Sběr dat pro mluvený korpus

Dana Hlaváčková
hlavack@fi.muni.cz

FI MU
2. patro, budova B
laboratoř B206 (prostřední)

Co probereme

- Technické vybavení
- Administrativa
- Nahrávání
- Přepis

Technické vybavení

- počítač s připojením k internetu
- nahrávač s USB připojením k počítači
- sluchátka
- Transcriber (ucnk.ff.cuni.cz/oral)!!!
- MConvertor (ucnk.ff.cuni.cz/oral)

Administrativa

- Prohlášení nahrávajícího 1x
- Dohoda o provedení práce (osobní údaje, číslo účtu, podpis) – každý měsíc, kdy odevzdáváte sondu
- Protokol o zapůjčení přístroje
- Přihlášení do databáze – heslo (žádost o heslo na mail)
- Vyplňování databáze http://trnka.ff.cuni.cz/mluvka/corpus – manuál
- Kontrola a schválení práce
- Výplata – pevná částka za nahrávku, pohyblivá za počet přepsaných znaků

Nahrávání

- přirozená, neformální situace
- informovat mluvčí o nahrávání
- dialog (2-3 mluvčí)
- dospělí od 18 let
- vyvážené sociolingvistické kategorie – muži, nad 30 let, ZŠ/SŠ vzdělání
- oblast nahrávání – celá Morava
- 1 mluvčí – pouze 15 000 slov
- délka nahrávky 20-30 min
- oříznutí nahrávky http://audacity.sourceforge.net
- ne děti, cizinci, telefonní hovor, poloformální a formální rozhovory, monolog

Přepis – obecné zásady

- přepisovat věrně, co slyšíte, nic nevynechat
- segmenty v Transcriberu (max. 15 slov)
- pauzová interpunkce – krátká pauza (.), střední pauza (..), *(odmlčení) – žádný mluvčí*
- otazník (*?*), vykřičník (*!*), neukončená (*...:*) nebo přerušená výpověď (*...*)
- poznámky – *(smích), (se smíchem), (mluví dítě), (zpěv)*...
- nesrozumitelný úsek ---
- anonymizační zkratky – samostatný segment
- simultánní úseky

Vlastní přepis

- co nejvíce se držte standardního zápisu
- varianty spisovné výslovnosti
- výslovnost v rozporu s českou výslovnostní normou
- příznakové rysy běžné mluvy, včetně regionálních
- ustálené varianty běžné mluvy
- znělostní asimilace – skupina *sh [sch], [zh]*, předložka *s, se*
- nářeční asimilace – *tfůj, zme, gvůli...*
- artikulační asimilace – *hežčí, ešče, pracovišče...*
- odlišná kvantita – *viš, brat, nějaky...*
- moravské realizace komparativů a superlativů – *pěkňéší, hlópjéší...*

Vlastní přepis

- cizí slova a cizí vlastní jména – původní pravopis, pokud existuje – počeštěná podoba
- nedořečené slovo – hvězdička – *to bylo vče**
- příklonné *s* – **s tam byl?*
- citoslovce – přesně tak, jak zazní
- responzní zvuky – přitakací *hmm*, odporovací *emem*
- hezitační zvuky – souhláskové *mmm*, vokalické *eee*
- zkratky – tak, jak byly vysloveny – *dé vé dé, aids, v š e*

Shrnutí

- nahrát
- přepsat (poslat vzorek)
- požádat o heslo
- vložit do databáze
- pokud je to nutné, opravit
- podepsat smlouvu a prohlášení
- čekat na výplatu a pořizovat další sondy