# Dėl nepatikimos informacijos Daivos Šveikauskienės straipsnyje „Lietuvių kalbos gramatikos informacinė sistema: I morfologija“, publikuotame elektroniniame žurnale *Lietuvių kalba* 2016 m. nr. 10.

**Andrius Utka, Agnė Bielinskienė, Loic Boizou, Erika Rimkutė, Jolanta Kovalevskaitė**
Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centras
K. Donelaičio g. 52
LT-44244 Kaunas
El. paštas *klc@hmf.vdu.lt*

D. Šveikauskienė šiame straipsnyje apžvelgia kitų kalbų ir lietuvių kalbai pritaikytas automatinės morfeminės ir morfologinės analizės bei sintezės programas, pristato Lietuvių kalbos institute planuojamą kurti lietuvių kalbos gramatikos informacinę sistemą. Šioje sistemoje bus pateikiama morfeminė ir morfologinė informacija.

Straipsnio autorė daug kartų pabrėžia, kad Vytauto Didžiojo universitete sukurtos programos, duomenų bazės pateikia klaidinančią informaciją, veikia netiksliai ir pan. D. Šveikauskienė siūlo gana neįprastus analizės ir vertinimo būdus. Be to, straipsnyje savo teiginiams pagrįsti pateikia realybės neatitinkančias internetinių puslapių ekrano kopijas (14, 15, 16 ir 17 pav.) bei neteisingas internetines nuorodas (publikacijos šaltinių sąrašo 12 interneto nuoroda). Norime atsakyti į daugeliu atvejų nepagrįstą kritiką, paaiškinti, kodėl negalima pasitikėti minėtame straipsnyje pateikta informacija ir vertinimais. Mūsų nuomone autorė straipsnį parašė nesusipažinusi su visa, viešai prieinama informacija jos straipsnio tema.

Straipsnio įvade (p. 1) D. Šveikauskienė rašo „Lietuvių kalbos institute pradėta kurti lietuvių kalbos gramatikos informacinė sistema. Ji apima dvi sritis – morfologiją ir sintaksę. Pirmo etapo metu bus paruošti morfologiniai duomenys. Pagrindinis tikslas – sukaupti išsamią gramatinę informaciją apie visų lietuvių kalbos žodžių visas formas. Vertinant jau atliktus lietuvių kalbos kompiuterizavimo darbus galima pasakyti, kad jie visi turi vieną bruožą – atspindi lietuvių kalbą fragmentiškai.“

Šią straipsnio ištrauką galima pakomentuoti taip: neįmanoma sukaupti informacijos apie visus žodžius, nes nuolatos atsiranda naujų. Kita vertus, turint gerą sintezavimo įrankį, galima susintezuoti visas įmanomas turimų žodžių formas. Bet kam to reikia? Kaip tada bus atrinktos realiai vartojamos ir tik teoriškai galimos formos? Apie tokį tik teoriškai galimų žodžių poveikį automatinės morfologinės analizės kokybei dar 2000 m. rašė Vytautas Zinkevičius.

Tame pačiame puslapyje straipsnio autorė teigia, kad „[d]augiausia morfemikos kompiuterizavimo srityje nuveikta Vytauto didžiojo universitete (VDU), kur atliekami darbai remiasi tekstynu. Tačiau ir kitų kalbų lingvistai kaip trūkumą nurodo, kad tokio pobūdžio tyrimai teapima tik tekstyno žodžius ir tegali atspindėti tik juose esančią leksiką. Tai ypač aktualu didelio kaitomumo kalboms, nes „net ir labai didelės apimties tekstynuose gali nebūti rečiau pasitaikančių formų“ (Paikens, Rituma, Pretkalnina 2013, 272). Ne kitokia padėtis ir su lietuvių kalba. VDU paruoštose duomenų bazėse – tiek morfemikos, tiek morfologijos – trūksta kai kurių žodžių formų. Morfemikos duomenų bazėje (1 interneto nuoroda), nėra labai įprastų, gerai visiems žinomų žodžių, pvz., *laikmenai*, *laikmeną*, *laikmenoms*, *laikmenomis*, *laikrodžiui*, *laikrodyje*, *laikrodžių*, *laikrodžiams*, *laikrodžiais*, *laikrodžiuose* ir tai tokie žodžiai, kurių negalima atmesti ir traktuoti juos kaip nevartojamus, t.y. archaizmus ar pan. Trūksta nutrumpėjusių formų, kurios ypač paplitusios šnekamojoje kalboje, pvz., *laikrody*, *laikrodžiuos*, *šnekamojoj*. Bandant gauti informaciją apie žodžio *šnekamojoj* morfemas, sistema nurodo, kad duomenų bazėje tokio žodžio nėra, o 2015 metais sukurta ir viešai internete prieinama Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema (2 interneto nuoroda) teigia netgi klaidinančią informaciją: žodžių junginiui *šnekamojoj kalboj* parašo: „Pateiktas tekstas yra ne lietuvių kalba arba gramatiškai neteisingas“.“

Galima sutikti su D. Šveikauskienės teiginiu, kad „tokio pobūdžio tyrimai teapima tik tekstyno žodžius ir tegali atspindėti tik juose esančią leksiką“. Kyla klausimas, o iš kur dar imti reprezentatyvius dabartinės kalbos duomenis, jei ne iš tekstyno? Vargu ar verta aiškinti, kad duomenų kiekis priklauso nuo tekstynų dydžio, tad nereikėtų stebėtis, kad aptariamo straipsnio autorė morfemikos duomenų bazėje, kuri sudaryta remiantis apie 310 tūkstančių žodžių tekstynu, nerado anksčiau minėtų žodžių. Kita vertus, reikia pabrėžti, kad net ir didesniame tekstyne ne visada galima rasti ne pačių dažniausių žodžių (*laikrodis, laikmena* nėra dažni lietuvių kalbos žodžiai) rečiau vartojamų linksnių, pvz., naudininko, įnagininko, vietininko (apie lietuvių kalbos linksnių dažnumą ir dažniausius žodžius žr. Rimkutė 2006; Dabašinskienė 2009; Žilinskienė 2005; Grumadienė et al. 1998).

Svarbu akcentuoti, kad ankstesnėje citatoje D. Šveikauskienė painioja du skirtingus dalykus: duomenų bazės nepilnumą (su tuo iš dalies galima sutikti, nes analizuotas gana nedidelis duomenų kiekis) ir duomenų bazės pritaikymą kalbos vartosenai nustatyti, nes iš anksčiau minėto teiginio galima suprasti, kad jei žodžio nėra morfemikos duomenų bazėje, tai žodis yra nevartotinas. Sudarant morfemikos duomenų bazę norėta nustatyti morfemų ribas, jų tipus analizuotuose žodžiuose, o ne pateikti visus įmanomus žodžius, juo labiau juos suklasifikuoti pagal kokius nors vartosenos požymius (vartotini ar nevartoti, dabartinės kalbos ar archajiški ir pan.).

Sutinkame su pastaba, kad *Lietuvių kalbos sintaksinės ir semantinės analizės informacinėje sistemoje* tais atvejais, kai dėl įvairių priežasčių nepavyksta išanalizuoti teksto, pateikiamas ne visiškai tinkamas paaiškinimas „Pateiktas tekstas yra ne lietuvių kalba arba gramatiškai neteisingas“. Pripažįstame, kad įspėjimo formuluotė yra klaidinanti, vis dėlto autorė turėjo pastebėti, kad visais atvejais sistema pateikia šį vienintelį standartinį įspėjimą ir tai nebūtinai reiškia, kad pateiktas ne lietuvių kalba parašytas ar gramatiškai netaisyklingas tekstas.

Įvade toliau rašoma: „Todėl nuspręsta kurti lietuvių kalbos gramatikos informacinę sistemą, kurios tikslas – pradžioje sukaupti išsamius ir labai aukšto patikimumo duomenis apie visų lietuvių kalbos žodžių gramatinius požymius, o ateityje įtraukti ir sintaksės duomenis.“ Kaip minėta anksčiau, neįmanoma sukaupti visų lietuvių kalbos žodžių. Būtina sukonkretinti, ką reiškia „aukšto patikimumo duomenys“ ir kas, kokiais metodais jį nustatys, ką turi mintyje teigdama „sintaksės duomenys“.

Tęsiant diskusiją reikėtų pabrėžti, kad VDU Kompiuterinės lingvistikos centro mokslininkai (ir turbūt kiti tekstynų bei kompiuterinės lingvistikos specialistai) skirtingai nei straipsnio autorė suvokia taikomosios kalbotyros tikslą. Mūsų manymu, kalbos analizei automatizuoti reikalingos su tam tikra paklaida veikiančios programos. Susidaro įspūdis, kad D. Šveikauskienės siekiamybė – sukurti be jokių klaidų veikiančią sistemą, tik tada neaišku, kiek duomenų tokia sistema gebėtų išanalizuoti. Išanalizavus kelis žodžius, kelis sakinius, galima sukurti tobulą automatinės morfologinės ir sintaksinės analizės sistemą, bet ar tokia sistema galės anotuoti šimtus tūkstančių ar milijonų žodžių? Kiek mums žinoma, net ir tų kalbų, kurių automatinės kalbos analizės sukurtos ir tobulinamos kelis dešimtmečius, automatinė analizė neveikia 100 procentų tikslumu.

P. 3 autorė išsamiai aprašo VDU parengtą morfemikos žodyną, pateikia pastabų dėl medžiagos vaizdavimo būdo: „2011 metais pasirodė viešai internete prieinamas morfemikos žodynas, kuris sukurtas Vytauto Didžiojo universitete tekstyno pagrindu ir teapima tik jame esančius žodžius. Sunku suprasti, kodėl buvo pasirinktas toks neinformatyvus morfemų vaizdavimo būdas – jos atskiriamos viena nuo kitos brūkšneliais, visai nepateikiant jokios informacijos apie morfemos tipą ir vienodai vaizduojant skirtingos morfeminės struktūros žodžius – kai gausu gerų pavyzdžių tiek kitų kalbų, tiek lietuvių kalbos žodžių skaidyme į morfemas jau buvo anksčiau. Neaišku, kodėl nepasidomėta ir nepasinaudota tikrai gera ir vertinga patirtimi. Vienintelė priežastis turbūt – ribotos kompiuterių galimybės šioje srityje.

VDU darbai atlikti tyrinėjant žodžių morfeminę struktūrą apima 310 000 žodžių analizę (Rimkutė, Kazlauskienė, Raškinis 2011a, 7). Rezultatai pateikiami trijų tomų žodyne (5, 6 ir 7 interneto nuorodos), kur žodžiai išskaidyti morfemomis, ir jos atskirtos viena nuo kitos brūkšneliais.

Taigi, išsamios informacijos jame trūksta. Kaip vieną iš pačių didžiausių trūkumų galima būtų paminėti informacijos apie morfemos tipą nebuvimą. Nors žodyno aprašyme sakoma, kad *-un-* laikoma priesaga žodyje *šunį* (Rimkutė, Kazlauskienė, Raškinis 2011, 7), tačiau žodyne jis pateikiamas tokios pat struktūros, kaip ir žodis *sutemos*: *š-un-s* (Rimkutė, Kazlauskienė, Raškinis 2011a, 686) ir *su-tem-os* (Rimkutė, Kazlauskienė, Raškinis 2011a, 665). Abu šie žodžiai sudaryti iš trijų morfemų, tačiau visai nėra informacijos apie tai, kad žodyje *šuns* pirma morfema yra šaknis, antra – priesaga, o žodyje *sutemos* pirma morfema yra priešdėlis, o antra – šaknis. Patys autoriai įvade nurodo, kad ateityje ketinama parengti daug išsamesnį žodyną. 2013 metais pasirodė viešai prieinama internete *Lietuvių kalbos morfemikos duomenų bazė* (1 Interneto nuoroda), tačiau autorių ketinimai nebuvo įvykdyti. Paruošta tik patogesnė paieška pateikiant morfemikos žodyne esančius duomenis, tačiau informacija nepasidarė nė kiek išsamesnė – tai, kas buvo žodynuose, perkelta į duomenų bazę, bet papildomai neatlikta nieko: žodis į morfemas skaidomas tuo pačiu principu – atskiriant jas brūkšneliais, kaip ir buvo žodyne. Pateikiamų duomenų apimtis taip pat išliko ta pati: žodžio išskaidymas morfemomis, jo lema, dažnumas ir gramatinė informacija (7 pav. ir 8 pav.). Tesiskiria tik informacijos išdėstymas ekrane, bet ne jos turinys. Žodžių kiekis taip pat nepadidėjo: tų žodžių, kurių nebuvo morfemikos žodyne, nėra ir morfemikos duomenų bazėje. Šiuos teiginius gerai pagrindžia pavyzdžiai. Žodyne yra šeši įrašai su žodžio *laikrodis* formomis (Rimkutė, Kazlauskienė, Raškinis 2011a, 332) (7 pav.). Duomenų bazėje taip pat tegalima gauti informaciją tik apie šias šešias žodžio formas. Tų formų, kurių nebuvo žodyne, pvz. *laikrodžiams* (7 pav.), nėra ir duomenų bazėje (1 Interneto nuoroda) (9 pav.). Ir negalima teigti, kad tai retai pasitaikantis žodis: sakinys, pvz., *Manoma, kad laikrodžiams prižiūrėti kasmet reikės iki 3 tūkst. litų* yra labai įprastas ir vartojamas, paimtas iš tekstyno (8 Interneto nuoroda).

<...> Ir tai yra viena priežasčių, kodėl buvo nuspręsta sukurti lietuvių kalbos gramatikos informacinę sistemą, apimančią išsamius gramatinius duomenis apie lietuvių kalbos žodžius bei sakinius. Ji bus laisvai prieinama internete ir skiriama plačiajam vartotojų ratui, todėl duomenys bus pateikiami populiariai: jais naudotis galės ir neturintys specialaus išsilavinimo žmonės. VDU išleistame morfemikos žodyne ir jo pagrindu paruoštoje morfemikos duomenų bazėje informacija gali būti naudinga tik gilias lituanistines žinias turintiems specialistams, kurie labai gerai žino žodžių skaidymą į morfemas. Tačiau neturintiems specialaus išsilavinimo žmonėms žodžių užrašymas atskiriant tam tikrus raidžių rinkinius brūkšneliais dažniausiai naudingos morfeminės informacijos nesuteikia.“

Mūsų komentaras: 2011 m. morfemikos žodynai sudaryti vykdant Lietuvos mokslo tarybos finansuotą projektą „Morfeminė lietuvių kalbos žodžių struktūra“. Šio projekto pagrindinis tikslas – nustatyti ir aprašyti dabartinės lietuvių kalbos žodžių morfeminės struktūros modelius, ypač kreipiant dėmesį į jų produktyvumą ir vartojimo dažnumą. Šiam tikslui pasiekti buvo sudarytas tekstynas, jame pavartoti žodžiai buvo morfemiškai suskaidyti: nustatytos morfemų ribos ir jų tipai. Projekte neįsipareigota viešai prieigai pateikti informaciją apie morfemų tipus. 2013 m. kelių KLC mokslininkų iniciatyva (be jokių finansinių įsipareigojimų) anksčiau minėtų morfemikos žodynų pagrindu parengta morfemikos duomenų bazė tam, kad būtų patogesnė prieiga prie duomenų. Taigi jei D. Šveikauskienė kai kurių žodžių nerado morfemikos žodynuose, jų negalėjo rasti ir morfemikos duomenų bazėje. Nebuvo numatyta ir niekam neįsipareigota šioje bazėje pateikti morfemų tipus. Remdamiesi išsamiais duomenimis apie lietuvių kalbos žodžių morfeminę struktūrą, VDU mokslininkai parengė kelis mokslo darbus (žr. Rimkutė et al. 2011; Kazlauskienė 2010; Kazlauskienė 2013; Kazlauskienė et al. 2013); VDU lituanistai parengė kelis bakalauro ir magistro darbus. Sutinkame, kad būtų patogu, jei interneto naudotojai galėtų matyti ne tik morfemų ribas, bet ir jų tipus. Tokiam darbui reikalingas atskiras finansavimas.

Trečiame skyriuje D. Šveikauskienė aptaria morfologinius analizatorius. Norime patikslinti keletą teiginių. P. 6 rašo, kad „[p]irmasis lietuvių kalbos morfologinis analizatorius sukurtas Matematikos ir informatikos institute 2000 metais. Tai lietuvių kalbos morfologinės analizės ir sintezės programinė įranga – lemuoklis (Zinkevičius 2000).“ Iš tikrųjų minėtame Z. Zinkevičiaus

straipsnyje (p. 245) rašoma: „Programa sukurta 2000 metais ir skirta Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro (KLC) tekstyno (Marcinkevičienė, 1997) moksliniams lingvistiniams tyrinėjimams automatizuoti.“

Toliau autorė teigia: „[d]idžiausias jo privalumas yra tai, kad jis atpažįsta ir pateikia informaciją apie visus lietuvių kalboje esančius žodžius. Kaip trūkumą galima paminėti pateikiamus perteklinius žodžius, kurių nėra lietuvių kalboje, pvz., žodžiui *blizgėjo* kaip trečias variantas nurodomas daiktavardžio **blizgėjas* kilmininko linksnis.“ Patiksliname: anksčiau minėtas anotatorius atpažįsta ne visus lietuvių kalbos žodžius, tuo labiau ne visus „lietuvių kalboje esančius žodžius“ (nes tai reiškia ir kitų kalbų žodžius, jei jie vartojami lietuviškuose tekstuose). 2007 m. nustatyta, kad ši programa neatpažįsta 5,6 proc. žodžių (žr. Rimkutė et al. 2007: 32). Anksčiau autorė minėjo, kad būtina analizuoti visus žodžius, o čia galima matyti, kokios „visų“ žodžių analizės pasekmės: atpažįstami, sugeneruojami realiai kalboje nevartojami žodžiai.

Tame pačiame puslapyje rašoma: „[š]ios programinės įrangos pagrindu Vytauto Didžiojo universitete 2008 m. sukurtas morfologinis analizatorius prieinamas viešai internete (12 interneto nuoroda), tačiau informacija pateikiama naudojant sutrumpinimus ir anglų kalbos žodžius, ir dėl to plačiajai visuomenei toks formatas nėra labai patogus naudotis.“ Straipsnyje pateikta neteisinga minėtos programos nuoroda, t. y. http://donelaitis.vdu.lt/NLP/nlp.php. Visiems interneto vartotojams prieinama programa yra pateikta čia: http://tekstynas.vdu.lt/page.xhtml?id=morphological-annotator.

P. 7 autorė toliau diskutuoja dėl teoriškai galimų žodžių: „Kita funkcija „Lemuoti“ pateikia visus galimus daugiareikšmių žodžių variantus, bet tuo pačiu ir lietuvių kalboje neegzistuojančius žodžius, tokius kaip, pvz., **blizgėjas* (16 pav.). Informacijoje apie žodį nurodytas statusas „teorinis“ lietuvių kalbos žodžiu jo nepadaro. Kad tam tikras raidžių rinkinys būtų kokios nors kalbos žodis, jis turi atitikti tris reikalavimus: a) jis turi turėti garsinę struktūrą, b) turi egzistuoti tikrovėje daiktas ar reiškinys, kurį tas žodis pavadina ir c) žmogaus sąmonėje turi būti to daikto ar reiškinio atspindys (Jakaitienė 1980, 16). Raidžių rinkinys „**blizgėjas*“ tenkina tik pirmąjį reikalavimą – turi garsinę struktūrą, kitų dviejų reikalavimų jis neatitinka: nėra nei daikto ar reiškinio tikrovėje, kurį jis pavadintų, nei jo atspindžio žmogaus sąmonėje. Nei tas lietuvis, kuris sako, t.y. ištaria garsų rinkinį „**blizgėjas*“, nei tas, kuris jį girdi, nežino, ką tai reiškia. Vadinasi, tai nėra lietuvių kalbos žodis.“ Nesiveldami į išsamesnes diskusijas apie tai, kas yra žodis, rekomenduotume skaitytojams paskaityti 2000 m. Z. Zinkevičiaus straipsnį, kuriame išsamiai paaiškina, kodėl atsirado tokio tipo žodžių, kaip *blizgėjas.*

P. 9 D. Šveikauskienė pateikia savo išvadą apie VDU KLC svetainėje (http://tekstynas.vdu.lt) ir svetainėje http://www.semantika.lt prieinamus morfologinius analizatorius: „Išvada galėtų būti tokia: panaikinant klaidas dėl perteklinių, lietuvių kalboje neegzistuojančių žodžių pateikimo, tuo pačiu prarandami milžiniški kiekiai ir naudingos informacijos. Sistemoje, kuri nebepateikia lietuvių kalboje nesančių žodžių, tokių kaip **blizgėjas* ar žodžio *susitikimas* būdvardžio varianto, labai daug taisyklingų ir dažnai vartojamų lietuvių kalbos žodžių pasidaro „tekstas ne lietuvių kalba“. Taigi lieka klausimas: kiek galima tikėti tokios sistemos duomenimis ir jos teikiama informacija? Apibendrinant galima būtų pasakyti: VDU darbai, tobulinant morfologinio analizatoriaus ir anotatoriaus veikimą, nepasiteisino, sintaksinės ir semantinės analizės informacinė sistema, sukurta 2015 m., daro dar didesnes klaidas nei analizatorius, pateiktas viešam naudojimui 2008 m. <...> Panašus atvejis dabar yra VDU morfologinė analizė: paskutinė (2015 metais pasirodžiusi) morfologinės analizės versija dirba žymiai blogiau nei ankstesnės (2008 metų).“

Pirmiausia paminėtina, kad *Lietuvių kalbos sintaksinės ir semantinės analizės informacinę sistemą* kūrė ne tik VDU, bet ir KTU mokslininkai. Antra, metodologiškai neteisinga įvertinti dvi sistemas remiantis dviejų žodžių (*blizgėjo* ir *susitikimas*) ir dviejų žodžių junginių (*mažas peliukas* ir *šnekamojoj kalboj*) analize. Kiek mums žinoma, norint įvertinti kelias morfologinės analizės programas, reikėtų išanalizuoti bent kelis tūkstančius žodžių (paprastai sistemos vertinamos remiantis kelių šimtų tūkstančių žodžių tekstynu). Dabartinė kompiuterinės lingvistika jau turi gerai aprašytus anotavimo kokybės vertinimo metodus, pvz., 1) aprašyti automatiniam sistemos mokymui

skirtus duomenis; 2) pirmasis įvertinimo ciklas naudojant dalį „auksinio standarto“ duomenų (t. y. patikimus, patikrintus, paprastai negausius duomenis); 3) įvertinimas naudojant visus „auksinio standarto“ duomenis; 4) galutinis sistemos įvertinimas (Paroubek et al. 2007: 11). Programų tikslumas skaičiuojamas tam tikrais įverčiais, pvz., *precision, Recall, F-Score.* Taigi kadangi nėra statistinio įvertinimo ir nepasirinktas tinkamas vertinimo metodas, galime teigti, kad šiame straipsnyje D. Šveikauskienės pateiktas anotatorių įvertinimas neturi jokios vertės.

Trečia, atsižvelgdami į realią vartoseną, nekuriame analizatorių, kurie sugeneruotų daug teorinių formų. Vidiniam naudojimui turime įrankį, vadinamąją *spėlioklę* (angl. *guesser*), kuri gali būti naudojama neatpažintiems žodžiams analizuoti.

Ketvirta, autorė KLC svetainėje ir http://www.semantika.lt svetainėje veikiančius morfologijos analizės komponentus traktuoja kaip vieno ir to paties komponento versijas. Tai yra visiškai skirtingi morfologijos analizatoriai. Ankstesnio analizatoriaus technologijos buvo atsisakyta dėl tam tikrų technologinių ribotumų ir, rengiant *Lietuvių kalbos sintaksinės ir semantinės analizės informacinės sistemos* komponentus, buvo sukurtas naujas morfologijos analizatorius su visiškai kitokiu technologiniu sprendimu. Šiame analizatoriuje yra galimybė papildyti leksinę bazę, pagerinti sutrumpintų formų ir įvardytųjų esybių atpažinimą. Analizatorius sukurtas kaip atvirojo kodo programa, pagrįstas plačiai naudojamais technologiniais sprendimais. Kol kas KLC neatliko rimto statistinio analizatorių palyginimo, todėl negalime skelbti patikimo įvertinimo.

P. 10 D. Šveikauskienė vėl grįžta prie morfeminės analizės. Moksliniame straipsnyje derėtų išsamesnis tiriamosios medžiagos pristatymas, nes p. 11 rašo: „[t]yrinėjant sudurtinius lietuvių kalbos žodžius didžiausias pastebėtas šaknų kiekis buvo trys šaknys – s*ienlaikraštis*.“ Neaišku, kur autorė tyrinėjo sudurtinius žodžius, kiek jų ištyrė. Rengiant anksčiau aptartą lietuvių kalbos morfemikos duomenų bazę, rastas net septynias šaknis turintis daiktavardis (žinoma, tai nėra įprastas ir populiarus žodis, bet, pvz., chemijos, medicinos srities terminai gali turėti daugiau negu tris šaknis). Taigi kyla abejonių dėl pasirinkto preskriptyviojo analizės principo: susidaro įspūdis, kad pirmiausia bandoma teoriškai aprašyti morfeminę struktūrą, o tada į ją bandyti „sukišti“ lietuvių kalbos žodžius.

Netikslus p. 11 parašytas šis teiginys: „[p]o šaknies esanti dalis apima priesagas, galūnę ir sangrąžos dalelytę *si*. Galūnė į atskirą žodžio dalį nebuvo išskirta, nes ji visada žodyje būna tik viena, ko negalima pasakyti apie priesagas.“ Kai kurie lietuvių kalbos žodžiai gali visai neturėti galūnių (pvz., *berniuk,* prieveiksmis *gerai,* bendratis *dirbti,* padalyvis *gyvenant*).

Tame pačiame puslapyje rašo: „Kuriant duomenų bazę siekiama kuo didesnio tikslumo ir duomenų patikimumo. Patikimumui užtikrinti naudojama daug žmogaus darbo. Duomenys apie pradinę žodžio formą (lemą) bus suvedami rankomis. Visos likusios formos generuojamos automatiškai panaudojant lietuvių kalbos žodžių morfologinės sintezės programinę įrangą, kuri 2000 metais buvo sukurta Matematikos ir informatikos institute Vilniuje (Zinkevičius 2000).“

Mums kyla klausimų, iš kur autorė gaus duomenų apie pradinę žodžio formą (beje, reikėtų vadinti pagrindine forma), kiek bus tų duomenų? Kodėl duomenis planuoja „suvesti rankomis“, o kodėl nesinaudoja LKI skaitmenizuotais žodynų duomenimis?

P. 12–13 pateikto teksto komentarai: „Kadangi informacinė sistema skirta plačiajai visuomenei, todėl pateikiant duomenis nebus naudojami sutrumpinimai – visa gramatinė informacija bus pateikiama pilnais žodžiais. Siekiant kuo didesnio vaizdumo atskiriems morfemų tipams naudojamos skirtingos spalvos. Įvertinant VDU *Lietuvių kalbos sintaksinės ir semantinės analizės informacinės sistemos* duomenų pateikimo būdą, reikia pasakyti, kad jis nėra optimalus. Kad vardininkas yra linksnis, o vienaskaita yra skaičius, reikia pasakyti kompiuteriui, bet ne žmogui. Žmogus šią informaciją ir taip žino. Todėl gramatikos informacinėje sistemoje pasirenkamas lakoniškesnis žodžių gramatinių požymių pateikimo būdas: jie surašomi ištisiniu tekstu, visi viename laukelyje ir svarbiausia – nenurodant požymių, kurių žodis neturi, pvz., netikslinga daiktavardžiui *medis* nurodyti, kad jis yra nesangrąžinis (22 pav.). Tai nereikalingas balastas. Lygiai taip pat netikslinga kiekvienam žodžiui nurodyti, kad jo šaknyje nėra infikso ar nevyksta balsių kaita. Tokio

tipo požymiai gramatikos informacinėje sistemoje bus pateikiami tik prie tų žodžių, kurie juos turi, pvz., žodžiui *nebeatsinešdavau* bus nurodyta, kad jis yra sangrąžinis. Pateikiant duomenis apie žodį bus nurodomi tik tie gramatiniai požymiai, kurie jam būdingi, o apie kitus, kurių analizuojamas žodis neturi, net neužsimenama. <...> Taigi toks morfologinių duomenų pateikimas, koks yra VDU sukurtoje *Lietuvių kalbos sintaksinės ir semantinės analizės informacinėje sistemoje*, plačiajam vartotojų ratui nėra optimalus. Todėl gramatikos informacinėje sistemoje siūlomas labiau vartotojui priimtinas morfologinių ir morfeminių duomenų apie žodį atvaizdavimo būdas.“

Dažniausiai anotavimo standartuose nurodomi visi tam tikrai kalbos daliai būdingi požymiai (plg. MULTEXT-East: Multilingual Text Tools and Corpora for Central and Eastern European Languages, žr. http://nl.ijs.si/ME/). Jie gali būti neišreikšti, tada žymimi kaip nuliniai arba kaip nežymėtasis tos gramatinės kategorijos narys (pvz., sangrąžos kategorijos nežymėtasis narys yra nesangrąžinė forma). Jei D. Šveikauskienė planuoja tai pačiai kalbos daliai pateikti skirtingą gramatinių kategorijų skaičių, tokią informaciją bus sunku apdoroti automatinės analizės programoms.

Galima diskutuoti dėl vaizdinio gramatinių kategorijų pateikimo. Norint rasti geriausią sprendimą, reikia apklausti statistiškai patikimą tam tikrų programų vartotojų skaičių, kad būtų galima daryti išvadas, koks formatas jiems priimtinesnis. Dabar išvados apie morfologinių duomenų pateikimo priimtinumą padarytos remiantis asmenine straipsnio autorės nuomone.

Taip pat norime atkreipti dėmesį į D. Šveikauskienės straipsnio kalbos kokybę. Daugelyje vietų straipsnis parašytas ne moksliniu stiliumi. Palikta kalbos, korektūros klaidų (jos ištaisytos skliaustuose, stiliaus klaidos netaisytos), pvz.:

*Tada nebelieka priemonių (=,) kaip jas atskirti nuo priesagų ar galūnės (pvz., kaipmat, tąsyk ir kt. – antra šaknis užimtų galūnės poziciją).* (p. 1)

*Kita funkcija „Lemuoti“ pateikia visus galimus daugiareikšmių žodžių variantus, bet tuo pačiu (=kartu) ir lietuvių kalboje neegzistuojančius žodžius, tokius kaip, pvz., *blizgėjas (16 pav.).* (p. 7)

*Nei tas lietuvis, kuris sako, t. y. ištaria garsų rinkinį „*blizgėjas“ (=*blizgėjas), nei tas, kuris jį girdi, nežino, ką tai reiškia.* (p. 7)

*Rusų kalbos morfologinė analizė pateikia visas nagrinėjamo žodžio formas (7 pav.). Kuriant lietuvių kalbos morfologinę duomenų bazę nuspręsta nepateikti ekrane visų formų, nes tik retais atvejais jos visos gali būti reikalingos, be to (=,) tai užima daug vietos. Šis klausimas sprendžiamas kiek kitu būdu: šalia pradinės formos talpinamas (=pateikiamas) mygtukas „VISOS FORMOS“, kuriuo atidaromas langas(=,) turintis visų nagrinėjamo žodžio formų sąrašą. Žodžio apyrankė visų formų lango pavyzdys pateiktas 24 paveikslėlyje.* (p. 15)

*Viešai prieinami šaltiniai, pvz., Vytauto Didžiojo universitete sudarytas morfemikos žodynas bei jo pagrindu sukurta morfemikos duomenų bazė(=,) nepateikia išsamios informacijos apie morfemos tipą, todėl gali būti naudingi tik kalbininkams, turintiems specialiąsias žinias žodžių skaidymo į morfemas srityje (=turintiems specialių morfeminės analizės žinių).* (p. 18)

Šiuo straipsniu norime atkreipti dėmesį į neobjektyvią, metodologiškai nepatikimą informaciją, klaidinančią skaitytojus, kalbos analizės programų ir duomenų bazių naudotojus.

**Literatūra**

Dabašinskienė I. 2009: Šnekamosios lietuvių kalbos morfologinės ypatybės. Acta *Linguistica Lithuanica* LX, 1–15.

Grumadienė L., Žilinskienė V. 1998: *Dažninis dabartinės rašomosios lietuvių kalbos žodynas.* Vilnius: Lietuvių kalbos institutas.

Kazlauskienė A. 2010: Lietuvių kalbos žodžių foneminės struktūros dėsningumai. *Žmogus ir žodis: mokslo darbai. Didaktinė lingvistika* 12 (1), 35–41.

Kazlauskienė A. 2012: Priebalsiai veiksmažodžio morfemų sandūroje. *Lietuvių kalba* 6.

Kazlauskienė A., Raškinis G. 2013: Lietuvių kalbos veiksmažodžio morfemų struktūra. *Respectus philologicus* 23 (28), 198–210.

Paroubek P., Chaudiron S., Hirschman L. 2007: Principles of Evaluation in Natural Language Processing. Traitement Automatique des Langues. *ATALA* 48 (1), 7–31. Prieiga internete https://halshs.archives-ouvertes.fr/hal-00502700/document.

Rimkutė E., Daudaravičius V. 2007: Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas. *Kalbų studijos* 11, 30–35.

Rimkutė E., Kazlauskienė A., Raškinis G. 2011: Lietuvių kalbos veiksmažodžių morfeminė struktūra. *Acta Linguistica Lithuanica* 64–65, 87–105.

Zinkevičius V. 2000: Lemuoklis – morfologinei analizei. *Darbai ir Dienos* 24, 245–273.

Žilinskienė V. 2005: Vardažodžių, įvardžių ir jų gramatinių formų vartojimas lietuvių kalbos stiliuose. *Lituanistca* 64 (4), 28–44.