

KOMPIUTERINĖ LINGVISTIKA/ COMPUTATIONAL LINGUISTICS

Automatizuotas lietuvių kalbos morfologinio daugiareikšmiškumo ribojimas

Erika Rimkutė, Gintarė Grigonytė

Anotacija. Straipsnyje pristatomi galimi morfologinio daugiareikšmiškumo (toliau žymima MD) ribojimo būdai. Tai svarbu padaryti, kad būtų galima pradėti automatinę sintaksinę analizę, kurti ir lietuvių kalbai pritaikyti mašininio vertimo sistemas. Daugumai kalbų yra sukurti morfologiniai analizatoriai, kurie ne tik pateikia lemas ir morfologines pažymas, bet kartu ir pasirenka reikiamą formą iš kelių galimų. Lietuvių kalbos morfologinis analizatorius Lemuoklis gali pateikti tik lemas ir morfologines pažymas, kurios dažnai yra daugiareikšmės. Taigi reikia ieškoti būdų MD-ui mažinti.

Vieni iš būdų MD-ui riboti yra automatiniai – tai dažniausiai statistiniai, loginiai metodai, kiti MD-o ribojimo būdai yra automatizuoti. Gauti gana geri automatinio MD-o ribojimo rezultatai – pavyko išspręsti apie 90% daugiareikšmių atvejų. Kitas vienareikšminimo būdas – automatizuotas MD-o ribojimas. Tai paprastai lingvistų sukurtų taisyklių pritaikymas kompiuterinėse programose, ribojančiose MD-ą. Automatizuotą MD-o ribojimą sudaro morfologiškai daugiareikšmių kaitomų ir nekaitomų kalbos dalių vienareikšminimas.

Riboti nekaitomų kalbos dalių MD-ą yra labai sudėtinga, nes problemiška skirti nekaitomas kalbos dalis. Norint sumažinti morfologiškai daugiareikšmių kalbos dalių skaičių, reikia taikyti įvairialypę analizę: reikalingos įvairių lygių ir įvairaus pobūdžio taisyklės. Labiausiai riboti nekaitomų kalbos dalių MD-ą padeda sakinio skyryba; statistiniai duomenys, gauti iš Dabartinės lietuvių kalbos tekstyno; sintaksinė sakinio analizė; semantika, išsamesnė viso teksto ar kelių gretimų sakinių analizė; pakeitimas kitais aiškiau vartojamais žodžiais; gretimų žodžių morfologinės pažymos; kaip vienas vienetas pažymėtos morfologinės samplaikos.

Kaitomų kalbos dalių MD-o ribojimas daugiausia pagrįstas sintaksine analize. Remiantis priklausomybių gramatikos principais buvo parašytos sintaksinės taisyklės ir pritaikytos kuriamame sintaksiniame analizatoriuje. Sintaksinės taisyklės skirstytos į dvi dalis: sakinio dalis ir sakinio dalių junginius apimančias taisykles. Prie taisyklių nurodyti būtini parametrai: priklausomybė, žodžių tvarka ir įsiterpimas. Remtasi morfologiniais ir kai kuriais semantiniais požymiais. Sintaksinės taisyklės naudojamos ne tik MD-ui riboti, bet bus pritaikytos automatinės sintaksinės analizės programoje.

Įvadas

20 a. atsiradusios naujos mokslo šakos – tekstynų lingvistika ir kompiuterinė lingvistika – paspartino įvairių kalbų kompiuterizacijos procesą. Lietuvoje gana neseniai pradėta sudaryti tekstynus (pvz., *Dabartinės lietuvių kalbos tekstynas* (DLKT), *morfologiškai anotuotas tekstynas* (toliau žymima MAT), *lygiagretusis tekstynas* (apie šiuos tekstynus daugiau žr. <http://donelaitis.vdu.lt>)), kurie leidžia analizuoti didelius kalbos kiekius pasirinktais aspektais (ir kiekybiškai, ir kokybiškai); taip pat imamasi kurti mašininio vertimo sistemą¹.

Šiame straipsnyje pristatomas vienas iš lietuvių kalbos automatinės analizės etapų – automatinė morfologinė analizė, pusiau automatiškai sudarytas morfologiškai anotuotas tekstynas ir jame išryškėjęs morfologinis daugiareikšmiškumas (MD). Apžvelgiami lietuvių kalbos MD-ui riboti pritaikyti įvairūs metodai. Išsamiau aprašomi automatizuoti MD-o ribojimo būdai.

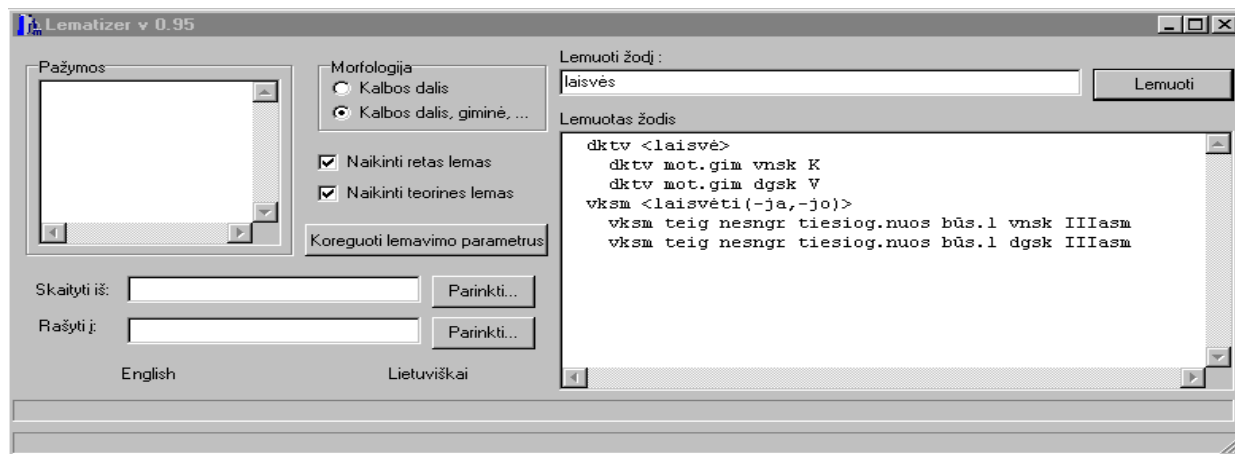
Morfologiškai anotuotas lietuvių kalbos tekstynas

MAT buvo rengtas Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centre 2000-2005 m. (2000-2003 m. MAT-o rengimą rėmė Valstybinė lietuvių kalbos komisija). Lietuvių kalbos MAT sudarytas pusiau automatiškai: naudota Vytauto Zinkevičiaus sukurta kompiuterinė programa *Lemuoklis*, pateikianti visas galimas lemas ir morfologines pažymas (plačiau žr. Zinkevičius 2000).

¹ Kompiuterinės lingvistikos centro vykdomas Lietuvos 2004-2006 metų bendrojo programavimo dokumento 3 prioriteto 3 priemonės „Informacinių technologijų paslaugų ir infrastruktūros plėtra“ projektas *Internetinė informacijos vertimo priemonė*; plačiau žr. http://donelaitis.vdu.lt/main.php?id=3&nr=1_3

Galima skirti tris pagrindinius automatinės morfologinės analizės etapus: pirmiausiai tekstas lemuojamas – pateikiama tekстыne pavartoto žodžio antraštinė forma, t. y. lema (pvz., daiktavardis *namas*, veiksmazodis *daryti* ir pan.). Antrojo etapo metu gali būti pateiktos žodžio formos morfologinės pažymos (pvz.: *namo* – vyriškosios giminės daiktavardžio vienaskaitos kilmininkas; *tikslų* – nelyginamojo laipsnio neįvardžiuotinio vyriškosios giminės būdvardžio vienaskaitos galininkas (lema *tikslus*) ir vyriškosios giminės daiktavardžio

daugiskaitos kilmininkas (lema *tikslas*)). Tolesnis etapas – kelių antraštinių pavidalus turinčių žodžių nustatymas, t. y. vienareikšminama (pvz., forma *laisvės* gali būti sulemuota ir kaip *laisvė*, ir kaip *laisvėti*). Tam tikslui reikalinga speciali programa, kurios pagrindinė funkcija būtų dviprasmybių panaikinimas (angliškas terminas – *ambiguity resolution*) (Zinkevičius et al. 2005). Dabartinėje *Lemuoklio* versijoje nėra įdiegta vienareikšminimo funkcija, todėl reikalingos atskiros MD-ą ribojančios programos.

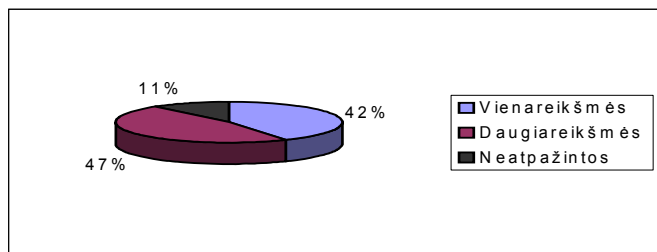


1 pav. Automatinės morfologinės analizės programos *Lemuoklis* langas

Morfologinis daugiareikšmiškumas

Analizuojant anotuotus tekstus matyti didžiulis ir sunkiai išsprendžiamas daugiareikšmiškumas, todėl ir vyksta diskusijos, ar apskritai verta anotuoti tekstynus, jei jų pažymos tokios daugiareikšmės ir dažnai klaidinančios. Yra manančių, pvz., J. Sinclairis, kad įterptos pažymos skaido tekstą, pažeidžia jo vientisumą (Sinclair 2000); kitų supratimu, pvz., N. Ide, L. Romary (Ide et al. 2004; Ide 1995; Erjavec et al. 1995), pažymos gali padėti tolesnei analizei.

Lietuvių kalbos MD yra didelis: MAT-e 47% žodžių yra morfologiškai daugiareikšmiai, 42% – vienareikšmiai, o 11% – visiškai morfologinio analizatoriaus neatpažinti žodžiai (plačiau apie MD-ą žr. Rimkutė 2003).



2 pav. Lietuvių kalbos vienareikšmės, daugiareikšmės ir neatpažintos žodžių ar žodžių formos morfologiškai anotuotame tekстыne

MD atsiranda dėl įvairių priežasčių: dėl *Lemuoklio* specifikos; dėl to, kad analizuojamos pavienės formos, o ne kontekstas; dėl žodynuose pateiktų leksinių ir gramatinių duome-

nų netikslumo ar trūkumo (išsamiau apie MD-o atsiradimo priežastis žr. Rimkutė et al. 2006).

Morfologinio daugiareikšmiškumo ribojimas

Ribojant MD-ą taikomi įvairūs metodai. Vienokios priemonės taikomos mažinant lemų, kitokios – morfologinių pažymų daugiareikšmiškumą. Tobulinant automatinę morfologinę lietuvių kalbos analizę, svarbu papildyti *Lemuoklio* leksikos bazę, ypač naujažodžiais ir sudurtiniais žodžiais. Lietuvių kalbos žodžius reikėtų suskirstyti į smulkesnes klases; taip pat reikėtų pakeisti kompiuterines žinias apie lietuvių kalbos darybą bei kaitybą, pvz., iš kokių žodžių daryti deminutyvai, kokioms formoms būdingi priešdėliniai vediniai ir pan. Taip pat svarbu nuosekliai aprašyti kai kurias kalbos dalis, gramatinės kategorijas, nes tai padės išvengti daugiareikšmiškumo, atsirandančio dėl netikslų kalbinių formuluočių, neaiškių apibrėžtų gramatinių požymių.

Lietuvių kalbos MD pradėtas riboti visai neseniai. Pirmieji MD-o ribojimo būdai buvo **automatiniai** – tai dažniausiai informatikų pritaikyti statistiniai ir loginiai metodai. Nuo 2001 m. pradėti taikyti statistiniai lietuvių kalbos MD-o ribojimo metodai: slaptieji Markovo modeliai (HMM), Viterbi algoritmas. Buvo pasiektas apytiksliai 85% efektyvumas (plačiau žr. Gričiūtė 2001; Gričiūtė et al. 2001). Vėliau (nuo 2002 m.) lietuvių kalbos MD-ui riboti imti taikyti ne tik statistiniai, bet ir kompiuterių loginio mokymosi metodai. Kaip loginis MD-o ribojimo metodas panaudotas ID3 algoritmas (plačiau žr. Bartkuvienė 2005; Rimkutė et al. 2004). Pritaikius minėtus

statistinius ir loginius metodus, pasiektas 90,69% lietuvių kalbos MD-o ribojimo tikslumas (Bartkuvienė 2005). Kol kas patikimiausiai veikia statistiniai metodai, nors juos dar reikia gerokai tobulinti; daugiau tyrimų reikalauja loginių metodų pritaikymas lietuvių kalbos MD-ui mažinti.

Automatiniu MD-o ribojimo būdu galima laikyti ir prieš morfologinę analizę atliekamą nerealiųjų homonimų panaikinimą. Nerealiaisiais homonimais vadinami tokie žodžiai, kuriuos automatiškai sugeneruoja *Lemuoklis* dėl specifinės analizės (pvz., daiktavardžiai *padarytis*, *norėtumas*, *galėjas*) ar dėl to, kad žodynuose yra pateikiami dabartinėje lietuvių kalboje jau nebevartojami žodžiai (pvz., daiktavardžiai *kokis*, *deja*).

Toliau šiame straipsnyje daugiausia dėmesio skirta **automatizuotiems** MD-o ribojimo būdams, kurie remiasi lingvistinėmis taisyklėmis pagrįstais metodais.

1. Automatizuoti morfologinio daugiareikšmiškumo ribojimo metodai

Automatizuotai ribojant MD-ą yra kuriamos taisyklės. Vienos iš jų pagrįstos statistika, kitos – morfologine ar sintaksine analize. MD-o rūšys yra nevienodo dažnumo (dažniausios yra veiksmažodžių trečiojo asmens, nekaitomų kalbos dalių sutapimai, vardažodžių linksnių sinkretizmas; plačiau žr. Rimkutė 2003), taigi skiriasi ir MD-o ribojimo efektyvumas. Kai kuriuos daugiareikšmius atvejus galima panaikinti gana greitai ir lengvai, o kai kurios morfologiškai daugiareikšmės formos lieka net ir pritaikius kelis metodus.

Ribojant kitų kalbų daugiareikšmiškumą, dažniausiai naudojamos šimto procentų tikslumu veikiančios lingvistų sukurtos vienareikšminančios taisyklės. Tokių taisyklių paprastai būna vos kelios, bet jos padeda išspręsti apie pusę visų kalbos dalies sutapimų (Nepil et al. 2001). Anglų kalbos MD-ui riboti buvo sukurta 3 300 taisyklių. Taip buvo pasiektas 77% tikslumas 1 mln. žodžių anototame *Brown*o tekстыne. Rengiant estų kalbos PMB-ą, buvo sudarytos 1 240 MD-ą ribojančios taisyklės ir apie 1 300 sintaksinius ryšius nustatančios taisyklės. Panaudojus minėtas taisykles, pavyko gauti 86,6% vienareikšminimo tikslumą (Bick et al. 2004, p. 127). Žinoma, skirtingoms kalboms reikia vis kitokių taisyklių. Geriausia tokias taisykles taikyti kartu su *n*-gramų metodu. Taip galima gerokai sumažinti MD-ą (Dutoit 1997, pp. 95-96).

Paprasčiausias statistiniais duomenimis pagrįstas automatizuotas MD-o ribojimo būdas yra nerealiųjų homoformų panaikinimas (žr. 1.1 skyrių). Sudėtingesni MD-o ribojimo metodai pritaikyti analizuojant nekaitomas ir kaitomas kalbos dalis. Pirmuoju atveju taikyta įvairialypė analizė (žr. 1.2 skyrių); kaitomų kalbos dalių MD-o ribojimas daugiausia pagrįstas sintaksine analize (žr. 1.3 skyrių).

Kituose šio skyriaus poskyriuose aprašomos lingvistinės taisyklės, pritaikytos ribojant MD-ą. Tos taisyklės yra įvairaus pobūdžio: vienos pritaikomos išsamiai išanalizavus tam tikrų žodžių vartoseną ir nusprendus, kad kai kuriais atvejais vienas ar kitas žodis gali eiti tik tam tikra kalbos dalimi arba gali būti vartojama tik viena kuri konkreti forma (tas ypač aktualu nagrinėjant nekaitomas kalbos dalis ir ribojant jų

MD-ą (žr. 2.2 skyrių)). Kitiems daugiareikšmiams atvejams panaikinti reikalingas aukštesnis lygmuo – sintaksinė analizė. Kai kurios formos vienareikšminamos tik atlikus išsamią morfologinę, sintaksinę ir semantinę analizę.

1.1. Automatizuotas nerealiųjų homoformų panaikinimas

Pats paprasčiausias automatizuotas MD-o ribojimo būdas – nerealiųjų homoformų, kurias sudaro 1) keli gana dažnai vartojami skirtingi žodžiai ar 2) kai kurios to paties žodžio formos, panaikinimas. Pirmuoju atveju būtų, pavyzdžiui, tokios homoformos:

Kartais – ši forma gali būtirieveiksmis (*interpretacija kartais atveda prie kraštutinumu*), dalyvis (*baisėjomės kartais, bet nepakartais žmonėmis*) ir daiktavardis (*Bailus atrodo dabar tas jos vyras, ne toks kaip anais kartais*). Pati dažniausiai tekстыne vartojama forma yrarieveiksmis, o dalyvis *kartais* galimas tik teoriškai (įprastesnis šios šaknies dalyvis su priešdėliu).

Jos – tai įvardžio (*jos emigravimo priežastys dažniausiai ekonominės*) arba veiksmažodžio (*iš ten su palyda jos į Merkinę*) forma. Daug dažnesnė įvardžio forma.

Motina – tai iš veiksmažodžio *moti* padaryto mot. gim. reikiamybės dalyvio vnsk. V, Įn ar beverd. gim. (*pamojo su motina skarele*) ir daiktavardžio *motina* vnsk. V, Įn ar Š forma (*štai mano motina*). Akivaizdu, kad dabartinėje lietuvių kalboje nevartojamas šis reikiamybės dalyvis.

Sutampančias to paties žodžio formas (antrasis nerealiųjų homoformų atvejis), iš kurių viena yra dažnai vartojama, o kita – retai, iliustruoja tokie pavyzdžiai:

Galima – neveik. r. esam. l. dalyvio mot. gim. vnsk. V bei Įn. ir beverd. gim. (*kas yra netiesa, klaida, iš kur ji kyla, kodėl yra galima ir koks jos kriterijus; Pavaizduoti erdvinio nuoseklumo principą galima schema; tokių bandymų galima atlikti ir kolbą pripylus kito skysčio*). Analizuojant anotuotus tekstus pastebėta, kad beveik visada *galima* – tai beverd. gim. dalyvio forma.

Būti – bendratis ir neveik. r. būt. l. dalyvio vyr. gim. dgsk. V (*negali būti; būti ir nebūti dalykai*). Dažniausiai *būti* vartojama kaip bendratis, o ne kaip dalyvis.

Panašaus pobūdžio pavyzdžių yra daugybė. Tokių nerealiųjų homoformų negalima visiškai panaikinti pačioje programoje. Pavyzdžiui, forma *kartu* yrarieveiksmis, dalyvis, būdvardis ir daiktavardis. Dažniausiai ši forma tekstuose vartojama kaiprieveiksmis; mažai tikėtina, kad *kartu* kokiame nors tekste bus dalyvis. Norint automatiškai panaikinti šią homoformų rūšį, reikėtų iš *Lemuoklio* leksikos duomenų išmesti veiksmažodį *karti*. Bet akivaizdu, kad kitos minėto veiksmažodžio formos tikrai yra vartojamos, todėl tokio pobūdžio MD sprendžiamas sudėtingiau. Reikia atsižvelgti į analizuojamo teksto specifiką – gali būti, kad tam tikrame tekste bus vartojamos mažai tikėtinos formos. Taigi daugiareikšmes formas galima pa-

naikinti pusiau automatiškai, t. y. Anotuotame tekste peržiūrimos ir panaikinamos konkrečios formos, o ne visa leksema.

Kai kuriais atvejais panaikinamos ne tik konkrečios žodžių formos, bet ir kai kurios gramatinės kategorijos ar tam tikros tų kategorijų dalys. Pavyzdžiui, automatinėje morfologinėje analizėje buvo atsisakyta veiksmažodžių veiksmo kategorijos, nes pastebėta, kad sunku skirti eigos ir įvykio veiksmo veiksmažodžius. Be to, ši kategorija nėra svarbi kituose automatinės morfologinės ar sintaksinės analizės etapuose. Taip pat nenurodomas šauksmininkas – taip pavyksta gerokai sumažinti linksnių sinkretizmą. Nenurodomos nekaitomų daiktavardžių ar būdvardžių giminės ir linksnio kategorijos.

Panašaus pobūdžio taisyklės, t. y. panaikinami retesni žodžiai ar žodžių formos, rečiau vartojamos gramatinės kategorijos, pvz., šauksmininkas, naudojamos ir kitoms kalboms. Panašios taisyklės buvo pritaikytos čekų kalbai kuriant MD-ą ribojančią programą, ir tai palengvino dalinę sintaksinę analizę, padėjo išvengti daugybės galimų klaidingų variantų (Žáčková 2002, 47 p.). Panašus retų žodžių ar jų formų filtravimas vyksta rusų kalbos IPPI tekстыne (Sičinava).

MD-ą taip pat automatizuotai galima sumažinti konkrečiame tekste. Analizuojant morfologiškai anotuotą tekstą į akis krenta tam tikram kalbos stiliui būdingos kolokacijos ar junginiai, kurie gana dažnai būna morfologiškai daugiareikšmiai. Pavyzdžiui, nagrinėjant morfologiškai anotuotus Lietuvos Respublikos Vyriausybės nutarimus, pastebėti labai dažni ir morfologiškai daugiareikšmiai junginiai: *Lietuvos Respublikos, Lietuvos Respublikos Vyriausybės, Lietuvos Respublika* ir t. t. Akivaizdu, kad šiuose junginiuose vartojamas daiktavardžio *respublika* vienaskaitos kilmininkas, o ne daugiskaitos vardininkas ar šauksmininkas. Taip pat daug dažnesnis žodžio *respublika* vienaskaitos vardininkas, o ne įnagininkas. Minėtuose junginiuose nevartojamas vienaskaitos šauksmininkas, todėl šio linksnio formas galima iš karto panaikinti visame tekste.

Taigi nerealiąsias homoformas morfologiškai anotuotuose tekstuose galima panaikinti automatizuotai. Po išsamesnės konteksto analizės nustatius, kurios morfologiškai daugiareikšmės formos yra dažnesnės, o kurios – visai nevartojamos tame kontekste, galima panaikinti gana didelę homoformų dalį.

1.2. Nekaitomų kalbos dalių morfologinio daugiareikšmiškumo ribojimas

Vienos iš dažniausių yra nekaitomų kalbos dalių homoformas, todėl daug dėmesio reikia skirti šiai MD-o rūšiai, jai riboti reikia įvairaus pobūdžio taisyklių. Skirti nekaitomas kalbos dalis labai problemiška, todėl norint nustatyti, kokia kalbos dalis pavartota, dažnai išsamiai reikia analizuoti vos ne kiekvieno nekaitomo žodžio vartoseną. Norint automatiškai išspręsti nekaitomų kalbos dalių daugiareikšmiškumą, reikalingos įvairių lygių ir įvairaus pobūdžio taisyklės. Riboti nekaitomų kalbos dalių MD-ą padeda: 1) saki-

nio skyryba; 2) viso DLKT-o statistiniai duomenys; 3) sintaksinė sakinių analizė; 4) semantika, išsamesnė viso teksto ar kelių gretimų sakinių analizė; 5) pakeitimas kitais aiškiau vartojamais žodžiais; 6) gretimų žodžių morfologinės pažymos; 7) kaip vienas vienetą pažymėtos morfologinės samplaikos.

Buvo bandyta taikyti taisykles, dažniausiai apimančias kelis minėtus lygmenis. Riboti kai kurių nekaitomų kalbos dalių MD-ą pavyksta automatiškai pažymint iš kelių, dažniausiai nekaitomų, žodžių sudarytas, reikšmės atžvilgiu neskaidomas frazes, vadinamąsias morfologines samplaikas, pvz., *kaip nors, bet kas* (žr. Rimkutė et al. 2005).

Nagrinėjant sutampančias nekaitomas kalbos dalis ir siekiant automatiškai nustatyti, kuri forma pavartota tam tikrame kontekste galima atsižvelgti į statistiką – toks daugiareikšmiškumo mažinimo būdas pritaikytas lenkų kalbos tekстыne (Vetulani et al. 1998, p. 89). Statistines taisykles būtų galima pritaikyti ir lietuvių kalbos nekaitomiems žodžiams, pavyzdžiui, dažniau vartojama *per* kaip prielinksnis, o ne kaip dalelytė; *kad* kaip jungtukas, o ne kaip dalelytė; *be* kaip prielinksnis, o ne kaip dalelytė. Apskritai pastebėta tendencija, kad jei tas pats žodis gali būti prielinksnis ir dalelytė, jungtukas ir dalelytė, prieveiksmis ir dalelytė ir pan., pirmasis variantas, t. y. prielinksnis, jungtukas, prieveiksmis, bus dažnesnis. Kita vertus, visiškai negalima pasikliauti šiomis taisyklėmis ir reikia ieškoti tikslesnių MD-o ribojimo būdų.

Daugiareikšmiai nekaitomi žodžiai dažnai būna prielinksniai ir dar kuri nors kita nekaitoma kalbos dalis, todėl tokio pobūdžio daugiareikšmiškumas gali būti panaikinamas nustatius sintaksinius ryšius, t. y. atsižvelgiant į tai, kokio linksnio reikalauja prielinksnis, pvz., prielinksnis *per* reikalauja galininko linksnio, o *per* kaip dalelytė vartojama su vardininku, pvz.: *kas per naktis! Kas per paukštis ten tupi?* (DLKŽ 2000). Prielinksniai *be, ant* reikalauja kilmininko, dalelytės *be, ant* dažniausiai eina su vardininku, pvz.: *kokia čia be kalba gali būti su juo; koks ten be darbas* (labai prastas); *ant ir atjoja šimtas brolelių* (DLKŽ 2000).

Taip pat nemažai padeda skyryba, pavyzdžiui, gana dažnai iš kelių sutampančių nekaitomų kalbos dalių prijungiamuosius ir kai kuriuos sujungiamuosius jungtukus galima atskirti, nes prieš juos rašomas kablelis. Skirti dalelytes padeda tai, kad tokio sakinio gale gali būti šauktukas arba klaustukas.

Nekaitomos kalbos dalys žodynuose ir gramatikose aprašytos nenuosekliai, dažnai sunku suvokti, kas lėmė vieno ar kito žodžio priskyrimą tam tikroms nekaitomoms kalbos dalimis. Todėl norint sumažinti nekaitomų kalbos dalių MD-ą, reikia išsiaiškinti: 1) kokie galimi nekaitomų žodžių sutapimai; 2) kokiomis kalbos dalimis laikytini tam tikri nekaitomi žodžiai; 3) ar analizuojamas nekaitomas žodis, ar kaitomos kalbos dalies forma. Be aiškių tokių teorinių nuostatų negalima pradėti rašyti MD-ą ribojančių taisyklių. Teorinės nekaitomų kalbos dalių klasifikacijos ir aprašymo

kalbiniuose darbuose nuostatos – ne šio straipsnio tikslas, todėl išsamiau nesigilinama į minėtas problemas.

Kelios nekaitomų kalbos dalių MD-ą ribojančios taisyklės:

1. Prieveiksmių ir prielinksnių sutapimas

a) tai grupei priskiriami žodžiai: *anapus, abipus, apsukui, arti, greta, įkypai, išilgai, netoli, pirma, skersai, šalia, viduj, vietoj, viršuj, viršum, įstrižai, pusiau, vidury, kitapus, antrapus, šiapus, arčiau, artyn, toliau, žemiau*

b) MD-o sprendimas: jei po šių žodžių eina (vienaskaitos) kilmininkas, daugiskaitos vardininkas ir daugiskaitos šauksmininkas, tada jie bus prielinksniai, pvz.: *stovėjau arti stalo; sklypas baigiasi šiapus tvoros*. Jei pavartoti kitų linksnių vardažodžiai arba veiksmažodžių formos, tada busrieveiksmai, pvz.: *laukėme viduj; stovi šalia*. Prieveiksmai bus tada, kai šalia nėra valdomų žodžių.

c) pastabos: ne visada įmanoma skirti anksčiau minėtus nekaitomus žodžius ir tam tikras kaitomų kalbos dalių formas. Kai kurie nekaitomi žodžiai dar nėra labai sustabarėję, aiškus jų ryšys su pamatiniais daiktavardžiais, todėl kartais neaišku, ar tam tikra forma laikytina jau sustabarėjusia, taigi nekaitomu žodžiu, ar daiktavardžiu. Pavyzdžiui, *gale, viduj/ viduje, šone, vietoj/ vietoje, viršuj/ viršuje, prieky/ priekyje* gali būti laikomi ne tik prieveiksmais ar prielinksniais, bet ir daiktavardžiais. Gal kiek aiškiau skiriami minėti nekaitomi žodžiai nuo daiktavardžių pasitelkus tokią taisyklę: kai analizuojamasis žodis reikalauja kito žodžio kilmininko, tai bus prielinksnis, pvz.: *sėdi gale stalo; dirba vietoj manęs; eina prieky visų*; jei šalia nėra valdomų žodžių, pvz.: *stovi šone; kabo viršuj*, tai bus prieveiksmai; jei tiriamasis žodis turi kokį nors pažyminį, tai bus daiktavardis, pvz.: *sėdi stalo gale; mano šone atsirado gumbas; gyvename toje pačioje vietoje*. Vadinas, reikalinga išsami sintaksinė analizė. Vis dėlto nepakanka vien panašaus pobūdžio taisyklių, nes, pvz., DLKŽ-e prie *vidury* kaip prielinksnio pateiktas toks pavyzdys: *paparčiai žydi vidury girelės*, o prie *vidury* kaip daiktavardžio randamas toks pavyzdys: *važiuok viduriu lauko* (DLKŽ 2000). Iš tikrųjų tokia vartoseną mažai kuo skiriasi, ir paskutiniame pavyzdyje forma *viduriu* laikoma daiktavardžiu, nors sintaksiškai ir semantiškai nesiskiria nuo *vidury* kaip prielinksnio. O *gale, šone, prieky/ priekyje* DLKŽ-e apskritai nelaikomi atskirais nekaitomais žodžiais.

2. Prieveiksmių, dalelių ir jungtukų sutapimas

a) tai grupei priskiriami žodžiai: *čia*

b) MD-o sprendimas: nesunku automatiškai atpažinti *čia* kaip jungtuką, nes jis jungia vienaarūšes sakinio dalis, pvz.: *čia*

rytas, čia vakaras. Dažniausiai vartojami du ar daugiau jungtukų *čia*. Dalelytė *čia* pabrėžia žodžio reikšmę, bet tik tekstą analizuojantis žmogus gali nustatyti, ar daugiareikšmis žodis pabrėžia, taigi laikytinas dalelyte, ar nurodo vietą, todėl galima laikyti prieveiksniu. Kol kas neįmanoma pritaikyti aiškios taisyklės šiam daugiareikšmiškumui panaikinti. Vienintelis aiškus kriterijus yra tai, kad *čia* kaip prieveiksmais vartojamas daug dažniau nei *čia* kaip dalelytė.

c) pastabos: reikalinga išsamesnė analizė, norint skirti *čia* kaip dalelytę ir prieveiksmį.

3. Prielinksnių ir dalelių sutapimas

a) tai grupei priskiriami žodžiai: *per*

b) MD-o sprendimas: *per* dažniausiai vartojamas kaip prielinksnis, bet nemažai atvejų, kai šis žodis sakinyje atlieka dalelytės funkcijas. Vienareikšminimo taisyklė galėtų būti tokia: *per* + vardažodžio galininkas = prielinksnis *per*, t. y. šią MD-o rūšį padės išspręsti automatinė sintaksinė analizė. Dalelytė *per* dažniausiai vartojama su vardažodžių vardininku ar tiesiog pabrėžia prieveiksnius. Vis dėlto šis daugiareikšmiškumas nėra toks paprastas: kartais sutampa vyriškosios giminės vardažodžių vienaskaitos vardininkas ir moteriškosios giminės daugiskaitos galininkas (plg. sakinį *kas jis per vienas*). Čia galima tiek *per* kaip prielinksnio (jei forma *vienas* laikoma būdvardžio, įvardžio ar skaitvardžio moteriškosios giminės daugiskaitos galininku), tiek *per* kaip dalelytės interpretacija (jei forma *vienas* laikoma būdvardžio, įvardžio ar skaitvardžio vyriškosios giminės vienaskaitos vardininku). Išspręsti šį sutapimą gali padėti skyrybos ženklas, nes sakinyje, kuriame pavartota *per* kaip dalelytė, gali baigtis šauktuku (pvz.: *kas tu per vyras!*) arba klaustuku (pvz.: *kas ten per paukštis?*).

c) pastabos: reikalinga išsamesnė morfologinė analizė. Daugumą šios MD-o rūšies atvejų turėtų padėti išspręsti sintaksinė analizė.

Nekaitomoms kalbos dalims taikant taisykles, kad būtų sumažintas MD, reikia aprašyti praktiškai kiekvieną žodį atskirai – netgi tai pačiai grupei, pvz., sutampantiems jungtukams ir dalelytėms, sunku rasti kelias bendras taisykles. Kaitomų kalbos dalių MD-o grupėms, pvz., veiksmažodžių trečiojo asmens sutapimui, vardažodžių linksnių sinkretizmui, jau galima taikyti bendresnio pobūdžio taisykles.

Nekaitomų kalbos dalių MD-ą ribojančios taisyklės ir pasiūlymai, kaip tiksliau ir nuosekliau analizuoti nekaitomas lietuvių kalbos dalis, jau yra pritaikyti praktiškai. Atsižvelgus į minėtus pasiūlymus, pakoreguota morfologinės analizės programos *Lemuoklis* leksinių duomenų bazė, t. y. atsisakyta kai kurių kalbos dalių, kurioms priskiriamas nekaitomas žodis, arba atvirkščiai: įtrauktos naujos kalbos dalys, kurių vartoseną išryškėjo MAT-e.

1.3. Kaitomų kalbos dalių morfologinio daugiareikšmiškumo ribojimas

Kaip vienas iš būdų riboti kaitomų kalbos dalių MD-ą pasirinkta sintaksinė analizė, tiksliau kompiuterių sintaksinės analizės mokymosi metodas. Šis būdas naudotas viena-reikšminant kitų kalbos dalių MD-ą, pvz., čekų (žr. Hajič 2004). Sintaksinė analizė taip pat padėjo išaiškinti ir kai kuriuos daugiareikšmius nekaitomų kalbos dalių atvejus.

Sintaksinė analizė nėra vien tik priemonė MD-ui riboti. Tai jau antrasis automatinės kalbos analizės lygis. Po automatinės morfologinės analizės turi būti atliekama sintaksinė analizė. Šio proceso metu nustatomi ryšiai tarp žodžių, kartu išsprendžiamas ir MD. Kita vertus, sintaksinė analizė be semantinės nėra pakankama ir gali būti netgi pavojinga, nes taikant gramatinės taisykles sakiniui ir neatsižvelgiant į jo reikšmę galima nustatyti daugybę žodžių ryšių, o tai vėl didina daugiareikšmiškumą (žr. tolesnį pavyzdį).

Automatinę sintaksinę lietuvių kalbos analizę apsunkina tai, kad lietuvių kalbos žodžių tvarka nėra griežta, kad dažnai tarp susijusių žodžių įsiterpia kitų, todėl nežinia, kaip

nustatyti ryšius tarp žodžių. Ne visada net tekstą analizuojantis lingvistas sugeba vienareikšmiškai nuspręsti, kas su kuo siejama ar derinama sakinyje, pavyzdžiui, sakinyse *skamba nuostabios didžiosios zylės giesmės* gali būti interpretuojamas dvejopai: būdvardžiai *nuostabios* ir *didžiosios* gali būti derinami su *giesmės* arba su *zylės*. Tik labiau išgilinus paaiškėja, kad *didžiosios* priklauso nuo *zylės* (*didžioji zylė* yra terminas), o ne nuo *giesmės*, bet formaliai tikėtinas ir pirmasis variantas. O su kuriuo daiktavardžiu (su *zylės* at/ ir *giesmės*) yra siejamas būdvardis *nuostabios*, dažnai gali pasakyti tik tą konkretų sakinių parašęs žmogus, todėl nėra ko tikėtis, kad automatiškai pavyks lengvai nustatyti ryšius tarp žodžių.

Automatinė sintaksinė lietuvių kalbos analizė kuriama priklausomybių gramatikos principu (plačiau žr. Grigonytė et al. 2005a). Sintaksinėse taisyklėse pateikiama informacija apie tai, kokios kalbos dalys ar tų kalbos dalių formos sudaro sakinio dalis – tai pirmojo lygio taisyklės; iš antrojo lygio sintaksinių taisyklių paaiškėja, kokie galimi ryšiai tarp sakinio dalių. Automatinei sintaksinei analizei dar svarbu, kuris junginio žodis yra valdantis, kas gali įsiterpti ir kokia galima to junginio žodžių tvarka.

Dalis sintaksinių taisyklių yra parašyta E. Rimkutės, kita dalis yra išgauta automatiškai. Taisyklės parašytos remiantis kalbos jausmu, lietuvių kalbos gramatikų duomenimis, taip pat naudotasi MAT-u ir DLKT-u. Nemaža dalis sintaksinių junginių (šiam straipsnyje vadinamų sintaksinėmis taisyklėmis) yra surašyta gramatikose ar sintaksei skirtuose vadovėliuose, bet jie nėra tinkamai susisteminti, yra pritaikyti ištisinį tekstą skaitančiam žmogui, sunkiai panaudojami kuriant automatinę sintaksinę analizę. Taigi reikėjo išrinkti ir automatinei sintaksinei analizei tinkamiausius suklasifikuoti įvairiuose lietuvių kalbotyros darbuose pateiktas sintaksines taisyklės, taip pat aprašyti tuos atvejus, kurie nebuvo niekur užfiksuoti.

Lietuvių kalbos sintaksinėms taisyklėms nustatyti automatinio būdu buvo panaudoti tokie kalbiniai ištekliai: MAT, DLKT ir morfologinis analizatorius *Lemuoklis*. Taisyklių išgavimas iš DLKT-o remiasi lietuvių kalbai adaptuotu modelių atpažinimo (angl. *pattern recognition*) metodu (žr. Theodoridis et al. 1999). Žodžių junginių atpažinimas ir sintaksinių taisyklių nustatymas susideda iš šių etapų: teksto vienetų atpažinimas, automatinė morfologinė žodžių analizė, taisyklių klasifikavimas ir jų įvertinimas (plačiau žr. Grigonytė et al. 2005b).

Morfologinis vienareikšminimas atliekamas iš kelių žodžio morfologinio sąrašo narių išrenkant vieną. Pastebėta, kad tam tikruose junginiuose beveik visada vartojama tik kuri nors viena konkreti veiksmažodžio forma, pvz., junginyje *galima daryti išvadą, kad...* forma *galima* yra neveikiamosios rūšies bevardės giminės dalyvis, o ne moteriškosios giminės dalyvio vienaskaitos vardininkas ar įnagininkas; junginyje *galimas daiktas* pavartotas vyriškosios giminės neveikiamosios rūšies dalyvio vienaskaitos vardininkas, o ne moteriškosios giminės daugiskaitos galininkas.

kas. Analizuojant pavienes formas automatiškai, negalima nustatyti, kuri forma yra tinkama tame kontekste.

Atpažįstant veiksmažodžių junginius, bandoma panaikinti netinkamus junginius, paliekant tik sintaksiškai taisyklingas struktūras. Pritaikius šį metodą tiriamiems junginiams, gaunami vienareikšmiai prasmingi ir sintaksiškai taisyklingi veiksmažodžių junginiai. Pavyksta išspręsti labai dažnus asmenuojamųjų veiksmažodžio formų ir daiktavardžių, bendraties ir dalyvio, būdvardžio ir dalyvio bei kitus sutapimus. Taigi yra panaikinama daugybė MD-o atvejų (plačiau žr. Rimkutė et al. 2006).

Toliau pateiktos kelios sintaksinės taisyklės², naudojamos automatinei sintaksinei analizei kurti ir kaitomų kalbos dalių MD-ui mažinti:

1. Sudėtinis subjektas

Sntmp + dktv. V

a) žodžių tvarka: griežta

b) įsiterpimas: +³ (gali įsiterpti bdvr., įvrd., sktv., dlv., prvks. dažniausiai prieš pirmą dėmenį, bet gali įsiterpti ir prieš antrą)

c) pavyzdys: *gerb. rašytojas*

d) priklausomybė: sntmp. ← dktv.

2. Sudėtinis predikatas

Vksm (vksm_mąstymo) + veik. r. dlv. V (vksm_modal) + bndr.

a) žodžių tvarka: 123, 312

b) įsiterpimas: + (po pirmo ar antro dėmens gali įsiterpti prvks., įvrd., dktv.)

c) pavyzdys: *mano (1) galis (2) pakeisti (3)*

d) priklausomybė: vksm. → dlv. → bndr.

3. Subjektas ir predikatas

Asmenuojamosios vksm. formos dgsk. III asm. + dktv. dgsk. K

a) žodžių tvarka: beveik laisva

b) įsiterpimas: + (prieš dktv. gali įsiterpti bdvr., dlv., dktv. K, sktv., įvrd., prieš vksm. – prvks.)

c) pavyzdys: *atvažiavo svečių*

d) priklausomybė: vksm. → dktv.

4. Subjektas ir atributas

Bdvr. (gim., sk., L) + dktv./ tikr. dktv. (gim., sk., L)⁴

a) žodžių tvarka: beveik griežta

b) įsiterpimas: + (prieš bdvr. gali įsiterpti prvks.; prieš bdvr. arba po jo – kitas bdvr., dlv., sktv., įvrd., dktv. K)

c) pavyzdys: *protingas vaikas; senasis Jonas*

d) priklausomybė: dktv. → bdvr.

² Prie sintaksinių taisyklių nurodyti ir būtinieji jų parametrai: žodžių tvarka, įsiterpimas ir priklausomybė.

³ Pliusas reiškia, kad tarp taisyklės dėmenų gali įsiterpti kitos kalbos dalys.

⁴ Šia taisykle nurodoma, kad subjekto ir atributo junginių sudaro gimine, skaičiumi ir linksniu suderinti būdvardžiai ir daiktavardžiai.

Išvados

Straipsnyje apžvelgtas MAT-o kūrimo procesas, šiame tekstyne išryškėjęs MD. Trumpai pristatyti automatiniai (statistiniai bei loginiai) ir automatizuoti lietuvių kalbos MD-o ribojimo metodai. Automatizuotus MD-o mažinimo būdus galima skirti į tris grupes: nerealiųjų homoformų panaikinimas, nekaitomų kalbos dalių MD-o ribojimas ir kaitomų kalbos dalių MD-o ribojimas.

Išsamiai paaiškinti įvairialypiai metodai, galintys sumažinti nekaitomų kalbos dalių daugiareikšmiškumą. Aprašyta, kaip sudarytos sintaksinės taisyklės, naudojamos sintaksiniam analizatoriui kurti, taip pat mažinančios kaitomų (tam tikrais atvejais ir nekaitomų) kalbos dalių MD-ą.

Literatūra

1. Bartkuvienė, A 2005, Morfologinio daugiareikšmiškumo problemos sprendimas ir jo įtaka kalbos technologijų produktų kūrimui, VDU Informatikos fakulteto magistro darbas.
2. Bick E, Uibo, H, Mütirise, K 2004, Arboret – a Growing treebank of Estonian. Nordic Language Technology, Museum Tusculanum Forlag København Universitet, København, pp. 125-142.
3. Dabartinės lietuvių kalbos gramatika (DLKG) 1996, (red. V. Ambrazas), Mokslo ir enciklopedijų leidykla, Vilnius.
4. Dabartinės lietuvių kalbos žodynas (DLKŽ) 2000, (vyr. red. S. Keinys), Mokslo ir enciklopedijų leidykla, Vilnius.
5. Dutoit, Th 1997, An introduction to text-to-speech synthesis, Kluwer Academic Publishers, Dordrecht, Boston, London.
6. Erjavec, T, Ide, N, Petkevič, V, Véronis, J 1995, MULTTEXT-EAST: Multilingual Text Tools and Corpora for Central and Eastern European Languages, Proceedings of the first TELRI European seminar „Language resources for language technology“, Hungary.
7. Griciūtė, V 2001, Morfologinio daugiareikšmiškumo problemų sprendimas. Statistinio modelio realizacija taikant Viterbi algoritmą, VDU Informatikos fakulteto bakalauro darbas.
8. Griciūtė, V, Pajarskaitė, G 2001, Paslėptų Markovo modelių (HMM) taikymas morfologinio daugiareikšmiškumo problemos sprendimui, 6-tosios magistrantų ir doktorantų konferencijos „Informacinė visuomenė ir universitetinės studijos“ pranešimų medžiaga, pp. 66-71.
9. Grigonytė, G, Rimkutė, E 2005a, Formal specifications for a dependency grammar of the Lithuanian language, tarptautinės konferencijos „The second Baltic conference on human language technologies“ pranešimų medžiaga, 2005, pp. 237-242.
10. Grigonytė, G, Rimkutė, E 2005b, Priklausomybių gramatika pagrįstų lietuvių kalbos sintaksinių taisyklių išgavimas iš Dabartinės lietuvių kalbos teksto, 10-osios tarpuniversitetinės magistrantų ir doktorantų konferencijos „Informacinės technologijos“ pranešimų medžiaga, Kaunas: Technologija, 2005, pp. 65-67.
11. Hajič, J 2004, Disambiguation of rich inflection. Computational morphology of Czech, Univerzita Karlova v Praze, Nakladatelství Karolinum, Praha.
12. Ide, N, Romary, L 2004, International standard for a linguistic annotation framework, Natural language engineering 10 (3/4), pp. 211–225.
13. Ide, N 1995, Encoding standards for linguistic corpora, Proceedings of the first TELRI European seminar „Language Resources for Language Technology“, Hungary.
14. Nepil, M, Popelinský, L 2001, Part-of-speech tagging by means of ILP and active learning, In Proceedings of the workshop on instance selection at ECML/PKDD 2001, Freiburg, pp. 25–31. Retrieved at <http://kd.cs.uni-magdeburg.de/~scheffler/ws01proceedings.html>
15. Paulauskienė, A (1994) Lietuvių kalbos morfologija (LKM). Paskaitos lituanistams, Mokslo ir enciklopedijų leidykla, Vilnius.
16. Rimkutė, E 2003, Morfologinio daugiareikšmiškumo tipologija, Lituanistica, 4 (56), pp. 60-78.
17. Rimkutė, E, Grybinaitė, A 2004, Dažniausios lietuvių kalbos morfologinio daugiareikšmiškumo rūšys ir jų automatinis vienareikšminimas, Kalbų studijos, no 5, pp. 74-78.
18. Rimkutė, E, Jarašiūnaitė, G, Homola, P 2005, Morfologinių samplaikų atpažinimas ir klasifikavimas, Lituanistica 2 (62), pp. 58-75.
19. Rimkutė, E, Grigonytė, G 2006, Statistiniai, loginiai ir kompiuterių mokymosi metodai lietuvių kalbos morfologiniam daugiareikšmiškumui riboti, konferencijos „Informacinės technologijos 2006“ pranešimų medžiaga, pp. 104-108.
20. Sičinava, DV: Сичинава, ДВ К задаче создания корпусов русского языка. Retrieved at <http://rscorpora.narod.ru/article.html>
21. Sinclair, J 2000, Net ir bilijoniniai tekstynai yra per maži. Pokalbis su prof. Johnu Sinclairiu, Darbai ir dienos, 24, pp. 297-299.
22. Theodoridis, S, Kountroumbas, K 1999, Pattern recognition, Academic Press, Athens.
23. Vetulani, Z, Martinek, J, Obreński, T, Vetulani, G 1998, Dictionary based methods and tools for language engineering, Wydawnictwo naukowe, Poznań.
24. Zinkevičius, V 2000, Lemuoklis – morfologinei analizei, Darbai ir dienos, 24, pp. 245-273.
25. Zinkevičius, V, Daudaravičius, V, Rimkutė, E 2005, The morphologically annotated Lithuanian corpus, tarptautinės konferencijos „The second Baltic conference on human language technologies“ pranešimų medžiaga, pp. 365-370.
26. Žáčková, E 2002, Parciální syntaktická analýza (Čeština), Disertační práce. Fakulta informatiky Masarykovy univerzity v Brně, Brno.

Erika Rimkute, Gintarė Grigonytė

Automated Disambiguation Of Lithuanian Morphological Ambiguity

Summary

We describe methods for disambiguation of Lithuanian morphological ambiguity. The methods we present can be automatic and automated. Automatic are statistical and logical methods, also the removal of unreal homonyms. Automated methods are: the removal of unreal homoforms, disambiguation of non-inflective and inflective parts of speech.

We need various rules for disambiguation of non-inflective parts of speech. Analysis of punctuation, statistical data from Contemporary Lithuanian Language Corpus, syntactic analysis, semantics, exhaustive analysis of all the text or a few nearest sentences, transformation in more clear words, morphological information, etc. can be useful for this disambiguation.

The best method for disambiguation of inflective parts of speech is an automatic syntactic analysis. Syntactic rules we present are based on Dependency Grammar. These rules consist of two levels: the level of word groups (lower level) and the level of the combinations of word groups (upper level). It is very important to recognize which word is the main, i.e. governing, and dependent word in automatic syntactic analysis. Other relation parameters are: word order, insertion and priority (in some cases). We also have described the methodology of syntactic rules extraction in this article.

Statistical and logical methods for disambiguation of morphological ambiguity gives good results – more than 90% of forms can be disambiguated correctly. At the meantime we have no concrete results what is an accuracy of automated disambiguation but it is clear that the automatic syntactic analysis can resolve many cases of morphological ambiguity of inflective parts of speech.

Autorės

Erika Rimkutė, dr., VDU Kompiuterinės lingvistikos centro jaunesnioji mokslo darbuotoja.

Mokslinių interesų sritys: tekstynų lingvistika, kompiuterinė lingvistika, automatinė morfologinė analizė bei sintezė, mašininis vertimas.

Adresas: K. Donelaičio g. 52-206, Kaunas, Lietuva

El. paštas: e.rimkute@hmf.vdu.lt

Gintarė Grigonytė, MA, VDU Kompiuterinės lingvistikos centro inžinierė-programuotoja.

Mokslinių interesų sritys: automatinė sintaksinė analizė, matematinė statistika, programų inžinerija, interaktyvioji multimedija.

Adresas: K. Donelaičio g. 52-206, Kaunas, Lietuva.

El. paštas: gintare_grigonyte@fc.vdu.lt

