

Lemuoklis – morfologinei analizei

ĮVADAS

Kompiuterizacijai skverbiantis į pačias įvairiausias gyvenimo sferas, per keletą pastarųjų dešimtmečių atsirado naujos mokslo tiriamųjų darbų kryptys, susijusios su kompiuterinių technologijų kūrimu ir taikymu specifinėse srityse. Viena tokių krypčių yra natūralios kalbos reiškinių apdorojimas kompiuteriais (*natural language processing, NLP*). Kalbant apie kalbos apdorojimą kompiuteriais, turimos omenyje automatiško rašytinės kalbos atpažinimo (analizės) bei generavimo (sintezės) technologijos, skirtos kompiuterinio vertimo, bendravimo su kompiuteriais natūralia kalba, kompiuterinio kalbos mokymo ir kitiems panašioms tikslams. Šios kalbinės technologijos (*language technologies*) panaudojamos ir pačiai kalbai tyrinėti kiekybiniu bei struktūriniu aspektais.

Straipsnyje supažindinsime su kompiuterine programa, automatiškai apibūdinančia lietuviškas rašytines žodžių formas gramatiniu (morfologiniu) aspektu ir nustatančia žodžiams antraštinius (žodyninius) pavidalus. Procesui, kurio metu kompiuteris nustato teksto žodžiams antraštinius pavidalus, pavadinti angliškoje literatūroje prigijo terminas *lemmatizing* (nuo žodžio *lemma*, dgsk. *lemmata* – antraštinė, žodyninė leksemos forma). Straipsnyje šis procesas vadinamas *lemavimu*, iš čia ir *lemavimą* atliekančios programos pavadinamas *Lemuoklis*. Programos, sužyminčios žodžių formų gramatinės reikšmės (gramatiniai anotatoriai), dažnai pavadinamos *taggers* (nuo angl. *tag*, reiškiančio etiketę, žymę). Internete galima rasti gana daug informacijos apie įvairių kalbų „*lemuoklius*“ bei gramatinius anotatorius (*tagerius*), jų paskirtį, veikimo principus, naudojimą (žr., pvz., nuorodas į AUTASYS, EUSLEM, SphinxSurvey).

Patikslinsime straipsnyje vartojamų terminų reikšmes. *Žodžių formomis* arba tiesiog *formomis* čia vadinami žodžių (leksemų) kaitybinių gramatinių formų rašytiniai (ortografiniai) pavidalai. *Žodžių antraštiniais pavidalais* vadinamos ortografinės išraiškos tokių jų gramatinių formų, kokiomis leksemos įtraukiamos į žodynus. Vardazodžiams tai paprastai vienaskaitos vardininko, veiksmazodžiams – bendraties forma. *Nustatyti* žodžiui jo *lemą* straipsnyje reiškia apibūdinti žodį antraštiniu pavidalu ir kalbos dalimi. Žodžio formos *gramatiniu apibūdinimu* vadinamas jos gramatinės reikšmės nusakymas turinčiomis jai prasmę gramatinėmis kategorijomis bei požymiais. *Lemuoti* žodžio formą – tai nustatyti jai visas galimas (hipotetines) lemas ir visus galimus gramatinius apibūdinimus.

KAS YRA LEMUOKLIS

Lemuoklis yra kompiuterinė programa, automatiškai lemuojanti lietuviškas žodžių formas iš pradinių tekstinių failų ir rezultatinis tekstinius failus. Programa sukurta 2000 metais ir skirta Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro (KLC) tekstyno (Marcinkevičienė, 1997) moksliniams lingvistiniams tyrinėjimams automatizuoti. Lemuoklis gali veikti IBM tipo personaliniuose kompiuteriuose, kuriuose įdiegta Microsoft Windows NT, Windows 95 ar aukštesnės versijos 32 bitų operacinė sistema. Straipsnio autorius atliko visus programavimo darbus bei sukūrė kompiuterinę žinių apie lietuvių kalbos leksiką ir gramatiką bazę, lingvistinės informacijos paieškos šioje bazėje ir jos panaudojimo automatiško lemuoimo procesui metodiką bei programinę įrangą. Šiame darbe taip pat dalyvavo KLC darbuotojai, išbandydami tarpinius Lemuoklio varian-

tus, analizuodami sulemuotus tekstus, registruodami klaidingo lemavimo atvejus bei dėsningumus, teikdami pasiūlymus dėl galimybių mažinti žodžių formų lemavimo rezultatų perteklinį daugiaprasmiškumą.

Lemuoklyje panaudoti visi šiuo metu autoriui žinomi ir prieinami kompiuteriniai lietuvių kalbos resursai (apie tai plačiau straipsnio skyrelyje „Lietuvių kalbos leksikos ir gramatikos duomenų bazė“). Taip pat panaudota ir ankstesnių programinių sistemų, automatiškai manipuliuojančių lietuvių kalbos gramatikos reiškiniais, kūrimo patirtis. Tarp tokių sistemų-prototipų galima paminėti autoriaus sukurtą programą MAN (Morfologinė Analizė ir Normalizacija; programa naudota „Dažniam dabartinės rašomosios lietuvių kalbos žodynui“ parengti (DDRLKŽ, 1997:X, 1998: XII)) bei į lietuviškąją programos Microsoft Office versiją įdiegti programinės įrangos ir informacinės bazės komponentus, skirtus automatiškai lietuviškų žodžių formų rašybos teisingumo kontrolei.

KAIP VARTOTOJAI DIRBA SU LEMUOKLIU

Lemuoklis lemuoja tekstinių failų žodžių formas, įrašydamas lemavimo rezultatus į tekstinius rezultatų failus. Vienos užduoties metu Lemuoklis sulemuoja vieną jam nurodytą pradinį failą į vieną rezultatų failą; po to vartoto-

jas gali užduoti lemuoti kitą pradinį tekstą. Lemuojamų tekstų apimtys neribojamos.

Ką ir kaip reikia sulemuoti, kur kompiuteryje turi būti įrašyti lemavimo rezultatai, vartotojas nurodo, naudodamas programos lango mygtukus ir laukelius. 1 paveiksle pateiktas Lemuoklio pagrindinio lango kompiuterio ekrane vaizdas.

Nuspaudus klavišą „Susirandam failą, kuri reikia sulemuoti“, atsiveria failų atidarymo dialogo langelis, ir su jo pagalba vartotojas, naršydamas po kompiuterį, suranda ir nurodo Lemuokliui lemuotiną failą. Panašiai, paspaudęs klavišą „Susirandam vietą, į kur įrašysim lemavimo rezultatus“, vartotojas nurodo Lemuokliui lemavimo rezultatų failo vardą ir jo įrašymo vietą.

Parengiant tekstus šiuolaikinėmis duomenų apdorojimo programomis, pastarosios gali specialiu formatu įrašyti į kompiuterinius tekstų failus specifinius duomenis apie teksto pastraipų formatus, apie naudojamų rašmenų fontus bei kitą tarnybinę informaciją. Lemuoklis šiuo metu nemoka tokios informacijos atpažinti ir ją pasinaudoti ir netikrina, ar į nurodomus lemuoti pradiniai failai yra įrašytas vien tik tekstas, ar ir minėti specifiniai duomenys. Todėl lemuotinuose failuose turėtų būti įrašyta vien tik simbolinė tekstų informacija. Be to, lemuojamas tekstas turėtų būti be žodžių perkėlimų į kitą eilutę, nes kol kas Lemuoklis nemoka prijungti perkeltosios dalies prie žodžio.



1 paveikslas.
Pagrindinis
Lemuoklio langas

Naudodamas rėmelyje „Nurodom, kaip lemuojam“ išdėstytus mygtukus ir laukelius, vartotojas nurodo lemuojamų tekstų bei lemavimo rezultatų ypatumus. Viename lemuojamame failе visos lietuviškos raidės turi būti užkoduotos pagal vieną kodavimo sistemą. Kol kas Lemuoklis tekstų failuose atpažįsta tik nekirčiuotas raides. Operacinėje sistemoje MS DOS paruoštuose failuose lietuviškų raidžių kodai turi atitikti KBL kodavimo lentelę, o sistemoje MS Windows paruoštuose failuose – standartinius Windows lietuviškų raidžių kodus, vieną teksto simbolių koduojant vienu baitu. Pasirinkties mygtukais „Windows Baltic RIM“ ir „DOS KBL“ vartotojas nurodo lietuviškų raidžių kodavimą lemuojamame ir rezultatų failuose. Rėmelyje „Hipotetinės žodžių formų gramatinės reikšmės“ esančiais pasirinkties mygtukais vartotojas gali pasirinkti, ar į lemavimo rezultatus turi būti įrašomos tik žodžių lemos (antraštiniai žodžių formų pavidalai ir jų kalbos dalis), ar dar ir tekste pavartotų žodžių formų gramatiniai apibūdinimai, reiškiami gramatinėmis kategorijomis bei požymiais. Pažymėjus laukelį rėmelyje „Skyrybos ženklai“ varnele, Lemuoklis, radęs prie teksto žodžio skyrybos ženklą, prirašys jį prie to žodžio ir lemavimo rezultatų failе; jei varnelės nebus, į rezultatų failą bus įrašomi žodžiai be skyrybos ženklų.

Klavišas „Nurodom parenkamus lemavimo parametrus“ skirtas patikslinti, kaip turi būti lemuojamos tam tikros specifinės žodžių formų kategorijos. Dauguma šių parenkamų parametrų (opcijų) susijusios su lemavimo rezultatų daugiaprasmiškumo ribojimu: jais nurodoma, ar tam tikrais specifiniais lemavimo atvejais Lemuokliui leidžiama naudoti jo turimas lemavimo daugiaprasmiškumą mažinančias priemones, ar ne. Šiuo metu šis klavišas dar neveikia, o minėtais specifiniais atvejais Lemuoklis lemuoja, naudodamas tokias parenkamų lemavimo parametrų reikšmes, kokios yra priskirtos programiškai „pagal nutylėjimą“ (*by default*). Straipsnio skyrelyje „Parenkami (opcioniai) lemavimo parametrai“ plačiau aprašoma, kaip šie parametrai veikia lemavimo rezultatų formavimą ir kokios jų reikšmės „pagal nutylėjimą“.

Programos naudotojas paleidžia lemavimo užduotį, nuspausdamas ekrane lango klavišą „LEMUOJAM“. Kol Lemuoklis lemuoja, lango apačioje slenkanti juostelė rodo, kiek dar truks procesas.

Lemuoklis kol kas dar neturi jokių priemonių automatiškai teksto sintaksinei ar semantinei analizei atlikti. Kiekvieną žodžio formą jis nagrinėja atskirai, atsietą nuo konteksto. Todėl lemuodamas teksto žodį, jis išrenka visas jo gramatinės traktuotes, kokias tik suranda savo kompiuterinėse žiniose apie lietuvių kalbos leksiką ir gramatiką, dažnai žodžio formai pateikdamas ne vieną, o kelias hipotetines lemas ir/ar kelis hipotetinius tos formos gramatinius apibūdinimus. Nevienareikšmio lemavimo atvejus Lemuoklis rezultatų failuose pažymi specialiais ženklais. Vartotojas, analizuodamas lemavimo rezultatų tekstą, turi pats atmesti klaidingas lemavimo hipotezes ir palikti teisingąsias. Žodžių formų lemavimo nevienareikšmiškumas šiuo metu yra didžiausias Lemuoklio trūkumas. Kol kas visiškai panaikinti automatiško lemavimo daugiaprasmiškumą, neturint automatiškos sintaksinių ryšių tarp teksto žodžių analizės priemonių, neįmanoma. Vis dėlto Lemuoklyje įdiegti tam tikri metodai, leidžiantys kai kuriais atvejais sumažinti perteklinį lemavimo rezultato daugiaprasmiškumą. Šie metodai aprašomi straipsnio skyrelyje „Lemavimo daugiaprasmiškumo mažinimas“. Lemuoklio pateikiama lemavimo rezultatuose leksinė ir gramatinė informacija aprašyta skyrelyje „Lemavimo rezultatų turinys ir pateikimo forma“, naudojama rezultatuose gramatinės informacijos žymėjimų sistema – skyrelyje „Gramatinės informacijos žymėjimas“.

KAIP LEMUOKLIS LEMUOJA IR KAIP PATEIKIA LEMAVIMO REZULTATUS

VIENO FAILO LEMAVIMO EIGA

Nurodytą lemuoti failą Lemuoklis apdoroja dviem etapais. **Pirmojo etapo** metu Lemuoklis skaito lemuojamą failą ir pasižymi jame simbolių sekas – virtinėles, galinčias būti žodžių formomis, t. y. sudarytasias vien iš raidžių ir neviršijančias tam tikro ilgio. Nors lotyniškų raidžių *q*, *x* ir *w* nėra lietuvių kalbos abėcėlėje, virtinėles su tokiomis raidėmis Lemuoklis traktuoja kaip galinčias būti žodžių

formomis, nes jo kalbinių duomenų bazėje yra informacijos apie santrumpas, akronimus, vardus, kurių rašyboje galimos ir tokios raidės. Lemuoklis taip pat pasižymi, kurios raidžių virtinėlių raidės didžiosios, jei tokių buvo, ir sekusius betarpiškai po raidžių virtinėlių skyrybos ženklus, jei tokių buvo. Šiuo metu kaip skyrybos ženklus Lemuoklis atpažįsta šauktuka, kablelį, tašką (bet ne daugtaškį), dvitaškį, kabliataškį, klaustuką ir brūkšnelį. Žodžių kėlimo į kitą eilutę brūkšnelį Lemuoklis traktuoja kaip paprastą skyrybos ženklą, t.y. nesujungia perkeltųjų dalių į vieną žodį, ir todėl tokias dalis sulemuoja atskirai kaip savarakiškus teksto žodžius. Neretai tekstuose pasitaikančius romėniškus skaičius, formaliai žiūrint, irgi sudaro vien tik raidės. Lemavimo proceso valdyme numatytas specialus parametras, nurodantis nelemuoti simbolių virtinėlių – kandidačių į romėniškus skaičius (apie šį bei kitus lemavimo rezultatų valdymo parametrus plačiau parašyta skyrelyje „Parenkamieji (opcioniai) lemavimo parametrai“).

Antrojo etapo metu Lemuoklis ima iš eilės po vieną ir lemuoja atrinktas pirmajame etape raidžių sekas – kandidates į žodžių formas. Vienos žodžio formos lemavimo procesą sudaro: 1) raidžių seką atitinkančios gramatinės informacijos paieška kalbinių duomenų bazėje, 2) šios paieškos rezultatų analizė ir hipotetinių lemų formavimas, 3) dėl lemavimo atsiradusio perteklinio daugiaprasmiškumo lemavimo rezultate analizė ir sumažinimas, 4) žodžio formos lemavimo rezultato įrašymas į rezultatų failą. 1 – 3 etapų aprašymas, t. y. kaip Lemuoklis suformuoja vienos žodžio formos lemavimo rezultatą, pateikiamas skyreliuose „Informacijos paieška kalbinių duomenų bazėje ir hipotetinių lemų formavimas“ bei „Lemavimo daugiaprasmiškumo mažinimas“. Toliau apibūdinsim, kas sudaro vienos žodžio formos lemavimo rezultatą.

LEMAVIMO REZULTATŲ TURINYS IR PATEIKIMO FORMA

Įrašydamas į rezultatų failą informaciją apie kiekvienos pirminio teksto žodžio formos lemavimą, Lemuoklis pirmiausia iš naujos eilutės įrašo šį žodį tokiu pavidalu, koku jis buvo užrašytas pirminiame tekste. Jei pirmi-

niame tekste betarpiškai po žodžio ėjo skyrybos ženklas, o laukelis „Skyrybos ženklai“ buvo pažymėtas varnele (žr. 1 pav.), Lemuoklis prirašo po žodžio ir tą skyrybos ženklą. Jei Lemuoklis savo kalbinių duomenų bazėje nerado raidžių sekai – kandidatėi į žodžio formą jokios informacijos (jos neatpažino), tai iš naujos eilutės po dviejų tarpelių įrašoma simbolių seka „<?>“, ir tuo lemavimo šiai raidžių sekai rezultato įrašymas baigiamas.

Leksinę bei gramatinę informaciją apie atpažintą žodžio formą Lemuoklis pateikia šia tvarka. Pirmiausia žiūrima, kiek skirtingų hipotetinių lemų buvo nustatyta lemuotajai žodžio formai. Lemuoklis laiko, kad dvi hipotetinės vienai formai nustatytosios lemos yra skirtingos, jei skiriasi tarpusavyje bent vienas iš dviejų lemą nusakančių komponentų, būtent kalbos dalis ar antraštinis žodžio pavidalas.

Jei lemuotajai žodžio formai Lemuoklis vienareikšmiškai nustatė vieną lemą, tai iš naujos eilutės po dviejų tarpelių įrašoma lemos kalbos dalis, po to, po vieno tarpelio, tarp kampukų „< >“, – lemos ortografinė išraiška – atstatytasis žodžio formai antraštinis pavidalas. Jei Lemuoklis nustatė, kad lemos kalbos dalis yra tikrinis daiktavardis, tai antraštinis pavidalas tarp kampukų visada įrašomas didžiąja raide (nesvarbu, ar žodžio forma tekste buvo parašyta didžiąja, ar mažąja raide), kitais atvejais antraštinis pavidalas visada įrašomas mažąja raide. Jei Lemuokliui buvo nurodyta nustatyti ne tik lemų žodžių formų lemas, bet dar ir tų žodžių formų gramatinius apibūdinimus (t. y. jei rėmelyje „Hipotetinės žodžių formų gramatinės reikšmės“ buvo nuspaustas pasirinkties mygtukas „Nustatom ir įrašom į rezultatų failą“, žr. pav. 1), tai Lemuoklis išveda visus hipotetinius lemuotosios žodžio formos gramatinius apibūdinimus, kiekvieną hipotetinį apibūdinimą – variantą rašydamas iš naujos eilutės po keturių tarpelių. Vieną gramatinį žodžio formos apibūdinimą sudaro gramatinių kategorijų, įskaitant ir kategoriją „kalbos dalis“, bei požymių, turinčių prasmę šiai žodžio formai, reikšmių seka.

Jei lemuotajai žodžio formai Lemuoklis nustatė ne vieną, o kelias hipotetines lemas, tai leksinę bei gramatinę informaciją įrašoma iš eilės apie kiekvieną iš hipotetinių lemų. Ši

informacija kiekvienai hipotetinei lemai pateikiama tokia pat tvarka, kaip ir tuo atveju, kai Lemuoklis nustato žodžio formai viena-reikšmiškai tik vieną lemą. Skirtumas tik toks, kad, prieš pradėdant rašyti informaciją apie bet kurią hipotetinę lemą, įrašomos dvi žvaigždutės „**“.

Veiksmažodžio bendraties, dalyvio, padalyvio ir pusdalyvio formų gramatiniuose apibūdinimuose Lemuoklis vietoj kalbos dalies nurodo šių formų pavadinimus: „bendraties“, „dalyvis“, „padalyvis“ ar „pusdalyvis“, o ne „veiksmažodis“. Kalbos dalis „veiksmažodis“ nurodoma tik asmenuojamų veiksmažodžių formų gramatiniuose apibūdinimuose. Veiksmažodžių lemu kalbos dalimi nurodomas ne „veiksmažodis“, bet „bendraties“.

Antraštinius žodžių pavidalus Lemuoklis atstato laikydamasis šių principų. Linksniuojamiems vardažodžiams suteikiamas viena-skaitos vardininko linksnis; kaitomiems gimine – vyriškoji giminė; jei vardažodis gali būti tiek įvardžiuotinė, tiek ir neįvardžiuotinė formos, pateikiama neįvardžiuotinė; jei laipsniuojamas – nelyginamasis laipsnis ir pan. Kitoks antraštinių pavidalas pateikiamas tik nekaitomiems žodžiams arba tais atvejais, jei kitaip žodis nevertojamas. Įrašydamas lemu – veiksmažodžių antraštinį pavidalą, Lemuoklis visada įrašo ne tik bendratį, bet ir kitas dvi pagrindines formas – esamojo bei būtojo kartinio laiko trečiuosius asmenis. Pastarosios formos įrašomos ne išsamiai, o tik jų pabaigos po brūkšnelio nuo tos vietos, kur jos nebesutampa su bendraties forma. Lemuoklis stengiasi atstatyti morfologiškai artimiausią antraštinę formą: išlaiko buvusius žodžio formoje prefiksus (dalelytes *ne (be)-*, *te (be)-*, *be-*, priešdėlius), sangražiškumą.

Antraštinių pavidalų atstatymui iš žodžių formų Lemuoklis naudoja specialius skaitmeninius lietuvių kalbos gramatinės kaitybės modelius. Antraštinių pavidalų generavimas šių modelių pagalba kai kurių žodžių kategorijoms turi tam tikrų ypatumų. Apie šiuos skaitmeninių kaitybės modelių ypatumus, kitas savybes bei galimybių ribas žr. skyrelyje „Lietuvių kalbos leksikos ir gramatikos duomenų bazė“.

Jei Lemuoklis atpažįsta raidžių sekoje žodžio formą, bet visiškai atstatyti antraštinį pavidalą ir/ar apibūdinti ją gramatiškai jam

neužtenka kalbinių duomenų bazėje turimų žinių, tada suformuojamas ir įrašomas nepilnas žodžio formos lemavimo rezultatas. Plačiau apie tokias situacijas ir kaip tokiais atvejais formuojamas lemavimo rezultatas, aprašyta skyrelyje „Informacijos paieška kalbinių duomenų bazėje ir hipotetinių lemu formavimas“ (5-oji, 6-oji, 7-oji ir 8-oji situacijos).

Žemiau pateikiamas šešių žodžių formų dvejojimo lemavimo pavyzdys. Pirmuoju atveju Lemuokliui nurodyta nustatyti tik lemuojamų žodžių formų lemas, antruoju atveju – ir formų gramatinius apibūdinimus. Visi čia ir toliau straipsnyje pateikiami lemavimo pavyzdžiai yra perrašyti iš leidinio E. Gudavičius. *Lietuvos istorija. Vilnius, 1999* lemavimo rezultatų failų.

Žodžių formų *valstybės*, *jos*, *odaliniių*, *karinės*, *vad*, *save* lemavimo pavyzdys:

1) kai Lemuokliui nurodyta nustatyti tik žodžių formų lemas

```
valstybės
dktv <valstybė>
jōs
** įvrd <jis>
** bndr <joti (-ja, -jo)>
odaliniių
<??>
karinės
** bdvr <karinis>
** bndr <karinėti (-ja, -jo)>
vad
sntmp <vad.>
save
įvrd <savęs>
```

2) kai Lemuokliui liepta nurodyti ne tik žodžių formų lemas, bet ir gramatinius apibūdinimus toms formoms

```
valstybės
dktv <valstybė>
dktv mot.gim vnsk K
dktv mot.gim dgsk V
dktv mot.gim dgsk Š
jōs
** įvrd <jis>
įvrd neįvardž mot.gim vnsk K
įvrd neįvardž mot.gim dgsk V
** bndr <joti (-ja, -jo)>
vksm nesngr tiesiog.nuos būs.l IIIasm
odaliniių
<??>
karinės
** bdvr <karinis>
bdvr nelygin.l neįvardž mot.gim vnsk K
bdvr nelygin.l neįvardž mot.gim dgsk V
** bndr <karinėti (-ja, -jo)>
```

vksm nesngr tiesiog.nuos būs.l IIIasm
 vad
 sntmp <vad.>
 sntmp
 save
 įvrd <savęs>
 įvrd G

Dabar plačiau aprašysim žymėjimų sistemą, Lemuoklio naudojamą žodžių ir jų formų gramatinėms reikšmėms nusakyti.

GRAMATINĖS INFORMACIJOS ŽYMĖJIMAS

Žodžio kalbos dalį Lemuoklis nusako kalbos dalių bei veiksmožodžio formų pavadinimų santrumpomis: daiktavardis - *dktv*, tikrinis daiktavardis - *tikr dktv*, būdvardis - *bdvr*, skaitvardis - *sktv*, įvardis - *įvrd*, veiksmožodis - *vksm*, bendratis - *bndr*, dalyvis - *dlv*, padalyvis - *padlv*, pusdalyvis - *psdlv*, būdinys - *būdñ*,rieveksmis - *prvks*, prielinksnis - *prln*, jungtukas - *jngt*, dalelytė - *dll*, jaustukas - *jstk*, ištiktukas - *ištck*.

Žodžių formų gramatinės reikšmės Lemuoklis apibūdina šiomis gramatinių kategorijų ir požymių pavadinimų santrumpomis: eigos veikslas - *eigos vksl*, įvykio veikslas - *įvykio vksl*, sangražinė forma - *sngr*, nesangražinė forma - *nesngr*, veikiamoji rūšis - *veik.r*, neveikiamoji rūšis - *neveik.r*, dalyvių reikiamybės rūšis - *reikiamyb.r*, tiesioginė nuosaka - *tiesiog.nuos*, liepiamoji nuosaka - *liep.nuos*, tariamoji nuosaka - *tariam.nuos*, esamasis laikas - *esam.l*, būtasis kartinis laikas - *būt.kart.l*, būtasis dažninis laikas - *būt.d.l*, būsimasis laikas - *būs.l*, kiekiniai (skaitvardžių skyrius) - *kiekin*, dauginiai (skaitvardžių skyrius) - *daugin*, kuopiniai (skaitvardžių skyrius) - *kuopin*, kelintiniai (skaitvardžių skyrius) - *kelintin*, nelyginamasis laipsnis - *nelygin.l*, aukštesnysis laipsnis - *aukštesn.l*, aukštėlesnysis laipsnis - *aukšč.l*, neįvardžiutinė forma - *neįvardž*, įvardžiutinė forma - *įvardž*, vyriškoji giminė - *vyr.gim*, moteriškoji giminė - *mot.gim*, bendroji giminė - *bendr.gim*, bevardė giminė - *bevrd.gim*, vienaskaita - *vnsk*, daugiskaita - *dgsk*, dviskaita - *dvisk*, vardininkas - *V*, kilmininkas - *K*, naudininkas - *N*, galininkas - *G*, įnagininkas - *Įn*, vietininkas - *Vt*,

šauksmininkas - *Š*, pirmasis asmuo - *Iasm*, antrasis asmuo - *IIasm*, trečiasis asmuo - *IIIasm*. Atpažinta kaip santrumpa ar akronimą raidžių seka Lemuoklis apibūdina žyme *sntmp*.

Žodžių formų atpažinimui ir jų gramatiniam apibūdinimui Lemuoklis naudoja skaitmeninius lietuvių kalbos gramatinės kaitybos modelius, plačiau aprašomus skyrelyje „Lietuvių kalbos leksikos ir gramatikos duomenų bazė“. Šie modeliai operuoja ne konkrečiais gramatinių kategorijų ir požymių reikšmių pavadinimais, o jų eilės numeriais. Tai leidžia ateityje suteikti galimybę Lemuoklio naudotojui pačiam laisvai pasirinkti jam priimtinius gramatinius žymėjimus.

Informacijos apie sulemuotus žodžius pateikimą ateityje galima būtų tobulinti ir kitomis kryptimis. Pavyzdžiui, galima numatyti lemavimo rezultatų įrašymą ne tik paprasto teksto pavidalu, bet ir vadinamuoju HTML (*HyperText Markup Language*) formatu. Šiuo atveju sulemuotas tekstas kompiuterio ekrane nesiskirtų nuo pirminio teksto, tačiau būtų prikimštas tik kompiuteriui matomų lemavimo rezultatų, automatiškai pasirodančių, užlipus ant teksto žodžio ir/ar spragtelėjus kompiuterio pelės klavišu. HTML formatas nepriklauso nuo techninės ir programinės kompiuterių konfigūracijos, jį supranta daugelis šiuolaikinių tekstais manipuliuojančių kompiuterinių programų.

Tekstų mokslinio tyrinėjimo praktikoje plačiai paplitusi dar viena tekstų kodavimo priemonė - SGML (*Standard Generalized Markup Language*). SGML yra formalus aparatas įvairiausio pobūdžio žymių aprašymui. Programos, mokačios SGML kalbą, pagal tokius joms pateikiamus aprašus žino, kaip interpretuoti įterptąsias į tekstą žymes. SGML priemonėmis aprašius lietuviškų gramatinių žymėjimų sistemą (*tagset*), sulemuotą į SGML formata, lietuvišką tekstą toliau būtų galima tyrinėti įvairiomis specializuotomis lingvistinės analizės programomis, kurioms jau nebesvarbu, kokia kalba parašyti pirminiai tiriamaieji tekstai. Apie kompiuterinių technologijų naudojimą lingvistiniuose tekstų tyrimuose, šių tyrimų tikslus bei galimybes rašyta (Marcinkevičienė, 1997, 2000).

LIETUVIŲ KALBOS LEKSIKOS
IR GRAMATIKOS DUOMENŲ BAZĖ

Lemuoklio žinios apie lietuvių kalbą įrašytos kompiuterinėje leksikos ir gramatikos duomenų bazėje. Kalbinių duomenų bazę sudarantys komponentai pagaminti transformuojant į skaitmenines struktūras iš įvairių šaltinių surinktą kalbą aprašančią medžiagą. Lemuodamas Lemuoklis naršo duomenų bazę specialiomis programinėmis informacijos paieškos ir išrinkimo procedūromis. Dabartinėje Lemuoklio versijoje vartotojas negali pasižiūrėti į lietuvių kalbos leksikos ir gramatikos žinias nei papildyti jas ar kaip kitaip keisti.

Kalbinių duomenų bazę sudaro šeši kompiuteriniai žodynai: *GF*, *G*, *TG*, *F*, *T* ir *S*. Svarbiausias Lemuoklio žodynas yra *GF*. Jo pagalba Lemuoklis gali atpažinti lietuviškų žodžių kaitybines formas ir apibūdinti jas gramatiškai. Tam naudojami į *GF* žodyno sudėti įeinantys skaitmeniniai žodžių gramatinės kaitybės modeliai, straipsnyje vadinami *kompiuterine morfologija*. Tačiau *GF* žodyno pagalba atpažįstamos ir gramatiškai apibūdinamos tik bendrinės lietuvių kalbos žodžių kaitybinės formos. Formų su nutrumpėjusiomis galūnėmis, taip pat pasenusių ar tarmiškų kaitybinių formų *GF* žodynas neatpažįsta. Taip yra todėl, kad *GF* kompiuterinė morfologija kol kas atspindi tik bendrinės kalbos dėsnius. Kitas *GF* gramatinės kaitybės modelių trūkumas – juos sudarant neatkreiptas dėmesys į tikrinius daiktavardžius kaip į atskirą svarbų daiktavardžių poskyrį. *GF* žodyno leksinėje dalyje yra iš įvairių šaltinių surinktų tikrinių daiktavardžių – asmenvardžių, geografinių vardų ir kt. Jų formos kaitybės modelių pagalba atpažįstamos ir apibūdinamos gramatiškai. Tačiau kalbos dalimi, jas atpažįstant, nurodomas „daiktavardis“, t. y. kompiuterio atpažintos tikrinės formos niekaip neatskiriamos nuo analogiško morfologinio tipo bendrinių daiktavardžių formų.

GF žodyno nesugebėjimą atpažinti formų nutrumpėjusiomis galūnėmis Lemuoklis iš dalies kompensuoja žodynais *F* ir *G*. Į *F* žodyną surašytos lietuviškuose tekstuose užfiksuotos netikrinių žodžių formos. Taigi *F* žodyne yra ir pasenusių, ir tarmiškų formų, ir formų

nutrumpėjusiomis galūnėmis. *G* žodyne surašytos netikrinių žodžių šaknys, o žodžių formas, esant reikalui, generuoja *G* kompiuterinė morfologija, jungdama prie kiekvienos šaknies jai priklausančius pagal žodžio morfologinį tipą afiksus. *G* žodyno morfologija generuoja ne tik pilnas žodžių formas, bet ir nutrumpėjusias ar pasenusias formas (pvz., *iliatyvus*). Tačiau *G* morfologija manipuliuoja tik afiksais, bet ne gramatinėmis reikšmėmis. Taigi *G* žodyno pagalba galima atpažinti raidžių sekose „*legalias*“ lietuviškų žodžių formas, bet informacijos apie tų formų gramatinės reikšmės šiame žodyne nėra.

GF žodyno nesugebėjimą atskirti tikrinius daiktavardžius nuo bendrinių iš dalies kompensuoja žodynai *T* ir *TG*. Į *T* žodyną surašytos lietuviškuose tekstuose užfiksuotos tikrinių žodžių formos. Žodyne *TG* surašytos tikrinių žodžių šaknys. Žodžių formų generavimui *TG* žodynas, panašiai kaip *G* žodynas, naudoja tikrinių žodžių kompiuterinę morfologiją. *TG* žodyno morfologija generuoja ir kaitybines tikrines formas, ir išvestines iš tikrinių žodžių darybines, tarp kurių gali būti tiek tikrinės, tiek bendrinės formos. Tačiau *TG* morfologija, kaip ir *G* žodyno morfologija, taip pat manipuliuoja tik afiksais, taigi *TG* žodynas, kaip ir *G* žodynas, negali suteikti informacijos apie aptiktą jame formų gramatinės reikšmės.

Į *S* žodyną surašytos santrumpos bei akronimai.

Žodynai *TG*, *F*, *T* ir *S* buvo sukurti ne lemavimui. Jie perkelti į Lemuoklio kalbinių duomenų bazę iš rašybos teisingumo kontrolės funkcija atliekančių programų (*spelerių*). Taip padaryta siekiant, kad Lemuoklis atpažintų kuo daugiau lietuviškų žodžių formų. Taigi lemavimo procesui naudojama kalbinė informacija yra išmėtyta po gana skirtingos sudėties ir galimybių žodynus; negana to, neišvengta šios informacijos dubliavimosi atskiruose žodynuose. Ateityje Lemuoklio kalbinių duomenų bazę turėtų būti perdirbta, paliekant vieną kompiuterinį žodyną. Šio žodyno pagrindą sudarys *GF* žodynas, papildžius jo kompiuterinę morfologiją galūnių nutrumpėjimo reiškiniais bei tikrinių daiktavardžių kaityba ir daryba. O kol kas Lemuoklis lemuodamas semiasi kalbinių

žinių visuose žodynuose. Apie Lemuoklio naudojamą strategiją analizuojant paieškos žodynuose rezultatus parašyta skyrelyje „Informacijos paieška kalbinių duomenų bazėje ir hipotetinių lemų formavimas“.

Dabar plačiau apibūdinsime Lemuoklio žodynų turinį ir sandarą; kokia kalbinė informacija juose atspindima, šios informacijos šaltinius, jos transformavimo į skaitmeninį invariantą metodus bei kompiuterinės paieškos žodynų skaitmeninėse duomenų struktūrose procedūras.

Žodynas GF. Tai svarbiausias Lemuoklio žodynas. Jo pagalba Lemuoklis atpažįsta lietuviškų žodžių kaitybines formas ir apibūdina jas gramatiškai. Trumpai aprašysim, kaip jis tai daro.

GF žodynas sudėtas iš dviejų komponentų. Pirmasis – tai skaitmeniniai lietuvių kalbos gramatinės kaitybos modeliai – kompiuterinė morfologija. Antrojo komponento turinį sudaro lietuviškų žodžių šaknų sąrašas; prie kiekvienos šaknies sąrašė yra jos *morfologinio tipo* rodiklis. Kiekvieną morfologinį tipą kompiuterinėje morfologijoje vienareikšmiškai atitinka taisyklių rinkinys, nusakantis, kaip ir kokius afiksus galima jungti prie tokio tipo žodžio šaknies ir kokias gramatinės reikšmės įgyja taip padaromos žodžio formos. Norėdamas sužinoti kurios nors žodžio formos gramatinę reikšmę (ar kelias galimas gramatinės reikšmės formų ortografinės homonimijos atvejais), Lemuoklis bando įvairius lemuojamos raidžių vartinėles skaidymo pagal schemą *prefiksai+šaknis+postfiksai* variantus. Jei po skaidymo gautai hipotetinei šakniai GF šaknų sąrašė randamas atitinkamas, tada kompiuterinėje morfologijoje žiūrima, kokie prefiksų ir postfiksų rinkiniai galimi nurodytajam morfologiniam tipui. Jei tarp galimų afiksų derinių Lemuoklis aptinka sutampantį su hipotetinio skaidymo afiksais, žodžio forma laikoma atpažinta, ir iš kompiuterinės morfologijos išrenkama informacija apie gramatinę reikšmę, kurią formai suteikia afiksų rinkinys.

Panašų algoritma, bandydamas suvokti lietuviškų žodžių formų gramatinę reikšmę, naudoja ir lietuviškai menkai temokantis (ar ir visai nemokantis) žmogus. Skaitydamas lietuvišką tekstą, jis gali pasitelkti dviejų rūšių

pagalbines priemones: 1) žodynus, kur surašyti lietuviški žodžiai ir jų kalbos dalys ir/ar kita žodžių morfologinius tipus nusakanti informacija, ir 2) lietuvių kalbos gramatikas, kur galima pasižiūrėti, kaip įvairių tipų žodžiai kaitomi, t. y. ką įvairios to žodžio formos reiškia gramatiškai.

Tokio leksinės informacijos atskyrimo nuo gramatinės principo laikomasi ir kuriant įvairių kitų fleksinių kalbų morfologinio žodžių formų atpažinimo bei generavimo kompiuterines technologijas. Juk surašyti į vieną sąrašą visas tokiose kalbose vartojamų žodžių kaitybines formas ir dar kartu su jų gramatinėmis reikšmėmis praktiškai neįmanoma. Todėl gramatiniam manipuliavimui fleksinių kalbų žodžiais kuriami specifiniai kompiuteriniai žodynai (*lexicons*). Į tokius *leksikonus* imamos tik kalbos leksemų šaknys ar kamienai, o kaitybinės formos generuojamos naudojant prie šaknų/kamienų pateiktą morfologinio pobūdžio informaciją. Pagrindine problema, realizuojant automatišką gramatinį žodžių atpažinimą ir/ar sintezę tokiu leksikoniniu principu, tampa kalbos morfologinių reiškinių formalizavimas, t. y. radimas būdų, kaip žodžių afiksinės kaitybos ir/ar darybos taisykles transformuoti į formalų matematinių skaitmeninių modelių pavidalą.

Dažniausiai morfologiniai kalbos reiškiniai formalizuojami TLM (*Two-Level Morphology*) metodu, kuri baigtinių automatų teorijos pagrindus sukūrė Koskenniemi 9-ojo dešimtmečio pradžioje (Koskenniemi, 1983). Į šiuo metodu organizuotų leksikonų leksemų pavidalus įterpiamos nuorodos į vadinamuosius FST (*finite state transducers*) morfologinius keitiklius. TLM metodas aprašo formalų aparatą tokių keitiklių nusakymui. TLM metodologijos principu sukurti morfologiniai analizatoriai teoriškai gali būti laikomi nepriklausomais nuo konkrečios natūralios kalbos, t. y. ta pati kompiuterinė programa gali būti naudojama bet kokios kalbos žodžių formų morfologinei analizei. Reikia tik prieš ją paleidžiant įvesti atitinkamos kalbos FST keitiklių aprašą bei pridėti tos kalbos leksemų leksikoną. Būtent dėl šio teorinio universalumo natūralių kalbų atžvilgiu TLM šiuo metu yra labiausiai paplitusi morfologinių reiškinių kompiuterizavimo metodologija. Kuriant

konkrečias kalbines technologijas, paaikškėjo tam tikri metodologijos trūkumai; ji buvo tobulinama, įvairiai modifikuojama bei plečiama (Kaplan, 1988, Ritchie, 1992).

Paties morfologinių reiškinių formalizavimo aparato sukūrimas ar parinkimas tėra pusė darbo kompiuterizuojant fleksinės kalbos morfologiją. Dar reikia žodžių gramatinės kaitybės ir/ar darybos taisykles bei dėsningumus, surašytus tradicinėse kalbos gramatikose, perrašyti naudojantis to formalaus aparato priemoneis. Tai nėra greitai padaromas darbas, turint omenyje, kad turtingos fleksijų sistemos kalbų morfologijos aprašymai tradicinėse gramatikose užima šimtus puslapių. Be to, būtina sukurti atitinkamą programinę įrangą, manipuliuosiančią aparato kategorijomis.

Trumpai apibūdinsime pagrindinius žodžių gramatinės kaitybės taisyklių transformavimo ir skaitmenines duomenų struktūras principus, kuriais remiantis buvo sukurta į *GF* žodyno sudėti įeinanti kompiuterinė lietuvių kalbos morfologija. Visa žodžių kaityboje dalyvaujanti gramatinė informacija sąlygiškai buvo suskirstyta į afiksus ir gramatinės reikšmes. Visas galimas tam tikram kaitybiniam tipui priklausančio žodžio gramatinės formos *GF* kompiuterinėje morfologijoje nusako to kaitybinio tipo paradigma. Kaitybinio tipo paradigma apibrėžiama dviem parametrais. Pirmasis parametras nusako visus iš eilės surikiuotus afiksus tos paradigmos formoms sudaryti. Paradigmos afiksų sekoje turi būti tiek afiksų, kiek paradigmoje yra gramatinių formų. Antrasis parametras nusako visus iš eilės surikiuotas gramatinės reikšmes, kurias reiškia paradigmos formos. Paradigmos gramatinių reikšmių sekoje taip pat turi būti tiek gramatinių reikšmių, kiek paradigmoje yra gramatinių formų. Taip paradigmos afiksų seka nusako visus kaitybinio tipo ortografinės formas, o paradigmos gramatinių reikšmių seka – visus kaitybinio tipo formų gramatinės reikšmes. Tokio paradigmos nusakymo būtina sąlyga: tiek paradigmos afiksai ir afiksų seka, tiek ir paradigmos gramatinės reikšmės ir gramatinių reikšmių seka turi būti surikiuoti pagal vieną ir tą patį sutvarkymo principą.

Žodžių formų gramatiniam atpažinimui reikia pagal duotą ortografinę formos išraišką pasakyti formos gramatinę reikšmę. Nusa-

kus kaitybinio tipo paradigmą taip, kaip aprašyta aukščiau, to tipo formos **gramatinio atpažinimo** uždavinio sprendimas (algoritmas) formaliai atrodytų šitaip: paradigmos afiksų sekoje randame afiksą, sutampantį su duotos ortografinės formos afiksų, ir pasižiūrime, koks to afiksų numeris sekoje. Formos gramatinę reikšmę parodys gramatinė reikšmė, paradigmos gramatinių reikšmių sekoje turinti tą patį eilės numerį.

Žodžių formų gramatinei sintezei, atvirkščiai, reikia pagal duotą pageidaujamą gramatinę formos reikšmę nusakyti formos ortografinę išraišką. Toks kaitybinio tipo formos **gramatinės sintezės** uždavinio sprendimas formaliai atrodytų šitaip: paradigmos gramatinių reikšmių sekoje randame gramatinę reikšmę, sutampantią su duotąja (pageidaujama), ir pasižiūrime, koks tos gramatinės reikšmės numeris sekoje. Formos ortografinę išraišką nusakys afiksas, paradigmos afiksų sekoje turintis tą patį eilės numerį.

Tokie būtų pagrindiniai *GF* kompiuterinės morfologijos sandaros ir veikimo principai. Aukščiau aprašytas kaitybinio tipo paradigmos nusakymo būdas reikalauja, kad dviejų vienam tipui priklausančių žodžių tiek afiksų sekos, tiek gramatinių reikšmių sekos idealiai sutaptų. Kadangi lietuvių kalbos gramatikoje paradigmos (linksniuotės, asmenuotės ir kt.) šiuo požiūriu aprašomos gerokai liberaliau, teko sudaryti lietuviškų žodžių morfologinių tipų klasifikatorių, griežčiau apibrėžiantį žodžių paradigmas nei tai padaryta tradicinėje gramatikoje. *GF* morfologijos klasifikatorius nusako apie 700 skirtingų morfologinių žodžių tipų. Suprantama, kad šitaip suklasifikuoti didelius kiekius žodžių rankiniu būdu vargiai įmanoma, be to, dėl klasifikavimo požymių gausos ir įvairovės atsiranda didelė klaidingo klasifikavimo tikimybė. Tuo labiau kad kuriant *GF* žodyną iš kiekvieno klasifikuojamo antraštinio žodžio reikėjo dar išskirti šaknį, kartais nusakant ir galimus šaknies alomorfus. Todėl buvo sukurta specialioji kompiuterinė programa, padedanti suklasifikuoti žodžius pusiau automatiškai. Klasifikuojant antraštinį žodį šia programa, kompiuteris pagal žodžio sandaros ypatumus formuluoja klausimus operatoriui ir priskiria žodžiui morfologinio tipo numerį priklausomai nuo operatoriaus atsa-

kymų. Autorius ypač dėkingas žmonai Rasai, šitaip suklasifikavusiai visus žodyną (DLKŽ, 1972) ir (TŽŽ, 1985) antraštinius žodžius.

Kuriant kompiuterinę lietuvių kalbos morfologiją, pagrindiniais, kanoniniais morfologija aprašančiais šaltiniais laikyti LKG, 1965 ir LKG, 1971. 1984–1990 m., kai buvo kuriamas kompiuterinė morfologija, tai buvo išsamiausi, akademiniai morfologijos darbai, naujesnieji (DLKG 1994, 1996, 1997) dar nebuvo pasirodę. Todėl dauguma morfologijos dalykų kompiuterinėje morfologijoje atspindimi būtent taip, kaip jie traktuojami LKG 1965 ir 1971.

Sukurtieji formalūs lietuvių kalbos morfologijos modeliai išsamiau dar nepublikuoti. Šiek tiek apie morfologijos formalizavimo principus, gramatinės sintezės ir analizės uždavinius, leksikoninių žodynų sudarymą rašyta (Zinkevičius, 1996); šiek tiek apie lietuviškų žodžių klasifikavimą pagal morfologinius tipus, klasifikavimo kriterijus, morfologinės informacijos elementų grupavimo ir struktūrizacijos principus – (Zinkevičius, 1996*).

Skaitmeniniai GF morfologijos modeliai buvo kuriami ne vien lemavimui. Tiesiog Lemuoklis naudoja GF morfologiją, išgaudamas iš jos ir pateikdamas gramatinę informaciją tokią ir taip, kaip to reikia lemavimo procesui. Lemuojami formai atstatomas morfologiškai artimiausias antraštinis pavidas. Tačiau GF morfologijos priemonėmis galima ir gilesnė morfologinė žodžių analizė, automatiškai atsekant, pavyzdžiui, tokias analizuojamų formų morfologinių ryšių grandinėles: *nusijuokė-nusijuokti-juoktis, išgėrinėdamas-išgėrinėti-išgerti-gerti, nedarbingųjų-nedarbingas-darbingas-darbas, vabalėliui-vabalėlis-vabalas*. Bendras teoriškai įmanomų ir GF žodyno pagalba atpažįstamų žodžių gramatinių formų skaičius siekia kelis milijardus. Kompiuterinę GF morfologiją galima naudoti ne tik žodžių formoms gramatiškai atpažinti (analizuoti), bet ir sintezuoti. Internetu yra pateikta programa, demonstruojanti automatišką lietuviškų žodžių formų analizę bei sintezę GF žodyno priemonėmis (žr. LexMorphDemo).

Atpažintoms GF žodyne žodžių formoms antraštinius pavidulus Lemuoklis nustato užduodamas GF morfologijai gramatinės sinte-

zės užduotį. Nustatydamas antraštinių pavidalą vardažodžiui, Lemuoklis liepia sintezuoti vienaskaitos vardininko linksnį atitinkančią ortografinę formą. Nustatydamas antraštinių pavidalą veiksmažodžiui, Lemuoklis liepia sintezuoti ne tik bendraties, bet ir kitas dvi pagrindines formas – esamojo bei būtojo kartinio laiko trečiuosius asmenis.

Kadangi antraštinius pavidulus Lemuoklis nustato būtent tokiu formos sintezės būdu, tai atstatomi antraštiniai pavidulai kai kuriais atvejais gali šiek tiek skirtis nuo tradicinėje leksikografijoje priimtų tokių atvejų traktuočių. Pvz., asmeninių įvardžių linksnių formos *manęs, tavęs, man, tau, mūsų* DLKŽ pateikiamos atskirais antraštiniais žodžiais. Lemuoklis, aptikęs tokias formas, nelaiko jų antraštinėmis; formoms *manęs, man* ir *mūsų* jis nustato antraštinių pavidalą *aš*, formoms *tavęs* ir *tau* – pavidalą *tu*. Atstatydamas įvardžių dviskaitos formų antraštinius pavidulus, Lemuoklis skrupulingai laikosi jam įdiegto supratimo apie antraštines formas ir dviskaitą verčia į vienaskaitą. Pvz., formoms *tiedu, tuodu, tiemviem* jis pateikia antraštinių pavidalą *tas*, formai *jiedviem* – *jis*, formai *muo* – *aš*. Skaitvardžių formų atveju Lemuoklis gal kiek ir persistengia, visoms joms visada nustatydamas antraštinių neįvardžiuotinės kiekinės formos pavidalą. Formai *keturioliktojoje* antraštinė forma bus *keturiolika* (ne *keturioliktas* ar *keturioliktasis*); formai *pirmojo* – *vienas* (ne *pirmasis* ir ne *pimas*). Ateityje kitose Lemuoklio versijose visus čia suminėtus su antraštinėmis formomis susijusius nukrypimus galima bus priartinti prie tradicinėje gramatikoje bei leksikografijoje priimto minėtų atvejų traktavimo.

GF morfologija neturi geidžiamosios nuosakos, kuriant lietuvių kalbos gramatinės kaitybės skaitmeninius modelius, ji tiesiog prasydo pro akis. Todėl veiksmažodžio geidžiamosios nuosakos formas su prefiksu *te-* Lemuoklis atpažįsta tik kaip tiesioginę nuosaką. Pvz., lemuodamas formą *tegyvuoja*, Lemuoklis nustatys jai tiesioginės nuosakos esamojo laiko gramatinę reikšmę ir antraštinių pavidalą *tegyvuoti* (*-uoja, -avo*).

GF žodyno šaknų sąrašo pagrindas buvo formuojamas iš šių šaltinių antraštinių žodžių: DLKŽ (1972), TŽŽ (1985), LKRŽ, iš pastarojo imtas vardų sąrašas. Vėliau GF šaknų

sąrašas papildytas nauju, tekstynuose aptiktų žodžių šaknimis. Šiuo metu *GF* žodynas atspindi maždaug 91 tūkstantį leksemų; iš jų apie 69 tūkst. – DLKŽ žodžiai ir apie 22 tūkst. – TŽŽ žodžiai. Šiems žodžiams *GF* žodyne atstovauja 58 tūkst. šaknų. Šaknų žodyne žymiai mažiau negu leksemų, kurioms jos atstovauja. Taip yra dėl trijų priežasčių:

1) Jei tas pats žodis pasitaikė ir DLKŽ, ir TŽŽ, į *GF* jis įtrauktas tik vieną kartą. Pvz., žodžiui *abažūras*, esančiam tiek DLKŽ, tiek TŽŽ, *GF* šaknų sąrašė atstovauja viena šaknis *abažūr*.

2) Jei iš skirtingą morfologinį tipą turinčių antraštinių žodžių išskirtos šaknys sutampa, tai *GF* šaknų sąrašė tokiai šakniai išvardijami visi morfologiniai tipai, bet pati šaknis į sąrašą įtraukta tik vieną kartą. Pavyzdžiui, daiktavardžiui *kalba*, būdvardžiui *kalbus* ir veiksmažodžiui *kalbėti* *GF* šaknų sąrašė atstovauja viena šaknis *kalb*, tik ties ją yra duotos tris skirtingus morfologinius tipus reiškiančios nuorodos.

3) Kadangi *GF* kompiuterinėje morfologijoje atspindimi ir kai kurie žodžių darybos reiškiniai, tai eama daug tokių atvejų, kai keli išvestiniai žodžiai, turintys vieną bendrą pamatinį, šaknų sąrašė atstovaujami viena kartą. Pavyzdžiui, dešimčiai DLKŽ žodžių – *mokslas, mokslingas, mokslingai, mokslingumas, mokslininkas, mokslininkė, mokslinis, mokslīškas, mokslīškai, mokslīškumas* – *GF* šaknų sąrašė atstovauja viena šaknis *moksl*.

Žodynas G. *G* žodynas, kaip ir *GF* žodynas, taip pat sudėtas iš dviejų komponentų: šaknų sąrašo ir kompiuterinės morfologijos. Tačiau, skirtingai nei *GF* morfologija, *G* žodyno morfologija yra bejėgė nustatyti generuojamų ortografinių formų gramatinės reikšmės. Taip yra todėl, kad *G* žodynas buvo specialiai kuriamas kaip duomenų bazė rašybos klaidas aptinkančioms programoms (*speleryms*). Tokių programų žodžių gramatinės reikšmės paprastai nedomina, joms pakanka, jei žodynas pasako, ar nagrinėjama raidžių seka gali būti kokio nors žodžio forma, ar ne. Tačiau *G* morfologija, lyginant ją su *GF* morfologija, turi vieną privalumą. Visų morfologinių tipų žodžiams ji generuoja ne tik bendrinės kalbos gramatika apibrėžiamas formas, bet ir formas su galūnių nutrupėjimais ar senesnės vartosenos formas (pvz., *iliatyvus*).

G žodyno šaknų sąrašas gautas iš *GF* žodyno šaknų sąrašo išmėčius visas tikrinių vardu šaknis (pastarosios buvo perkeltos į *TG* žodyną, žr. toliau). Šiuo metu *G* žodynas atspindi maždaug 84 tūkstantį leksemų; iš jų apie 67 tūkst. – DLKŽ žodžiai ir apie 27 tūkst. – TŽŽ žodžiai. Šiems žodžiams *G* žodyne atstovauja 48 tūkst. šaknų. Žodyne šaknų žymiai mažiau negu leksemų, kurioms jos atstovauja, dėl tu pačių priežasčių, kurios paaiškintos aukščiau, aprašant *GF* žodyno sandarą.

Žodynas TG. Tai tikrinių vardažodžių šaknų žodynas, kiekvienai šakniai yra nuoroda į kompiuterinę tikrinių daiktavardžių morfologiją. *TG* morfologijos pagalba Lemuoklis žino, kokius afiksus ir kaip reikia jungti prie tikrinių žodžių šaknų (ar kamienų), darant jų kaitybines ir darybines formas. *TG* žodyno pagalba Lemuoklis atpažįsta tikrinių vardu kaitybines ir darybines formas ir dar žino, kurios jų tikrinės, o kurios bendrinės, bet apibūdinti jų gramatiškai negali.

TG morfologijos pagalba generuojamos ne tik kaitybinės, bet ir įvairios išvestinės vardu formos. Pavyzdžiui, iš vyriškų pavardžių šaknų ar kamienų išvedamos moteriškų pavardžių formos, iš vietovardžių – bendriniai priesagų *-iškis, -ietis* vediniai ir pan.

TG žodyno šaknų (kamienų) sąrašas sukompiliuotas iš įvairių šaltinių. Tai DLKŽ, TŽŽ pasitaikę tikriniai vardai, LKRŽ žodyne pateikti vardai. Į *TG* taip pat įtraukti ir tekstynuose aptikti tikriniai daiktavardžiai. *TG* žodynas buvo sukurtas ne lemavimui. Jis perkeltas į Lemuoklio kalbinių duomenų bazę iš rašybos teisingumo kontrolės funkcija atliekančių programų (*speleryų*). *TG* žodynas atspindi apie 15 tūkst. tikrinių vardu, šioms vardams atstovauja apie 9 tūkst. šaknų/kamienų.

Žodynas F. Tai netikrinių žodžių formų sąrašas. Jo sudarymui panaudotos iš KLC tekstyno (Marcinkevičienė, 1997) ir iš DDRLKŽ imčių tekstyno paimtos formos. *F* žodyne įrašyta apie 132000 formų.

Žodynas T. Tai tikrinių žodžių formų sąrašas. Jam sudaryti taip pat naudotasi KLC tekstyno ir iš DDRLKŽ imčių tekstyno formų sąrašais. *T* žodyne įrašyta apie 74 tūkst. formų.

Žodynas S. Tai santrumpų ir akronimų sąrašas, sudarytas panaudojant įvairius šaltinius, tarp jų ir tekstynų analizės rezultatus.

Žodyne įrašyta apie 230 santrumpų ir akronimų. Kiekvienai santrumpai ar akronimui žodyne priskirtas kodas, iš kurio Lemuoklis gali spręsti apie santrumpos ar akronimo rašybą: ar būtinas taškas po santrumpos, kurias santrumpos ar akronimo raidės reikėtų rašyti didžiosiomis.

Visi aprašytieji kompiuteriniai žodynai turi vieningą loginę struktūrą. Tai reiškia, kad ieškodamas kalbinės informacijos Lemuoklis naršo po visus žodynus ir ištraukia iš jų duomenis naudodamas bendras visiems žodynams programines procedūras. Visi žodynai suformuoti kaip *medžio* pavidalo duomenų struktūros. Kompiuterinė informacijos paieška ir jos išrinkimas tokiose struktūrose vyksta labai greitai. Medžių šakas sudaro raidžių sekos, o ieškoma informacija įrašyta jų terminalinėse viršūnėse – lapuose. Žodynuose *GF*, *G* ir *TG* tokia informacija yra nuorodos į atitinkamas kompiuterines morfologijas; *S* žodyne – santrumpų ar akronimų kodai; *F* ir *T* žodynų-medžių lapuose jokios papildomos informacijos nėra.

Žodynų – medžių struktūros loginė organizacija yra pastovi, ne dinaminė. Todėl, norint pakeisti kalbinę žodynų informaciją (pvz., papildyti žodyną naujomis žodžių šaknimis ar formomis), visas žodynas – medis turi būti formuojamas iš naujo. Šaknų ir formų alfabetinių sąrašų transformavimui į medžio struktūras sukurta speciali programinė įranga. Šios programinės įrangos pagalba bet koks alfabetinis sąrašas automatiškai išanalizuojamas (sudaromos sąrašo šakojimosi matricos, paskaičiuojami šakojimosi mazgų parametrai) ir po to perkeliama į medį. Tokia žodynų – medžių formavimo programinė įranga į Lemuoklio sudėtį neįeina, ji tik buvo panaudota rengiant jo kalbinių duomenų bazę.

INFORMACIJOS PAIEŠKA KALBINIŲ DUOMENŲ BAZĖJE IR HIPOTETINIŲ LEMŲ FORMAVIMAS

Lemuodamas kurią nors žodžio formą, Lemuoklis ieško apie ją informacijos kalbinių duomenų bazėje, kreipdamasis paeiliui į žodynus *GF*, *TG*, *F*, *G*, *T* ir *S*. Kreipimosi į žodynus argumentas yra žodžio formą sudaran-

čių raidžių seka. Jei raidžių seką atitinkančios informacijos eiliniame žodyne nėra, gautas neigiamas atsakymas, jei yra, teigiamą atsakymą žodynai suformuluoja skirtingai. Žodynai *F*, *G* ir *T* jokios papildomos informacijos daugiau nepateikia. Žodynas *GF* pateikia gramatinę informaciją apie visas duotąja raidžių seką ortografiškai reiškiamas homoformas. Žodynas *TG* pasako, ar atpažintoji raidžių sekoje žodžio forma yra tikrinis daiktavardis, ar ne, daugiau nepateikdamas jokios kitos gramatinės informacijos. Žodynas *S* apibūdina atpažintosios santrumpos ar akronimo rūšį: ar reikia po santrumpos taško, ar ją būtina rašyti didžiosiomis raidėmis ir pan.

Išnagrinėjęs žodynų atsakymus (pasižiūrėjęs, kurie žodynai į raidžių sekos paiešką atsakė teigiamai, o kurie neigiamai), Lemuoklis pasirenka vieną iš dešimties būdų, kaip formuoti hipotetines raidžių seką atitinkančias lemas.

Šis procesas pavaizduotas *1* *schemoje*.

Paieškos žodynuose rezultatų situacijos *schemoje* sunumeruotos skaičiais rutuliukuose. Aprašysime, kaip Lemuoklis formuoja lemas kiekviename šių situacijų.

1-oji situacija: raidžių seką radom žodyne *GF*, žodyne *T* neradom, žodyne *TG* neradom. Jei žodyne *GF* informacija apie raidžių seką aptikta, o žodynuose *T* ir *TG* tokios informacijos nebuvo, tai į likusių trijų žodynų atsakymus Lemuoklis nebežiūri. Iš žodyno *GF* išrenkama gramatinė bei leksinė informacija apie visas duotąja raidžių seką ortografiškai išreikštas homoformas. Informacijos apie raidžių seką nebuvimas žodynuose *T* ir *TG* Lemuokliui parodo, kad žodžio forma nėra tikrinė. Formuodamas lemapavimo rezultatus, homoformų antraštinį pavidalą Lemuoklis įrašo mažąja raide.

Formų naujas, seniausieji, kalba, radikalus, permainingai, geri, politika, itakos, kurias lemuojant susiklosto aptariamoji situacija, lemapavimo pavyzdys:

naujas

bdvr <naujas>

bdvr nelygin.1 neįvardž vyr.gim vnsk V

bdvr nelygin.1 neįvardž mot.gim dgsk G

seniausieji

bdvr <senas>

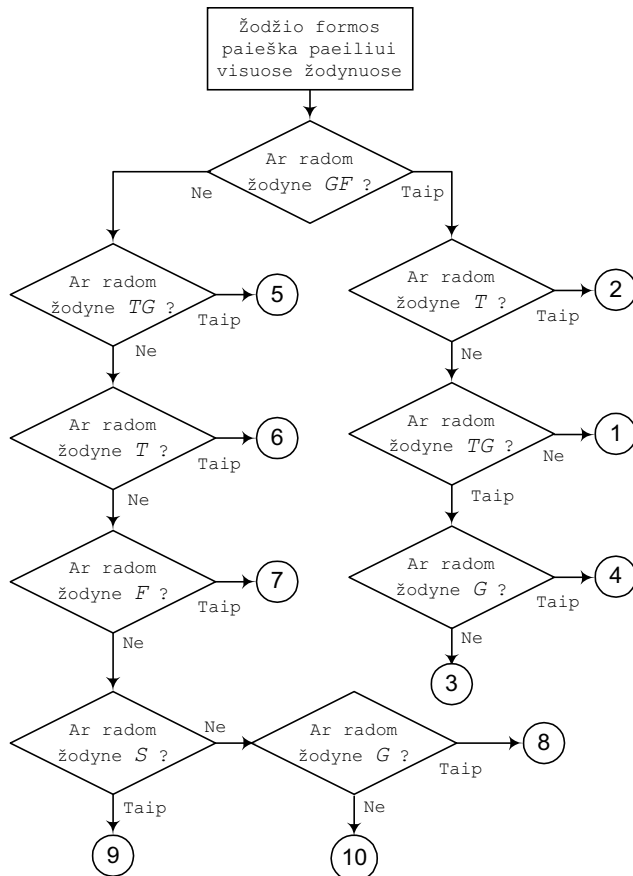
bdvr aukšč.1 įvardž vyr.gim dgsk V

kalba
 ** dktv <kalba>
 dktv mot.gim vnsk V
 dktv mot.gim vnsk In
 dktv mot.gim vnsk Š
 ** bndr <kalbėti (-a, -ėjo)>
 vksm nesngr tiesiog.nuos esam.l IIIasm
 radikalus
 ** dktv <radikalas>
 dktv vyr.gim dgsk G
 ** bdvr <radikalus>
 bdvr nelygin.l neįvardž vyr.gim vnsk V
 permainingai
 ** prvks <permainingai>
 prvks nelygin.l
 ** bdvr <permainingas>
 bdvr nelygin.l neįvardž mot.gim vnsk N
 geri
 ** bdvr <geras>
 bdvr nelygin.l neįvardž vyr.gim dgsk V
 ** bndr <gerti (-eria, -ėrė)>

vksm nesngr tiesiog.nuos esam.l vnsk IIasm
 politika
 ** dktv <politikas>
 dktv vyr.gim vnsk G
 ** dktv <politika>
 dktv mot.gim vnsk G
 įtakos
 ** dktv <įtaka>
 dktv mot.gim vnsk K
 dktv mot.gim dgsk V
 dktv mot.gim dgsk Š
 ** bndr <įtakoti (-ja, -jo)>
 vksm nesngr tiesiog.nuos būs.l IIIasm

2-oji situacija: raidžių seka radom žodyne GF ir žodyne T. Žodžių formų *Baltijos, Rusijoje, Merkinės, Neris, Palestinoje, Alpiu*, kurias lemujant susiklosto tokia situacija, lemavimo pavyzdys:

Baltijos
 tikr dktv <Baltija>
 tikr dktv mot.gim vnsk K



1 schema. Hipotetinių lemų formavimo būdo pasirinkimas priklausomai nuo paieškos žodynuose rezultatų

Rusijoje

tikr dktv <Rusija>

tikr dktv mot.gim vnsk Vt

Merkinės

** bndr <merkinėti (-ja, -jo)>

vksm nesngr tiesiog.nuos būs.l IIIasm

** tikr dktv <Merkinė>

tikr dktv mot.gim vnsk K

Neries

** tikr dktv <Neris>

tikr dktv mot.gim vnsk K

** bndr <nerieti (-ja, -jo)>

vksm nesngr tiesiog.nuos būs.l IIIasm

** bndr <neriesti (-čia, -tė)>

vksm nesngr tiesiog.nuos būs.l IIIasm

Palestinėje

** tikr dktv <Palestina>

tikr dktv mot.gim vnsk Vt

** bndr <palesti (-a, -ė)>

dlv nesngr reikiamyb. r neįvardž mot.gim vnsk Vt

Alpių

** tikr dktv <Alpės>

tikr dktv mot.gim dgsk K

** bđvr <alpus>

bđvr nelygin.l neįvardž vyr.gim dgsk K

bđvr nelygin.l neįvardž mot.gim dgsk K

Žodynas *GF* Lemuokliui pateikia visų įmanomų homoformų, išreikštų duotąja ortografinė forma, išsamius gramatinius apibūdinimus, išskyrus nomastinį aspektą. Raidžių sekos aptikimas žodyne *T*, kuris yra iš įvairių šaltinių surinktų tikrinių vardų kaitybinių formų sąrašas, Lemuokliui signalizuoja, kad šią raidžių virtinėle atitinkančios homoformos, kurioms žodynas *GF* nustatė kalbos dalį „daiktavardis“, yra tikrinės. Formuodamas lemavimo rezultatus, tokių homoformų antraštinį pavidalą Lemuoklis įrašo didžiąja raide ir kalbos dalį keičia į „tikrinis daiktavardis“. Taip gautos lemos – tikriniai daiktavardžiai <Baltija>, <Rusija>, <Merkinė>, <Neries>, <Palestina> ir <Alpės>.

Jei tarp homoformų, kurias žodynas *GF* atpažino raidžių sekoje, yra ir ne daiktavardžių, tai tokių homoformų antraštinį pavidalą Lemuoklis įrašo mažąja raide, o kalbos dalį nurodo tokia, kokią jai nustatė žodynas *GF*. Taip gautos lemos bndr <merkinėti (-ja, -jo)>, bndr <nerieti (-ja, -jo)>, bndr <neriesti (-čia, -tė)>, bndr <palesti (-a, -ė)>, bđvr <alpus>. Panašu, kad dažniausiai šitaip šioje situacijoje gaunamos hipotetinės lemos būna klaidingos. Jei jau Lemuoklis aptiko žodžio formą tikrinių formų sąrašė *T*, tai tikriausiai

ji ir yra tikrinis daiktavardis, o kitokios žodyno *GF* siūlomos gramatinės traktuotės iškyla dėl atsitiktinės ortografinės homonimijos ir greičiausiai yra neteisingos. Tačiau visiškai atsisakyti šioje situacijoje šių homonimų ir automatiškai išmesti jas iš lemavimo rezultatų negalime. Juk visai galimi tokie pasakymai: *Neries ji man čia nosies; Alpių rugpjūčio vakarų prisiminimai*.

Galima būtų Lemuokliui liepti nedaiktavardines traktuotes palikinti tik atvejais, kai žodis tekste buvo užrašytas mažąja raide. Bet vėlgi, dėl visiško korektiškumo tada dar reikėtų tikrinti, ar žodis nėra pirmasis sakinyje (juk pirmas paprastai visada rašomas didžiąja raide, nors ir ne tikrinis); be to, reikėtų numatyti atskirą šių situacijų traktavimą atvejams, kai lemuojamas ne rišlus tekstas, o pavienių žodžių formų sąrašai, ar kai lemuojamas vien didžiosioms raidėms parašytas tekstas. Kol kas viso to Lemuoklis nedaro ir pateikia visas hipotetines žodžių gramatinių reikšmių versijas, laikydamasis principo „geriau per daug, negu per mažai“. Visiškai automatiška teisingų homoformų atranka šiais atvejais bus įmanoma, ko gero, tik išmokius Lemuoklį atlikti sintaksinių ryšių tarp rišlaus teksto žodžių analizę, o tokia lietuviško teksto kompiuterinė analizė dar tik ateities planuose.

3-ioji situacija: raidžių seką radom žodyne *GF*, žodyne *T* neradom, žodyne *TG* radom, žodyne *G* neradom. Raidžių sekos aptikimas žodyne *TG* Lemuokliui rodo, kad tai arba tikrinis daiktavardis, arba nebūtinai tikrinis vedinys iš tikrinio. Jei tai tikrinio varbo forma, tai greičiausiai gana reta, nes žodyne *T*, iš įvairių šaltinių surinktų tikrinių vardų formų sąrašė, jos neaptikom.

G žodyne tikrinių formų nėra. Todėl žodžio formos neaptikimas *G* žodyne Lemuokliui rodo, kad jei *TG* žodynas apibūdina aptiktąją jame žodžio formą kaip tikrinę, tai ji tokia ir yra, – netikrinių jos homoformų nėra.

Žemiau pateikiamas žodžių formų *Durbės, rygiečių, Alšėnų, Tveriškis, Volgos, maskvietiška-jai, Deltuvos*, kurias lemuojant susiklosto tokia situacija, lemavimo pavyzdys:

Durbės

tikr dktv <Durbė>

tikr dktv mot.gim vnsk K

rygiečių

** dktv <rygietis>
dktv vyr. gim dgsk K
** dktv <rygietė>
dktv mot. gim dgsk K
Alšėnu
tikr dktv <Alšėnai>
tikr dktv vyr. gim dgsk K
Tveriškis
dktv <tveriškis>
dktv vyr. gim vnsk V
Volgos
tikr dktv <Volga>
tikr dktv mot. gim vnsk K
maskvietiškajai
bdvr <maskvietiškas>
bdvr nelygin. l įvardž mot. gim vnsk N
Deltuvos
tikr dktv <Deltuva>
tikr dktv mot. gim vnsk K

Pažiūrėkime, kaip Lemuoklis formuoja lemas šioje situacijoje.

Jei žodynas *GF* atpažino žodžio formą kaip ne daiktavardį, tai tokios žodžio formos antraštinį pavidalą Lemuoklis įrašo mažąja raide, o kalbos dalį nurodo tokia, kokią jai nustatė žodynas *GF*. Taip gauta lema bdvr <maskvietiškas>.

Kalbos dalį žodžio formai, kurią žodynas *GF* atpažino kaip daiktavardį, Lemuoklis nustato tokiu būdu. Jei *TG* žodynas rastąją jame raidžių seką apibūdino kaip tikrinę, tai tokios žodžio formos antraštinį pavidalą Lemuoklis įrašo didžiąja raide, o kalbos dalimi nurodo tikrinį daiktavardį. Taip gautos lemos-tikriniai daiktavardžiai <Durbė>, <Alšėnai>, <Volga> ir <Deltuva>. Jei žodynas *TG* rastąją jame raidžių seką apibūdino kaip netikrinę, tai tokios žodžio formos antraštinį pavidalą Lemuoklis įrašo mažąja raide, o kalbos dalimi nurodo daiktavardį. Taip gautos lemos – daiktavardžiai <rygietis>, <rygietė> ir <tveriškis>. Kitų gramatinių kategorijų reikšmes homoformų gramatiniams apibūdinimams Lemuoklis įrašo taip, kaip jas pateikia žodynas *GF*.

Šiame pavyzdyje Lemuoklis tik su forma *Tveriškis* kiek prašovė pro šalį. Formą *Tveriškis* Lemuoklis *GF* žodyne rado kaip išvestinį iš žodžio *Tverė* daiktavardį. Žodynas *TG*, į kurį taip pat įtrauktas *Tverė*, formą *Tveriškis* apibūdino kaip padarytą iš tikrinio žodžio bendrinę. Lemuotajame tekste buvo *Demetrius Tveriškis*. Kad *Tveriškis* yra tikrinis as-

menvardis, Lemuoklis nežino, nes informacijos apie tokį asmenvardį jo žodynuose nėra, kaip, tarp kitko, ir informacijos apie *Demetrijų* – šio žodžio formos Lemuoklis apskritai neatpažino.

4-oji situacija: raidžių seką radom žodyne *GF*, žodyne *T* neradom, žodyne *TG* ir žodyne *G* radom. Situacija skiriasi nuo 3-iosios tuo, kad raidžių seką aptikta ir *G* žodyne. *G* žodyne tikrinių formų nėra, tik bendrinės. Todėl, kai šioje situacijoje *TG* žodynas rastąją jame raidžių seką apibūdina kaip tikrinę, tai reiškia, kad lemuojama ortografinė forma atitinka ir tikrinę, ir bendrinę gramatinę homoformą. Vienodą ortografinę išraišką turinčios tikrinė ir bendrinė homoformos šiuo atveju gali priklausyti semantiškai artimoms leksemoms (plg. bendrinis *gintaras* ir tikrinis *Gintaras*, bendrinis *turgeliai* ir tikrinis *Turgeliai*). Ortografiškai sutapti gali ir semantiškai tolimų leksemų formos, t. y. toks sutapimas gali būti ir atsitiktinis (plg. *jonas* ir *Jonas*). Gali šioje situacijoje ortografiškai sutapti ir skirtingų kalbos dalių homoformos (pvz.: prievėksmis *gana* – tikrinis daiktavardis *Gana*, prievėksmis *greta* – tikrinis daiktavardis *Greta*, skaitvardis *viena* – tikrinis daiktavardis *Viena*).

Žodžių formų *Joną*, *gintaro*, *gana*, *greta*, *Turgelių*, *viena* lemavimo pavyzdys:

Joną
tikr dktv <Jonas>
tikr dktv vyr. gim vnsk G
gintaro
tikr dktv <Gintaras>
tikr dktv vyr. gim vnsk K
gana
** tikr dktv <Gana>
tikr dktv mot. gim vnsk V
tikr dktv mot. gim vnsk Įn
tikr dktv mot. gim vnsk Š
** prvks <gana>
prvks
greta
** tikr dktv <Greta>
tikr dktv mot. gim vnsk V
tikr dktv mot. gim vnsk Įn
tikr dktv mot. gim vnsk Š
** prvks <greta>
prvks
Turgelių
tikr dktv <Turgelis>
tikr dktv vyr. gim dgsk K

viena

** tikr dktv <Viena>

tikr dktv mot.gim vnsk V

tikr dktv mot.gim vnsk Įn

tikr dktv mot.gim vnsk Š

** bđvr <vienas>

bđvr nelygin.1 neįvardž mot.gim vnsk V

bđvr nelygin.1 neįvardž mot.gim vnsk Įn

bđvr nelygin.1 neįvardž bevrđ.gim

** sktv <vienas>

sktv kiekis mot.gim vnsk V

sktv kiekis mot.gim vnsk Įn

sktv kiekis bevrđ.gim

** įvrd <vienas>

įvrd mot.gim vnsk V

įvrd mot.gim vnsk Įn

įvrd bevrđ.gim

Nors ši situacija kiek skiriasi nuo 3-iosios, lemas joje Lemuoklis formuoja lygiai taip pat, kaip ir 3-iojoje situacijoje. Tačiau jei 3-iojoje situacijoje buvo galima gana patikimai nuspręsti, kokia – tikrinė ar bendrinė – yra lemuojama forma, tai šioje situacijoje Lemuokliui tai padaryti daug sunkiau. Pateiktame lemavimo pavyzdyje Lemuoklis teoriškai įmanomų bendrinių daiktavardžių lemų <jonas>, <gintaras>, <greta>, <turgelis> nepateikia; *gintaro* atveju tai jau neteisingo lemavimo atvejis, nes lemuotajame tekste ši žodžio forma buvo bendrinė. Gali atrodyti, kad šioje situacijoje Lemuokliui derėtų ginčytino onomastiškumo daiktavardžių lemas išvesti abiem variantais – ir kaip tikrinius, ir kaip bendrinius daiktavardžius. Tačiau tokiu atveju atsirastų klaidingos lemos – bendriniai daiktavardžiai <viena> ir <gana>. Jei šitoje situacijoje Lemuoklis nereaguotų į onomastiškumą ir formotų tik bendrines lemas, būtų dar blogiau: ne tik atsirastų klaidingos lemos – bendriniai daiktavardžiai <viena> ir <gana>, bet ir prarastume potencialiai teisingas tikrines lemas <Jonas>, <Gintaras>, <Greta>, <Turgelis>. Beje, dėl pastarosios: į *TG* žodyną *Turgeliai* įtrauktas kaip daugiskaitinis vietovardis, tačiau atstatyti jam antraštinę formą *TG* morfologija, kaip minėta, nemoka. Už antraštinių formų atstatymą atsakingas yra *GF* žodynas. Kadangi į *GF* žodyną *turgelis* įtrauktas ne kaip daugiskaitinis, tai *GF* kompiuterinė morfologija formai *Turgelių* atstato antraštinę vienaskaitos vardininko formą.

Šiaip ar taip, visa ši painiava su tikrinių formų lemavimu ateityje išsispres, atitinka-

mai patobulinus *GF* žodyną. Kai į žodyno kompiuterinę morfologiją įtrauksime daiktavardžių onomastiškumo požymį, formas *Joną*, *gintaro*, *Turgelių* Lemuoklis lemuos dvejojai – ir kaip tikrinius, ir kaip bendrinius daiktavardžius, formą *gana* – kaip tikrinį daiktavardį ir kaip prievieksmį, *greta* – kaip bendrinį daiktavardį, kaip tikrinį daiktavardį ir kaip prievieksmį, *viena* – kaip tikrinį daiktavardį, būdvardį, skaitvardį ir įvardį. Beje, formai *viena* priskiriamų kalbos dalių įvairovė gali pasirodyti keista, tačiau čia Lemuoklis nekaltas, tokias galimas kalbos dalis žodžiui *vienas* nurodo DLKŽ.

5-oji situacija: raidžių sekos žodyne *GF* neradom, bet radom žodyne *TG*. Kadangi žodžio forma neaptikta žodyne *GF*, apie jos gramatinę reikšmę Lemuoklis nieko pasakyti negali, kaip ir atstatyti jai antraštinio pavaldalo. Tačiau iš žodyno *TG* teigiamo atsakymo Lemuoklis ši tą išpeša ir suformuoja nepilną lemavimo rezultata, kuri pailiustruosim formų *Tacitas*, *alanų*, *Horodlės*, *naugardiečiams*, *Vorsklos*, *pskoviečių* lemavimo pavyzdžiu:

Tacitas

tikr dktv <Tacit..?>

gr char ..?

alanų

tikr dktv <Alan..?>

gr char ..?

Horodlės

tikr dktv <Horodl..?>

gr char ..?

naugardiečiams

vrdž arba prvksm iš tikr dktv <naugard..?>

gr char ..?

Vorsklos

tikr dktv <Vorsk..?>

gr char ..?

pskoviečių

vrdž arba prvksm iš tikr dktv <pskov..?>

gr char ..?

Žodyno *TG* kompiuterinė morfologija, kaip minėta, operuoja tik afiksais, bet ne gramatinėmis reikšmėmis. Tačiau žodynas *TG* vienareikšmiškai nurodo, ar rastoji jame forma tikrinė, ar bendrinė. Jei *TG* žodynas rastąją jame formą apibūdina kaip išvestinę iš tikrinio vardo bendrinę, tai teoriškai ji gali būti arba daiktavardis (pvz., *naugardiečių*), arba būdvardis (pvz., *naugardietiškas*), arba būdvardis/prievieksmis (pvz., *naugardietiška*). To-

kiems nepilno kalbos dalies nurodymo atvejams Lemuoklis naudoja gana grioždišką formuluotę „vardažodis arba prieveiksmis iš tikrinio daiktavardžio“.

Atstatyti antraštinį žodžio pavidalą tokioje situacijoje Lemuoklis gali taip pat tik nepilnai. *TG* morfologija jam tik nurodo, nuo kurios formos vietos prasideda darybinė priėmimo (jei tokia yra) su galūne. Įrašęs formą iki tos vietos, toliau Lemuoklis deda daugtaškį.

Pateiktame žodžių formų nepilno lemavimo pavyzdyje forma *alanų* sulemuota ne visai teisingai. Lemuoklis formą apibūdino kaip tikrinę (nuo asmenvardžio *Alanas*), nors tekste ji reiškė bendrinę genties pavadinimą. Taip atsitiko dėl to, kad vardas *Alanas*, tekstuose pasitaikantis gerokai dažniau už tautovardį *alanai*, pateko į *TG* žodyną, o pastarasis neužfiksuotas nė viename Lemuoklio žodyne.

6-oji situacija: raidžių sekos žodyne *GF* neradom, žodyne *TG* neradom, žodyne *T* radom. Kadangi žodžio forma neaptikta žodyne *GF*, apie jos gramatinę reikšmę Lemuoklis nieko pasakyti negali, kaip ir atstatyti jai antraštinio pavidalo. Iš žodyno *T* teigiamo atsakymo Lemuoklis tik žino, kad tai tikrinio daiktavardžio forma.

Formos *Pizos* lemavimo pavyzdys:

Pizos
tikr dktv <P..?>
gr char ..?

Aukščiau aptartojoje 5-ojoje, taip pat nepilno lemavimo situacijoje *TG* žodynas nurodė Lemuokliui, nuo kurios formos vietos prasideda kintanti jos dalis, ir Lemuoklis galėjo formuoti nepilną antraštinį formos pavidalą, perrašydamas formą iki tos vietos. Tačiau šioje situacijoje forma *TG* žodyne neaptikta, tik *T* žodyne; *T* žodynas kompiuterinės morfologijos neturi, ji sudaro paprastas tikrinių formų sąrašas. Kadangi Lemuoklis nežino, kurioje lemujamos formos vietoje prasideda jos fleksija, tai vietoj antraštinio pavidalo išrašo tik pirmąją didžiąją formos raidę ir daugtaškį.

7-oji situacija: raidžių sekos žodynuose *GF*, *TG* ir *T* neradom, bet radom žodyne *F*. Žodžio forma *GF* žodyne nerasta, tad apie jos gramatinę reikšmę Lemuoklis nieko pasakyti negali, negali ir atstatyti jai antrašti-

nio pavidalo. Iš žodžių *TG* ir *T* neigiamo atsakymo Lemuoklis žino, kad forma ne tikrinė. *GF* žodynas atpažįsta bendrinės lietuvių kalbos žodžių formas, išskyrus formas su nutrupėjusiomis galūnėmis. Kadangi žodyne *F* (į jį sudėtos formos iš tekstynų) forma aptikta, tai Lemuoklis daro išvadą, kad lemuotas žodis turėjo nutrupėjusią galūnę arba buvo pavartotas kokia pasenusia ar tarmiška gramatine forma. Šią išvadą Lemuoklis rezultatuose pažymi formuluote „...? galbūt nutrupėjusi galūnė“, įrašyta vietoj antraštinio formos pavidalo.

Žodžių formų *trim*, *valdžion* lemavimo pavyzdys:

trim
klb d ..? <..? galbūt nutrupėjusi galūnė>
gr char ..?
valdžion
klb d ..? <..? galbūt nutrupėjusi galūnė>
gr char ..?

8-oji situacija: raidžių sekos žodynuose *GF*, *TG*, *T*, *F* ir *S* neradom, bet radom žodyne *G*. Situacija labai panaši į 7-ąją. Skiriasi tik tuo, kad forma rasta ne *F* žodyne, kaip 7-osios situacijos atveju, bet *G* žodyne. Tiek žodyne *F*, tiek ir žodyne *G* laikomos tik netikrinės žodžių formos be jų gramatinių reikšmių. Skiriasi tik šių žodžių organizavimo būdas. Į žodyną *F* surašytos ortografinės žodžių formos, jas išvardijant. Žodyne *G* surašytos žodžių šaknys, o formų sąrašą generuoja *G* žodyno morfologija, prijungdama prie kiekvienos šaknies jai priklausančius pagal morfologinį tipą afiksus. Ar formą Lemuoklis aptinka *F* žodyne, ar *G* žodyne, abiem atvejais jis daro tą pačią išvadą. Todėl šioje situacijoje Lemuoklis elgiasi taip pat kaip ir 7-ojoje.

Formos *vykdytojom* lemavimo pavyzdys:

vykdytojom
klb d ..? <..? galbūt nutrupėjusi galūnė>
gr char ..?

9-oji situacija: raidžių sekos žodynuose *GF*, *TG*, *T* ir *F* neradom, bet radom žodyne *S*. *S* žodynas Lemuokliui nurodo atpažintosios jame santrumpos ar akronimo rūšį: ar reikia po santrumpos taško, ar visos raidės būtinai didžiosios ir pan. Vietoj gramatinės reikšmės žymėjimo Lemuoklis išrašo žymę „sntnmp“. Į antraštinio pavidalo vietą Le-

muoklis perrašo pačią santrumpą, bet jau vadovaudamasis žodyno *S* nurodytosiomis santrumpos ar akronimo rašybos taisyklėmis.

Pavyzdžiui, *tūkst* lemavimas:

tūkst

sntmp <tūkst.>

sntmp

10-oji situacija: raidžių sekos neradomė viename žodynų. Neatpažinęs raidžių sekoje lietuviško žodžio formos ar santrumpos, Lemuoklis įrašo į rezultatų failą sutartinę tokios nesėkmės žymėjimą.

Pavyzdžiui, formų *submonarcho*, *Rusdorfas*, *antimindauginė*, *apostazijos*, *didvalstybė*, *Radoškovičių*, *kontrspaudima*, *plačiaveidis* lemavimas:

submonarcho

<??>

Rusdorfas

<??>

antimindauginė

<??>

apostazijos

<??>

didvalstybė

<??>

Radoškovičių

<??>

kontrspaudima

<??>

plačiaveidis

<??>

Sudurtinių lietuviškų žodžių darybos modeliai išnagrinėti ir aprašyti pakankamai plačiai (žr., pvz., Būda, 1994). Todėl, ateityje papildžius *GF* žodyno kompiuterinę morfologiją sudurtinės žodžių darybos taisyklėmis, galima būtų išmokyti Lemuoklį lemuoti ir tokias dabar neatpažįstamas formas *submonarcho*, *antimindauginė*, *didvalstybė*, *kontrspaudima*, *plačiaveidis*.

LEMAVIMO DAUGIAPRASMIŠKUMO MAŽINIMAS

Silpnoji Lemuoklio vieta yra lemavimo nevienareikšmiškumas. Išnagrinėjęs žodžio formą, Lemuoklis dažnai pateikia ne vieną, o kelis galimus (hipotetinius) jos gramatinius apibūdinimus. Lemavimo daugiaprasmiškumo priežastys glūdi ortografinės homonimijos reiškinyje. Dažnai kelios skirtingų gramatinių reikšmių žodžių formos turi vieną ir

tą pačią ortografinę (rašytinę) išraišką. Taip sutapti gali skirtingos vieno žodžio gramatinės formos. Pvz., daiktavardžio *motina* vienas kitą kilmininkas, daugiskaitos vardininkas ir daugiskaitos šauksmininkas ortografiškai reiškiami ta pačia išraiška *motinos*. Gali sutapti ir skirtingų žodžių formos, plg. sutampančias formas frazėse *įnirtingai laužo šakas* ir *sėdėjo laužo šviesoje*; *mes dirbame* ir *jis mes akmenį į langą*. Skyrelyje „Informacijos paieška kalbinių duomenų bazėje ir hipotetinių lemų formavimas“ pateiktuose lemavimo pavyzdžiuose apstu pačių įvairiausių skirtingų gramatinių formų ortografinio sutapimo atvejų.

Su panašia problema susiduria ir kitų kalbų lemuokliai bei gramatiniai anotatoriai. Lemavimo nevienareikšmiškumas juose dažnai išsprendžiamas ar sumažinamas (*morphological disambiguation*, *ambiguity resolution*) statistiniais-tikimybiniais metodais, panaudojus duomenis apie leksemų ir/ar gramatinių reikšmių vartosenos dažnius. Lemuoklyje tokie problemos sprendimo metodai nenaudojami dėl šių priežasčių:

1) Šiuo metu lietuvių kalbos gramatinių reikšmių dažninių charakteristikų nėra iš kur paimti. Dažniniuose lietuvių kalbos žodynuose (DDRLKŽ, 1997 ir 1998) pateikti žodžiai, jų kalbos dalys ir suminiai visų jų kaitybinių formų vartojimo dažniai, informacijos apie gramatinių reikšmių dažnius šiuose žodynuose nėra.

2) Akivaizdu, kad vien žodžių gramatinių formų vartosenos dažninių charakteristikų žinojimas ortografinio homonimiškumo sukeltam lemavimo problemų neišspręs. Pvz., žinom, kad ortografinės formos *kalba* daiktavardinė homoforma yra gerokai dažnesnė už veiksmazodinę, bet be gilesnės konteksto analizės vis tiek negalim vienareikšmiškai nuspręsti, ar *kalba* tekste yra veiksmazodžio esamojo laiko III asmens forma, ar daiktavardžio vienas kitą vardininkas.

Autoriaus nuomone, visiškai panaikinti automatiško lemavimo daugiaprasmiškumą įmanoma tik išmokus Lemuoklį nagrinėti ryšius tarp lietuviškų žodžių sakiniuose. Šiuo metu dar nėra kompiuterinių technologijų, kurias šiam tikslui būtų galima įdiegti Lemuoklyje. Pasišovusiųjų sukurti programines procedūras, nagrinėsiančias sintaksinius ryšius tarp

rišlauso lietuviško teksto žodžių, laukia ne-
lengvas darbas. Svarbiausias uždavinys čia
turbūt būtų sintaksinių ryšių tarp žodžių dė-
sningumų formalizavimas.

Taigi šiuo metu Lemuoklis pateikia gra-
matinę informaciją apie visas teoriškai įma-
nomas lemuojamų ortografinių formų homo-
formas. Vis dėlto jame įdiegti tam tikri meto-
dai, leidžiantys kai kuriais atvejais sumažinti
lemavimo rezultato daugiaprasmiškumą.
Kiekvieną iš šių metodų Lemuoklis gali pa-
naudoti ir nepanaudoti. Metodų naudojimas
ar nenaudojimas valdomas specialiai paren-
kamais (opciniiais) lelavimo parametrais. At-
eityje Lemuoklio naudotojas lelavimo para-
metrus galės keisti per specialų meniu, nu-
spaudęs klavišą „Nurodom parenkamus le-
lavimo parametrus“ (žr. 1 pav.). Kaip jau
minėta, kol kas šis klavišas dar neveikia, ir
Lemuoklis lemuoja, naudodamas programiš-
kai „pagal nutylėjimą“ priskirtas šių lelavimo
parametrų reikšmes.

Aprašysim kiekvieną iš šių lelavimo dau-
giaprasmiškumo mažinimo metodų: kaip jis
veikia, kaip keičia lelavimo rezultata, ar Le-
muoklis taiko jį „pagal nutylėjimą“.

Visus lelavimo daugiaprasmiškumo ma-
žinimo metodus galima suskirstyti į dvi grupes.
Pirmąją grupę sudarytų metodai, ku-
riuos naudodamas Lemuoklis gali ignoruoti
tam tikras vienos leksemos homoformas. Tai
metodai, toliau straipsnyje sąlygiškai pava-
dinti „šauksmininko užtušavimas“, „nekaitomų
vardažodžių skaičiaus, linksnio (giminės) užtu-
šavimas“ ir „III asmens skaičiaus užtušavimas“.
Antrojoje grupėje būtų metodai, kuriuos nau-
dodamas Lemuoklis gali ignoruoti kai kurias
lemas iš nustatytųjų vienai žodžio formai hi-
potetinių lemų sąrašo. Tie metodai toliau va-
dinami „sudaiktavardėję dalyviai ir būdvardžiai“,
„subūdvardėję dalyviai“, „padaryčiai“, „yra-bū-
na-būva“, „žodėlyčiai“, „su“, „iš“, „nors“, „mano,
tavo, savo“, „mūsai“, „mūsas“, „pats“, „visti“,
„viską“ ir „abu“.

Šauksmininko užtušavimas

Jei žodžio forma yra būdvardis, skaitvardis,
įvardis ar dalyvis (bet ne daiktavardis) ir jei
jai buvo nustatyti du hipotetiniai linksniai –
vardininkas ir šauksmininkas, tai Lemuoklis
tokiais atvejais gali palikti abu arba palikti
tik vardininką.

„Pagal nutylėjimą“ tokiais atvejais palie-
kamas tik vardininkas, pavyzdžiui:

```
elitinės
bdvr <elitinis>
bdvr nelygin.1 neįvardž mot.gim vnsk K
bdvr nelygin.1 neįvardž mot.gim dgsk V
pimas
sktv <vienas>
sktv kelintin nelygin.1 neįvardž vyr.gim vnsk V
sktv kelintin nelygin.1 neįvardž mot.gim dgsk G
visi
įvrd <visas>
įvrd vyr.gim dgsk V
laikyti
bndr <laikyti (-o, -ė)>
bndr nesngr
dlv nesngr neveik.r būt.kart.1 neįvardž vyr.gim
dgsk V
```

Tų pačių žodžių formų lelavimas ne „pa-
gal nutylėjimą“:

```
elitinės
bdvr <elitinis>
bdvr nelygin.1 neįvardž mot.gim vnsk K
bdvr nelygin.1 neįvardž mot.gim dgsk V
bdvr nelygin.1 neįvardž mot.gim dgsk Š
pimas
sktv <vienas>
sktv kelintin nelygin.1 neįvardž vyr.gim vnsk V
sktv kelintin nelygin.1 neįvardž vyr.gim vnsk Š
sktv kelintin nelygin.1 neįvardž mot.gim dgsk G
visi
įvrd <visas>
įvrd vyr.gim dgsk V
įvrd vyr.gim dgsk Š
laikyti
bndr <laikyti (-o, -ė)>
bndr nesngr
dlv nesngr neveik.r būt.kart.1 neįvardž vyr.gim
dgsk V
dlv nesngr neveik.r būt.kart.1 neįvardž vyr.gim
dgsk Š
```

Nekaitomų vardažodžių skaičiaus,
linksnio (giminės) užtušavimas
Daiktavardžiai lietuvių kalboje yra linksniuo-
jami ir gali būti kaitomi skaičiumi. Pasisko-
linti iš kitų kalbų nekaitomi daiktavardžiai,
tokie kaip *taksi*, *ledi*, Lemuoklio naudojamų
skaitmeninių kaitybos modelių supratimu, tu-
ri, kaip ir visi daiktavardžiai, linksnio ir gi-
minės kategorijas. Todėl, lemuodamas ne-
kaitomą daiktavardį, Lemuoklis jam automa-
tiškai generuoja 14 hipotetinių gramatinių
reikšmių, t.y. 7 vienaskaitos ir 7 daugiskai-
tos linksnius. „Pagal nutylėjimą“ paliekama

tik viena gramatinė reikšmė, užtušavus skaičiaus ir linksnio kategorijas kaip neapibrėžtas tokiems daiktavardžiams.

Panašiai ir nekaitomiems būdvardžiams, tokiems kaip *bordo*, *bruto*, Lemuoklis generuoja 28 hipotetines gramatines reikšmes: 7 vienaskaitos linksnius plius 7 daugiskaitos ir visa tai dukart – atskirai vyriškajai giminei ir atskirai moteriškajai. „Pagal nutylėjimą“ paliekama tik viena, užtušavus giminės, skaičiaus ir linksnio kategorijas kaip neapibrėžtas.

Žemiau pateikiamas žodžio formos *amplua* dvejetainis lemavimas.

„Pagal nutylėjimą“ neapibrėžtos šiam daiktavardžiui kategorijos užtušuojamos:

amplua

dktv <*amplua*>

dktv vyr. gim

Jei neužtušuojama:

amplua

dktv <*amplua*>

dktv vyr. gim vnsk V

dktv vyr. gim vnsk K

dktv vyr. gim vnsk N

dktv vyr. gim vnsk G

dktv vyr. gim vnsk Įn

dktv vyr. gim vnsk Vt

dktv vyr. gim vnsk Š

dktv vyr. gim dgsk V

dktv vyr. gim dgsk K

dktv vyr. gim dgsk N

dktv vyr. gim dgsk G

dktv vyr. gim dgsk Įn

dktv vyr. gim dgsk Vt

dktv vyr. gim dgsk Š

III asmens skaičiaus užtušavimas

Jei žodžio forma yra tiesioginės arba tariamosios nuosakos veiksmazodis, ir jai nustatomas trečiasis asmuo (pvz., *dirba*), tai Lemuoklis tokią žodžio formą apibūdina dvejetainis: kaip vienaskaitą ir kaip daugiskaitą. Kadangi lietuvių kalboje tiesioginės ir tariamosios nuosakos veiksmažodžių trečiojo asmens vienaskaitos ir daugiskaitos formos sutampa, „pagal nutylėjimą“ skaičiaus reikšmė užtušuojama, ir paliekamas tik vienas gramatinis apibūdinimas be skaičiaus kategorijos.

Žemiau pateikiamas žodžių formų *atsirado*, *lentu*, *susikūrė* dvejetainis lemavimas.

„Pagal nutylėjimą“ skaičiaus užtušavimas atliekamas:

atsirado

bndr <*atsirasti* (-nda, -do)>

vksm sngr tiesiog.nuos būt.kart.1 IIIIasm

lentu

** bđvr <*lentas*>

bđvr nelygin.1 neįvardž vyr.gim dgsk K

bđvr nelygin.1 neįvardž mot.gim dgsk K

** bndr <*lenti* (-emia, -ėnė)>

vksm nesngr tariam.nuos IIIIasm

dlv nesngr neveik.r būt.kart.1 neįvardž vyr.gim dgsk K

dlv nesngr neveik.r būt.kart.1 neįvardž mot.gim dgsk K

susikūrė

bndr <*susikurti* (-uria, -ūrė)>

vksm sngr tiesiog.nuos būt.kart.1 IIIIasm

Tų pačių žodžių formų lemavimas, jei užtušavimas nenaudojamas (ne „pagal nutylėjimą“):

atsirado

bndr <*atsirasti* (-nda, -do)>

vksm sngr tiesiog.nuos būt.kart.1 vnsk IIIIasm

vksm sngr tiesiog.nuos būt.kart.1 dgsk IIIIasm

lentu

** bđvr <*lentas*>

bđvr nelygin.1 neįvardž vyr.gim dgsk K

bđvr nelygin.1 neįvardž mot.gim dgsk K

** bndr <*lenti* (-emia, -ėnė)>

vksm nesngr tariam.nuos vnsk IIIIasm

vksm nesngr tariam.nuos dgsk IIIIasm

dlv nesngr neveik.r būt.kart.1 neįvardž vyr.gim dgsk K

dlv nesngr neveik.r būt.kart.1 neįvardž mot.gim dgsk K

susikūrė

bndr <*susikurti* (-uria, -ūrė)>

vksm sngr tiesiog.nuos būt.kart.1 vnsk IIIIasm

vksm sngr tiesiog.nuos būt.kart.1 dgsk IIIIasm

Sudaiktavardėję dalyviai ir būdvardžiai

Kilusius iš dalyvių ar būdvardžių daiktavardžius (*nelabasis*, *pėstysis*, *pažįstamas*, *miegamas*, *laukiamasis*, *mylimasis*, *dirbantysis*, *suaukęs*, *jaunasis*, *sėjamoji*, *lygiosios* ir pan.) Lemuoklis atpažįsta kaip daiktavardžius tais atvejais, jei jie kaip daiktavardžiai buvo įtraukti į (DLKŽ, 1972) žodyną. Kadangi dažniausiai žodyne DLKŽ yra ir atitinkami veiksmažodžiai ar būdvardžiai, iš kurių kilę šie daiktavardžiai, tai tokių žodžių formas Lemuoklis atpažįsta dvejetainis – ir kaip daiktavardžius, ir kaip dalyvius ir/ar būdvardžius. Todėl Lemuoklis tokioms formoms nustato

kelias hipotetines lemas: daiktavardį plius atitinkamą veiksmąžodį ir/ar būvdvardį. Lemuoklyje numatyta galimybė tokias dalyvių ir/ar būvdvardžių formas, nors pagal DLKŽ ir sudaiktavardėjusias, t. y. atliekančias daiktavardžių funkcijas, visada laikyti dalyviais ar būvdvardžiais – atmesti jų daiktavardiškąją lemą.

„Pagal nutylėjimą“ toks atmetimas nedaromas, ir tokios formos apibūdinamos dvejo-pai – ir kaip daiktavardžiai, ir kaip dalyviai / būvdvardžiai. Čia pastebėsime, kad Lemuoklio prototipas MAN, naudotas rengiant „Dažninį dabartinės rašomosios lietuvių kalbos žodyną“ (DDRLKŽ, 1997:X ir 1998:XII), tokį atmetimą darė.

Žemiau pateikiamas žodžio formos *pažįstamų* dvejopas lemavimas.

„Pagal nutylėjimą“ toks daiktavardžių atmetimas nedaromas:

pažįstamų

** dktv <pažįstamas>

dktv vyr.gim dgsk K

** dktv <pažįstama>

dktv mot.gim dgsk K

** bđvr <pažįstamas>

bđvr nelygin.l neįvardž vyr.gim dgsk K

bđvr nelygin.l neįvardž mot.gim dgsk K

** bndr <pažinti (-įsta, -ino)>

dlv nesngr neveik.r esam.l neįvardž vyr.gim dgsk K

dlv nesngr neveik.r esam.l neįvardž mot.gim dgsk K

Jei toks daiktavardžių atmetimas atliekamas (ne „pagal nutylėjimą“):

pažįstamų

** bđvr <pažįstamas>

bđvr nelygin.l neįvardž vyr.gim dgsk K

bđvr nelygin.l neįvardž mot.gim dgsk K

** bndr <pažinti (-įsta, -ino)>

dlv nesngr neveik.r esam.l neįvardž vyr.gim dgsk K

dlv nesngr neveik.r esam.l neįvardž mot.gim dgsk K

Subūvdvardėję dalyviai

Kilusius iš dalyvių būvdvardžius (*sukalbamas, užkrečiamas, mylimas, neatidėliotinas, pažymimasis, suaugęs* ir pan.) Lemuoklis atpažįsta kaip būvdvardžius tais atvejais, jei tie žodžiai yra būvdvardžiais įtraukti į (DLKŽ, 1972) žodyną. Dažniausiai žodyne DLKŽ yra ir atitinkami veiksmąžodžiai, iš kurių kilę šie da-

lyviški būvdvardžiai. Todėl tokių žodžių formos Lemuoklis atpažįsta dvejo-pai – ir kaip būvdvardžius, ir kaip dalyvius; taigi joms Lemuoklis nustatys po dvi hipotetines lemas: būvdvardį plius atitinkamą veiksmąžodį. Lemuoklyje numatyta galimybė tokias veiksmąžodžių formas, nors pagal DLKŽ ir subūvdvardėjusias, t. y. atliekančias būvdvardžių funkcijas, visada laikyti dalyviais – atmesti būvdvardiškąją lemą.

„Pagal nutylėjimą“ toks atmetimas nedaromas, ir tokios formos apibūdinamos dvejo-pai – ir kaip veiksmąžodžio formos – dalyviai, ir kaip būvdvardžiai. Čia reikia pasakyti, kad Lemuoklio prototipas MAN, naudotas rengiant „Dažninį dabartinės rašomosios lietuvių kalbos žodyną“ (DDRLKŽ, 1997:X ir 1998:XII), tokį atmetimą darė.

Žemiau pateikiamas žodžio formos *sukalbama* dvejopas lemavimas.

„Pagal nutylėjimą“ toks būvdvardžių atmetimas nedaromas:

sukalbama

** bđvr <sukalbamas>

bđvr nelygin.l neįvardž vyr.gim vnsk G

bđvr nelygin.l neįvardž mot.gim vnsk G

** bndr <sukalbėti (-a, -ėjo)>

dlv nesngr neveik.r esam.l neįvardž vyr.gim vnsk G

dlv nesngr neveik.r esam.l neįvardž mot.gim vnsk G

Jei toks būvdvardžių atmetimas daromas (ne „pagal nutylėjimą“):

sukalbama

bndr <sukalbėti (-a, -ėjo)>

dlv nesngr neveik.r esam.l neįvardž vyr.gim vnsk G

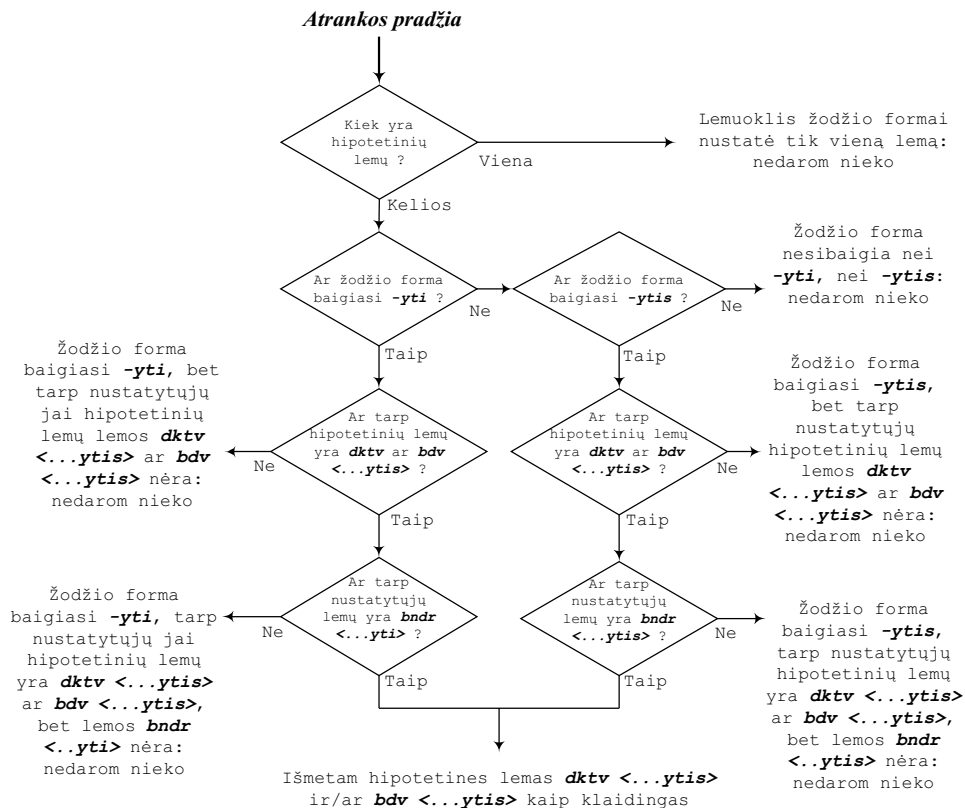
dlv nesngr neveik.r esam.l neįvardž mot.gim vnsk G

Padaryčiai

Trumpai paaiškinsime *padaryčių* vardu pavadinantą problemą. Kaip jau aprašyta, žodžių formų atpažinimui ir gramatiniam apibūdinimui Lemuoklis turi *GF* žodyną, kurį sudaro lietuviškų žodžių šaknys ir skaitmeniniai gramatinės kaitybės modeliai, t.y. kompiuterinė morfologija. *GF* šaknų sąrašo pagrindą sudaro žodžiai iš DLKŽ. Žodžių kaitība kompiuterinėje morfologijoje vienu specifiniu aspektu kiek skiriasi nuo tradicinėse lietuvių kalbos gramatikose aprašytosios žodžių kaitybės. Lemuoklio kompiuterinėje mor-

fologijoje kaitybai priskirti ir kai kurie patys bendriausi, labiausiai paplitę lietuviškų žodžių darybos reiškiniai. Šitaip darybą pakeisti į kaitybą privertė DLKŽ sandara. Į šį žodyną neįtrauktos ištisos didelės išvestinių žodžių grupės laikant, kad žodyno naudotojams tokių žodžių reikšmės bus savaime suprantamos iš pateikiamų pamatinių žodžių aprašymo. Šitaip, pvz., į DLKŽ žodyną nedėtos daiktavardžių mažybinės formos (*stalas*→*stalelis*, *staliukas*), veiksmazodžių priesagos *-inėti* vediniai (*bėgti*→*bėginėti*). Todėl kompiuterinėje GF žodyno morfologijoje tokius reguliariosios darybos reiškinius teko įtraukti į atitinkamų pamatinių žodžių kaitybą, antraip Lemuoklis nepažintų daug lietuviškų žodžių formų. Apie tai, būtent kokie darybos atvejai priskirti kaitybai ir kaip tai padaryta, čia plačiau neaprašinėsimė, dabar tik aptarsime *padaryčių* problemą, iškilusią dėl Lemuoklio sugebėjimo atpažinti įvairiausias teoriškai įmanomas mažybinės vardažodžių formas.

Vyriškosios giminės daiktavardžių ir būdvardžių mažybinų priesagos *-ytis* vedinių (iš *sūnus* – *sūnytis*, iš *bernas* – *bernytis*, iš *mažas* – *mažytis*) vienskaitos vardininkai gale turi *-ytis*, o šauksmininkai *-yti*. Šitie šauksmininkai ortografiškai gali sutapti su atsitiktinai panašių veiksmazodžių bendratimis. Pavyzdžiui, daiktavardžių *aplankyti* (mažybinis nuo *aplankas*), *aprašyti* (nuo *aprašas*), *blaškyti* (nuo dktv. *blaškas*, žr. DLKŽ – „išblokštų rugių pėdas“), *ginčyti* (nuo *ginčas*) šauksmininkų formos, gale turėdamos *-ti*, ortografiškai sutampa su panašių veiksmazodžių (*aplankyti*, *aprašyti*, *blaškyti*, *ginčyti*) bendratimi. Kai kada dar ir tokių mažybinų vardažodžių vardininkai, gale turėdami *-tis*, sutampa su panašių sangražinių veiksmazodžių bendratimi (*blaškytis*, *ginčytis*). Todėl tokioms ortografinėms formoms Lemuoklis nustato po dvi hipotetines lemas: vieną „daiktavardis ...*ytis*, o kita „veiksmazodis ...*yti*“ arba „veiksmazodis ...*ytis*“. Tačiau Lemuoklis tokius pa-



2 schema. Vienos žodžio formos hipotetinių lemu atranka pagal kriterijų *-yti (s)*

daryčius gali neutralizuoti, teikdamas pirmenybę lemai-veiksmažodžiui ir ignoruodamas kaip klaidingą lema – daiktavardį. *Padaryčių* neutralizavimo algoritmas pavaizduotas 2 *chemoje*.

Žemiau pateikiamas žodžių formų *padaryti*, *laužyti*, *krikštytis*, *laikytis*, *ginčytis* dvejo-
pas lemavimas.

„Pagal nutylėjimą“ *padaryčiai* neutralizuojami:

padaryti

bndr <padaryti (-o, -ė)>

bndr nesngr

dlv nesngr neveik. r būt. kart. l neįvardž vyr. gim

dgs V

laužyti

bndr <laužyti (-o, -ė)>

bndr nesngr

dlv nesngr neveik. r būt. kart. l neįvardž vyr. gim

dgs V

krikštytis

bndr <krikštytis (-ijasi, -ijosi)>

bndr sngr

laikytis

bndr <laikytis (-osi, -ėsi)>

bndr sngr

ginčytis

bndr <ginčytis (-ijasi, -ijosi)>

bndr sngr

Ne „pagal nutylėjimą“ *padaryčiai* paliekami:

padaryti

** dktv <padarytis>

dktv vyr. gim vnsk š

** bndr <padaryti (-o, -ė)>

bndr nesngr

dlv nesngr neveik. r būt. kart. l neįvardž vyr. gim

dgs V

laužyti

** dktv <laužytis>

dktv vyr. gim vnsk š

** bndr <laužyti (-o, -ė)>

bndr nesngr

dlv nesngr neveik. r būt. kart. l neįvardž vyr. gim

dgs V

krikštytis

** dktv <krikštytis>

dktv vyr. gim vnsk V

** bndr <krikštytis (-ijasi, -ijosi)>

bndr sngr

ginčytis

** dktv <ginčytis>

dktv vyr. gim vnsk V

** bndr <ginčytis (-ijasi, -ijosi)>

bndr sngr

Yra - būna - būva

Veiksmažodis *būti*, Lemuoklio žiniomis, gali turėti tris skirtingas paradigmas: 1) *būti-yra-buvo*, 2) *būti-būna-buvo* ir 3) *būti-būva-buvo*. Tokiu būdu kaitybinės *būti* formos, iš kurių nesimato, kuriai paradigmai jos priklausytų (t. y. iš kurių negalima spręsti apie esamojo laiko formą), gali priklausyti bet kuriai iš šių trijų lemų: 1) *būti (yra, buvo)*, 2) *būti (-ūna, -uvo)* ir 3) *būti (-ūva, -uvo)*. Tokiems atvejams, kada iš lemuojamos veiksmažodžio *būti* formos apie esamojo laiko formą spręsti negalima, Lemuoklyje numatyta galimybė pateikti tik pirmąją *būti* lema, ignoruojant dvi kitas teoriškai įmanomas.

„Pagal nutylėjimą“ (2) ir (3) lemos tokiais atvejais nepateikiamos, pavyzdžiui:

būti

** bndr <būti (-yra, -buvo)>

bndr nesngr

dlv nesngr neveik. r būt. kart. l neįvardž vyr. gim

dgs V

** bndr <būtas>

bndr nelygin. l neįvardž vyr. gim dgs V

yra

** bndr <irti (-yra, -iro)>

vksm nesngr tiesiog. nuos esam. l IIIasm

** bndr <būti (-yra, -buvo)>

vksm nesngr tiesiog. nuos esam. l IIIasm

buves

bndr <būti (-yra, -buvo)>

dlv nesngr veik. r būt. kart. l neįvardž vyr. gim vnsk V

Ne „pagal nutylėjimą“ tokiais atvejais pateikiamos visos trys *būti* lemos:

būti

** bndr <būti (-ūna, -uvo)>

bndr nesngr

dlv nesngr neveik. r būt. kart. l neįvardž vyr. gim

dgs V

** bndr <būti (-ūva, -uvo)>

bndr nesngr

dlv nesngr neveik. r būt. kart. l neįvardž vyr. gim

dgs V

** bndr <būti (-yra, -buvo)>

bndr nesngr

dlv nesngr neveik. r būt. kart. l neįvardž vyr. gim

dgs V

yra

** bndr <irti (-yra, -iro)>

vksm nesngr tiesiog. nuos esam. l IIIasm

** bndr <būti (-yra, -buvo)>

vksm nesngr tiesiog. nuos esam. l IIIasm

buves

** bndr <būti (-ūna, -uvo)>

dlv nesngr veik.r būt.kart.l neįvardž vyr.gim
vnsk V

** bndr <būti (-ūva, -uvo)>

dlv nesngr veik.r būt.kart.l neįvardž vyr.gim
vnsk V

** bndr <būti (-yra, -buvo)>

dlv nesngr veik.r būt.kart.l neįvardž vyr.gim
vnsk V

Žodelyčiai

Iš kelių žodžių sudarytų junginių, kurie į DLKŽ žodyną įtraukti kaip savarankiški leksiniai vienetai, Lemuoklis nelemuoja. Aptikęs tekste tokius žodžių junginius kaip *kaži (n) kas, iš anksto*, Lemuoklis lemuoja atskirai *kaži (n)*, atskirai *kas*, atskirai *iš*, atskirai *anksto*. Taip yra dėl to, kad Lemuoklis, kaip jau minėta, kol kas apskritai teksto žodžių formas lemuoja tik autonomiškai po vieną.

Visokių smulkių žodelyčių lemavimą apsunkena tai, kad, viena vertus, dauguma jų priklauso dažniausiai vartojamų kalbos žodžių kategorijai, o kita vertus, tokie žodelyčiai dažnai gali turėti kelias kalbos dalies kategorijas. Pavyzdžiui, žodynas DLKŽ 1972 nurodo, kad žodelyčiai *ir, čia, kaip* gali eiti dalelyte, jungtuku arba prieveiksniu; žodelyčiai *lig, ligi* – dalelyte, jungtuku ir prielinksniu; žodelyčiai *jau, vis* – dalelyte ir prieveiksniu, žodelyčiai *nors, ne, lyg* – dalelyte ir jungtuku.

Lemuoklio prototipas MAN, naudotas rengiant „Dažnių dabartinės rašomosios lietuvių kalbos žodyną“ (DDRKŽ, 1997:X ir 1998:XII), tokių žodelyčių kalbos dalies daugiareikšmiškumą išsprendė trimis gana griežtomis taisyklėmis:

1) Žodelyčius *jau, dar, vėl, beveik, ypač, vis, kažin, kaži* traktavo tik kaip dalelytes. Atkreipsime dėmesį, kad *dar, vėl, beveik, ypač* į žodyną (DLKŽ, 1972) įtraukti tik kaip prieveiksniai.

2) Žodelyčių *ir, čia, kaip* netraktavo kaip prieveiksnių.

3) Visais kitais žodelyčių kalbos dalies nevienareikšmiškumo atvejais kalbos dalies kategorija parenkama šitaip:

- jei žodelytis gali būti ir prieveiksniu, ir dalelyte, paliekama tik dalelyte,

- jei žodelytis gali būti ir prielinksniu, ir dalelyte, paliekamas tik prielinksniu,

- jei žodelytis gali būti ir prieveiksniu, ir prielinksniu, paliekamas ir tas, ir tas.

Šią pastarąją (3) taisyklę taiko ir Lemuoklis „pagal nutylėjimą“.

Žodelyčių *ir, čia, kaip, jau, vis* lemavimas „pagal nutylėjimą“:

```
ir
** dll <ir>
** jngt <ir>
čia
** dll <čia>
** jngt <čia>
kaip
** dll <kaip>
** jngt <kaip>
jau
dll <jau>
vis
dll <vis>
```

Žodelyčių *ir, čia, kaip, jau, vis* lemavimas ne „pagal nutylėjimą“:

```
ir
** prvks <ir>
** dll <ir>
** jngt <ir>
čia
** prvks <čia>
** dll <čia>
** jngt <čia>
kaip
** prvks <kaip>
** dll <kaip>
** jngt <kaip>
jau
** prvks <jau>
** dll <jau>
vis
** prvks <vis>
** dll <vis>
```

Su

Raidžių seka *su* Lemuoklis atpažįsta ir kaip prielinksnių, ir kaip daiktavardį (*su*: prancūzų smulkūs pinigėliai).

„Pagal nutylėjimą“ *su* negali būti daiktavardis:

```
su
prln <su>
prln
```

Lemuojant ne „pagal nutylėjimą“:

```
su
** dktv <su>
dktv vyr.gim
** prln <su>
prln
```

Iš

Raidžių seką *iš* Lemuoklis atpažįsta ir kaip prielinksni, ir kaip veiksmažodžio būsimąjį laiką nuo *ižti*.

„Pagal nutylėjimą“ *iš* negali būti veiksmažodžiu:

š
prln <iš>
prln

Lemuojant ne „pagal nutylėjimą“:

š
** bndr <ižti (-yžta, -ižo)>
vksm nesngr tiesiog.nuos būs.l IIIasm
** prln <iš>
prln

Nors

Raidžių seką *nors* Lemuoklis atpažįsta ir kaip dalelytę/jungtuką, ir kaip veiksmažodžio būsimąjį laiką nuo *norti*.

„Pagal nutylėjimą *nors* visada dalelytė arba jungtukas:

nors
** dll <nors>
dll
** jngt <nors>
jngt

Lemuojant ne „pagal nutylėjimą“:

nors
** bndr <norti (-sta, -o)>
vksm nesngr tiesiog.nuos būs.l IIIasm
** dll <nors>
dll
** jngt <nors>
jngt

Mano, tavo, savo

Formas *mano, tavo, savo* Lemuoklis atpažįsta dvejopai: 1) kaip savybinių įvardžių kilmininkus; antraštiniai pavidalai atitinkamai būtų *mano tavo savo*; 2) kaip įvardžių vyriškosios giminės vienaskaitos kilmininkus; antraštiniai pavidalai atitinkamai būtų *manas tavas savas*.

„Pagal nutylėjimą“ Lemuoklis ignoruoja (2) variantą ir palieka tik pirmąjį:

savo
įvrd <savo>
įvrd K

Lemuojant ne „pagal nutylėjimą“:

savo
** įvrd <savas>

įvrd neįvardž vyr.gim vnsk K

** įvrd <savo>
įvrd K

Mūsai

Raidžių seką *mūsų* Lemuoklis atpažįsta ne tik kaip įvardį, bet ir kaip daiktavardžio kilmininką (yra tokie *mūsai*).

„Pagal nutylėjimą“ tokiais atvejais *mūsų* tik įvardis:

mūsų
įvrd <aš>
įvrd dgsk K

Lemuojant ne „pagal nutylėjimą“:

mūsų
** įvrd <aš>
įvrd dgsk K
** dktv <mūsai>
dktv vyr.gim dgsk K

Mūsas

Raidžių seką *mūsų* Lemuoklis atpažįsta ne tik kaip įvardžio *aš* formą, bet ir kaip įvardžio *mūsas* (kaip *tavas, manas*) formą.

„Pagal nutylėjimą“, jei Lemuoklis aptinka, kad žodžio formos *lema* gali būti arba įvardis *aš* arba įvardis *mūsas*, tai palieka tik *lema aš*

mūsų
įvrd <aš>
įvrd dgsk K

Lemuojant ne „pagal nutylėjimą“:

mūsų
** įvrd <aš>
įvrd dgsk K
** įvrd <mūsas>
įvrd neįvardž vyr.gim dgsk K
įvrd neįvardž mot.gim dgsk K

Pats

Pats gali būti ir įvardis, ir daiktavardis.

„Pagal nutylėjimą“ Lemuoklis *lema-daiktavardį* ignoruoja:

pats
įvrd <pats>
įvrd neįvardž vyr.gim vnsk V

Lemuojant ne „pagal nutylėjimą“:

pats
** dktv <pats>
dktv vyr.gim vnsk V
** įvrd <pats>
įvrd neįvardž vyr.gim vnsk V

Visti

Gana reto veiksmožodžio *visti* (*vysta, viso*) kai kurios būtojo kartinio ir būsimojo laiko formos savo ortografinė išraiška sutampa su žymiai dažnesniais įvardžiu,rieveiskmiu ir dalelyte *visi, visai, vis*.

„Pagal nutylėjimą“ *visi, visai vis* negali būti veiksmožodžiu, pavyzdžiui:

visai

** įvrd <visas>

įvrd mot.gim vnsk N

** prvks <visai>

prvks

Lemuojant tą patį *visai* ne „pagal nutylėjimą“:

visai

** įvrd <visas>

įvrd mot.gim vnsk N

** bndr <visti (-ysta, -iso)>

vksm nesngr tiesiog.nuos būt.kart.l vnsk II-asm

** prvks <visai>

prvks

Viską

„Pagal nutylėjimą“ *viską* – visada įvardis, ignoruojant, kad gali būti ir esamojo laiko dalyvis nuo *viskėti*:

viską

įvrd <viskas>

įvrd G

Lemuojant ne „pagal nutylėjimą“:

viską

** įvrd <viskas>

įvrd G

** bndr <viskėti (-a, -ėjo)>

dlv nesngr veik.r esam.l neįvardž vyr.gim djsk V

dlv nesngr veik.r esam.l neįvardž bevrd.gim

Abu

Žodžio formai *abu* Lemuoklis nustato tiek lemą skaitvardį, tiek lemą daiktavardį (žr. TŽŽ, 1985 – *abu*: [arab. tėvas], musulmonų šalyse – valdytojas, turto turėtojas).

„Pagal nutylėjimą“ daiktavardis *abu* ignoruojamas:

abu

sktv <abu>

sktv kiekin vyr.gim dvisk V

sktv kiekin vyr.gim dvisk G

Lemuojant ne „pagal nutylėjimą“:

abu

** sktv <abu>

sktv kiekin vyr.gim dvisk V

sktv kiekin vyr.gim dvisk G

** dktv <abu>

dktv vyr.gim

PARENKAMIEJI (OPCINIAI)

LEMAVIMO PARAMETRAI

Lemavimo parametrais Lemuokliui nurodoma, kaip lemuoti tam tikrų specifinių kategorijų lietuviškas žodžių formas. Šie nurodymai gali būti keičiami (parenkami) pagal poreikius. Kol kas šiems parametrms priskirtos jų reikšmės „pagal nutylėjimą“, ir naudotojai jų keisti neturi galimybės. Kitose Lemuoklio versijose prieš lemuojant naują failą šiuos parametrus bus galima kaitaloti.

Kandidatai į romėniškus skaitmenis Kandidatais į romėniškus skaitmenis čia pavadintos simbolių sekos, susidedančios vien iš didžiųjų raidžių V, I ir X. Lemuoklis gali tokias raidžių sekas ignoruoti – nelaikyti jų žodžių formomis, jų nelemuoti ir nerašyti į rezultatų failą. „Pagal nutylėjimą“ kandidatus į romėniškus skaitmenis Lemuoklis ignoruoja. Pavyzdžiui, jei parametras bus nustatytas „neignoruoti“ (ne „pagal nutylėjimą“), tai raidžių seką XII Lemuoklis sulemus šitaip:

XII

<??>

Jei parametras nustatytas „ignoruoti“, kaip kad yra „pagal nutylėjimą“, tai, aptikęs tokia raidžių seką, Lemuoklis jos netraktuos kaip galimos žodžio formos ir todėl iš viso nelemuos.

Veikslo kategorijos užtušavimas

Veiksmožodžių veikslas lietuvių kalboje morfologinės raiškos neturi. DLKŽ žodynuose veikslas daugumai veiksmožodžių nenudomas, ir tai savaime suprantama. Daugelis veiksmožodžių gali turėti ir eigos, ir įvykio veikslo reikšmes. Be to, tas pats veiksmožodis esamajame laike gali būti eigos veikslo, o būtajame – įvykio, plg. *ateina-atėjo, laimi-laimėjo* (DLKG, 1997:288-290). Dėl šių veikslo savybių Lemuoklis, lemuodamas kiekvieną žodžio formą atskirai, veiksmožodžio formų veikslą nustato nepatikimai. Todėl Lemuok-

lyje yra numatyta galimybė lemuojant veikslų kategoriją ignoruoti, tas ir daroma pagal nutylėjimą.

Jei nusprendžiama neignoruoti, tai veikslų kategorija Lemuoklis bando nustatyti gana mechaniškai, pavyzdžiui, visiems priešdėliniams veiksmazodžiams priskirdamas įvykio veikslą ir pan.

Žemiau pateikiamas žodžių formų *susiklostant*, *nutraukti*, *puolė*, *veržimusi* dvejetainis lemovimas.

„Pagal nutylėjimą“ veikslas ignoruojamas:

susiklostant

bndr <susiklostyti (-to, -tė)>

padlv sngr esam.l

nutraukti

bndr <nutraukti (-ia, -ė)>

bndr nesngr

dlv nesngr neveik.r būt.kart.l neįvardž vyr.gim dgsk V

puolė

bndr <pulti (-uola, -uolė)>

vksm nesngr tiesiog.nuos būt.kart.l IIIasm

veržimusi

dktv <veržimasis>

dktv sngr vyr.gim vnsk Įn

Jei veikslas neignoruojamas (ne „pagal nutylėjimą“):

susiklostant

bndr <susiklostyti (-to, -tė)>

padlv įvykio vksl sngr esam.l

nutraukti

bndr <nutraukti (-ia, -ė)>

bndr įvykio vksl nesngr

dlv įvykio vksl nesngr neveik.r būt.kart.l neįvardž vyr.gim dgsk V

puolė

bndr <pulti (-uola, -uolė)>

vksm eigos vksl nesngr tiesiog.nuos būt.kart.l IIIasm

veržimusi

dktv <veržimasis>

dktv eigos vksl sngr vyr.gim vnsk Įn

Neįvardžiuotinumo, nelyginamojo laipsnio, nesangražiškumo ir tiesioginės nuosakos užtušavimas

Lemuoklio gramatiniuose apibūdinimuose pateikiama informacija apie nelyginamąjį laipsnį, tiesioginę nuosaką, neįvardžiuotinę ar nesangražinę formas gali pasirodyti bereikalinga, perteklinė. Lemuoklis gali šią informaciją užtušuoti. Tada įvardžiuotinumo požymis bus nurodomas tik įvardžiuotinėms formoms, laips-

nio kategorija – tik laipsnį turinčioms formoms, sangražiškumas – tik sangražinėms formoms ir nuosakos kategorija – tik tariamosios ir liepiamosios nuosakos formoms. „Pagal nutylėjimą“ toks gramatinių reikšmių ir požymių užtušavimas neatliekamas.

Pavyzdžiui, *Virvelinės*, *būdavo* lemovimas „pagal nutylėjimą“:

Virvelinės

bdvr <virvelinis>

bdvr nelygin.l neįvardž mot.gim vnsk K

bdvr nelygin.l neįvardž mot.gim dgsk V

būdavo

bndr <būti (-yra, -buvo)>

vksm nesngr tiesiog.nuos būt.d.l IIIasm

Tų pačių žodžių formų lemovimas ne „pagal nutylėjimą“, ignoruojant:

Virvelinės

bdvr <virvelinis>

bdvr mot.gim vnsk K

bdvr mot.gim dgsk V

būdavo

bndr <būti (-yra, -buvo)>

vksm būt.d.l IIIasm

Visais kitais parenkamais parametrais Lemuokliui užduodami lemovimo daugiaprasmiškumo mažinimo metodų naudojimo režimai. Šie parametrai aprašyti ankstesniajame skyrelyje „Lemavimo daugiaprasmiškumo mažinimas“.

TECHNINĖS LEMUOKLIO CHARAKTERISTIKOS

Lemuoklis skirtas IBM tipo personaliniams kompiuteriams, kuriuose įdiegta Microsoft Windows NT, Windows 95 ar aukštesnės versijos 32 bitų operacinė sistema. Lemuoklį sudaro trys kompiuteriniai failai: programa (45 KB exe tipo failas), lemovimui reikalingų programinių funkcijų biblioteka (400 KB dll tipo failas) ir lietuvių kalbos leksikos bei gramatikos duomenų bazė (2 MB lex tipo failas).

Lemavimo greitis daugiausia priklauso nuo kompiuterio procesoriaus greičio, taip pat šiek tiek nuo pasirinkto lemovimo režimo (ar nustatom tik lemuojamų žodžių formų antraštinius pavidalus, ar ir formų gramatinius apibūdinimus). Žemiau pateikiamas Lemuoklio lemovimo greitis, išmatuotas lemuojant įvairius tekstus trim skirtingo galingumo kompiuteriais:

1. Procesorius Intel Pentium (100 MHz), 64 MB RAM: 13–17 tūkst. žodžių formų per minutę;
2. Procesorius Intel PII 233MMX (233 MHz), 96 MB RAM: 18–20 tūkst. žodžių formų per minutę;
3. Procesorius ATX Intel Celeron (433 MHz), 64 MB RAM: 30–32 tūkst. žodžių formų per minutę.

LEMUOKLIO TOBULINIMO PERSPEKTYVOS

Kuriant šią pirmąją Lemuoklio versiją, stengiasi kuo greičiau turėti realiai dirbantį produktą. Kai kuriais techniniais aspektais programa liko ne iki galo išbaigta, ir jos praktinis naudojimas sukelia tam tikrų nepatogumų. Lemuoklis dar neturi instaliavimo dalies. Norėdamas įdiegti Lemuoklį kompiuteryje, vartotojas turi pats rankiniu būdu sukurti reikiamus katalogus kompiuterio kietajame diske ir ten įrašyti Lemuoklio failus, taip pat patikrinti, ar kompiuterio Windows sistemoje netrūksta sisteminų failų, kurie būtini Lemuoklio funkcionavimui, ir pan. Paleidus failo lemavimo užduotį, Lemuoklis dirba, kol sulemuoja visą failą, sustabdyti ar nutraukti šį procesą dabartinėje versijoje vartotojas neturi galimybių. Šiuos ir kitus panašius techninius Lemuoklio neišbaigtumus naujose versijose galima lengvai pašalinti.

Kaip minėta, lemuojamuose tekstuose Lemuoklis atpažįsta tik nekirčiuotas raides, be to, nemoka sujunti perkeltųjų žodžio dalių į vieną žodį. Šias problemas ateityje taip pat nesunku išspręsti. Kiek sudėtingiau būtų Lemuoklį išmokyti lemuoti įvairesnius failus, pvz., doc tipo ar failus, paruoštus duomenų bazių valdymo sistemų priemonėmis. Galbūt tokių Lemuoklio sugebėjimų kol kas ne labai ir reikia – juk visada galima norimą tekstą perrašyti vadinamuoju gryno teksto (*text-only*) formatu, kurį Lemuoklis supranta.

Lemavimo rezultatų pateikimo srityje nesunkiai galima realizuoti galimybę vartotojui pačiam laisvai pasirinkti jam priimtinus gramatinius žymėjimus, taip pat lemavimo rezultatų įrašymą ne tik gryno teksto pavidalu, bet ir HTML bei SGML formatais. Ateityje, praktiškai eksploatuojant Lemuoklį, be

abejo, gali kilti minčių ir apie kitokius panašaus pobūdžio Lemuoklio patobulinimus.

Didžiausias Lemuoklio lingvistinių sugebėjimų trūkumas yra lemavimo nevienareikšmiškumas. Atsiradus kompiuterinėms technologijoms, manipuliuosiančioms nors ir pačiomis paprasčiausiomis lietuviškų žodžių derinimo jų junginiuose taisyklėmis, būtų galima žymiai sumažinti lemavimo perteklinį daugiaprasmiškumą. Kol kas tokios technologijos dar nesukurtos. Autoriaus nuomone, norint sukurti pakankamai galingas sakinių sintaksinės struktūros kompiuterinės analizės priemones, bus tiesiog būtina įdiegti kompiuteriui sugebėjimą ortografinėse žodžių formose atpažinti visas teoriškai įmanomas gramatinės homoformas. Lemuoklis *GF* su žodynu ir remdamasis morfologija, pagalba tą daro jau dabar. Taigi lemavimo daugiaprasmiškumas, kuris dabar gali būti traktuojamas kaip trūkumas, ateities lietuvių rašytinės kalbos kompiuterinės analizės sistemose išvirs į privalumą.

Lemuoklio kalbinių žinių bazė ateityje turėtų būti perdirbta, paliekant vieną kompiuterinį žodyną – leksikoną. Šio žodyno kompiuterinę morfologiją reikėtų papildyti galūnių nutrupėjimo reiškiniiais, tikrinių daiktavardžių kaityla ir daryba, taip pat pašalinti dabar joje esančių trūkumus. Toks kompiuterinės morfologijos pertvarkymas negali būti ir nėra greitai padaromas dalykas. Dabar *GF* morfologijos kartotekinį variantą sudaro apie 6000 kortelių, kuriose yra apie 60 tūkst. skaitmenų. Visa ši kartoteka modifikuojant morfologiją turi būti atitinkamai perskaičiuota ir pernumeruota, jos duomenys perrašyti į atitinkamus failus ir tik po to jau galima specialia programine įranga sugeneruoti naują kompiuterinę morfologiją. Taigi Lemuoklio morfologinių žinių bazės modernizavimas gali ir užtrukti.

Straipsnyje trumpai supažindinta su kompiuterinio manipuliavimo lietuviškų žodžių gramatinėmis formomis galimybėmis ir problematika. Šios galimybės ir problematika aptartos žodžių lemavimo aspektu. Kompiuterinis lemavimas daugiausia turbūt taikomas tekstų lingvistikoje kalbos tyrinėjimams. Tačiau automatiška žodžių formų gramatinė analizė bei sintezė reikalinga, be abejo, ne tik tekstų lingvistikai. Šiuo metu pasaulyje intensyviai kuriamos naujos informacinės

technologijos, kurioms tokia analizė bei sintezė yra būtina. Tai mašininio vertimo, bendravimo su kompiuteriais natūralia kalba, teksto supratimo bei informacijos iš jo gavimo, garsinės kalbos atpažinimo bei sinteza-

vimo ir kitos panašios technologijos. Lemuoklyje įdiegtas leksikos ir morfologijos žinių kompiuterizavimas, autoriaus nuomone, gali būti panaudotas ir kuriant tokias technologijas lietuvių kalbai.

LITERATŪRA

- AUTASYS – AUTASYS – *A Fully Automatic English Wordclass Analysis System*. Interneto puslapio adresas <http://www.phon.ucl.ac.uk/home/alex/project/tagging/tagging.htm>
- Būda, 1994 – Būda V. P. *Sudurtiniai ir priešdėlinės darybos žodžiai su tarptautiniais dėmenimis*. Vilnius, 1994.
- DDRLKŽ, 1997 – Grumadienė L., Žilinskienė V. *Dažninis dabartinės rašomosios lietuvių kalbos žodynas (mažėjančio dažnio tvarka)*. Vilnius, 1997.
- DDRLKŽ, 1998 – Grumadienė L., Žilinskienė V. *Dažninis dabartinės rašomosios lietuvių kalbos žodynas (abėcėlės tvarka)*. Vilnius, 1998.
- DLKG, 1994 – *Dabartinės lietuvių kalbos gramatika*. Vilnius, 1994.
- DLKG, 1996 – *Dabartinės lietuvių kalbos gramatika / Antrasis pataisytas leidimas*. Vilnius, 1996.
- DLKG, 1997 – *Dabartinės lietuvių kalbos gramatika / Trečiasis pataisytas leidimas*. Vilnius, 1997.
- DLKŽ, 1972 – *Dabartinės lietuvių kalbos žodynas / II papildytas leidimas*. Vilnius, 1972.
- EUSLEM – EUSLEM. *A lemmatizer/tagger for Basque*. Interneto puslapio adresas <http://ixa.si.edu.es/ingeles/dokument/EUSLEM.html>
- Kaplan, 1988 – Kaplan R. M. “Regular models of phonological rule systems”. *Alvey Workshop on Parsing and Pattern Recognition*. Oxford University, April 1988.
- Koskenniemi, 1983 – Koskenniemi K. *Two-Level Morphology: A General Computational Model for Word-Form Recognition and Production*. University of Helsinki, Department of General Linguistics. Publication No. 11. 1983.
- LexMorphDemo – *LexMorphDemo: Kompiuterinė programa, demonstruojanti automatišką lietuvių kalbos morfologinę analizę ir sintezę*. Interneto puslapis adresu <http://donelaitis.vdu.lt/LexMorphDemo>
- LKG, 1965 – *Lietuvių kalbos gramatika. I tomas*. Fonetika ir morfologija (daiktavardis, būdvardis, skaitvardis, įvardis). Vilnius, 1965.
- LKG, 1971 – *Lietuvių kalbos gramatika. II tomas*. Morfologija (veiksmažodis, prievieksmis, dalelytė, prielinksnis, jungtukas, jaustukas, iš-tiktukas). Vilnius, 1971.
- LKRŽ, 1948 – *Lietuvių kalbos rašybos žodynas*. Vilnius, 1948.
- Marcinkevičienė, 1997 – Marcinkevičienė R. „Tekstynų lingvistika ir lietuvių kalbos tekstynas“. *Lituanistica*, 1997. Nr. 1. 58–78 p.
- Marcinkevičienė, 2000 – Marcinkevičienė R. „Terminografija ir tekstynas“. *Terminologija*, 2000. Nr. 6. 5–22 p.
- Ritchie, 1992 – Ritchie Graeme D. “Languages Generated by Two-level Morphological Rules”. *Computational Linguistics*. 1992, 18, 1:41–59.
- SphinxSurvey – *SphinxSurvey – Lemmatizer Module*. Interneto puslapis adresu <http://www.lesphinx-developpement.fr/en/Products/Lemmatizer.htm>
- TŽŽ, 1985 – *Tarptautinių žodžių žodynas*. Vilnius, 1985.
- Zinkevičius, 1996 – Zinkevičius V. „Lietuvių kalbos morfologinių reiškinių kompiuterizavimas“. *Lietuvių katalikų mokslo akademijos suvažiavimo darbai*. 1996. XVI tomas, 155–162 p.
- Zinkevičius, 1996* – Zinkevičius V. „Lietuvių kalbos morfologinių reiškinių kompiuterizavimo lingvistiniai aspektai“. *Lietuvių katalikų mokslo akademijos suvažiavimo darbai*. 1996. XVI tomas, 148–154 p.

Gauta 2000 10 12

Parenta 2000 10 19

Vytautas ZINKEVIČIUS

MORPHOLOGICAL ANALYSIS WITH LEMUOKLIS

Abstract

Lemuoklis is a morphological analyzer, lemmatizer and tagger for Lithuanian. It assigns its lemma (or several hypothetical lemmas) to each token in a text and performs its morphological analysis. A

word form is characterized grammatically by a combination of properties with respect to 13 categories: part of speech, aspect, reflexiveness, voice, mood, tense, group, degree, definiteness, gender,

number, case and person. The program processes over 30,000 tokens per minute on ATX Intel Celeron (433 MHz, 64 MB RAM).

The database of lexical and grammatical information of the program consists of six lexicons. All lexicons are organized as letter trees with some information on the leaf nodes of the trees. Three of the lexicons store roots of Lithuanian words with the pointers to appropriate morphological rules. Two other lexicons store word forms without any morphological information. The last lexicon contains a list of abbreviations and acronyms.

In Lemuoklis, morphological rules are expressed in the form of digital tables. The tables represent graph structures that define both collections of affixes and collections of grammatical properties. Using morphological rules together with word-root lexicons enables us to analyse milliards of theoretically available Lithuanian written forms. In case when a surface form is homonymous, i. e.

it has several grammatical meanings, the programme gives a full grammatical characteristic for each possible homoform of the surface form. The author views grammatical disambiguation of forms as the subject of syntactic analysis which has not been performed so far. However, some methods are used to reduce the ambiguity without taking into account the context. Algorithm of disambiguation between diminutive nouns that have the inflectional ending *-yti(s)* and the respective verbal infinitive forms are presented. The disambiguation between proper and common nouns is performed using special lexicons that contain proper noun forms from Lithuanian corpora and other sources. Forms with shortened endings are quite common in Lithuanian texts. These forms are recognized in Lemuoklis by means of special lexicons that primarily were designed for the needs of Lithuanian spell-checking. The article gives a lot of examples of various categories of Lithuanian word forms tagged and lemmatized with Lemuoklis.