

Korpusová lingvistika – 6

Budování korpusů

Mluvené korpusy

Mgr. Dana Hlaváčková, Ph.D.

CJBB105

PRZA009

Budování korpusů

- specifikace **cíle a účelu** korpusu – proč korpus chceme?
- specifikace cílové skupiny **uživatelů** – kdo ho bude používat?
- specifikace **typu** korpusu – zařazení korpusu do typologie
- **výběr** a **sběr** jazykového materiálu (texty, nahrávky + přepisy)
- **autorská práva** – smlouvy s dodavateli textů, u mluvených korpusů informovaný souhlas mluvčích a anonymizace osobních informací ve zvuku i transkripci
- **zpracování textů** – vertikál, tokenizace, jednotné kódování a jednotný formát
- **vnitřní** a **vnější** značkování (atributy, metadata, tagging)
- **softwarové vybavení** – korpusový manažer, příp. nástroje pro automatickou tvorbu korpusů

Budování korpusů podle jejich typu

Různé typy korpusů se mohou budovat odlišným způsobem:

- tradiční synchronní psané korpusy
- webové synchronní psané korpusy
- specializované korpusy
- mluvené korpusy

Budování tradičních korpusů

- **Český národní korpus** (korpus.cz)
- výběr textů – vyváženost a reprezentativnost korpusu
- dohody s poskytovateli textů (autorská práva)
- texty v elektronické podobě
 - odstranění netextového obsahu (obrázky, grafy, tabulky)
 - sjednocení kódování a formátu
 - záznam metatextových informací (autor, dílo, rok vydání ad.)
 - vertikál – tokenizace
 - atributy – **poziční** (*word, lemma, tag*) a **strukturní** (*opus, doc, s*)
 - lemmatizace a morfologické (příp. syntaktické) značkování

Budování webových korpusů

- Centrum zpracování přirozeného jazyka FI MU, firma **Lexical Computing**
- např. české korpusy **czTenTen12, Czech Web 2017**
- korpusový manažer **Sketch Engine** (<https://www.sketchengine.eu/>)
- autorská práva – veřejně dostupné texty na Internetu

- **Nástroje pro automatickou tvorbu korpusů**
- **web crawler** (*SpiderLing*) – stahování textů, nástroj prochází internet, hledá souvislé texty v češtině (obsahující celé věty) a stahuje je do korpusu
- odstranění **boilerplate** (netextového obsahu – obrázky, reklamy, záhlaví a zápatí atd.) z webové stránky (*jusText*)

Budování webových korpusů

- **odstranění duplikátů**, řada textů se na internetu nachází několikrát (*onion – ONE Instance ONLY*)
 - nástroj funguje podobně jako aplikace Vejce vejci v IS pro kontrolu plagiátů, měří se podobnost textů
- **sjednocení kódování**, pro řadu jazyků (*chared – character encoding*)
- stejně jako u tradičních korpusů následuje tokenizace (vertikál), vnitřní a vnější značkování
- metadata jsou velmi stručná:
 - webová adresa
 - nadřazená doména
 - rok stažení
 - jazyk

Vybudování vlastního korpusu

- ve **Sketch Engine**
 - **Corpus Architect** – tvorba korpusů
 - nahrávání textů v elektronické podobě uživatelem
 - uživatel si sám připraví texty, ze kterých chce sestavit korpus
 - **WebBootCaT** – najde automaticky texty z webu
 - uživatel zadá **seed words** (klíčová slova)
 - nebo URLs (adresy webových stránek)
- funkce **NEW CORPUS**
 - každý uživatel má k dispozici 1 mil. slov a možnost požádat o zvětšení prostoru pro vybudování svého korpusu
 - zvážít, kdy má budování vlastního korpusu smysl

Budování mluvených korpusů – nahrávka

- specifikace **typu zaznamenané promluvy** = základní typologie mluvených korpusů
 - monolog – dialog
 - formální – neformální (poloformální – např. odpovědi mluvčích na předem připravené otázky)
 - připravená – nepřipravená
 - prezenční – distanční
 - jejich kombinace – např. předem připravený formální monolog
- specifikace **délky nahrávky**
 - u dialogů obvykle max. 30 min, nejlépe však cca 15 min – obtížně se přepisují
 - monology mohou být delší, např. 60 min přednášky

Budování mluvených korpusů – nahrávka

- specifikace mluvčích a **sociolingvistických kategorií**
 - pohlaví, věk, vzdělání, teritoriální zařazení
 - na základě nich pak vyváženost korpusu
- **autorská práva**
 - není možné nahrávat tajně a použít nahrávku bez souhlasu mluvčího
 - je možné nahrávat tajně a pak si souhlas mluvčího vyžádat
 - informovaný souhlas (podepisuje mluvčí) nebo prohlášení nahrávajícího (podepisuje nahrávající a ručí za to, že mluvčího informoval)
 - anonymizace citlivých údajů – jména, adresy, telefonní čísla apod., ve zvuku pípnutí, v přepisu anonymizační zkratky
- pořizení **kvalitní digitální nahrávky**
 - diktafony, mobilní telefony, příp. úprava nahrávky (oříznutí, např. v nástroji Audacity)
 - např. je dobré při spontánním dialogu nechat mluvčí trochu rozmluvit, rozpačité začátky nahrávek se pak můžou odstříhnout

Budování mluvených korpusů – přepis

- např. korpusy řady **ORAL** – spontánní dialogy
- přepis nahrávek podle předem stanovených pravidel
- nástroj ELAN, dříve Transcriber (volně dostupné nástroje)
- **segmentace** přepisu a synchronizace segmentů
 - v běžně mluveném jazyce není možné identifikovat začátky a konce vět, volí se segmenty ohraničené např. pauzami, nádechy apod.
 - v nástrojích se automaticky synchronizuje zvuk s přepisem
- **ortografický** přepis
 - zásada „přepiš, co slyšíš“, s určitými pravidly, která zachycují např. nespisovnou výslovnost
- **fonetický** přepis (ORTOFON) – klasická fonetická transkripce
 - prováděla se částečně automaticky převedením z ortografického přepisu a následnou ruční editací

Budování mluvených korpusů – přepis

- **pauzová interpunkce**
 - nezaznamenává se interpunkce ve shodě s pravopisem, ale pauzy, které mluvčí dělá
- dále se zachycují především
 - hezitační a jiné zvuky, přeřeknutí, smích, citoslovce, nesrozumitelné úseky, neverbální zvuky, nedořečená slova, příklonné *s (...knihu Babička *s četl?..)*, otazník značí tázací intonaci
- **simultánní úseky**
 - mluvčí mluví současně, skáčou si do řeči
 - v dialogu běžný jev, zaznamenává se přesně, kde se promluvy kryjí
- ideál – doplnění videonahrávkou, vidíme i mimiku a gesta
 - **multimediální korpus**, např. DIALOG (ÚJČ)

Ukázka přepisu v nástroji ELAN. Zprávy z Českého rozhlasu Brno. Ve stopě **ort** je vidět segmentace textu, z oscilogramu (záznam zvuku) jsou patrné pauzy – rovná čára. Byla použita klasická interpunkce, jde převážně o čtení textu ve spisovné češtině.

The screenshot displays the ELAN software interface for audio transcription. At the top, the title bar reads "ELAN - zpráva1_08.eaf". Below it is a menu bar with options: File, Edit, Annotation, Tier, Type, Search, View, Options, Window, Help. A toolbar contains various editing and playback controls. The main workspace is divided into several sections:

- Volume:** A slider set to 100.
- Rate:** A slider set to 100.
- Selection:** A time range indicator showing "00:00:00.000 - 00:00:00.000".
- Waveform:** A dual-track audio waveform showing amplitude over time. A specific time point, "00:00:11.580", is highlighted on the top track.
- Transcription Tiers:** A list of transcription layers at the bottom, including:
 - 0 ort [S]**: The main transcription tier, containing the text: "Bylo šest hodin. Český rozhlas Brno, zprávy. Návrh zákona, který by městské policie zbavil možnosti měřit rychlost aut, se mnohým městům nelíbí. Řediteli zlí..."
 - 0 fon [Q]**: A phonetic transcription tier.
 - 0 meta [Q]**: A metadata tier.
 - 1 ort [Q]**: A second transcription tier.
 - 1 fon [Q]**: A second phonetic transcription tier.

Pražský mluvený korpus

- **první korpus mluvené češtiny, 675 tis. slov**
- autentická mluvená čeština, tematicky nesespecializovaná
- z městského prostředí Prahy a jejího okolí
- neformální dialogy, poloformální řízený rozhovor (dotazník)
- magnetofonové nahrávky (**304**), přepis do MS Word
- z let **1988–1996**, odrážejí jazyk jak konce předchozího společenského období, tak začátek nového
- pravidla přepisu ortografická, pro obecnou češtinu
- větná interpunkce
- bez záznamu překryvů (simultánních úseků)

Brněnský mluvený korpus

- **první korpus mluvené češtiny z oblasti Moravy, 490 tis. slov**
- běžně mluvený jazyk z městského prostředí Brna
- **250** anonymních magnetofonových nahrávek z let **1994–1999**, **294** mluvčích
- prolíná se středomoravský interdialekt s obecnou češtinou
- v oblasti slovní zásoby zbytky někdejšího soužití brněnské češtiny s německým jazykem a vliv brněnského slangu (hantecu)
- neformální a poloformální dialogy
- v pravidlech přepisu zohledněna specifika brněnské mluvy (např. zachycení nářečních asimilací)
- pauzová interpunkce
- zachyceny překryvy

Korpusy řady ORAL

- **ORAL2006** – mluvená čeština z celé oblasti českých nářečí
- 221 nahrávek z let 2002–2006
- pouze neformální dialogy, přátelský vztah mezi mluvčími
- 111,5 hodin, 1 000 798 slov od 754 mluvčích
- **ORAL2008** – plně vyvážený v základních sociolingvistických kategoriích (pohlaví, věk, vzdělání, oblast pobytu v dětství)
- 297 nahrávek z let 2002–2007
- výhradně neformální situace
- 115 hodin, 1 000 097 slov od 995 mluvčích

Korpusy řady ORAL

- **ORAL2013** – nahrávky pořízeny v Čechách, na Moravě i ve Slezsku
- 835 nahrávek z let 2008–2011
- 2 785 189 textových slov, tj. celkem 3 285 508 pozic
- 2 544 mluvčích, z toho 1 297 unikátních
- délka téměř 300 hodin

Korpusy řady ORAL

- **ORAL** v. 1 – sjednocuje předchozí korpusy
- (+ ORAL-Z)
- nevyvážený, referenční
- 5,4 mil. slov
- snaha o sjednocení transkripce
- lemmatizace a morfologické značkování
 - problematické, vysoká homonymie
 - *byt/být, sem/jsem*

Korpus ORTOFON

- plně vyvážený
- Čechy, Morava, Slezsko
- ortografická a fonetická transkripce
- lemmatizace a morfologické značkování
- cca 1 mil. slov z let 2012–2017

Korpus ORATOR

- předem připravený monologický projev
- rodilí mluvčí češtiny
- za přítomnosti publika, bez čtení (ale malá část nahrávek čtená a bez publika)
- 12 typů situací
- ortografická transkripce
- lemmatizace a tagging
- nevyvážený
- 318 nahrávek z let 2005–2019 od 332 mluvčích
- verze 2 – 1,2 mil. slov

Korpus DIALEKT

- 100 tis. slov, referenční
- dialektologická a ortografická transkripce
- celé území ČR, dělení dle Mapy nářečních oblastí ČR
- starší část 50.–80. léta 20. st.
- novější část 90. léta – současnost
- starší generace (nad 60. let)
- řízený rozhovor (osvědčená témata)
- lemmatizace a morfologické značkování
- spojení s aplikací Mapka