Prof. Dr. Jolanta Gelumbeckaitė
Institut für Vergleichende Sprachwissenschaft
Goethe-Universität Frankfurt am Main
eMail: gelumbeckaite@em.uni-frankfurt.de

Tarptautinis mokslinis seminaras
A.Bezzenbergeris - etnografas ir akademinės baltistikos pradininkas
(Juodkrantė, Neringa 2011 09 9–10)

Senosios lietuvių kalbos referencinis korpusas (SLIEKKAS). Lingvistinė anotacija

Mieli kolegos ir bičiuliai! Esu Jums labai dėkinga už pakvietimą į Adalbertui Bezzenbergeriui skirtą seminarą.

Pirmasis publikavęs lietuvių kalbos senuosius tekstus ir pirmasis pradėjęs analizuoti senąją lietuvių kalbą, Bezzenbergeris iš esmės padėjo pagrindus visiems tolesniems XVI–XVII amžiaus lietuvių kalbos tyrimams. Deja, dar negalime teigti, kad tradiciškai senąja lietuvių kalba vadinamą laikotarpį jau esame ištyrę išsamiai, nuodugniai ir sistemingai. Norisi tikėtis, kad iki 2077 metų, kai sukaks 200 metų nuo Bezzenbergerio *Beiträge zur Geschichte der litauischen Sprache* (*Auf Grund litauischer Texte des XVI. und XVII. Jahrhunderts*, Göttingen: Robert Peppmüller, 1877), šių dviejų šimtmečių lietuvių kalba bus detaliai ir sistemingai aprašyta, turėsime lietuvių kalbos istorinį žodyną ir senosios lietuvių kalbos gramatiką.

Siekdami sistemingo, o ne paremto atsitiktiniais pavieniais duomenimis tyrimo, pirmiausia turime nuspręsti, kaip parengti tokią sistemą, kuri leistų mums rasti patikimus atsakymus į įvairius lingvistinius klausimus. Pirmiausia, žinoma, reikia turėti visus tiriamo laikotarpio tekstus. Čia nepralenkiama yra Lietuvių kalbos institute rengiama bei nuolat pildoma *Senųjų raštų duomenų bazė*, apimanti tekstus nuo 1573 iki 1816 metų. Ji yra didžiausias ir tekstų dokumentinio tekstų perteikimo atžvilgiu patikimiausias senosios lietuvių kalbos tekstynas. Kitas etapas yra suskaitmenintus tekstus parengti tolesnei analizei. Šiuo atžvilgiu pavyzdiniai yra Frankfurto prie Maino Goethe's universiteto TITUS (*Thesaurus Indogermanischer Text- und Sprachmaterialien*) duomenų bazėje esantys lietuviški tekstai. Šioje bazėje įvairių tekstų žodžių formos yra pagal tam tikrą indeksacijos sistemą susietos vienos su kitomis ir iš dalies su vertimo šaltiniais.

Modernūs kalbos istorijos tyrimai reikalauja ne tik priėjimo prie patikimų empirinių duomenų, bet ir kokybiškos tų duomenų paieškos galimybės pagal įvairius lingvistinius kriterijus. Taigi suskaitmenintus ir filologiškai užtikrintus tekstus, kaip empirinių rašytinių senosios lietuvių kalbos faktų rinkinį, reikia paruošti tolesniems moksliniams tyrimams, visų pirma lingvistiniams, bet taip pat ir istoriniams plačiąja prasme.

Senosios **LIE**tuvių **K**albos **K**orpus**AS** ⇨ **SLIEKKAS** ir siekia parengti tokio tyrimo mokslinius bei techninius pagrindus. Referencinio senosios lietuvių kalbos korpuso tikslas yra reprezentuoti šio laikotarpio kalbą kaip visumą bei pateikti išsamią informaciją apie kalbą, jos variantus, žodyną ir gramatiką. SLIEKKAS turi padėti ir

pradėti įgyvendinti du didžiausius diachroninės lituanistikos darbus: senosios lietuvių kalbos gramatikos ir istorinio lietuvių kalbos žodyno rengimą.

SLIEKKAS susideda iš trijų pagrindinių dalių: iš metainformacijos apie tekstus, iš pačių tekstų ir iš jų lingvistinės anotacijos. Visas tris dalis kartu ar kiekvieną atskirai turi būti galima tirti pagal pasirenkamus paieškos kriterijus:

metaduomenys	paieškos įrankiai
rašytiniai tekstai (empiriniai kalbos faktai)	
lingvistinė anotacija	

SLIEKKAS yra daugiapakopės (vad. *Stand-off*) architektūros struktūra. Ją sudaro įvairūs metalingvistinės, teksto ir lingvistinės informacijos anotaciniai sluoksniai. Pats tekstas ir kiekvienas jo anotacinis sluoksnis faktiškai yra atskiras ir integralus dokumentas. Tokią architektūrą įgalina sukurti XML (*Extensible Markup Language*) duomenų struktūrų kalbos formatas. Kiekvieną sluoksnį galima plėsti ir pildyti nepažeidžiant kitų sluoksnių, tačiau visi sluoksniai yra tarpusavyje sinchronizuoti ir susieti. Toks formatas leidžia pagal įvairius kriterijus koduoti pirminius teksto duomenis ir pateikti visas reikalingas tekstologines bei lingvistines teksto anotacijas. Paieškos kriterijus galima modeliuoti pagal kiekvieną anotacijos sluoksnį ir pagal kelių sluoksnių kombinacijas.

Tekstai su daugiapakopėmis anotacijomis perauga „tekstynu“, t. y. tekstų rinkiniu, vadinamas ribas. Todėl ne iš išpaikimo, bet sąmoningai vartojamas terminas „korpusas“.

Čia nekalbėsiu apie korpuso techninį konceptą, apie anotacines programas ir apie vizualizacijos bei paieškos sistemas. Šita korpuso pusė yra ne tiek lingvistikos, kiek lingvistinės informatikos dalykas. Taip pat nedetalizuosiu tekstologinės anotacijos, kuri tiesiogiai susijusi su rankraščio ar spaudinio kritiniu parengimu ir atspindi arba atstoja kritinį tekstologinį aparatą.

Šiuo atveju mums svarbiausias yra vieningais principais paremtõs lingvistinės anotacijos sistemos kūrimas. Kaip ir rašant gramatiką ar sudarant žodyną, lingvistinė anotacija yra pagrįsta hierarchijos principu. Lingvistinė informacija pateikiama sluoksniais, pradedant nuo stambiausių kategorijų ir detalizuojant iki smulkiausių fleksinės morfologijos požymių.

LEMATIZAVIMAS, GLOSAVIMAS. Kiekviena korpuso teksto „žodžio forma“ (šis terminas, vad. *Token*, apima ne tik atskirų žodžių formas, bet ir skaitmenis, skyrybos ženklus, grafinius simbolius bei santrumpas) yra lematizuojama. Pirmiausia pateikiama standartizuota žodžio forma, tokia, kokia ji būtų atstatyta istoriniame žodyne, ir antraštinis žodis, lema, t. y. pagrindinė forma istoriniame žodyne. Toliau antraštinis žodis (lema) glosuojamas, t. y. užrašomas dabartine bendrine kalba ir pateikiamos jo reikšmės. Taip pat glosuojama konkreti pavartota žodžio forma, pateikiama jos reikšmė konkrečiame aprašomame kontekste. Kaip tai atrodo, matyti pirmoje lentelėje:

1 lentelė. Standartizuota forma, lematizavimas, glosavimas, vertimas

Anotacijos sluoksniai	*Reikšmė*	*Pavyzdys*	
Rankraštis, spaudinys	forma rankraštyje ar spaudinyje	ſchwaiſdes	Gregorius
... (tekstologinės anotacijos sluoksniai)			
Standartizuota forma	standartizuota forma (atstatyta kaitoma forma istoriniame žodyne)	žvaizdės	Gregorijus
Lema	antraštinis žodis (pagrindinė forma istoriniame žodyne)	žvaizdė	Grigalius
Lemos glosavimas	antraštinio žodžio glosavimas: užrašymas bendrine kalba ir reikšmės, vertimas (skolinių, tarmybių ir nebevartojamų, dabartiniam vartotojui nebeatpažįstamų žodžių, kitomis kalbomis parašytų žodžių)	žvaigždė	Grigalius
Konkrečiõs formos glosavimas	konkrečiõs žodžio formos glosavimas: užrašymas bendrine kalba ir reikšmė (vertimas) konkrečiame kontekste	Aušrinė žvaigždė, Veneros planeta	Grigalius Didysis

LINGVISTINĖ ANOTACIJA: *POST* (*Part of Speech-Tagging*). Tolesni korpuso sluoksniai skirti lingvistinei anotacijai. SLIEKKO lingvistinė anotacija iš esmės yra morfologinė. Morfologinė anotacija apima kalbos dalių (gramatinių klasių) žymėjimą (vad. *Part of Speech-Tagging, POS*) ir morfologinius žodžių formų aprašus. Anotuojama hierarchijos principu, pirmiausia lema, tada konkreti pavartota žodžio forma. Visas SLIEKKO pažymų kompleksas (vad. *Tagset*) suskirstytas taip pat hierarchiškai, pradedant bendra informacija apie kalbos dalį ir toliau skirstant į smulkesnius morfologinius poklasius.

Lemos anotuojamos pagal 10 gramatinių klasių.

2 lentelė. Kalbos dalys (gramatinės klasės) ir jų pažymos

1. **ADJ** – būdvardis (adiectivum)
2. **ADV** – prieveiksmis (adverbium)
3. **AP** – prielinksnis, polinksnis (adpositio)
4. **CARD** – kiekinis (cardinale) skaitvardis, **ADJO** kelintinis (ordinale) skaitvardis
5. **ITJ** – jaustukas (interjectio), ištiktukas (onomatopoeticum)
6. **KO** – jungtukas (conjunctio)
7. **N** – daiktavardis (nomen): **NA** bendrinis (apeliatyvinis), **NT** tikrinis
8. **P** – įvardis (pronomen): **PD** demonstratyvinis, **PI** indefinitinis, **PK** klausiamasis (interogatyvinis), **PPER** asmeninis
9. **PTK** – dalelytė (particula)
10. **V** – veiksmažodis (verbum): **VA** pagalbinis, **VV** pagrindinis

Kiekvienõs žodžio formos gramatinė klasė anotuojama dvejopai: pagal lemą (antraštinio žodžio gramatinė klasė) ir pagal konkrečią pavartotą žodžio formą (konkrečiõs žodžio formos pavartojimas). Tokiu būdu lemos pagrindinė gramatinė klasė atskiriama nuo konkrečios formos vartojimo. Šis skyrimas padeda:

- žymėti gramatinių klasių kitimus, pvz., būdvardžių daiktavardėjimą, dalyvių būdvardėjimą bei daiktavardėjimą; daiktavardžių ir prieveiksmių virtimą adpozicija (prielinksniu ar polinksniu), daiktavardžių, būdvardžių, dalyvių prieveiksmėjimą ir pan.;

- apibūdinti ambivalentiškus atvejus, kai neaišku, kuriai kalbos daliai priskirti konkrečią žodžio formą, pvz., sustabarėjusių linksnių prieveiksmėjimo atveju konkreti pavartota forma yra prieveiksmis *atėjo vidunakčiu* (ADV), o pati lema *vidunaktis* – daiktavardis (NA);
- adpozicijos (prielinksnio) klasėje skirti prielinksnius (prepozicijas) ir polinksnius (postpozicijas), pvz.: *link* (APPR) *namų* ir *namų link* (APPO);
- pateikti aiškią neasmenuojamų veiksmažodžio formų definiciją, pvz., *kalbėdamas tarė* konkreti pavartota forma yra predikatyvinis pusdalyvis/semiparticipijas (VVSPP), o lema *kalbėti* – veiksmažodis (VV). Toks lemos ir konkrečios formos skyrimas svarbus aprašant bendratį, dalyvius, gerundyvus (reikiamybės dalyvius), pusdalyvius (semiparticipijus), padalyvius (gerundijus) ir supiną.

Kiekviena iš lemos gramatinių klasių turi įvairaus detalumo subklasifikaciją, kuri priklauso nuo konkrečiõs pavartotos žodžio formos morfologijos.

Tais atvejais, kada morfologinė išraiška kinta priklausomai nuo sintaksinės funkcijos, anotacija apima ir sintaksinės morfologijos lygmenį. Morfosintaksinis anotavimas leidžia išryškinti skirtybes, kurios neatsispindi grynai morfologiniame apraše. Tai svarbu šiais atvejais:

- įvairuojanti būtinojo (vardažodinis tarinys) ir laisvojo (tarininis pažyminys) predikatyvo morfologinė forma: vardininkas arba įnagininkas, pvz., *Jonas buvo pranašas/pranašu*. Konkrečių formų morfologinėse pažymose skiriamas predikatyvinis daiktavardis, būdvardis, skaitvardis, įvardis, dalyvis (+gerundyvas), pusdalyvis;
- pagalbinio veiksmažodžio ir pagrindinio veiksmažodžio išskyrimas reikalingas aprašant tokius veiksmažodinius kompleksus kaip sudėtinės laikų formos. Todėl tikslinga lemų morfologinėse pažymose veiksmažodį skirti į pagrindinį (VV) ir pagalbinį (VA). Konkrečių formų morfologinėse pažymose svarbu skirti atributinę ir predikatyvinę bei adverbialinę neasmenuojamų veiksmažodžio formų vartoseną.

Plg. 3 lentelę:

3 lentelė. Lemos ir konkrečios žodžio formos POS

Anotacijos sluoksniai	***Reikšmė***	***Pavyzdys***
M1a lema	antraštinio žodžio (lemos) kalbos dalis	[Jonas tapo] *pranašu* NA (bendrinis daiktavardis)
M1b konkreti žodžio forma	konkrečiõs žodžio formos morfologinė (šiuo atveju morfosintaksinė) charakteristika	NAP (bendrinis daiktavardis, predikatyvinis)

MORFOLOGINĖ ANOTACIJA. Tolesnė morfologinė anotacija taip pat yra hierarchinė. Ji susideda iš trijų sluoksnių. Pirmas ir antras sluoksnis žymi tokias nekintamas, inherentines kalbos dalių morfologines kategorijas kaip kaitybos klasė (kaitybos kamienas) ir giminė (giminėmis nekaitomų klasių: daiktavardžių ir daiktavardiškų skaitvardžių). Kaip ir anotuojant kalbos dalis (POS), apibūdinamos pirmiausia lemos (antraštinio žodžio), o tada konkrečiõs žodžio formos kategorijos. Toks lemos ir žodžio formos skyrimas leidžia fiksuoti kaitybos klasės ir giminės įvairavimą, pvz.:

- lema – *amžius* (ju_Masc) ir tekste vartojamos įvairių kamienų formos *amžias* (ja_Masc), *amžas* (a_Masc), *amžis* (i_Masc);
- lema – *gražus* pagrindinė būdvardžio forma nelyginamuoju laipsniu (*u,jo* kamienas, vyriškoji ir moteriškoji giminė), o konkreti žodžio forma *gražiausias* aukščiausiuoju laipsniu (*ja* kamienas);
- lema – *neklaužada* yra bendrosios giminės (Com), o konkreti žodžio forma junginyje *baisus <u>neklaužada</u>* yra vyriškosios giminės (Masc).

Trečias sluoksnis žymi kintamus fleksinės morfologijos požymius: giminę (giminėmis kaitomų klasių), skaičių, linksnį, laipsnį, asmenį, nuosaką, laiką, diatezę (rūšį), emfazę (apibrėžtumą, kitaip, įvardžiuotines formas). Plg. 4 lentelę:

4 lentelė. Lemos ir konkrečios žodžio formos morfologinė anotacija

Anotacijos sluoksniai	***Reikšmė***	***Pavyzdys***		
M2a lemos fleksija	antraštinio žodžio (lemos) nekintamos morfologinės kategorijos (kaitybos klasė ir giminėmis nekaitomų klasių giminė)	*gyvename* ti-Inf	*gražiausiame* u,jo	*amže* ju_Masc
M2b konkrečiõs žodžio formos fleksija	konkrečiõs žodžio formos nekintamos morfologinės kategorijos (kaitybos klasė ir giminėmis nekaitomų klasių giminė)	a	ja	a_Masc
M2c konkrečiõs žodžio formos fleksinė morfologija	konkrečiõs žodžio formos fleksinės morfologijos požymiai	Act_Ind_Pres_Pl_1	Sup_Masc_Sg_Loc	Sg_Loc

Daugiasluoksnė hierarchinė korpuso anotacijos struktūra ne tik įgalina detaliai aprašyti kiekvieno teksto morfologinę sandarą, bet ir gali būti diachroninio žodyno bei senosios lietuvių kalbos gramatikos pagrindas. Toliau galima ir reikia numatyti sintaksinės anotacijos sluoksnius, eksplicitinių ir implicitinių citatų žymėjimą, susiejimą su tiksliai nustatytais ir lingvistiškai anotuotais vertimo šaltiniais. Bet tai jau atskiro pranešimo reikalinga tema.

SLIEKKO iniciatyvoje ir parengiamuosiuose darbuose šiuo metu bendradarbiauja Frankfurto prie Maino Goethe's universitetas (Vokietija), Lietuvių kalbos institutas (Vilnius) ir Pisos universitetas (Italija). 2010 metais SLIEKKĄ rėmė Lietuvos Respublikos švietimo ir mokslo ministerija pagal programą „Mokslo ir studijų modernizavimas". Viliamės, kad tolesni korpuso darbai bus remiami bendromis tarptautinėmis jėgomis.

Ačiū Jums už kantrybę ir dėmesį! Būsiu labai dėkinga už Jūsų pastabas, komentarus ir pasvarstymus, kuriuos labai prašyčiau man atsiųsti elektroniniu paštu.