LAIMA KALĖDIENĖ

Lietuvių kalbos institutas
kaledienei@gmail.com

Multimodalioji Tarmyno vizija

Tarminių duomenų bazių būklė. LKI *Tarmių archyvo* duomenų bazėje(http://www.lki.lt/LKI_LT/index.php?option=com_content&view=article&id=78) saugoma apie 6800 val. garso įrašų iš visų lietuvių kalbos tarmių nuo 7–8 praėjusio amžiaus dešimtmečio iki dabar. Tai ne vienintelė Lietuvoje *tarmių garso įrašų bazė* (žr. www.tarmes.lt; Šiaurės rytų aukštaičių vilniškių tarminės medžiagos duomenų bazė ir kt.) Palyginti nedidelė dalis LKI *Bazės* garso įrašų jau yra šifruota, transkribuota ir paskelbta internetinėje prieigoje *Tarmių tekstyno duomenų bazė* (http://www.lki.lt/tarmiuarchyvas/pradinis.php?sutrump=bnd). Nemažai, bet santykiškai nedaug transkribuotų tarminių tekstų yra publikuota atskiromis knygomis, dažniausiai – su plokštelėmis.

***Tarmyno* modernizavimo etapai.** 1) skaitmeninimas; 2) šifravimas ir transkribavimas; 3) transkribuotų tekstų transponavimas; 4) konkordancijų sąrašo rengimas;5) gramatinių požymių anotavimas – lematizavimas ir glosavimas.

Skaitmeninimas didžiąja dalimi jau baigtas.

Ypač darbo imlus **šifravimas** reikalauja naujo požiūrio: tradicinis detalusis transkribavimas priklausomai nuo darbo tikslų gali būti taikomas rečiau (kaip vyksta taikant CHILDES tipo programas); be to, tiesiog neįmanoma kompiuterinė paieška. Siekiant kuo efektyviau naudotis šiuolaikiškais *Praat*, *Transcriber*, *EXMARaLDA* ir panašiais fonetinės teksto analizės įrankiais, spręstinos transkribavimo principo problemos: 1) susijusios su lietuviškąja transkripcija užrašytų tekstų pritaikymu apdoroti naujausiomis technologijomis; 2) su perėjimu prie *Tarptautinės fonetinės abėcėlės* (IPA). Dalis keblumų tiesiog atgaminant seniau transkribuotus tekstus gali būti išspręsta programiškai perkoduojant juos *Palemonu.*

Tolimesnei kompiuterinei analizei reikalingas **konkordancijų sąrašas**. Konkordancijos būtų abėcėliškai sutvarkyta teksto žodžių formų rodyklė su tekstų ir jų pateikėjų charakteristikomis – 1) punkto numeris, 2) tarmė (šnekta), 3) vietovė, 4) pateikėjas, 5) jo gimimo metai, 6) garso įrašo metai ir kt. Konkordancijose išskirtoji žodžio forma pateikiama su aplinkiniu konktekstu – pvz., teksto ištraukoje po 10 žodžių iš abiejų pusių. Jos rengiamos programiškai iš skaitmenine forma pateiktų tekstų: jie apdorojami programinių modulių sistemomis, o tada generuojamos konkordancijos. Konkordancijos gali būti rikiuojamos tiesiogine tvarka (abėcėliškai atsižvelgiant į žodžio pradžią) ir atvirkštine (abėcėliškai atsižvelgiant į žodžio pabaigą).

Kita įmanoma išgauti *Tarmyno* metalingvistinė informacija yra **gramatinių požymių anotavimas** prie kiekvienos užrašytos žodžio formos. Šitokio *lingvistinio anotavimo* esmė yra kalbos dalių ir morfologinių žodžių formų aprašų žymėjimas (angl. *POST – Part of Speech-Tagging*). Šiam anotavimui atlikti reikalingi procesai yra *lemavimas* ir *glosavimas*.

Perspektyva. Toks *Tarmynas* būtų daugiapakopė (angl. *stand-off*) struktūra, o kiekvienas anotuotas jo sluoksnis – ir integralus, ir savarankiškas, nes galimas plėsti ir pildyti nepriklausomai nuo kitų. Tuo pačiu visus sluoksnius tarpusavyje sietų XML (*Extensible Markup Language*) duomenų struktūros formatas, leidžiantis pagal įvairius kriterijus koduoti pirminius teksto duomenis ir pateikti visas reikalingas tekstologines bei lingvistines teksto anotacijas. Paieškos kriterijus galima būtų modeliuoti tiek pagal kiekvieną anotacijos sluoksnį, tiek ir pagal kelių sluoksnių kombinacijas.