Specializované korpusy mluveného jazyka - jejich tvorba a využití

Karolína Vyskočilová

34. Žďárek, Poděbrady, 2.–4. května 2014

vyskoczilova@seznam.cz

Obsah příspěvku

- korpusy
  - čeština v zahraničí
    - BANÁT
    - Texas Czech Dialect Archive
  - akviziční korpusy AKCES
    - ROMi 1.0
    - CZESL-LONG/CZEFL-LONG
- proč budovat specilizované korpusy?
- zaznamenávané údaje k materiálu
- co tyto korpusy nabízejí?
- problémy při tvorbě korpusů
- a co by měly nabízet?

Korpusy – češtiny v zahraničí

- BANÁT
  - sběr 2011 + 2014 (Bígr)
  - 150 tis. slov + ? (cca 40 hod., 30 mluvčích, 35-92 let)
  - připravuje ke zveřejnění (letos)
  - nepřipravené projevy, nahrávající integrován do rozhovoru

- Texas Czech Dialect Archive / Texas Czech Legacy Project (The University of Texas at Austin)
  - sběr 1970-2000
  - velikost ?
    - Svatava Pírková-Jakobsonová – cca. 97 hodin, 1970-80
    - Karel Kučera ?
    - Lida Cope – cca 327 hod., 90. léta
    - John Tomecek – cca 150 hod., 50 mluvčích (kol. r. 2000)
  - zveřejnění průběžně
  - nepřipravené projevy, polořízené rozhovory, jaz. úkoly, vyprávění apod.

1-1-032_A_06 Farming and Picking Cotton

Download 032_A_mp3.mp3

Hide English | Hide Czech

Interviewer (tk)	A	..	vy	ste	vy	ste	farmovali?	
Interviewer (tk-trans)	And	..	you	Aux2ndPl	you	Aux2ndPl	farmed?	
Interviewer (free-trans)	And did you farm?							
Speaker 1 (tk)							JA sem	...
Speaker 1 (tk-trans)							I Aux1stSg	...
Speaker 1 (free-trans)							I was ...	

Korpusy - akviziční korpusy AKCES

- ÚČJTK + další ústavy (ÚFAL, ÚČNK, TUL)

- ROMi 1.0
  - korpus mluvených projevů romských dětí a mládeže
  - připravuje se k zveřejnění (letos)
  - 50 nahrávek, 120 tis. slov
  - 143 mluvčích (škola + neziskovky), 13-24 let
  - často řízený rozhovor (set otázek/témat)
- CZESL-LONG/CZEFL-LONG
  - longitudinální korpus mluvených projevů žáků založený na čtyřletých sběrech
  - sběr několikrát za rok
  - prozatím 4 gymnázia (3 Praha + Kroměříž), střední škola, UNI v Káhiře
  - nahrávky vyučovacích hodin + „volná“ diskuse několika žáků ve skupinách s nahrávajícím

Proč budovat spec. korpusy?

- čeština v zahraničí
  - záchrana jazyka/dialektu, který se ztratí (+ záchrana existujících nahrávek)
    - srovnání vývoje, stability jaz. rysů apod.
  - zdokumentovat historii, příběhy a tradice (BANATxTCDP)
  - vzbudit zájem o jazyky v zahraničí a získat tak podporu na jejich zachování

- akviziční korpusy
  - studium procesů akvizice jazyka, resp. studiu užívání jazyka mluvčími ve fázi jazykové akvizice a pozdějšího jazykového vývoje nebo ve spojení s ní (např. při vyučování)
  - plánování jazykové výuky a přípravě materiálů
  - vytváření slovníků, mluvnic, přípravě testů
  - i při vlastní výuce jako zdroj.

Zaznamenáváné údaje k materiálu

- místo sběru
  - město (+ dial. oblast)
  - místo (restaurace, doma, škola + typ školy a její provozovatel, příp. zájmová skupina apod.)
  - sociální informace o lokalitě
- okolnosti sběru
  - připravenost/nepřipravenost
  - situace (návštěva, rozhovor s přáteli apod. referát, četba apod.)

- mluvčí/respondent
  - věk,pohlaví
  - vzdělání (+ třída/rok)
  - místo narození (+ pobyt v dětství)
  - současný pobyt
  - znalost dalšího jazyka, první jazyk, jazyk používaný doma, způsoby osvojování apod.
  - vztah mluvčích (známost, rodina, přátelskost)
  - zaměstnání
  - zvl. poznámky
  - počet hodin jazyka týdně, učebnice a spol.
- sběrač/tazatel
- přepisující

Co zpřístupněné korpusy nabízejí?

- přepis – folkloristický, ale…
- nahrávky – (mp3/KonText)

- „chyby“ odlišnosti od jazykové normy
- lemmatizace vs. více rovin přepisu
- překlad do aj
- morfologický komentář

Problémy při tvorbě korpusů: způsob přepisu

- používat interpunkci?
  - ano – ROMi (obvyklá ve spis., na pauzy se nehledí, porušování větné vazby se značí čátkou?)
  - větné celky - TCDA
  - ne – BANÁT (pauzová interpunkce)
  - všichni ale značí otázku či rozkaz
- Bude folkloristický překlad stačit?
  - nezachycují se asimilace (přitom mohou být realizovány různě), zjednodušené realizace (*kamennej, dcera)*
  - co odlišná kvantita – fologická vs. váhání?
  - ale třeba ROMi - jiná než standardní výslovnost *(dybyz mohl, uňiverzita*)
  - banatismy, reginální slova (mušim vs. musim apod.)
  - znační chyb? – do textu, v další rovině? vůbec?
- Vyřešila by to další rovina?
  - Fonetická? Značit přízvuk?
  - Ortografická?
    - Co např. protetické *v* apod.? *Pudu* vs. *půjdu, mohla bysem* vs. mohla bych

Problémy při tvorbě korpusů: další problémy

- anonymizace – zveřejnění nahrávek
  - zvuk není anonymní
  - které údaje jsou již nutné anonymizovat? (jméno, příjmení, město, situace?)
    - nahrazovat zkratkou?
    - nahrazovat jiným jménem shodného deklinačního typu?
    - ?
- problém s přepisem cizích jmen a názvů
- značení nejazykových okolností (štěkání psa, zvuky v okolí, hovor v pozadí)
- parazitní zvuky? Je k něčemu je zachycovat?
- cizí jazyky v nahrávce, code-mixing
  - přepisovat?
  - překládat?
- přítomnost nahrávajícího, který není ze skupiny
- Je třeba komentář k nahrávkám či korpusům a pokud ano, tak jak rozsáhlý?

Děkuji za pozornost.

… a co by měly nabízet?

- Co z toho jsme schopni využít?
  - zaznamenávané údaje
  - interpunkce
  - roviny přepisu
  - anonymizace
  - cizí jazyky
  - komentář?
  - …?

Zdroje k jednotlivým korpusům

- korpusy AKCES http://akces.ff.cuni.cz/
- ROMi 1.0 http://ufal.mff.cuni.cz/romi-10/
- Texas Czech Legacy Project http://blogs.utexas.edu/txczech/
- Texas Czehc Dialectic Archive http://blogs.utexas.edu/txczech/texas-czech-dialect-archive/

- BEDŘICHOVÁ, Z. – ŠEBESTA, K. – ŠKODOVÁ, S. – ŠORMOVÁ, K. (2011): Podoba a využití korpusu jinojazyčných a romských mluvčích češtiny: CZESL a ROMi. In: Fr. Čermák (ed.), *Korpusová lingvistika Praha 2011*. *Svazek 2 – Výzkum a výstavba korpusů. p*. 93–104.
- ŠEBESTA, K., ed. a kol. *Čeština - cílový jazyk a korpusy*. Vyd. 1. Liberec: Technická univerzita v Liberci, 2012. 166 s. ISBN 978-80-7372-848-9.
- VYSKOČILOVÁ, K. (2012): *Syntaktická analýza projevů českých mluvčích v rumunském Banátu.* Praha. Bakalářská práce. Univerzita Karlova v Praze, Filozofická fakulta, Ústav českého jazyka a teorie komunikace.