP Olomouc, dr. Petr Pořízka), ORTOFON
    - **syntax** – SYN2015, SYN2020, PDT (ÚFAL MFF UK Praha)

- **verzované korpusy** – SYN v. 12, InterCorp v. 16, DIAKORP v. 6

Vyváženost a reprezentativnost korpusů

- **vyváženost a reprezentativnost**
  - z pohledu produkce a recepce textů
  - rovnoměrná
  - z hlediska pokrytí variability textů v daném jazyce
  - u mluvených korpusů vyvážené **sociolingvistické** kategorie
- **nevyvážené korpusy**
  - vyváženost není cílem
  - webové

Ústav Českého národního korpusu

- **ÚČNK FF UK**, Panská ul., Praha, www.korpus.cz
- založen 1994, ředitel prof. F. Čermák do r. 2013
- spolupráce s Ústavem teoretické a komputační lingvistiky FF UK a Ústavem formální a aplikované lingvistiky MFF UK
- budování **ČNK**
- publikační činnost
  - **Frekvenční slovník češtiny**, 2004, na korpusu FSC2000
- 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 5 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 %
  - **SYN2020** – 2015 – 2019, 100 mil. slov
    - beletrie 33 %, odborná lit. 33 %, publicistika 33 %

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
- **NET** – polooficiální komunikace na internetu (diskuzní fóra, blogy)

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)
- návrh na využití tzv. hyperlemmat (kůň – kuoň, kóň)

Český národní korpus

- paralelní korpus **InterCorp**,
  - 25 mil. slov (verze 0, 2008, 19 jazyků
  - 5,3 mld. slov (verze 16, 2023), 61 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, 24 jazyků, stejná velikost, stejná metodika a technologie tvorby, slovakocentrický
- 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** – „koľko sa podarí“, 7 mld. jen pro češtinu
- např. Araneum Bohemicum, Araneum Germanicum, Araneum Francogallicum Helveticum

Korpusy od Lexical Computing – Sketch Engine

- **Czech Web 2023 (csTenTen23)** – 2023
  - **5,4 mld.** tokenů, **4,5 mld.** slov (word)
- **Czech Web 2017 (csTenTen17)** – 2017
  - **12,6 mld.** tokenů, **10,5 mld.** slov (word)
- **OPUS Czech** (Open Parallel Corpus) – česká část paralelního webového korpusu
- **Gutenberg Czech 2020** – e-books z Gutenberg Database (29 jazyků)
- **Czech parliamentary debates** – projekt Parla Mint (17 srovnatelných korpusů)

Přístup ke korpusům

- Český národní korpus http://wiki.korpus.cz/doku.php/cnk:uvod
- Sketch Engine https://www.sketchengine.eu/
- Aranea http://unesco.uniba.sk/