

SLDR

následní programu prvního TLS 19. března 2010

1. Struktura kurzu
 - přednášky o praktické lingvisticce v terénu, o archivaci dat apod.
 - tutoriály s nahráváním a elicitováním mluvčích
 - praktická práce s programy pro anotaci a další zpracování dat
 - otázky prohlubování spolupráce s komunitami – výukové programy, tematické slovníky (ryby, architektura atd.), multimédia (CD pro nativní obyvatele), výukové materiály (parsers a jejich využití pro e-learning)
 - o sociolingvistické situaci sámštiny
2. Dokumentace
 - co je projekt DOBES
 - základní cíle dokumentace (zajistit a archivovat data z ohrožených jazyků – aby byla srozumitelná nejen pro lingvisty, ale i pro další vědce bez znalosti jazyka; dokumentace musí být akceptována komunitou, možnosti revitalizace apod.)
 - základním principem dokumentace je otevřenost – dávají se data k dispozici ve výchozí podobě, aby si je mohl každý přehledně » archivace (viz v závěru bod E)
3. Metody elicitace: nepřekladová (problémový překladu)
 - word-to-text method: nejprve si nechat ukázat na předměty/činy/činnosti, z toho zjistíme slova a pak si nechat tvořit celé věty, neznáme elementy mezi tím budou funkční slova
 - substitute: měnit větu například přidáním slova pro „včera“
 - parafrázování: otázky typu: Co bys řekl, kdyby...?
4. Praktická příprava – nahrávání
 - příprava techniky a správná práce s technikou jako základ úspěchu
 - nahrávací zařízení: minimálně Edirol + externí mikrofon
 - otázky: chceme vytvořit laboratorní situaci nebo slyšet okolní ruchy? podle toho zvolit různé typy mikrofonu podle způsobu snímání
 - technicky: digitálně nahrávat stereo, WAV
 - archivace: nemazat nahrávky z SD karet, ale archivovat v obálkách

5. Anotování

- A. Vytvořit soubor s metadaty – zaznamenat všechny relevantní informace o nahrávce ve standardizovaném formátu (viz ukázka xml), ukládat v adresářích standardizovaně označovaných » ukázka toho xml souboru (M)
- B. ELAN – program pro praktickou anotaci, data jsou vztahována přímo k výchozí nahrávce, možno nahrávat i video na dvě kamery, a k tomu vztáhnout všechny anotace (velmi praktické např. pro znakový jazyk)
 - audio-video nahrávka
 - transkripcie (v IPA je vhodná jen pro část korpusu, zbytek lepší v praktické ortografii akceptovatelné komunitou)
 - překlady
 - tři roviny – doslovný / interlineární glosy / volný
 - do více jazyků; varianty: do jazyka vědeckého (angličtina), lingua franca, národního jazyka
 - vlastní komentáře (klíčová slova)
 - ukázka toho souboru (J)
- C. TOOLBOX – program zejména pro vytváření lexikálních databází a automatických interlineárních překladů; možný export do textového formátu (jakoby slovník).
Obvyklé vrstvy hesla:
 - lexém (head word) [ale také morfém]
 - slovní druh
 - definice, překlad
 - kolokace, idiomy
 - příklady» ukázka souborů z Joshuova tutoriálu – švédština a pitejská sámština; jak to umí vytvořit interlineární překlad (ALT-1)
Limity programu: vytvořila ho SIL pro dokumentaci jazyků v Pacifiku, které mají velmi jednoduchou (lineární) morfológii, chceme-li zajistit funkci interlineárního překladu pro flektivní jazyky, je nutné program obházet zaváděním podřízených slovníků, kde jsou obsaženy odchýlné tvary (+ v jedné vrstvě odkaz k lexému v hlavním slovníku). Program nemá zrovna uživatelsky příjemný.
6. Po získání dostatečného korpusu (viz program na další program TLS – získávání korpusu) je v další fázi možné napsat gramatickou skeč, která obsahuje fonologii, ortografii, slovní druhy, gramatické kategorie, příklady, typologický profil jazyka.
+ souběžně spolupráce s komunitou na dalších výstupech (slovníky, učebnice apod.)
7. Archivace: Max-Planck, Nijmegen – server, který přijímá k archivaci veškerá jazyková data od kohokoliv; řádně opatřená metadaty: http://corpus1.mpi.nl/ds/imdi_browser/

Magda
Jan