Asta Kazlauskienė, Erika Rimkutė, Andrius Utka

Vytauto Didžiojo universitetas, Lietuvių kalbos katedra

Kiekybiniai tyrimai kalbotyroje

Įvadas. Statistika ir lingvistika

Seniai niekas neabejoja, kad kiekybiniai dėsningumai kalboje iš tikrųjų egzistuoja, todėl šių dienų kalbotyra jau negali būti tik kokybės mokslas. Tarp kiekybinių ir kokybinių reiškinių dažnai esti ryšys (pavyzdžiui, tarp skiemenų ar garsų skaičiaus žodyje ir žodžio vartojimo dažnumo). Remiantis statistika galima greičiau įžvelgti kalbos dėsningumus (ar bent tendencijas). Vis dėlto statistika kokybinių ypatybių neatskleidžia, nors gali jas paaiškinti. Todėl ji kalbotyroje turi būti taikoma prasmingai, o ne dėl to, kad tyrimas neva atrodytų moksliškesnis.

Ne naujiena statistiniai metodai ir lietuvių kalbotyroje. Tokie tyrimai buvo (o dažnai dar ir dabar yra) labai įvairiai įvardijami: *matematinė lingvistika, statistinė lingvistika, statistika lingvistikoje, kalbos statistika, lingvostatistika, tekstometriniai tyrimai*. Pirmąjį terminą (*matematinė lingvistika*) reikėtų vartoti atsargiai, nes jis paprastai apima kalbos procesų matematinį modeliavimą. Todėl ten, kur stengiamasi nustatyti kiekybinius kalbos elementų struktūros ar vartosenos dėsningumus, kur bandoma įsitikinti, ar gauti skirtumai reikšmingi, vartotini kiti paminėti terminai.

Statistiniai metodai jau seniai įprasti eksperimentinėje fonetikoje matuojant, analizuojant, lyginant garsinio signalo akustinius požymius, vertinant paklaidas. Tokie yra daugelio lietuvių fonetikų eksperimentininkų darbai: jau klasika tapę Valerijos Vaitkevičiūtės, Alekso Girdenio, Antano Pakerio veikalai ir jaunesniosios kartos fonetikų tyrimai. Tai glaudžiausiai su tiksliaisiais mokslais susijusi kalbotyros šaka, kai kada netgi nelaikoma kalbotyra (tarkime, garsusis Davido Crystalo žodynas pavadintas „Kalbotyros ir fonetikos žodynas“ (angl. „A Dictionary of Linguistics & Phonetics“). Statistikos naudojimas eksperimentinėje fonetikoje nėra koks nors specifinis, išskirtinis (skaičiuojami vidurkiai, kai kurie statistiniai kriterijai). Tačiau pačios medžiagos surinkimas ir apdorojimas gana sudėtingas: reikia turėti ir mokėti

dirbti su specialiomis garso analizės programomis. Todėl tokių tyrimų čia plačiau nekomentuosime.

XX a. pabaigoje atsirado nemažai ir vadinamųjų tekstometrinių tyrimų, kuriais siekiama išsiaiškinti kalbos elementų vartojimo dėsningumus įvairiuose tekstuose. Tokiems tyrimams reikia didelės apimties šaltinio (konkrečių rašytinių tekstų ar sakytinės kalbos įrašų ir jų rašytinių variantų). Turėdami reprezentatyvią medžiagą mokslininkai gali susirasti, suskaičiuoti daug įvairių rūpimų dalykų. Tik visada svarbu neišleisti iš akių to, kad absoliutūs skaičiuojamų elementų kiekiai labai priklauso nuo medžiagos (vadinamosios imties) dydžio: kuo daugiau medžiagos, tuo daugiau ir atskirų vienetų. Todėl geriau pateikti santykinius dydžius (pvz., procentinę išraišką) ar kitaip įvardyti santykius (pvz., triskiemeniai žodžiai *du kartus* dažnesni už keturių skiemenų žodžius).

Rezultatai priklauso ne tik nuo medžiagos apimties, bet ir nuo jos pobūdžio. Pavyzdžiui, moksliniame tekste visada rasime daugiau tarptautinių žodžių, terminų, daug labai ilgų sakinių, o šnekamojoje kalboje – trumpųjų formų, nepilnų sakinių. Dėl to rengdami medžiagą tyrimui mokslininkai turi atidžiai atrinkti tekstus ir jų kiekį.

Gana dažnai tokie tyrimai atliekami su labai didelėmis duomenų sankaupomis, kurių niekaip rankomis neapdorosi, akimis rezultatų neaprėpsi. Tai iš dalies lėmė ir rimtų tekstometrinių tyrimų pasirodymą tik tada, kai atsirado galimybė pasinaudoti techninėmis priemonėmis (pirmosiomis skaičiavimo mašinomis ar moderniais kompiuteriais).

Didelės elektroninės tekstų (tiek sakytinės, tiek rašytinės kalbos) sankaupos vadinamos tekstynais. Tekstynai paprastai naudojami natūralios kalbos vartosenai tirti, leidžia daryti objektyvias išvadas (plačiau apie tekstynus žr. http://donelaitis.vdu.lt). Jais remiantis tiriamas žodžių dažnumas, sudaromi dažniniai žodynai, analizuojama žodžių kontekstinė aplinka.

Fonotaktikos (garsų junginių, skiemens struktūros, jų vartosenos dėsningumų) tyrimai apskritai neįmanomi be statistinių duomenų (minėtini A. Girdenio, Vidos Karosienės, Irenos Kruopienės, Astos Kazlauskienės ir Gailiaus Raškinio bei kt. darbai). Toks statistikos naudojimo aspektas nesvetimas ir literatūrologams (Juozo Girdzijausko, Leonardo Saukos eilėdaros, Valerijos Ramonaitės prozos ritmo tyrimai).

Tekstometriniais tyrimais naudojasi ir kalbos ekspertai, norėdami nustatyti autoriaus stiliaus ypatybes ar net autorystę. Tačiau tam jau reikia rimtos specializacijos, tokio pobūdžio darbų principai ir daugelis aspektų dėl suprantamų priežasčių retokai viešinami (bendrą supratimą galima susidaryti paskaičius Renatos Ryngevič (2007), Anelės Žalkauskienės (2005), Gintarės Žalkauskaitės (2011) straipsnius).

Pastaraisiais dešimtmečiais labai suaktyvėjo įvairaus pobūdžio lingvistinės apklausos (sociolingvistiniai tyrimai). Klausiama nuomonės apie kalbos reiškinius, bandoma išsiaiškinti, ką vartoja, kam teikia pirmenybę apklausiamieji (respondentai). Anketų duomenys turi būti apdoroti ir įvertinti. Tam irgi dažnai praverčia įvairūs statistikos metodai. Iš lietuvių autorių plačiausiai, nuosekliausiai ir humanitarams suprantamiausiai apie statistikos panaudojimo galimybes filologijoje rašo Petras

Skirmantas (jo paskaitų studentams medžiagą „Statistika filologams – teorija ir pratybos“ galima rasti internete http://www.flf.vu.lt/index.php?id=33).

Atliekant kiekybinius skaičiavimus vertėtų atsakyti sau į klausimą, ar tyrime tik naudojami statistiniai duomenys (įvairūs tam tikrų reiškinių skaičiavimai), ar taikoma aprašomoji statistika (pirminis skaitmeninių duomenų apdorojimas, pvz., vidurkis, procentinė išraiška), ar pasitelkiami statistiniai metodai (pvz., koreliacinė analizė). Dažnai paprastas žodžių suskaičiavimas ir skaičių palyginimas nepagrįstai yra įvardijamas kaip statistinis metodas.

Jūsų dėmesiui skiriame kelis straipsnius, kurių tikslas – supažindinti su kiekybinių kalbos tyrimų specifika, įvairove ir jų interpretavimu. Čia bus pristatomi tokie tyrimai, kuriems atlikti nereikia nei išklausyti atskiro statistikos kurso, nei specializuotų statistinių programų. Todėl tikimės, kad ne vienas mokytojas jais paįvairins pamokas ir paskatins mokinius atlikti mažą kalbos tyrimą.

Šį kartą išsamiau aptarsime tik vieną kiekybinių tyrimų aspektą – žodžių foneminės ir morfeminės struktūros analizę.

Foneminės žodžių struktūros dėsningumai

Ne vienam įdomu sužinoti, ar ilgi lietuvių kalbos žodžiai. Todėl V. Karosienė ir A. Girdenis (1993) pabandė nustatyti, kiek fonemų (garsų) sudaro rišlios kalbos žodžius. Jie analizavo beveik 14,5 tūkst. rišlaus teksto fonologinių žodžių (tai žodžiai su prisišliejusiais nekaitomais, paprastai vienskiemeniais, žodžiais). Rezultatai rodo, kad žodžiai gali būti nuo 1 iki 18 fonemų. Būdingiausi yra 4–8 fonemų žodžiai, jie sudaro beveik tris ketvirtadalius teksto, pvz.: *buvo, darbo, didelių, gyvybei, pavojaus*.

Vėliau mėginta išsiaiškinti, kokios struktūros žodžiai vyrauja (Kazlauskienė; 2010). Tai padaryti labai sunku, nes žodžio fonetinės struktūros modelių įvairovė didelė, gal net nesuskaičiuojama. Tačiau tyrimas parodė, kad nedideliame tekstyne, kurį sudarė 109 tūkst. žodžių, dažniausi 85 modeliai apima net tris ketvirtadalius visų tirtų žodžių. Dažniausiems modeliams būdinga nesudėtinga foneminė struktūra (priebalsis, balsis, priebalsis, balsis ir t. t.), pvz.: *gyvenimo, negalima, vadinasi*. Lietuvių kalbos žodžiai dažniausiai (keturi penktadaliai) pradedami priebalsiu ir baigiami balsiu (du trečdaliai žodžių), pvz.: *gali, būti, savo, galima, duomenų, sudaro, darbo, kalba, kalbų*. Vengiama ilgesnių nei du nariai priebalsinių samplaikų ir žodžio pradžioje, ir viduje, ir gale (sudaro apie 2 %). Tarp balsių žodžio viduje dažniausiai (73 %) esti tik vienas priebalsis, kaip ir žodžio pradžioje (85 %) ar gale (93 %) (plg. anksčiau paminėtus pavyzdžius).

Foneminę (garsinę) žodžių struktūrą nustatyti sunku: reikia žodžius transkribuoti, t. y. raides perrašyti garsais. Tačiau paanalizuoti, kiek raidžių ar skiemenų sudaro žodžius, gali ir mokiniai. Visai nesudėtingas galėtų būti mokslinio tekstelio ir, tarkim, grožinio kūrinio palyginimas tokiu aspektu.

Morfeminės struktūros modelių įvairovė ir dažniausi modeliai

Lietuvių kalbos žodžių struktūrą galima analizuoti ne tik garsų (ar skiemenų) atžvilgiu, bet ir pagal morfemas.

Dar iš mokyklinių gramatikų kiekvienas žino, kad žodį sudaro įvairios sudėtinės dalys (svarbiausia šaknis ir galūnė, prie jų gali prisišlieti priešdėlių, priesagų, sangrąžos afiksas). Tačiau retai pasvarstome, kokios struktūros žodžiai vyrauja lietuvių kalboje.

Lietuvių kalba yra fleksinė sintetinė kalba. Tai reiškia, kad gramatiniai žodžių santykiai reiškiami galūnėmis ir jos gali turėti ne vieną reikšmę, pavyzdžiui, žodyje *geruose* galūnė *-uose* rodo ir vyriškąją giminę, ir daugiskaitą, ir vietininką. Vadinasi, daugelis žodžių būtinai turi turėti šaknį ir galūnę, o priešdėliai ir priesagos dažniausiai yra darybiniai afiksai, su jais sudaromi nauji žodžiai. Žinoma, lietuvių kalboje yra ir kaitybinių priesagų. Tokios yra išvestinių veiksmažodžio formų priesagos, pavyzdžiui, *skaito* ir *skaitantis*. Čia esamojo laiko dalyvis turi kaitybinę priesagą *-ant-*.

Lietuvos mokslo tarybos 2010–2011 m. finansuoto projekto „Morfeminė lietuvių kalbos žodžių struktūra“ metu išsiaiškinta, kad žodžių morfeminės struktūros modelių gali būti labai daug.

Teoriškai, tarkim, viename žodyje gali būti: trys priešdėliai, sangrąžos afiksas, dvi šaknys, tarp jų dar jungiamasis balsis, trys priesagos ir galūnė (keturių priešdėlių ar priesagų nepavyko rasti nė viename žodyje). Žodžiai gali būti sudaryti net iš dešimties morfemų, bet ir be išsamių tyrimų galima pasakyti, kad tokių žodžių kalboje negali būti daug. Net ir klasikinis, visiems gerai žinomas ilgiausio žodžio pavyzdys *nebeprisikiškiakopūsteliaudamas* nėra teoriškai maksimalios struktūros (trūksta dar vienos priesagos ir priešdėlio, jei *nebe-* laikome vienu priešdėliu).

Mūsų tyrimas parodė, kad morfeminės struktūros įvairovė didžiulė: nuo 10 modelių skaitvardžių ir įvardžių iki 116 veiksmažodžių. Jų produktyvumas labai skiriasi: kai kurių modelių tėra tik po vieną žodį, pavyzdžiui, septynias šaknis turintis daiktavardis *fibroezofagogastroduodenoskopija*, dvi šaknis, priešdėlį, dvi priesagas ir įvardžiuotinę galūnę turintis būdvardis *ikiindoeuropietiškųjų*, iš trijų priešdėlių, sangrąžos afikso, vienos kaitybinės (liepiamosios nuosakos) ir dviejų darybinių priesagų, šaknies ir galūnės sudarytas veiksmažodis *neįsipareigokite*. Tai vieni iš daugiausiai morfemų turinčių žodžių, dėl to ir rečiausiai vartojamų.

Kitus morfeminės struktūros modelius reprezentuoja šimtai žodžių, pvz., galime pasakyti šimtus ar net tūkstančius iš šaknies ir galūnės sudarytų žodžių (*namas, laukia, geras, jis, vienas* ir t. t.). Taigi galima daryti išvadą, kad kuo žodžio morfeminės struktūros modelis sudėtingesnis, tuo jis retesnis (pvz., lietuvių kalboje ilgesni nei keturių morfemų žodžiai sudaro mažiau nei dešimtadalį visų žodžių).

Išanalizavus beveik 312 tūkst. kaitomų žodžių, nustatyta, kad lietuvių kalbai būdingiausias žodžių modelis *šaknis + galūnė*. Tokios struktūros yra penktadalis veiksmažodžių, pusė daiktavardžių, trečdalis būdvardžių, net du trečdaliai skaitvardžių, keturi penktadaliai įvardžių (žr. pavyzdžius, pateiktus ankstesnėje pastraipoje). Vargu ar tai turi stebinti, žinant, kad lietuvių kalboje, kaip jau minėta, afiksas turi ne

vieną reikšmę (beveik visos gramatinės formos išreiškiamos viena morfema – galūne). Vadinasi, remdamiesi kiekybinio tyrimo rezultatais galime daryti išvadą, kad fleksinei kalbai būdingesni nesudėtingos struktūros žodžiai (kiekybinis tyrimas padeda atskleisti, paaiškinti kokybinę ypatybę).

Kita vertus, lietuvių kalba, skirtingai nei, pavyzdžiui, anglų, turi labai daug darybos afiksų. Todėl darinių kalboje taip pat apstu. Tą liudija ir mūsų tyrimas. Tačiau reikia pasakyti, kad kalboje akivaizdi tendencija vartoti žodžius tik su vienu darybos afiksu (priešdėliu arba priesaga). Antrasis pagal bendrąjį modelių dažnumą yra *šaknis* + *priesaga* + *galūnė* modelis. Tokių yra šiek tiek daugiau nei dešimtadalis veiksmažodžių, penktadalis daiktavardžių, net du penktadaliai būdvardžių, po dešimtadalį skaitvardžių ir įvardžių (pvz.: *galvoti, gerumas, didingas, penktas, toks*). Trečiasis, bet jau gerokai retesnis modelis, yra *priešdėlis* + *šaknis* + *galūnė*: šiek tiek daugiau nei dešimtadalis veiksmažodžių, beveik dešimtadalis daiktavardžių ir būdvardžių (pvz.: *nemyli, apsauga, nemažas*).

Taigi matyti, kad lietuvių kalbai būdingiausi pirminiai žodžiai ir žodžiai su vienu darybiniu afiksu. Į tokį akivaizdų kalbos polinkį turėtų atkreipti dėmesį kalbos normintojai, kurdami naujus svetimybių atitikmenis.

Morfeminės analizės žodynai

Remiantis šio tyrimo medžiaga sudaryti elektroniniai internete laisvai prieinami morfemų žodynai: abėcėlinis, dažninis ir atgalinis, kuriuose pateikiama: konkreti tekstyne pavartota forma, morfeminė žodžio analizė (t. y. žodis suskaidytas į tiek dalių, kiek jų galima įžvelgti dabartinės kalbos požiūriu), pavartojimo dažnumas, lema (dar kitaip vadinama antraštine, žodynine forma, pvz., veiksmažodžių lema yra bendratis, daiktavardžių – vienaskaitos vardininkas, būdvardžių – vyriškosios giminės vienaskaitos vardininkas ir pan.), morfologiniai požymiai (pvz., nuosaka, laikas, giminė, skaičius, laipsnis). Žodynus galima rasti VDU Lietuvių kalbos katedros svetainėje (http://donelaitis.vdu.lt/lkk). Šiais žodynais gali naudotis mokytojai, dėstytojai, rengiantys užduotis.

Nustatant morfemų ribas laikytasi teorinių principų, pateiktų Vinco Urbučio „Žodžių darybos teorijoje“ (2009). Žodžiai suskaidyti į tiek morfemų (mažiausių reikšminių žodžio dalių), kiek dabartinės kalbos požiūriu jų galima nustatyti. Išskiriama žodžio dalis laikyta morfema, jeigu ji pasikartoja kituose žodžiuose: a) lengviausia, kai abi (arba visos) žodžio dalys (įtariamos morfemos) pasikartoja kituose žodžiuose (pvz.: *nam-el-is – namas, naminis* ir *stalelis, vaikelis*); b) tais atvejais, kai tik viena kuri nors žodžio dalis – šaknis ar afiksas – pasikartoja kituose žodžiuose (pvz.: *dilet-ant-as – debiutantas, praktikantas*).

Fleksinės sintetinės kalbos turi dar vieną svarbią ypatybę: labai išblukusias ribas tarp morfemų. Todėl dabartinėje lietuvių kalboje yra nemažai žodžių, kurių net nelaikome priesagų vediniais, pvz.: *antklodė, paklodė, įklodė* yra padaryti iš *kloti* su

priesaga *-dė* (su atitinkamais priešdėliais). Šiuose žodžiuose priesagas įžvelgti dar nėra taip sunku. Kitas dalykas tokie žodžiai kaip *miltai, butas, raktas, turtas, šlaitas* ir kt., kurie turi priesagą *-t-*. Čia tikrai ne visada lengva rasti pamatinį žodį dabartinėje kalboje, ypač turint galvoje, kad kai kurių žodžių šaknyje yra įvykusi garsų kaita. Tokie atvejai fiksuoti šiuose žodynuose, todėl reikalui esant čia galima pasitikslinti rūpimo žodžio morfeminę skaidą.

Norime atkreipti mokytojų dėmesį, kad ne viskas, kas yra pateikta šiuose žodynuose, gali būti naudojama rengiant mokiniams užduotis. Gana dažnai jiems gali būti sunku nustatyti morfemas. Pavyzdžiui, *nagrinėti* siejama su *nagrus*, nes toks būdvardis dabartinėje bendrinėje kalboje greičiausiai nevartojamas ir mokiniai jo nežino. Veiksmažodyje *bombarduoti* šaknies morfema laikyta *bombard-*, nes šis veiksmažodis sietinas su prancūzų kalbos *bombarder,* taigi nėra pagrindo skirti morfemos *-ard*.

Atgaliniai sąrašai gali praversti ne tik sudarant morfemikos užduotis, bet ir mokant žodžių darybos, nes žodžiai čia išrikiuoti pagal paskutinę raidę. Tad galima išsirinkti visus vienos priesagos vedinius, pvz.: su priesaga *-ybė*: *savivaldybė, realybė, ramybė, priešybė, platybė, nekaltybė, savybė*.

Šiuose žodynuose pateiktas labai didelis kiekis morfemiškai suskaidytų kaitomų žodžių, todėl remdamiesi žodynais mokiniai galėtų atlikti ir savo mažą tyrimą: nustatyti pasirinkto teksto fragmento žodžių morfeminę struktūrą.

Baigiamoji pastaba

Čia pateikėme tik pačias svarbiausias statistikos metodų panaudojimo kalbotyroje galimybes ir porą žodžių struktūros tyrimų. Plačiau apie kitus kiekybinius kalbos tyrimus rašysime tolesniuose straipsniuose.

Literatūra

1. *Karosienė V., Girdenis A.* 1993: Bendrinės lietuvių kalbos fonemų dažnumas // Kalbotyra, t. 42 (1), p. 28–38.
2. *Kazlauskienė A.* 2010: Lietuvių kalbos žodžių foneminės struktūros dėsningumai // Žmogus ir žodis, t. 12, nr. 1, p. 35–41.
3. *Ryngevič R.* 2007: Anoniminio teksto autoriaus lyties nustatymo klausimai // Jurisprudencija, t. 11 (101), p. 86–90.
4. *Urbutis V.* 2009: Žodžių darybos teorija. Vilnius: Mokslo ir enciklopedijų institutas.
5. *Žalkauskaitė G.* 2011: Idiolektų požymiai elektroninių laiškų skyryboje. *Lietuvių kalba*. http://www.lietuviukalba.lt/index.php?id=183 (žiūrėta 2011-06-30)
6. *Žalkauskienė A.* 2005: Lietuviško teksto autoriaus nustatymo metodikos pagrindai // Jurisprudencija, t. 18 (10), p. 113–121.