

LIETUVIŲ KALBOS TEKSTYNO MORFOLOGINĖS ANALIZĖS AUTOMATIZAVIMAS

Erika Rimkutė

*Vytauto Didžiojo universitetas, Humanitarinis fakultetas
Daukanto 28, LT-3000 Kaunas, Lietuva*

Šiame pranešime bus paaiškinta, kas yra tekstynas, kaip jis gali būti žymimas, kaip gramatiškai analizuojami kompiuteriniai tekstai. Didžiausias dėmesys skiriamas jau veikiančiai automatinei morfologinei analizei, kurią atlieka kompiuterinė programa Lemuoklis. Bus pristatyti šios programos privalumai ir trūkumai, problemų sprendimo būdai.

1 Įvadas

XX a. pabaigoje ištobulėjus kompiuteriams didelių pakeitimų įvyko ir gana stabilioje mokslo šakoje – kalbotyroje, kurioje tekstynų lingvistika, galima sakyti, padarė perversmą: leido visiškai kitaip pažvelgti į kalbą, ją analizuoti objektyviai, pateikti tikslesnius duomenis, kadangi atsiribojama nuo subjektyvios analizės, dažniausiai pagrįstos kalbos jausmu ir intuicija. Tekstynų lingvistika parodė visai kitą kalbos vaizdą, nes didelis ir ganėtinai reprezentatyvus kompiuteriu tvarkomas ir nuolat papildomas tekstynas atskleidžia kur kas didesnę kalbos vienetų vartojimo įvairovę, o statistika leidžia nustatyti būdingiausius, dažniausiai vartojamus atvejus ir atskirti juos nuo retesnių [2].

Vytauto Didžiojo universiteto dabartinės lietuvių kalbos tekstynas suartino gana tolimų specialybių žmones – informatikus ir lingvistus. Akivaizdu, kad tekstynų lingvistikos pažanga yra tiesiogiai susijusi su informatikos mokslo raida ir su efektyvesniu informatikų ir lingvistų bendradarbiavimu,- juk norint geriau aprėpti ir išanalizuoti didėjančią informacijos kiekį reikia naudoti vis pažangesnes lingvistines kompiuterines programas [6].

Glaudus informatikų ir lingvistų bendradarbiavimas ypač akivaizdus nagrinėjant automatiškai morfologiškai sužymėtą tekstą. Informatikų sukurtos programos gali anotuoti tekstus, bet be lingvistų kalbinės analizės ir taisyčių tokie tekstai nesuteikia daug informacijos, nelabai padeda gramatinei analizei, kadangi bent jau kol kas neįmanoma automatiškai panaikinti daugiaprasmybių, kurios trukdo analizei.

Darbas su morfologiškai tekstą pažyminčia kompiuterine programa paskatino parodyti, kad mūsų kalba nėra nedviprasmiška, lengvai suklasifikuojama į kalbos dalis. Toks sunkiai sugrupuojamų kalbos vienetų vaizdas ypač išryškėja tada, kai kompiuterinė programa, nesuvokdama konteksto, ryšių tarp žodžių, pateikia daugybę hipotetinių variantų. Todėl ir norisi, remiantis savo darbo patirtimi su morfologiškai anotuoju tekstu, parodyti, kokių iškyla problemų ir kaip jos galėtų būti sprendžiamos.

2 Tekstyno samprata

Tekstynas – bet koks, kad ir pats mažiausias, elektroninę formą turinčių tekstų rinkinys. Tačiau vienas žymiausių tekstynų lingvistikos atstovų J. Sinclairis tekstyną siūlo vadinti tik tokį tekstų rinkinį, kuris yra pakankamai didelis, matuojant pagal šių dienų kompiuterinių technologijų galimybes, ir sudarytas ne dėl kokio specialaus tyrimo, bet nepriklausomai nuo jo panaudojimo tikslų. Galima skirti net tris šio žodžio reikšmes: pati bendriausia tekstyno reikšmė yra „bet koks tekstų rinkinys“, dažniausia – „elektroninių tekstų rinkinys“, o griežčiausia, terminologiškiausia - „baigtinis elektroninių tekstų rinkinys, sudarytas taip, kad kuo geriau atspindėtų kalbą ar jos atmainą [3].

Tekstyną su tekstu sieja tai, kad jie sudaryti iš tekstų; gali sutapti teksto ir tekstyno apimtys. Tačiau tekstas analizuojamas išties, jis turi struktūrą: pradžią, vidurį, pabaigą, yra daugiau ar mažiau rišlus ir vientisas, o tekstynas neturi struktūros, tik sandarą. Jo neverta ir neįmanoma tirti tiesiogiai, skaityti taip, kaip teksto, o tik su programinėmis priemonėmis, įvairiais įrankiais. Taigi esminė teksto ir tekstyno skirtybė ne sandara, ne kalbų kiekis, net ne dydis ar reprezentatyvumas, bet su kompiliacine teksto prigimtimi susijusi jo tyrimo metodologija [3, 5].

Vytauto Didžiojo universiteto tekstynas buvo sumanytas kaip didelis neanotuotas ir nekoduotas, į skaitytoją orientuotas, daugiausiai išties periodikos ir knygų tekstų, daugiau bendro pobūdžio nei specialus, didelės temų ir kitokios įvairovės autentiškos rašytinės lietuvių kalbos tekstynas, kuris dabar jau yra pasiekęs 60 mln. žodžių apimtį ir toliau didinamas [3].

Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centras ketina ne tik didinti dabar esantį tekstyną, bet ir sukurti kelis mažesnius, iš kurių vienas bus morfologiškai anotuotas. Dėl tekstynų kodavimo ir anotavimo kyla daug lingvistų nesutarimų: ar teksto pažymos padeda analizuoti, ar trukdo.

3 Tekstyno žymėjimas

Tekstynas gali būti pažymimas ir anotuojamas teksto formatavimo ar analizės labui. Teksto žymėjimas arba kodavimas atsirado tekstynų gyvavimo pradžioje dėl to, kad senieji kompiuteriai nesugebėjo apdoroti teksto kaip teksto, bet jie mokėjo atpažinti pažymas. Dabartinės technologijos leidžia apdoroti tekstus jau be pažymų, bet senoji žymėjimo praktika išliko.

Koduoiant pažymimi tokie formalieji teksto struktūros elementai kaip nuorodos apie rašytinio teksto citatas, sąrašus, vardo raidžių ar kitokias santrumpas, akronimus, knygų įžangas ir apendiksus, skyrius, atskirų teksto dalių ir kitokius pavadinimus, tikrinius vardus. Sakytinės kalbos atveju galima žymėti pertraukinėjimus, kalbėjimą vienu metu, pauzes ir pan. [3]. Pavyzdžiui, Vytauto Didžiojo universiteto tekstyne yra pažymimimi užsienio kalbų žodžiai, pvz.:

```
<foreign lang=lot>
curruculum vitae
</foreign>
```

Toks pažymėjimas reiškia, kad šitas žodžių junginys yra iš lotynų kalbos. Taip pat žymimi Lemuoklio išskaidyti, nors prasmės atžvilgiu neskaidomi junginiai tokie kaip *be abejo*, *iš tikrųjų*, *kas nors*, *tam tikras*, *taip pat* ir panašūs. Taigi šie žymėjimai tvarkomame tekste atrodo taip:

```
<phr type=iiterp>          <phr type=prvks>
be abejo                  arba:  kada nors
</phr>                    </phr>
```

Dar žymimi tikriniai pavadinimai, pvz., *Juodoji jūra*, *Ivanas Rūstusis*, *Lietuvos ir Lenkijos karalystė* ir t. t., kadangi Lemuokliui išskaidžius tokius junginius negalima suprasti, kad tai tikrinis pavadinimas, pvz., junginį *Juodoji jūra* sulemavus į antraštines formas būdvardį *juodas* ir daiktavardį *jūra*, neparodoma, jog tai tikrinis pavadinimas, todėl šio junginio vientisumas tvarkomame sulemuotame tekste atrodo taip:

```
<name type=place>
Juodoji jūra
</name>
```

Taigi kodavimas daugiau susijęs su struktūriniais teksto elementais, tačiau jis apima ir interpretacijos atvejus, kai pažymos vietą ir pobūdį lemia subjektyvi tyrėjo nuomonė. Dar daugiau interpretacijos esama gramatinėse anotacijose, nes paprastai jos remiasi tradicine, daugeliu atvejų subjektyvia nuomone ir susitarimu paremta gramatika (*consensus grammar*). Gavęs gramatiškai anototą tekstą, kompiuteris netikrina gramatinų kategorijų, bet priima jas tokias, kokios jos yra, taigi ir analizės rezultatai daugeliu atvejų yra iš anksto nulemti, nes kompiuteris dirba su pažymomis ir ignoroja pačią kalbą. Kartais tokios pažymos gali padėti analizei, bet kur kas dažniau jai gali pakenkti. Be to, pažymos pažeidžia teksto vientisumą, jos yra žmogaus intervencija į tekstą. Tik per pažymas žvelgiama į tekstyno kalbą, o tai, kas nesužymėta, yra prarandama. Dar viena anotacijų yda – tekstas perkraunamas pažymomis, jį tampa sunkiau perdirbti, nes apimtis patrigubėja ar padidėja dar daugiau kartų. Anotauto teksto privalumas yra tas, kad jei anotacijos yra neklaidinančios, jos labai palengvina paiešką ir bet kokią lingvistinę tekstyno analizę [3, 4].

Tekstyno žymėjimą galima suskirstyti į tris etapus: pirmiausia grynas tekstas lemuojamas – pateikiama tekstyne pavartoto žodžio antraštinė forma, t. y. lema, pvz., forma *namą* sulemuojama į antraštinį pavidalą *namas*, t. y. vienaskaitos vardininko linksnį. Antrojo etapo metu gali būti pateikti žodžių formų gramatiniai apibūdinimai, reiškiami gramatinėmis kategorijomis, pvz., nurodoma, kad forma *namą* - tai daiktavardis, vyriškoji giminė, vienaskaitos galininkas. Šiuos žymėjimus gali atlikti ir jau veikianti morfologinės analizės programa Lmuoklis. Toliau turėtų būti išaiškinami keleriopas lemas turintys atvejai, nes, pvz., forma *laimės* gali būti sulemuota ir kaip *laimė* ir kaip *laimėti*. Tam tikslui reikalinga speciali programa, kurios pagrindinė funkcija būtų dviprasmybių panaikinimas (angliškas terminas – *ambiguity resolution*). Deja, tokios programos Lietuvoje dar nėra.

Tekstyno anotavimas neapsiriboja vien morfologija, todėl dar esama sintaksinių, semantinių, nusakančių skirtingas daugiareikšmio žodžio reikšmes, ir net tekstinių ar diskursinių žymių. Sudėtingiausias teksto apdorojimo etapas po anotavimo – sintaksinė analizė, kurią sudaro sintaksinius ryšius vaizduojančio medžio kūrimas kiekvienam teksto sakiniui. Ji atliekama su parseriu. Eama ir dar sudėtingesnių darbo su tekstu etapų ir juos atitinkančių priemonių: teksto generavimo, vertimo, santraukų automatinio rengimo sistemų, bet jos – jau nebe tekstynų, o kompiuterinės lingvistikos sritis [3].

4 Morfologinės analizės programa Lemuoklis

Šiuo metu Lietuvoje veikia tik morfologinės analizės programa Lemuoklis, automatiškai lemuojanti lietuviškas žodžių formas iš pradinių tekstinų failų į rezultatinius tekstinius failus. Programą sukūrė Vytautas Zinkevičius ir ji yra skirta VDU Kompiuterinės lingvistikos centro moksliniams lingvistiniams tyrinėjimams

automatizuoti. Programos pavadinimas yra atsiradęs nuo angliško termino *lemmatizing* (nuo žodžio *lemma*, daugiskaita *lemmata* – antraštinė, žodyninė leksemos forma) [7].

Lemuoklis žodžių formoms, kuriomis vadinami žodžių (leksemų) kaitybinių gramatinių formų rašytiniai pavidalai, nustato antraštinius pavidalus, t. y. tokias ortografines išraiškas, kuriomis leksemos įtraukiamos į žodynus. Vardažodžiams tai paprastai vienaskaitos vardininko, veiksmažodžiams – pagrindinės (bendratis, esamojo, būtojo kartinio laiko trečio asmens) formos. Lemuojant žodžio formą yra nustatomos visos galimos (hipotetinės) lemos ir visi galimi gramatiniai apibūdinimai.

Visus programavimo darbus atliko V. Zinkevičius. Jis sukūrė kompiuterinę žinių apie lietuvių kalbos leksiką ir gramatiką bazę, lingvistinės informacijos paieškos šioje bazėje ir jos panaudojimo automatiško lemavimo procesui metodiką bei programinę įrangą [7].

Štai taip atrodo nedidelė ištrauka iš Lemuoklio apdoroto teksto:

```
Tai
  įvrd <tas>
    įvrd neįvardž mot.gim vnsk N
    įvrd neįvardž bevrđ.gim
nebuvo
** vksm <nebūti(-ėra,-ebuvo)>
  vksm nesngr tiesiog.nuos būt.kart.1 IIIasm
** vksm <nebūti(-ūna,-uvo)>
  vksm nesngr tiesiog.nuos būt.kart.1 IIIasm
** vksm <nebūti(-ūva,-uvo)>
  vksm nesngr tiesiog.nuos būt.kart.1 IIIasm
** vksm <nebūti(-yra,-buvo)>
  vksm nesngr tiesiog.nuos būt.kart.1 IIIasm
vien
  dll <vien>
  dll
operos.
  dktv <opera>
  dktv mot.gim vnsk K
  dktv mot.gim dgsk V
  dktv mot.gim dgsk Š
Chorai,
  dktv <choras>
  dktv vyr.gim dgsk V
  dktv vyr.gim dgsk Š
dainos,
  dktv <daina>
  dktv mot.gim vnsk K
  dktv mot.gim dgsk V
  dktv mot.gim dgsk Š
šokiai,
** dktv <šokis>
  dktv vyr.gim dgsk V
  dktv vyr.gim dgsk Š
** bdvr <šokus>
  bdvr nelygin.1 neįvardž mot.gim vnsk N
** prvks <šokiai>
  prvks nelygin.1
```

5 Lemuoklio trūkumai

Iš pateiktos sulemuoto ir gramatiškai anotuoto teksto ištraukos matyti, kad Lemuokliui nepavyksta visiškai tiksliai morfologiškai sužymėti teksto. Problemų kyla dėl to, kad Lemuoklis nepažįsta į kitą eilutę perkeltų žodžių ir juos sulemuoja kaip atskirus žodžius; atpažįsta tik nekirčiuotas raides. Ši kompiuterinė programa kol kas dar neturi jokių priemonių automatiškai sintaksinei ar semantinei teksto analizei atlikti. Kiekvieną žodžio formą Lemuoklis 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. Taigi Lemuoklis negali išvengti dviprasmybių ir dažnai žodžio formai pateikia ne vieną, o kelias hipotetines lemas ir/ar kelis hipotetinius tos formos gramatinius apibūdinimus, pvz., formą *valstybės* gramatiškai apibūdina kaip mot. gim. vnsk kilminką (pvz., *valstybės* prezidentas), dgsk vardininką (pvz., kelios *valstybės* suteikė paramą) ir dgsk šauksmininką (pvz., mano mielosios *valstybės*) arba žodžio forma *jos*: tai mot. gim. įvardžio vnsk kilmininkas (pvz., *jos* namas) arba dgsk vardininkas (pvz., *jos* šiandien vėluoja); taip pat forma *jos* gali būti veiksmažodžio *joti* būsimojo laiko III asmuo (pvz., brolis *jos* į karą).

Kuriant kompiuterinę lietuvių kalbos morfologiją, pagrindiniais morfologiją aprašančiais šaltiniais laikyti Lietuvių kalbos gramatika (1965 ir 1971 metų leidimai). 1984-1990 m., kai buvo kuriama kompiuterinė morfologija, tai buvo išsamiausi akademiniai morfologijos darbai. Todėl dauguma morfologijos dalykų kompiuterinėje morfologijoje atspindimi būtent taip, kaip jie traktuojami tų metų gramatikose [7]. Tad kai kurių problemų kyla ir dėl to, kad šiek tiek pasikeitė dabartinės gramatikos ir žodynai, kiek kitaip traktuojamos kai kurios kalbos dalys, atsirado naujų žodžių ir pan.

Tačiau pagrindinis Lemuoklio trūkumas yra nesugebėjimas atskirti homonimijos, daugiaprasmybių, pvz., jei ši programa atpažįsta tikrinį daiktavardį ir jei gali sutapti šio daiktavardžio ir kito žodžio formos, tai Lemuoklis pateikia visus galimus variantus, pvz., analizuodamas tikrinį daiktavardį *Vieną* (pvz., apilankiau Austrijos sostinę – *Vieną*), nurodo, kad *vieną* dar gali būti ir būdvardis, ir skaitvardis (pvz., tai atsitiko *vieną* dieną), ir įvardis (pvz., pamačiau *vieną* žmogų). Forma *Palestinoje* gali būti ne tik tikrinio daiktavardžio *Palestina* vietininkas (pvz., *Palestinoje* vyksta neramumai), bet ir reikiamybės rūšies moteriškosios giminės dalyvio vietininkas (pvz., paukščiui *palestinoje* košėje buvo kelių rūšių kruopos). Teoriškai įmanoma, kad iš veiksmožodžio *palesti*, *palesa*, *palesė* kas nors sumanys padaryti reikiamybės dalyvį. Taigi Lemuoklis laikosi principo „geriau per daug, negu per mažai“ ir pateikia menkai tikėtinus variantus.

Dažniausiai kyla keblumų, kai sutampa keli to paties vardažodžio linksniai, t. y. dėl linksnių sinkretizmo, ar skirtingų kalbos dalių žodžių vienodai tariamos atskiros kaitybos formos, taigi kai Lemuoklis susiduria su homonimais ir homoformomis. Dirbant su šia programa išryškėjo tokios pagrindinės problemos:

5.1 Linksnių sinkretizmas

Labai dažnas moteriškosios giminės vardažodžių vnsk kilmininko ir dgsk vardininko bei šauksmininko formų sutapimas, pvz., mus žavi *naujos politinės situacijos* – dgsk vardininkas ir *naujos politinės situacijos* ekspertai – vnsk kilmininkas.

Neretai sutampa ir būdvardžių ar neveikiamosios rūšies dalyvių moteriškosios giminės vnsk vardininko, įnagininko ir bevardės giminės formos, pvz., *maža* mergaitė – vardininkas, su *maža* mergaite – įnagininkas, čia *maža* žmonių – bevardė giminė; arba *galima* sutartis – vardininkas, su *galima* sutartimi – įnagininkas, čia *galima* rūkyti – bevardė giminė.

Tokias formas gana dažnai galima rasti ir ištisiniame nesulemuotame tekste, kur jos atskiriamos remiantis kontekstu. Šnekamojoje kalboje tokių dviprasmybių išvengiama, kadangi dažnai vienaskaitos vardininko, įnagininko ir bevardės giminės ar vienaskaitos kilmininko ir daugiskaitos vardininko formos skiriasi kirčio vieta.

5.2 Homoformos

Daug įdomių, tiesiog neįtikėtinų homoformų galima rasti morfologiškai anotuotame tekste. Ko gero, daugelis ir be konteksto pamatę formą *kartu*, pasakys, jog tai prieveiksmis, reiškiantis *drauge*, pvz., į kiną eisime *kartu* su draugais, bet Lemuoklis nesitenkina nurodydamas tik vieną variantą; jis nurodo, kad *kartu* dar gali būti daiktavardžio *kartas* vnsk įnagininkas (pvz., džiaugėmės tuo *kartu*) arba būdvardžio *kartus* bevardė giminė (pvz., negaliu valgyti – *kartu*). Bet čia dar nesibaigia kalbiniai Lemuoklio resursai: menkai įtikėtina, bet įmanoma, kad iš veiksmožodžio *karti*, *karia*, *korė* gali būti padarytas būtojo laiko neveikiamosios rūšies dalyvis, kurio įnagininkas (pvz., *(pa)kartu* žmogumi visi baisėjosi), išskyrus kirčio vietą, sutaps su prieveiksniu *kartu*, būdvardžio *kartus* bevarde gimine ar daiktavardžio *kartas* įnagininku.

Panašių pavyzdžių sulemuotame tekste galima rasti daugybę, pvz., *metu* – tai daiktavardžio *metas* vnsk įnagininkas (pvz., tuo *metu* vyko karas) ar būdvardžio *metus*, reiškiančio *tolį metantis, kuris numeta nuo savęs jojiką*, pvz., *metus* arklys, [1] vnsk įnagininkas. *Metu* gali būti ir veiksmožodžio *mesti*, *meta*, *metė* tiesioginės nuosakos esamojo laiko vnsk I asmuo (pvz., *metu* tau kamuolį).

Formos *valdžią* lema nurodomas ne tik daiktavardis *valdžia* (pvz., *valdžią* paėmė jėga), bet ir būdvardis *valdus*: jo moteriškosios giminės vnsk galininkas visiškai sutampa su nurodyta forma (pvz., vargas turėti *valdžią* žmoną). Pamatęs formą *mažiau*, vargu ar kas pasakys, jog čia gali būti daiktavardžio *mažius* vnsk šauksmininkas (pvz., ateik čia, *mažiau*), nes daugiau tikimybių, kad *mažiau* yra prieveiksminio *mažai* aukštesnysis laipsnis (pvz., šį kartą išleidau *mažiau* pinigų).

Tik skaitytojas gali pasirinkti teisingą variantą, nes Lemuoklis nepajėgus atskirti, ar *staigiai* yra prieveiksmis (pvz., *staigiai* pašoko), ar būdvardžio *staigus* moteriškosios giminės vnsk naudininko forma (pvz., buvau nepasiruošęs *staigiai* reakcijai). Lemuoklis neatkirs *pergalės* – daiktavardžio *pergalė* vnsk kilmininko (pvz., mes siekėme *pergalės*) ir veiksmožodžio *pergalėti*, *pergali*, *pergalėjo* būsimąjo laiko III asmens (pvz., manau, jis tave *pergalės*).

Labai dažnai pasitaiko, kad sutampa daiktavardžio ir veiksmažodžio formos. Be jau minėtų, gana dažnai tekste randama sutapusių daiktavardžio *kova* vnsk kilmininkas (pvz., *kovos* įkarštis) ir veiksmažodžio *kovoti*, *kovoja*, *kovojo* būsimąjo laiko III asmens formos (pvz., jis *kovos* už laisvę) arba daiktavardžio *atvejas* vnsk įnagininkas (pvz., šiuo *atveju* niekas nepadės) ir veiksmažodžio *atvyti*, *atveja*, *atvijo* esamojo laiko I asmuo (pvz., iš miško *atveju* kiški).

5.3 Homografai

Kai kurios sutampančios formos skiriasi kirčio vieta, bet Lemuokliui tai nieko nepadeda – jis vis tiek neskiria *karaliaus* (pvz., *karaliaus* sostas) – daiktavardžio *karalius* vnsk kilmininko ir *karaliaus* (pvz., valdovas dar ilgai *karaliaus*) – veiksmažodžio *karaliauti*, *karaliauja*, *karaliavo* būsimąjo laiko III asmens; *seniai* (pvz., susėdo *seniai* ant suoliuko) – daiktavardžio *senis* dgsk vardininko ir *seniai* (pvz., tai įvyko *seniai*) – prieveiksmio; *perėjo* (pvz., vaikas *perėjo* per gatvę) – veiksmažodžio *pereiti*, *pereina*, *perėjo* ir *perėjo* (pvz., višta ilgai *perėjo* kiaušinių) – veiksmažodžio *perėti*, *peri*, *perėjo*.

5.4 Retų žodžių pateikimas

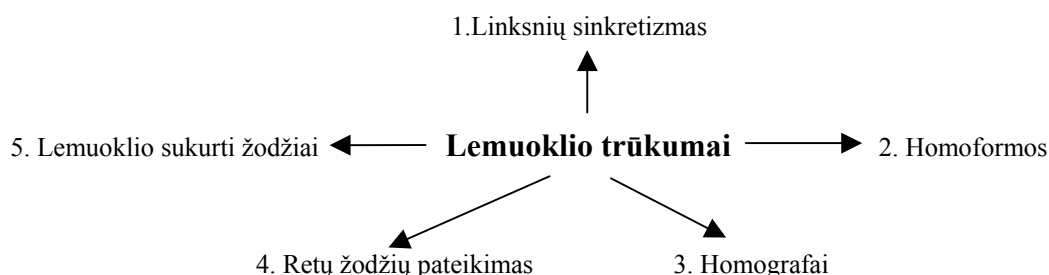
Kadangi Lemuoklis naudojasi visais į jį įdėtais žodyno resursais, tai pateikia be galo neįtikėtinų ir retų variantų. Vargu ar kas nurodys, kad forma *šlėktų*, t. y. daiktavardžio *šlėkta* dgsk kilmininkas (pvz., *šlėktų* nepasitenkinimas), gali būti ir veiksmažodžio *šlėkti*, *šlėkia*, *šlėkė*, reiškiančio *lieti*, *laistyti* [1], tariamoji nuosaka (pvz., *šlėktų* su vandeniu, jei būtų karšta). Kažin ar kam ateis į galvą, kad formos *rytų* lema gali būti ne tik daiktavardis *rytai* (pvz., *rytų* Europa), bet ir veiksmažodis *ryti*, *ryja*, *rijo*, nes *rytų* gali būti šio veiksmažodžio tariamosios nuosakos III asmuo (pvz., vilkas mašto, kad *(pra)rytų* Raudonkepuraitę). Lemuoklis nurodo, kad *pati* ne tik įvardžio *pats* moteriškosios giminės vnsk vardininkas (pvz., mergaitė jau *pati* apsirengia), bet ir daiktavardis *pati*, reiškiantis *žmona* (pvz., jo *pati* buvo ragana). Forma *sunkiau* – tai arba prieveiksmio *sunkiai* aukštesnysis laipsnis (pvz., žiemą *sunkiau* anksti keltis), arba veiksmažodžio *sunkti*, *sunkia*, *sunkė* būtojo kartinio laiko vnsk I asmuo (pvz., vasarą *sunkiau* sultis).

5.5 Lemuoklio sukurti žodžiai

Dar daugiau neaiškumų sukuria ir pats Lemuoklis, nes jis kartais padaro neegzistuojančių daiktavardžių, pvz., šalia prieveiksmio *aiškiai* (pvz., *aiškiai* pasakė) nurodo daiktavardį *aiškis*, šalia *puikiai* (pvz., *puikiai* atrodo) – daiktavardį *puikis*. Lemuoklis yra sukūręs tokių daiktavardžių kaip *reikėjas*, *išryškėjas*, *nenorėjas*, *laba*, *kokis*, *artis*, *pastovis*, *spartis*, *remtis* ir pan.

Kai kuriuos dviprasmiškumus Lemuoklis gali sumažinti. „Pagal nutylėjimą“ Lemuoklis ignoruoja tam tikras vienos leksemos homoforas: jis gali „užtušuoti“ šauksmininką, nekaitomų vardažodžių skaičių, linksnį, giminę; veiksmažodžių trečiojo asmens skaičių; atsako „padaryčių“, t. y. teoriškai įmanomų, bet praktiškai nevarojamų mažybinių daiktavardžių, kurių šauksmininkas sutampa su veiksmažodžių bendratimi, pvz., koks gražus tu, *padaryti* (nuo daiktavardžio *padarytis*), ar galiu įeiti į vidų, *pastatyti* (nuo daiktavardžio *pastatytis*), kodėl nešildai, *laužyti* (nuo daiktavardžio *laužytis*) ir pan.

Pagrindinius Lemuoklio trūkumus galima pateikti tokia schema:



6 Problemų spręsdimo būdai

Su panašiomis problemomis susiduria ir kitų kalbų lemuokliai bei gramatiniai anotatoriai. Lemavimo nevienareikšmiškumas juose dažnai išsprendžiamas ar sumažinamas statistiniais tikimybiniais metodais,

panaudojus duomenis apie leksemų ir/ar gramatinių reikšmių vartosenos dažnius. Bet V. Zinkevičiaus Lemuoklyje tokie problemos sprendimo metodai nenaudojami dėl to, kad šiuo metu lietuvių kalbos gramatinių reikšmių dažninių charakteristikų nėra iš kur paimti. Be to, vien žodžių gramatinių formų vartosenos dažninių charakteristikų žymėjimas ortografinio homonimiškumo sukeliama lemavimo problemų neišspręs, pvz., ortografinės formos *kalba* daiktavardinė homoforma (pvz., lietuvių *kalba*) yra gerokai dažnesnė už veiksmąžodinę (pvz., jis *kalba* labai garsiai), bet be gilesnės konteksto analizės vis tiek negalima vienareikšmiškai nuspręsti, ar *kalba* tekste yra veiksmąžodžio *kalbėti*, *kalba*, *kalbėjo* esamojo laiko III asmuo, ar daiktavardžio *kalba* vardininkas [7].

Geriausiai daugiaprasmiskumą padėtų išspręsti sintaksinė analizė, kuriai reikia lingvistų nustatytų taisyklių visumos, pvz., jei iš kairės eina prielinksnis *su*, tai forma *žmona* bus vienaskaitos įnagininkas, o ne vardininkas; jei šalia *metus* iš dešinės yra žodis *arklys*, vadinasi, tai bus būdvardžio, o ne daiktavardžio *metai* forma; šalia formos *bylos* esantis neigiamas veiksmąžodis, pvz., *bylos* nepavyko laimėti, gali signalizuoti, kad tai daiktavardžio *byla* kilmininkas, o jei iš kairės eina vardažodis, pvz., *jis*, *draugas*, *pažįstamas*, iš dešinės prielinksnis, pvz., *apie*, pvz., *jis bylos* apie senovę, tai, galima manyti, kad šiuo atveju bus veiksmąžodžio *byloti*, *byloja*, *bylojo* būsimojo laiko III asmuo. Didžiausia problema yra ta, kad lietuvių kalbos žodžių tvarka nėra griežta, todėl sunku nustatyti taisykles. Visas šias taisykles informatikai turėtų paversti kompiuteriui suprantama kalba. Taigi būtinas lingvistų ir informatikų bendradarbiavimas sintaksinės analizės srityje, kuri labai padėtų gramatinei automatinei tekstų analizei. Šioje perspektyvioje ir reikalingoje veiklos srityje gali daug nuveikti ir lingvistai, ir informatikai, tačiau jų bendras darbas būtų didelis indėlis į lietuvių kalbos tekstyno gramatinės analizės automatizavimą.

7 Išvados

Sraipsnyje trumpai supažindinta su nauja, perspektyvia kalbotyros šaka – tekstynų lingvistika, paaiškinta, kas yra tekstynas, kokia jo sandara, kaip gali būti žymimi tekstynai, kokie jų žymėjimų būdai, privalumai ir trūkumai. Pristatyta V. Zinkevičiaus sukurta kompiuterinė programa Lemuoklis, automatiškai lemuojanti ir anotuojanti lietuvių kalbos kompiuterinius tekstus. Paminėti ir suklasifikuoti homonimiški ir daugiaprasmiai kalbos vienetai, kurie išryškėja tik po automatinės morfologinės analizės. Pabandyta pateikti galimus šių problemų sprendimo būdus, iš kurių svarbiausiai yra sintaksinė analizė, kuriai būtinas lingvistų ir informatikų bendradarbiavimas.

Literatūros sąrašas

- [1] St. Keinys, J. Klimavičius, J. Paulauskas, J. Pikčilingis, N. Sližienė, K. Ulvydas, V. Vitkauskas Dabartinės lietuvių kalbos žodynas. *Mokslo ir enciklopedijų leidybos institutas*, 2000.
- [2] R. Marcinkevičienė Klausimas dėl klausimo, arba ką gali kompiuterinis tekstynas. *Darbai ir Dienos, Vytauto Didžiojo universiteto leidykla*, 1997, Nr. 5, P. 19-37.
- [3] R. Marcinkevičienė Tekstynų lingvistika (teorija ir praktika). *Darbai ir Dienos, Vytauto Didžiojo universiteto leidykla*, 2000, Nr. 24, P. 7-64.
- [4] Pokalbis su J. Sinclairiu Net ir bilijoniniai tekstynai yra per maži. *Darbai ir Dienos, Vytauto Didžiojo universiteto leidykla*, 2000, Nr. 24, P. 297-299.
- [5] E. Tognini Boneli Corpus Classroom Currency. *Darbai ir Dienos, Vytauto Didžiojo universiteto leidykla*, 2000, Nr. 24, P. 205-243.
- [6] A. Utka Kalbinė įranga ir jos galimybės. *Darbai ir Dienos, Vytauto Didžiojo universiteto leidykla*, 2000, Nr. 24, P. 275-285.
- [7] V. Zinkevičius Lemuoklis – morfologinei analizei. *Darbai ir Dienos, Vytauto Didžiojo universiteto leidykla*, 2000, Nr. 24, P. 246-273.

Summary

Automation of Lithuanian corpus morphological analysis

The article deals with a new trend – corpus linguistics. Some corpora can be annotated. In Lithuania a morphological analyzer Lemuoklis already analyses, lemmatizes Lithuanian words forms. The main problem is that this programme gives a full grammatical characteristic for each possible homoform in case when a word form is homonymous. Lemuoklis is not able to avoid homonymous and ambiguous forms. Syntactic analysis could help to solve this problem. It is necessary that computer specialists and linguists would work together in order to make automatic morphological analysis of Lithuanian corpus better.