Central and Eastern European Online Library

Klitikų paieškos lietuviškame tekste algoritmai

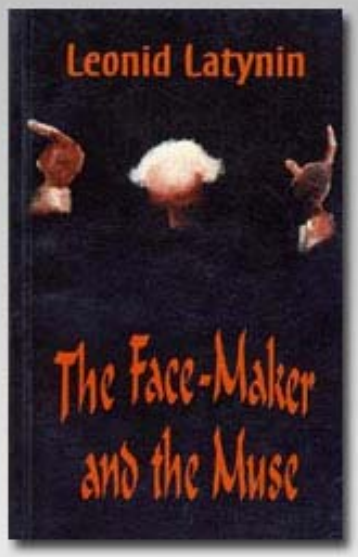
«Algorithms for Detecting Clitics in the Lithuanian Text»

by Pijus Kasparaitis; Tomas Anbinderis

Source:
Studies About Languages (Kalbų Studijos), issue: 10 / 2007, pages: 30-37, on www.ceeol.com.

The following ad supports maintaining our C.E.E.O.L. service

eBooks on Central, East and Southeast Europe



The Face-Maker and the Muse
Glas New Russian Writing, Moscow, 1999, 240 p.
By **Leonid Latynin**

The Face-Maker and the Muse is an anti-utopia about the society where work is a giant machinery mutilating human beings, where everybody desperately tries to climb the social ladder as high as he/she can, but the success depends on the degree of the person's likeness to the Model Face, created by the Chief Face-Maker.

more on:
www.dibido.eu

KOMPIUTERINĖ LINGVISTIKA/ COMPUTATIONAL LINGUISTICS

Klitikų paieškos lietuviškame tekste algoritmai

Tomas Anbinderis, Pijus Kasparaitis

Anotacija. Darbe nagrinėjamas lietuvių kalbos teksto automatinio kirčiavimo uždavinys. Kirčiuotas tekstas gali būti naudojamas kalbos mokymui, balso sintezėje iš teksto ir pan. Šnekamojoje kalboje kai kurie žodžiai (vadinamieji klitikai) netenka kirčio, t. y. prišlyja prie gretimo žodžio. Kalbotyros darbuose galima rasti tik bendras pastabas apie žodžių akcentinio šlijimo tendencijas, o klitikų paieškos lietuviškame tekste algoritmai, reikalingi kalbinėms technologijoms, – dar visai nenagrinėta tema. Darbe apžvelgiami šlijimą lemiantys faktoriai ir siūlomi metodai, kaip tekste rasti žodžius, kurie gali būti bekirčiai. Metodai remiasi šiais principais: 1) samplaikinių formų atpažinimu, 2) statistiniu žodžio kirčiavimo/ nekirčiavimo dažniu, 3) kai kuriomis gramatikos taisyklėmis bei 4) gretimų žodžių kirčių pasiskirstymu (ritmika). Antrasis metodas yra itin paprastas ir duoda neblogų rezultatų, tačiau darbe parodyta, kad kai kurioms žodžių klasėms geresni rezultatai gaunami taikant trečiąjį ir ketvirtąjį metodus. Kiekvienam metodui apibrėžtos žodžių klasės, kurioms jis geriausiai tinka. Paaiškinta, kaip visus metodus sujungti į vieną algoritmą. Kuriant algoritmą buvo stengiamasi minimizuoti pirmosios ir antrosios rūšies klaidų sumą. Pritaikius šį algoritmą testavimo duomenims, klaidų ir visų žodžių santykis buvo 4,1%, o klaidų ir nekirčiuotų žodžių santykis – 18,8%.

Raktiniai žodžiai: *klitikai, proklitikai, enklitikai, akcentinis šlijimas, samplaikinės formos, algoritmas.*

Įvadas

Tekstą, kuriame sudėti kirčio ženklai, patogu naudoti diktoriams skaitant žinias, mokantis kalbos ir pan., tačiau būtinybė kirčiuoti tekstą automatiškai visų pirma iškyla kompiuterinėse balso sintezės iš teksto sistemose. Sintezuotas balsas skamba suprantamiau ir natūraliau, jei žodžiai tinkamai kirčiuojami. Tam tikrų lietuvių kalbos dalių kirčiavimo algoritmus galima rasti šiuose straipsniuose (Kasparaitis 2000), (Kasparaitis 2001), (Kazlauskienė ir kt. 2004), (Norkevičius ir kt. 2004). Tačiau, jei į tekstą žiūrėsime ne kaip į pavienius žodžius, o kaip į visumą, tai apie 20% žodžių tekste yra nekirčiuojami. Nekirčiuoti žodžiai prišlyja prie kirčiuotų ir virsta pastarųjų skiemenimis. Jų nedera pauzėmis atskirti nuo kirtį turinčių žodžių. Bekirčiai žodžiai vadinami *klitikais* (DLKG 1996: p38). Jeigu jie prišlyja prie kirčiuoto žodžio pradžios, vadinami *proklitikais*, jei prie galo – *enklitikais*. Savarankiško žodžio ir klitiko junginys, vienijamas bendro kirčio, laikomas vienu *fonetiniu žodžiu*. Žodžio tapimą *klitiku* vadinsime *akcentiniu šlijimu*.

Kalbotyros darbuose galima rasti tik bendras žodžių akcentinio šlijimo tendencijas, o klitikų paieškos lietuviškame tekste algoritmai – dar visai nenagrinėta tema. Be to, tai gana sudėtinga problema, nes tas pats žodis viename kontekste gali turėti kirtį, o kitame jis gali netekti kirčio.

Šiame darbe apžvelgiami pagrindiniai žodžių akcentinio šlijimo dėsningumai ir jais remiantis pasiūlytos keturios metodų grupės klitikams rasti.

1. Akcentinį šlijimą lemiantys veiksniai

Ar žodis yra klitikas, priklauso nuo įvairių faktorių: nagrinėjamo žodžio skiemenų skaičiaus, jo funkcinio svorio tekste, atstumo tarp kirčiuotų skiemenų, nekirčiuotų skiemenų skaičiaus frazės pradžioje ir pabaigoje, frazės ilgio, loginio kirčio (Mauricaitė 1985; Stundžia 1991).

1.1. Skiemenų skaičiaus ir funkcinio reikšmingumo įtaka

Skiemenų skaičius turi tiesioginę įtaką žodžių akcentiniam šlijimui, t. y. dažniausiai šlyja vienskiemeniai žodžiai, rečiau dviskiemeniai, dar rečiau triskiemeniai ir t. t. Šio straipsnio 4 skyriuje nagrinėtuose tekstuose 95,3% klitikų buvo vienskiemeniai. Todėl toliau kalbėsime tik apie vienskiemenius ir dviskiemenius žodžius.

Darbe (Mauricaitė 1994) teigiama, kad dažniausiai šlyja vienskiemeniai tarnybiniai žodžiai (prielinksniai, jungtukai, dalelytės), funkciškai reikšmingesni vienskiemeniai žodžiai (įvardžiai, prieveiksmiai, skaitvardžiai) dažniau yra kirčiuojami. Kartais gali šlyti ir dviskiemeniai tarnybiniai žodžiai bei veiksmažodžiai – tarinio jungtys.

Darbe (Mauricaitė 1986) pateikiami tikrųjų (jie neturėtų būti kirčiuojami) ir potencialiųjų klitikų pavyzdžiai. Tikraisiais klitikais laikomi:

1. Vienskiemeniai prielinksniai: *ant*, *be*, *dėl*, *į*, *iš*, *lig*, *link*, *nuo*, *pas*, *per*, *po*, *prie*, *pro*, *su*, *tarp*, *už*.

2. Vienskiemeniai jungtukai: *ar*, *bei*, *bet*, *ir*, *jog*, *kad*, *kai*, *kaip*, *kiek*, *lyg*, *nes*, *o*.

3. Vienskiemenės dalelytės: *ar*, *be*, *bet*, *dar*, *gal*, *gi*, *ir*, *jau*, *juk*, *kai*, *kaip*, *ne*, *tik*.

Potencialieji klitikai:

1. Vienskiemeniai įvardžiai: *aš*, *jis*, *ji*, *kas*, *šis*, *ši*, *tas*, *ta*, *tu*.

2. Vienskiemeniai prieveiksmiai: *čia*, *kur*, *ten*, *tuoj*.

3. Vienskiemeniai skaitvardžiai: *du*, *dvi*, *trys*.

4. Veiksmažodžiai – tarinio jungtys: *bus*.

5. Dviskiemeniai prielinksniai: *apie*, *iki*.

6. Dviskiemenė dalelytė *nebe*.

Reikia atkreipti dėmesį į tai, kad čia laikomasi nuostatos, jog iš anksto žinoma, kokia kalbos dalis yra tam tikras žodis, pvz., *o* – jungtukas, tačiau apdorojant tekstą kompiuteriu kalbos dalis nėra žinoma, ją dar reikia nustatyti.

1.2. Gretimų skiemenų kirčių įtaka

Darbe (Mauricaitė 1994) tyrinėta, kokią įtaką šlijimui turi gretimo skiemens kirčiavimas. Potencialus klitikas greta kirčiuoto skiemens dažnai bus nekirčiuotas, o greta nekirčiuoto skiemens arba tarp nekirčiuotų skiemenų – kirčiuotas. Nagrinėti tie žodžiai, kurių kirčiuotų ir nekirčiuotų rasta maždaug po lygiai. Minėta hipotezė pasiteisino, gretimo skiemens kirčiuotumas labiausiai veikia įvardžių, kiek mažiau dalelyčių ir prieveiksmių kirčiavimą. Kalbant apie konkrečius žodžius, minėta taisyklė dažniausiai tinka (tikimybės mažėjimo tvarka) įvardžiams *tas*, *ta*, *aš*, *tu*, *jis*, *ji*, prieveiksmiui *čia*, dalelytei *vis*.

Siekiant dar geriau įsivaizduoti kirčių išsidėstymą lietuvių kalbos frazėse galima pasinaudoti (Stundžia 1991) pateiktais statistiniais duomenimis apie atstumus tarp kirčiuotų skiemenų ir kirčio padėtį frazės pradžioje ir pabaigoje. Dažniausiai tarp kirčiuotų skiemenų būna 2 nekirčiuoti skiemenys, dažniausiai kirčiuojamas antras skiemuo nuo frazės pradžios ir pabaigos. Siekiant išlaikyti atstumus tarp kirčiuotų skiemenų, kai kurie žodžiai paliekami nekirčiuoti.

Šiomis idėjomis remiantis ir bus bandoma sukurti klitikų atpažinimo algoritmą. Frazės ilgio, loginio kirčio ir kalbėjimo tempo įtaka nebus nagrinėjama.

2. Klitikai samplaikinėse formose

Verta atskirai pakalbėti apie kai kurias kalbos dalis (įvardžius, prieveiksmius, jungtukus, dalelytes), kurios sudarytos iš kelių žodžių ir vadinamos samplaikinėmis formomis. Samplaikinių formų radimas ir priskyrimas morfologinėms kategorijoms buvo nagrinėtas (Rimkutė ir kt. 2005), šiame darbe nagrinėjamas jų kirčiavimas. Daugelyje samplaikinių formų vienas žodis yra nekirčiuojamas ir prišlyja prie kito samplaikinės formos žodžio (Mauricaitė 1987). Daugumos šių samplaikinių formų kirčiavimas yra nusistovėjęs, t. y. žodynuose galima rasti, kurį samplaikinės formos žodį kirčiuoti, o kurio ne.

2.1. Samplaikinių formų atpažinimas

Siekiant sukurti samplaikinių formų atpažinimo algoritmą, pirmiausiai remiantis šiais leidiniais (DLKŽ 1993) ir (Mauricaitė 1987) buvo sudarytas samplaikinių formų sąrašas. Jame 95 samplaikinės formos priklauso nekaitomoms kalbos dalims (prieveiksmiai, jungtukai, dalelytės) ir 35 kaitomoms (įvardžiai). Įvardžiai turi 6 linksnių galūnes, be to, kai kurie iš jų turi vienaskaitą ir daugiskaitą, vyriškąją ir moteriškąją giminę. Taigi iš viso gautos 604 įvardžių formos. Trys samplaikiniai įvardžiai nebuvo nagrinėjami, nes abu juos sudarantys žodžiai yra kirčiuojami. Taigi gautas 699 samplaikinių formų sąrašas.

Dauguma į samplaikines formas įeinančių žodžių gali būti vartojami ir atskirai, todėl ne visada lengva atpažinti samplaikines formas. Toliau šiame darbe buvo bandoma išsiaiškinti, kaip dažnai žodžių grupė, kuri gali sudaryti samplaikinę formą, tėra tik atskiri žodžiai. Šiam tikslui buvo atliekama žodžių grupės paieška lietuvių kalbos tekstyne (http://donelaitis.vdu.lt) ir/arba internete (http://www.google.lt), o rasti rezultatai rankiniu būdu išnagrinėti bei užsirašyta statistika, kiek kartų žodžių grupė sudaro samplaikinę formą, o kiek kartų tai yra atskiri žodžiai. Radus daugiau nei 100 pavyzdžių, buvo nagrinėti tik 100 pirmųjų pavyzdžių. Taip pat reikia pastebėti, kad daugeliu atvejų (564 iš 699) žodžių grupė kirčiuojama vienodai, nesvarbu, ar ji sudaro samplaikinę formą, ar ne. Tokios žodžių grupės nebuvo nagrinėjamos, o paprastumo dėlei laikoma, kad grupė sudaro samplaikinę formą. Likusioms žodžių grupėms išnagrinėjus apie 10000 pavyzdžių nustatyti tokie dėsningumai, kurie bus panaudoti kuriant algoritmą:

1) žodžių grupė, galinti sudaryti samplaikinę formą, dažniausiai ją ir sudaro, išskyrus žodžių grupes *kaip ir*, *kas sykį*, *kam kita*, kurios dažniau samplaikinių formų nesudaro. Šias žodžių grupes paprasčiausia tiesiog pašalinti iš samplaikinių formų sąrašo;

2) žodžių grupės, kurios baigiasi žodžiu *nors*, nesudaro samplaikinių formų eidamos prieš žodį *kiek*, prieš kitus žodžius samplaikines formas sudaro. Paprasčiausias problemos sprendimo būdas, papildyti sąrašą nauja samplaikine forma *nors kiek* (nors tokia forma nebuvo minima nei žodyne (DLKŽ 1993), nei straipsnyje (Mauricaitė 1987)), o samplaikinių formų paiešką pradėti nuo sakinio galo (kodėl paieška atliekama nuo sakinio pabaigos, paaiškinta 3.2. skyrelyje).

Kiek išsamiau panagrinėta žodžių grupė *ką tik*. Ji samplaikinę formą sudaro apie 86% atvejų. Tarp likusių apie 10000 nagrinėtų pavyzdžių, žodžių grupės nesudarė samplaikinių formų tik 10 atvejų (0,1%).

2.2. Žodžių grupių skaidymas į samplaikines formas

Tarkime, kad turime trijų žodžių grupę *bet kas nors*. Samplaikinę formą gali sudaryti tiek žodžiai *bet kas*, tiek žodžiai *kas nors*. Siekiant išsiaiškinti, kaip spręsti tokias skaidymo problemas, buvo surastos samplaikinių formų poros, kur vienos formos paskutinis žodis sutampa su kitos formos pirmuoju, ir sudarytos visos galimos iš dalies persidengiančių samplaikinių formų žodžių grupės. Su jomis atlikta analogiška paieška tekstyne arba internete ir rasti rezultatai rankiniu būdu išnagrinėti bei užsirašyta vienokio ar kitokio skaidymo žodžių grupėmis statistika. Įvardžių, kurie gali turėti vienaskaitą ir daugiskaitą, buvo nagrinėjama tik vie-

naskaita. Iš viso buvo nagrinėtos 328 grupės, 9 iš jų rasta daugiau nei po 100 pavyzdžių, 48 rasta nuo 1 iki 99 pavyzdžių, kitų grupių pavyzdžių nerasta. Iš viso buvo išanalizuota apie 1500 pavyzdžių ir sudarytos tokios taisyklės, kurios bus panaudotos kuriant algoritmą:

1) Pirmiausiai tikrinama, ar grupėje yra samplaikinių formų (visi linksniai) *kas ne kas*, *koks ne koks*, *kuris ne kuris*, *kada ne kada*, *kame ne kame*, *kur ne kur*, *vos ne vos*. Šios formos niekada neskaidomos.

2) Jei grupė prasideda žodžiu *bet* po skyrybos ženklo, samplaikinę formą sudaro paskutiniai du žodžiai, pvz., *bet | kas nors*, priešingu atveju pirmieji du, pvz., *bet kas | nors*.

3) Toliau išvardintas grupes skaidyti taip: *vis vien | tik*, *vos ne | tik*, *daug kas | kita*.

4) Kitais atvejais samplaikinė forma yra grupės pabaigoje, t. y. nuo grupės pabaigos atskeliama samplaikinė forma, o likusiai grupei vėl gali būti taikoma ši taisyklė.

Buvo rasti tik 5 pavyzdžiai, kurie netenkina šių taisyklių. Kadangi daugelyje atvejų tenka taikyti 4 taisyklę, todėl samplaikinių formų paiešką patogiausia atlikti nuo sakinio galo.

Toliau pateiktas pavyzdys, kuriame yra šešios samplaikines formas galinčių sudaryti žodžių grupės, jos paryškintos. Skaičiai skliausteliuose rodo samplaikinių formų ribas bei taisyklę, kuria remiantis ta riba nustatyta. Pvz.: ***Kada nors (4) ką nors (4) kiek nors*** *yra mylėjęs* ***bet kuris (2) nors*** *ir pats paprasčiausias žmogus ir todėl žino, jei* ***ką (4) nors kiek*** *myli, o bandai tai neigti,* ***vis vien (3) tik*** *apie jį galvoji,* ***bet (2) kas nors*** *gali to nežinoti, o* ***kas ne kas (1) nors*** *ir žinodamas to nepripažinti.*

3. Klitikų paieška remiantis statistiniu žodžio kirčiavimo/ nekirčiavimo dažniu

Tolimesniuose eksperimentuose buvo naudojami lietuvių kalbos tarties mokymuisi skirti kirčiuoti tekstai, paimti iš (Pakerys ir Pupkis 1976). Šiuose tekstuose klitikai palikti nekirčiuoti. Buvo naudojami tik prozos tekstai, nes poezijoje kartais kirtis paslenkamas siekiant išlaikyti ritmą. Tolimesniuose eksperimentuose buvo naudojama tekstų dalis sudaryta iš 8397 žodžių, iš jų 1842 (21,9%) nekirčiuoti.

Toliau visuose eksperimentuose pirmiausiai surandamos ir sukirčiuojamos samplaikinės formos. Samplaikinėms formoms priklausantiems žodžiams kiti algoritmai nebetaikomi. Tokių formų rasta 49 (98 žodžiai), kirčiuojant padaryta viena klaida (2 žodžiai).

Toliau buvo bandoma išsiaiškinti, kaip sėkmingai galima aptikti nekirčiuojamus žodžius remiantis vien tik statistika, t. y. remiantis tekstais statistiškai nustatyti, kad tam tikras žodis dažniau yra nekirčiuotas, nei kirčiuotas, ir iš tokių žodžių sudaryti sąrašą, kurio žodžius laikysime nekirčiuojamais. Analogiškai galima būtų sudaryti ir kirčiuojamų žodžių sąrašą, bet galima tiesiog visus likusius žodžius laikyti kirčiuojamais. Tekstai buvo padalinti į dvi maždaug lygias dalis, kurios pakaitomis buvo naudojamos sąrašui sudaryti ir testuoti. Eksperimentų metu gali būti padaromos dviejų tipų klaidos: 1) žodis, kurį reikėtų kirčiuoti, paliekamas nekirčiuotas; 2) sukirčiuojamas žodis, kurio galima būtų nekirčiuoti. Visuose tolimesniuose eksperimentuose bus stengiamasi minimizuoti abiejų tipų klaidų sumą. Rezultatai pateikti 1 lentelėje.

Taigi sąrašui sudaryti ir testuoti naudojant tuos pačius duomenis pasiektas 12,7% tikslumas, o naudojant skirtingus duomenis – 19,2% tikslumas.

1 lentelė. Klitikų paieškos, remiantis nekirčiuojamų žodžių sąrašu, rezultatai.

Sąrašas sudarytas iš	Kirčiuotų žodžių + nekirčiuotų žodžių = suma, **(dalis tarp visų žodžių)**	**Sąrašo ilgis**	Pirmojo tipo klaidų + antrojo tipo klaidų = suma *(klaidų ir nekirčiuotų žodžių santykis)*	
			1 dalis testavimui	**2 dalis testavimui**
1 dalies	3225 + 910 = 4135 **(49,8%)**	69	60 + 56 = 116 *(12,7%)*	83 + 96 = 179 *(19,7%)*
2 dalies	3281 + 883 = 4164 **(50,2%)**	74	76 + 90 = 166 *(18,8%)*	45 + 67 = 112 *(12,7%)*

4. Klitikų radimas remiantis gramatika

Išanalizavus 4 skyriuje gautus rezultatus bus bandoma remiantis lietuvių kalbos gramatika sukurti taisykles, taikomas visai žodžių grupei, o gauti rezultatai bus lyginami su kiekvienam žodžiui atskirai pasiektais statistinio metodo rezultatais.

4.1. Kalbos dalys – potencialūs klitikai

Akcentinis žodžio šlijimas labai priklauso nuo to, kokia tai kalbos dalis (Mauricaitė 1994), todėl pirmiausiai apibrėžkime, tarp kokių kalbos dalių ieškosime klitikų. Dažniausi klitikai yra prielinksniai, jungtukai, dalelytės, rečiau įvardžiai, prieveiksmiai, skaitvardžiai, veiksmažodžiai (sakinyje atliekantys tarinio jungties funkciją). Kitas kalbos dalis laikysime visada kirčiuojamomis. Turimuose tekstuose buvo kirčiuoti visi skaitvardžiai, todėl juos taip pat laikysime visada kirčiuojamais, nors darbe (Mauricaitė 1986) yra pateikta pavyzdžių su nekirčiuotais skaitvardžiais.

Be to, turimuose tekstuose buvo rasti 6 žodžiai, kurie nepriklauso kalbos dalims, tarp kurių ieškosime klitikų. Dažniausiai tai žodžių skiemenys, parašyti atskirai siekiant parodyti tarimą skiemenimis.

4.2. Nekaitomos kalbos dalys

Remiantis 4 skyriaus rezultatais galima tikėtis, kad kai kuriems nekaitomiems žodžiams (prielinksniams, jungtukams, dalelytėms) galima sėkmingai taikyti statistinį metodą. Iš ankstesniame skyrelyje aprašyto nekirčiuojamų žodžių sąrašo buvo išrinkti tie nekaitomi žodžiai, kurie daugiau nei du kartus dažniau buvo nekirčiuoti, nei kirčiuoti. Gautas toks 41 žodžio sąrašas:

prielinksniai: *ant*, *apie*, *be*, *dėl*, *į*, *iki*, *iš*, *ligi*, *nuo*, *pas*, *per*, *po*, *prie*, *pro*, *su*, *tarp*, *ties*, *už*;

jungtukai: *ar*, *bet*, *ir*, *jog*, *kad*, *kai*, *lyg*, *lygu*, *negu*, *nei*, *nes*, *nors*, *o*, *tad*, *tai*;

dalelytės: *gi*, *jau*, *juk*, *kaži*, *nė*, *ne*, *nebe*;

prieveiksmis *tik*.

Čia nurodyta tik ta kalbos dalis, kuria atitinkamas žodis dažniausiai vartojamas. Lyginant su 2.1. skyrelyje pateiktu sąrašu, sąrašas ne tik papildytas kai kuriais vienskiemeniais ir dviskiemeniais žodžiais (*apie*, *bei*, *iki*, *kaži*, *ligi*, *lygu*, *nė*, *nebe*, *negu*, *nei*, *nors*, *tad*, *tai*, *ties*), tačiau buvo pastebėta, kad kai kurie žodžiai gan dažnai buvo kirčiuoti (*kaip*, *kiek*, *link*), todėl reikės sudėtingesnių taisyklių, o kai kuriuos geriau iš viso laikyti visada kirčiuojamais (pvz., žodis *dar* 36 kartus aptiktas kirčiuotas ir nė karto bekirtis, žodis *gal* – 11 kartų kirčiuotas ir 1 nekirčiuotas).

Kai kurie žodžiai, pvz., *ligi*, gali būti laikomi keliomis kalbos dalimis, tačiau tiek būdamas prielinksniu, tiek jungtuku, tiek dalelyte, šis žodis linkęs netekti kirčio. Tačiau yra žodžių, kurie būdami viena kalbos dalimi linkę netekti kirčio, o kita kalbos dalimi – kirčiuojami. Pvz., žodis *ant* dažniausiai būna nekirčiuojamas kaip prielinksnis, tačiau gali būti vartojamas ir kaip dalelytė (santrumpa nuo *antai*), kuri linkusi būti kirčiuojama. Analogiškai žodis *ir* dažniausiai būna nekirčiuojamu jungtuku, bet gali būti ir kirčiuotu prieveiksmiu (santrumpa nuo *irgi*). Tačiau pastarieji vartojimo atvejai yra itin reti, todėl laikysime, kad *ant* ir *ir* visada nekirčiuojami.

Turimuose tekstuose anksčiau minėtam 41 nekirčiuojamo žodžio sąrašui priklausantys žodžiai buvo rasti 1354 kartus, 16 kartų tekstuose šie žodžiai buvo kirčiuoti. Taigi minėtas sąrašas padengia (1354-16)/ 1842=72,6% nekirčiuojamų žodžių, o klaidos sudaro tik 16/ 1354=1,2%.

4.3. Skyrybos ženklo panaudojimas

Kai kirčiuoto ir nekirčiuoto varianto dažniai nedaug skiriasi, kalbos daliai nustatyti tenka taikyti kiek sudėtingesnius algoritmus. Pvz., žodis *o* būdamas jungtukas yra nekirčiuojamas, o jaustukas – kirčiuojamas. Tačiau jaustukas paprastai eina prieš skyrybos ženklą, o jungtukas – ne. Analogiškai dalelytės *ne* ir *nebe* paprastai kirčiuojamos prieš skyrybos ženklą (neigiamai atsakant į klausimą) ir nekirčiuojamos kitais atvejais. Apibrėžkime aibę O&NE = {*o*, *ne*, *nebe*} ir šios aibės kirčiavimui suformuluokim tokią taisyklę:

1 taisyklė. *Aibės O&NE žodį kirčiuoti prieš skyrybos ženklą, priešingu atveju nekirčiuoti.*

Toliau aibės O&NE žodžiai buvo išimti iš nekirčiuojamų žodžių sąrašo, kuriame lieka 1266 žodžiai (padaroma 14 klaidų), ir kirčiuojami naudojant 1 taisyklę. Rezultatai pateikti 4 lentelėje 1 eilutėje. Be to, 4 lentelėje palyginimui pateikti ir duomenys, kiek klaidų būtų padaroma kiekvienam grupės žodžiui atskirai taikant statistinį metodą.

4.4. Prielinksnių atpažinimas

Kai kurie žodžiai, pvz., *prieš*, *viršum*, *šalia*, gali būti ir prielinksniai (paprastai nekirčiuojami), ir prieveiksmiai (paprastai kirčiuojami). Be to, kai kurie žodžiai, pvz., *link*, *dėlei*, gali būti prielinksniai ir polinksniai (paprastai kirčiuojami). Šiais atvejais prielinksnis atpažįstamas pagal tai, kad po jo eina atitinkamo linksnio žodis. Paprastas, nors nepakankamai patikimas žodžio linksnio atpažinimo būdas – tai pasinaudoti kiekvieno linksnio galūnių sąrašais. Apibrėžkime tokias aibes PRIEL_K= {*viršum*, *šalia*, *link*, *dėlei*} ir PRIEL_G= {*prieš*}, ir šioms aibėms suformuluokime tokią taisyklę:

2 taisyklė. *Aibės PRIEL_K žodį nekirčiuoti, jei po jo eina žodis kilmininko linksniu, o aibės PRIEL_G žodį nekirčiuoti, jei po jo eina žodis galininko linksniu, priešingu atveju kirčiuoti.*

Rezultatai pateikti 4 lentelėje 2 eilutėje.

4.5. Parodomieji įvardžiai

Dabartinės lietuvių kalbos gramatikoje (DLKG 1996: p252 psl.) teigiama, kad įvardžiai *tas*, *ta*, *šis*, *ši* gali būti vartojami daiktavardiškai (pvz., *Tàs ir sãko.*) arba būdvardiškai (pvz., *Tas põnas ir sãko.*). Buvo pastebėta, kad daiktavardiškai vartojami šie įvardžiai paprastai yra kirčiuojami, o būdvardiškai – nekirčiuojami. Kai žodis vartojamas būdvardiškai, po jo eina to paties linksnio žodis. Apibrėžkime parodomųjų įvardžių aibę PAROD_IV = {*tas*, *to*, *tam*, *tą*, *tuo*, *tame*, *tie*, *tų*, *tiems*, *tuos*, *tais*, *tuose*, *ta*, *tos*, *toje*, *toms*, *tomis*, *tose*, *šis*, *šio*, *šiam*, *šį*, *šiuo*, *šiame*, *šie*, *šių*, *šiems*, *šiuos*, *šiais*, *šiuose*, *ši*, *šios*, *šiai*, *šią*, *šia*, *šioje*, *šioms*, *šias*, *šiomis*, *šiose*}. Aibės PAROD_IV žodžius galima sugrupuoti pagal linksnius. Aibėje nėra įvardžio *tai*, nes šis žodis daug dažniau vartojamas kaip jungtukas. Aibės PAROD_IV žodžiams kirčiuoti naudosime taisyklę:

3 taisyklė. *Aibės PAROD_IV žodį nekirčiuoti, jei po jo eina to paties linksnio žodis, priešingu atveju kirčiuoti.*

Rezultatai pateikti 4 lentelėje 3 eilutė. Naudojant paprastą linksnių galūnių paiešką padaroma 14 klaidų, jei pavyktų visur teisingai atpažinti linksnius, klaidų liktų 11.

5. Klitikų nustatymas remiantis konteksto kirčiavimu

Kai kurių žodžių kirčiavimas labai priklauso nuo to, ar kirčiuoti greta esantys žodžiai. Darbe (Mauricaitė 1994) tarp tokių žodžių minimi įvardžiai *aš*, *tu*, *jis*, *ji*, prieveiksmis *čia*, dalelytė *vis*. Peržvelgus turimus pavyzdžius, dar buvo nuspręsta panagrinėti veiksmažodžio *būti* formas, įvardžių *koks*, *kokia*, *kuris*, *kuri* grupę, įvardžių *šis*, *ši*, *tas*, *ta* grupę, įvardžių ir prieveiksmių *kas*, *kada*, *kaip*, *kiek* grupę.

Buvo atlikti tokie eksperimentai: pasirenkama tam tikra žodžių aibė, kuriai bus bandoma sukurti vieną taisyklę. Surenkama statistika (turimuose daugiau kaip 8000 žodžių tekstuose), kiek kartų šios grupės žodžiai buvo kirčiuoti/ nekirčiuoti tam tikrame kontekste. Buvo nagrinėti tokie kontekstai: 1) po skyrybos ženklo, 2) po nekirčiuoto žodžio, 3) po kirčiuoto skiemens, 4) po vieno nekirčiuoto skiemens kirčiuotame žodyje, 5) po dviejų ir daugiau nekirčiuotų skiemenų kirčiuotame žodyje, 6) prieš skyrybos ženklą, 7) prieš

nekirčiuotą žodį, 8) prieš kirčiuotą skiemenį, 9) prieš vieną nekirčiuotą skiemenį kirčiuotame žodyje, 10) prieš du nekirčiuotus skiemenis kirčiuotame žodyje, 11) prieš tris ir daugiau nekirčiuotus skiemenis kirčiuotame žodyje. Nekirčiuotas vienskiemenis žodis ir vienas nekirčiuotas kirčiuoto žodžio skiemuo gali turėti visai skirtingą įtaką, todėl tai buvo laikoma skirtingais kontekstais. Statistika surašoma į lentelę. Remiantis lentele sukuriama taisyklė, kiekvienai kontekstų „po“ ir „prieš“ porai nusakanti, kirčiuoti ar nekirčiuoti aibei priklausančius žodžius.

5.1. Asmeniniai įvardžiai

Pirmiausiai buvo atlikti eksperimentai su įvardžių *aš*, *tu*, *jis*, *ji* įvairiomis gramatinėmis formomis (iš viso 61 skirtingas žodis). Iš karto buvo pastebėta, kad kai kurios gramatinės formos elgiasi visiškai skirtingai, pvz., *aš*, *tu* lyginant su *man*, *tau*, todėl buvo bandoma jas skaidyti į dvi grupes ir sukurti dvi taisykles. Siekiant, kad taisyklės būtų kiek galima bendresnės, mažiau priklausytų nuo konkrečių duomenų, vienai grupei buvo priskiriami visi tam tikrą linksnį turintys ar tam tikra savybe pasižymintys žodžiai.

Perrinkus įvairius grupavimo variantus nustatyta, kad mažiausia klaidų padaroma, kai į vieną grupę priskiriami visi vienskiemeniai žodžiai, išskyrus naudininko linksnį. Apibrėžkim asmeninių įvardžių aibę iš 20 žodžių: ASM_IV = {*aš*, *mes*, *mūs*, *mus*, *tu*, *jūs*, *jus*, *jis*, *jo*, *jį*, *juo*, *jie*, *jų*, *juos*, *jais*, *ji*, *jos*, *ją*, *ja*, *jas*}. Turimuose duomenyse rasta 84 kirčiuoti ir 152 nekirčiuoti šios grupės žodžiai. Jų kirčiavimo statistika įvairiuose kontekstuose pateikta 2 lentelėje.

2 lentelė. Asmeninių įvardžių kirčiavimo priklausomybė nuo konteksto kirčiavimo.

	prieš skyrybos ženklą	**prieš nekirčiuotą žodį**	**prieš kirčiuotą skiemenį**	**prieš 1 nekirčiuotą skiemenį**	**prieš 2 nekirčiuotus skiemenis**	**prieš 3 ir daugiau nekirčiuotų skiemenų**
po skyrybos ženklo	**1/ 3**	*9/ 2*	**0/ 30**	**3/ 14**	*3/ 2*	*1/ 0*
po nekirčiuoto žodžio	*4/ 0*	*9/ 0*	*11/ 9*	*17/ 9*	*7/ 1*	-
po kirčiuoto skiemens	**0/ 3**	**2/ 6**	**2/ 16**	**2/ 17**	**1/ 2**	**0/ 1**
po 1 nekirčiuoto skiemens	**1/ 2**	4/ 4	**1/ 13**	**1/ 10**	**1/ 3**	-
po 2 ir daugiau nekirčiuotų skiemenų	–	**1/ 2**	**0/ 1**	*2/ 1*	1/ 1	-

2 lentelėje pirmas skaičius rodo, kiek kirčiuotų, antras – kiek nekirčiuotų įvardžių. Jei daugiau nekirčiuotų, skaičiai paryškinti, jei daugiau kirčiuotų – parašyta kursyvu. Siekdami kiek supaprastinti taisyklę, laikykim, kad antroje eilutėje visus reikia kirčiuoti, o paskutinėse trijose eilutėse – visus nekirčiuoti. Dabar taisyklė atrodo taip:

4 taisyklė. *Aibės ASM_IV žodžius kirčiuoti po skyrybos ženklo prieš nekirčiuotą žodį arba prieš du ar daugiau nekirčiuotų skiemenų kirčiuotame žodyje, taip pat kirčiuoti po nekirčiuoto žodžio, kitais atvejais nekirčiuoti.*

Rezultatus žr. 4 lentelėje 4 eilutė. Likusius asmeninius įvardžius (41 gramatinę formą) geriausia visada kirčiuoti. Jų rasta 88 ir padaromos 9 klaidos.

5.2. Veiksmažodis būti

Kita nagrinėta grupė buvo žodžio *būti* gramatinės formos: BUTI = {*buvo*, *bus*, *buvau*, *buvai*, *yra*, *esu*, *esi*, *nesu*, *nesi*, *nėra*}. Rasta 64 kirčiuoti ir 30 nekirčiuotų grupės žodžių, jiems gauta statistika pateikta 3 lentelėje.

3 lentelė. Veiksmažodžio *būti* formų kirčiavimo priklausomybė nuo konteksto kirčiavimo.

	prieš skyrybos ženklą	**prieš nekirčiuotą žodį**	**prieš kirčiuotą skiemenį**	**prieš 1 nekirčiuotą skiemenį**	**prieš 2 nekirčiuotus skiemenis**	**prieš 3 ir daugiau nekirčiuotų skiemenų**
po skyrybos ženklo	*1/ 0*	*1/ 0*	*3/ 0*	*3/ 1*	-	-
po nekirčiuoto žodžio	*5/ 0*	*5/ 0*	*3/ 1*	*5/ 0*	*1/ 0*	*2/ 0*
po kirčiuoto skiemens	*6/ 0*	*3/ 1*	**3/ 5**	**2/ 3**	**2/ 3**	-
po 1 nekirčiuoto skiemens	*4/ 0*	*1/ 0*	**3/ 8**	**5/ 6**	**1/ 2**	*1/ 0*
po 2 ir daugiau nekirčiuotų skiemenų	*1/ 0*	-	*2/ 0*	*1/ 0*	-	-

Pagal ją galima užrašyti taisyklę:

5 taisyklė. *Aibės BUTI žodžius nekirčiuoti po kirčiuoto skiemens arba po vieno nekirčiuoto skiemens kirčiuotame žodyje prieš kirčiuotą skiemenį, 1 arba 2 nekirčiuotus skiemenis, kitais atvejais kirčiuoti.*

Rezultatai 4 lentelėje 5 eilutė.

5.3. Kitos žodžių grupės

Analogiški eksperimentai atlikti su klausiamųjų įvardžių grupe KLAUS_IV = {*koks*, *kokio*, *kokiam*, *kokį*, *kokiu*, *kokie*, *kokių*, *kokiems*, *kokius*, *kokiais*, *kokia*, *kokios*, *kokiai*, *kokią*, *kokioms*, *kokias*, *kurs*, *kuris*, *kurio*, *kuriam*, *kurį*, *kuriuo*, *kurie*, *kurių*, *kuriems*, *kuriuos*, *kuriais*, *kuri*, *kurios*, *kuriai*, *kurią*, *kuria*, *kurioms*, *kurias*}, apimančia 34 skirtingus vienskiemenius ir dviskiemenius žodžius, taip pat prieveiksmiu *čia*, dalelyte *vis*. Gautos analogiškos statistikos lentelės, jomis remiantis užrašytos tokios trys taisyklės:

6 taisyklė. *Aibės KLAUS_IV žodžius nekirčiuoti prieš kirčiuotą skiemenį po skyrybos ženklo arba kirčiuoto skiemens, kitais atvejais kirčiuoti.*

7 taisyklė. *Žodį čia kirčiuoti po skyrybos ženklo prieš nekirčiuotą žodį arba kirčiuotą skiemenį, po nekirčiuoto žodžio prieš skyrybos ženklą, nekirčiuotą žodį, vieną ar daugiau nekirčiuotų skiemenų, kitais atvejais nekirčiuoti.*

8 taisyklė. *Žodį vis nekirčiuoti prieš nekirčiuotą žodį po kirčiuoto žodžio, prieš kirčiuotą skiemenį po vieno ar daugiau nekirčiuotų skiemenų.*

Rezultatai 4 lentelėje atitinkamai 6, 7 ir 8 eilutėse.

Analogiški eksperimentai atlikti ir su įvardžių *šis*, *ši*, *tas*, *ta* gramatinėmis formomis, tačiau jiems šis metodas nepasiteisino, nes buvo gauta 21 klaida, o tai daugiau, nei taikant 5.5. skyrelyje aprašytą metodą (kur buvo padaryta 14 klaidų).

5.4. Klausiamieji įvardžiai ir prieveiksmiai

Dar sudėtingesnės taisyklės reikia klausiamųjų įvardžių ir prieveiksmių grupei KI&P = {*kas*, *ko*, *kieno*, *kam*, *ką*, *kuo*, *kur*, *kame*, *kada*, *kaip*, *kiek*, *kodėl*}. Pagal parengtą statistikos lentelę užrašius taisyklę padaroma 40 klaidų. Buvo pastebėta, kad minėti žodžiai gali būti vartojami klausimuose, ir šiuo atveju jie paprastai būna kirčiuoti, todėl daliai KI&P aibės žodžių suformuluota tokia taisyklė:

9a taisyklė. *Jei KI&P aibės žodis eina po skyrybos ženklo, o sakinys baigiasi klaustuku, žodis kirčiuojamas.*

Ši taisyklė apima 32 žodžius, padarytos 2 klaidos (2 žodžiai turėjo likti nekirčiuoti).

Likusiems (9a taisyklei nepriklausantiems) 152 žodžiams buvo sudaryta statistikos lentelė ir pagal ją užrašyta tokia taisyklė:

9b taisyklė. *Aibės KI&P žodis kirčiuojamas prieš skyrybos ženklą ir nekirčiuotą žodį, taip pat po nekirčiuoto žodžio prieš du ar daugiau nekirčiuotų skiemenų.*

Taikant 9b taisyklę padaromos 26 klaidos. Taisyklės 9a ir 9b buvo sujungtos į vieną 9 taisyklę. Rezultatus žr. 4 lentelėje 9 eilutė.

4 lentelė. Sudarytų taisyklių ir statistinio metodo palyginimas.

Taisyklės Nr.	Žodžių aibė (*pavyzdžiai*)	Kirčiuotų žodžių + nekirčiuotų žodžių = suma, **(dalis tarp visų žodžių)**	Pirmo tipo klaidų + antro tipo klaidų = suma *(klaidų ir žodžių santykis)*	
			Taikyta taisyklė žodžių grupei	**Taikyta statistika atskiriems žodžiams**
1	O&NE (*o*, *ne*, *nebe*)	2 + 86 = 88 **(11,3%)**	0 + 0 = 0 *(0,0%)*	2 + 0 = 2 *(2,3%)*
2	PRIEL_K (*link*), PRIEL_G (*prieš*)	2 + 6 = 8 **(1,0%)**	1 + 0 = 1 *(12,5%)*	0 + 2 = 2 *(25,0%)*
3	PAROD_IV (*tas*, *šis*)	35 + 56 = 91 **(11,7%)**	9 + 5 = 14 *(15,4%)*	19 + 3 = 22 *(24,2%)*
4	ASM_IV (*aš*, *tu*, *jis*)	84 + 152 = 236 **(30,3%)**	23 + 23 = 46 *(19,5%)*	68 + 1 = 69 *(29,2%)*
5	BUTI (*yra*, *esu*)	64 + 30 = 94 **(12,1%)**	16 + 3 = 19 *(20,2%)*	0 + 30 = 30 *(31,9%)*
6	KLAUS_IV (*koks*, *kuris*)	19 + 20 = 39 **(5,0%)**	2 + 7 = 9 *(23,1%)*	3 + 3 = 6 *(15,4%)*
7	*čia*	10 + 23 = 33 **(4,2%)**	2 + 2 = 4 *(12,1%)*	10 + 0 = 10 *(30,3%)*
8	*vis*	6 + 13 = 19 **(2,4%)**	3 + 0 = 3 *(15,8%)*	6 + 0 = 6 *(31,6%)*
9	KI&P (*kas*, *kaip*)	65 + 105 = 170 **(21,9%)**	13 + 13 = 26 *(15,3%)*	16 + 26 = 42 *(24,7%)*
Iš viso		287 + 491 = 778 **(100%)**	69 + 53 = 122 *(15,7%)*	124 + 65 = 189 *(24,3%)*

Taigi taikant specializuotas taisykles žodžių grupėms gauta 15,7% klaidų, o tai 8,7% mažiau, nei remiantis atskirai kiekvienam žodžiui apskaičiuotu kirčiavimo/ nekirčiavimo dažniu.

5.5. Nuo konteksto priklausomų taisyklių sąveika

Sudarydami šiame skyriuje aprašytas taisykles laikėmės nuostatos, kad žodžio konteksto kirčiavimas jau yra žinomas, ir tereikia nustatyti tik vieno žodžio kirčiavimą. Iš tikrųjų gali greta eiti keli žodžiai, kirčiuojami pagal čia aprašytas taisykles, ir tokiu atveju konteksto kirčiavimas dar nėra žinomas. Taigi paskutinis klausimas, kurį liko išnagrinėti, tai šiame skyriuje aprašytų taisyklių sąveiką. Turimuose daugiau kaip 8000 žodžių tekstuose buvo rasti 52 atvejai (apimantys 104 žodžius), kai greta eina pagal anksčiau aprašytas taisykles kirčiuojami žodžiai. Su jais išbandžius žodžių peržiūrėjimą iš kairės ir iš dešinės, o greta esantį nežinomo kirčiavimo žodį laikant kirčiuotu, po to nekirčiuotu, buvo nustatyta, kad geriausia tokias žodžių sekas peržiūrėti iš kairės, aibių ASM_IV, KI&P žodžius bei žodžius *čia* ir *vis* laikyti nekirčiuotais, o kitus kirčiuotais. Yra dvi išimtys: 1) aibės ASM_IV žodžius po KI&P žodžių laikyti kir-

čiuotais. 2) radus aibės BUTI žodį, po kurio eina aibės KI&P žodis, pirmiau sukirčiuoti KI&P žodį (aibės BUTI žodį laikyti kirčiuotu), tada grįžti prie aibės BUTI žodžio.

Panagrinėkim tokį pavyzdį: *Aš jį čia buvau matęs.* Pateiktame pavyzdyje *Aš* yra po skyrybos ženklo prieš nekirčiuotą žodį (*jį* laikom nekirčiuotu), todėl *Aš* kirčiuojam (4 taisyklė), dabar žodis *jį* yra po kirčiuoto skiemens prieš nekirčiuotą žodį (*čia* laikom nekirčiuotu), todėl nekirčiuojam (4 taisyklė). Žodis *čia* yra po nekirčiuoto žodžio prieš kirčiuotą žodį (žodį *buvau* laikom kirčiuotu, bet pirmas skiemuo nekirčiuotas, o antras kirčiuotas), todėl kirčiuojam (7 taisyklė). Pagaliau *buvau* eina po kirčiuoto skiemens prieš kirčiuotą skiemenį, todėl nekirčiuojam (5 taisyklė). Taigi visą sakinį siūloma tarti taip: *Àš jį čià buvau mãtęs.*

6. Bendras algoritmas, testavimo rezultatai, tobulinimas

Dabar bendrą klitikų paieškos algoritmą galima nusakyti tokiais 4 žingsniais:

1. Remiantis samplaikinių formų sąrašu rasti samplaikines formas ir jas sukirčiuoti.

2. Remiantis nekirčiuojamų žodžių sąrašu rasti visus nekirčiuojamus žodžius. Sukirčiuoti žodžius remiantis 1-3 taisyklėmis.

3. Pasižymėti žodžius, kuriems bus taikomos 4-9 taisyklės, visus kitus žodžius kirčiuoti.

4. Pritaikyti 4-9 taisykles.

Remiantis anksčiau aprašytomis taisyklėmis bei jų sąveika buvo sukurtos kompiuterinės programos. Testavimas buvo atliktas su anksčiau taisyklėms sudaryti naudotais tekstais (daugiau kaip 8000 žodžių) bei tekstais (beveik 1000 žodžių), kurie iki šiol dar nebuvo naudoti. Testavimo su abiem tekstais rezultatai pateikti 5 lentelėje.

5 lentelė. Testavimo rezultatai.

Taisyklės Nr.	Žodžių aibė (*pavyzdžiai*)	Tie patys tekstai taisyklių sudarymui ir testavimui		Skirtingi tekstai taisyklių sudarymui ir testavimui	
		Kirčiuotų žodžių + nekirčiuotų žodžių = suma, **(dalis tarp visų žodžių)**	Pirmo tipo klaidų + antro tipo klaidų = suma *(klaidų ir žodžių santykis)*	Kirčiuotų žodžių + nekirčiuotų žodžių = suma, **(dalis tarp visų žodžių)**	Pirmo tipo klaidų + antro tipo klaidų = suma *(klaidų ir žodžių santykis)*
1	O&NE (*o*, *ne*, *nebe*)	2 + 86 = 88	0 + 0 = 0 *(0,0%)*	0 + 12 = 12	0 + 0 = 0 *(0,0%)*
2	PRIEL_K (*link*), PRIEL_G (*prieš*)	2 + 6 = 8	1 + 0 = 1 *(12,5%)*	0 + 3 = 3	0 + 1 = 1 *(33,3%)*
3	PAROD_IV (*tas*, *šis*)	35 + 56 = 91	9 + 5 = 14 *(15,4%)*	5 + 2 = 7	1 + 0 = 1 *(14,3%)*
4	ASM_IV (*aš*, *tu*, *jis*)	84 + 152 = 236	23 + 26 = 49 *(20,8%)*	11 + 27 = 38	7 + 7 = 14 *(36,8%)*
5	BUTI (*yra*, *esu*)	64 + 30 = 94	15 + 4 = 19 *(20,2%)*	8 + 2 = 10	2 + 1 = 3 *(30,0%)*
6	KLAUS_IV (*koks*, *kuris*)	19 + 20 = 39	2 + 7 = 9 *(23,1%)*	3 + 1 = 4	1 + 0 = 1 *(25,0%)*
7	*čia*	10 + 23 = 33	2 + 0 = 2 *(6,1%)*	2 + 4 = 6	0 + 0 = 0 *(0,0%)*
8	*vis*	6 + 13 = 19	3 + 0 = 3 *(15,8%)*	0 + 1 = 1	0 + 0 = 0 *(0,0%)*
9	KI&P (*kas*, *kaip*)	65 + 105 = 170	17 + 14 = 31 *(18,2%)*	5 + 12 = 17	4 + 2 = 6 *(35,3%)*
Iš viso žodžių taisyklėse		287 + 491 = 778 **(9,3%)**	72+56 = 128 *(16,5%)*	34 + 64 = 98 **(10,2%)**	15 + 11 = 26 *(26,5%)*
Samplaikinės formos		49 + 49 = 98 **(1,2%)**	1 + 1 = 2 *(2,0%)*	9 + 9 = 18 **(1,9%)**	0 + 0 = 0 *(0,0%)*
Nekirč. žodžių sąrašas		14 + 1252 = 1266 **(15,1%)**	14 + 0 = 14 *(1,1%)*	3 + 124 = 127 **(13,3%)**	3 + 0 = 3 *(2,4%)*
Kiti žodžiai		6205 + 50 = 6255 **(74,5%)**	0 + 50 = 50 *(0,8%)*	702 + 12 = 714 **(74,6%)**	0 + 12 = 12 *(1,7%)*
Iš viso		6555 + 1842 = 8397 **(100%)**	87+107 = 194 *(2,3%)*	745+207 = 957 **(100%)**	18 + 23 = 41 *(4,3%)*

Remiantis 5 lentele galima pastebėti, kad dėl taisyklių sąveikos atsirado 6 papildomos klaidos. Atliekant testavimą su tais pačiais duomenimis, kuriais remiantis buvo kuriamos taisyklės, gauta 2,3% klaidų, o klaidų santykis su nekirčiuojamais žodžiais yra 10,5%. Atlikus testavimą su dar nenaudotais duomenimis gauta 4,3% klaidų, o klaidų santykis su nekirčiuojamais žodžiais yra 19,8%. Geriausius rezultatus davė 1 taisyklė, samplaikinės formos ir nekirčiuojamų žodžių sąrašas. Be to, atrinkus žodžius visoms taisyklėms, lieka labai nedaug nekirčiuojamų žodžių.

5 lentelėje pateiktas žodžių, kuriems buvo pritaikytos tam tikros grupės taisyklės, ir visų žodžių santykis rodo taisyklių grupės naudojimo dažnį. Nekirčiuoti žodžiai dažniausiai randami remiantis nekirčiuojamų žodžių sąrašu.

Kalbant apie sukurto algoritmo tobulinimą, pirmiausia reikia atkreipti dėmesį į tai, kad sudarant taisykles ir nekaitomų žodžių sąrašą naudojami tik tie žodžiai, kurie pasitaikė tekstuose. Svarbu yra tai, kad nekaitomų tarnybinių žodžių skaičius lietuvių kalboje yra baigtinis, todėl (DLKŽ 1993) galima rasti visus nekaitomus tarnybinius lietuvių kalbos žodžius. Toliau apsiribosime tik vienskiemeniais ir dviskiemeniais žodžiais. Kadangi remiantis naudotais tekstais

visi prielinksniai pakliuvo į nekirčiuojamų žodžių sąrašą, todėl ir kitus prielinksnius laikysime nekirčiuojamais. Remiantis (DLKŽ 1993) nekirčiuojamų žodžių sąrašą galima papildyti tokiais 14 prielinksnių: *anot*, *apsuk*, *aukščiau*, *lig*, *pagal*, *palei*, *pasak*, *pirm*, *sulig*, *užu*, *vidur*, *virš*, *viršuj*, *žemiau*. Jungtukai taip pat dažniausiai nekirčiuojami (išskyrus jungtuką *jei* ir jam sinonimišką jungtuką *jeigu*), todėl nekirčiuojamų žodžių sąrašą papildome 19 jungtukų: *arba*, *begu*, *bei*, *betgi*, *būtent*, *idant*, *ik*, *nebent*, *neg*, *norint*, *pakol*, *tačiau*, *tartum*, *tegu*, *tegul*, *tiktai*, *užuot*, *vienok*, *visgi*. Taip pat galima papildyti ir 2 taisykle apibūdinamus žodžius, kurie gali būti ir prielinksniai, ir kitos kalbos dalys. Aibę PRIEL_K papildome tokiais 6 žodžiais: *arti*, *greta*, *paskiau*, *pusiau*, *šiapus*, *vietoj*, o aibę PRIEL_G tokiais 4 žodžiais: *aplink*, *paskui*, *paskum*, *priešais*. Sudėtingesnė situacija su likusiomis (maždaug 65) dalelytėmis. Be išsamesnių tyrimų negalima nustatyti, kurios dalelytės linkusios netekti kirčio, o kurios jį išlaiko.

Anksčiau aprašytu būdu papildžius sukurto algoritmo žodžių sąrašus ir patikrinus su testavimo duomenimis klaidų skaičius sumažėjo iki 39, t. y. 4,1% klaidų, o klaidų santykis su nekirčiuojamais žodžiais 18,8%.

Išvados

Šiame darbe nekirčiuojamų žodžių paieškai lietuvių kalbos tekste pasiūlyti keturių tipų metodai: pagrįsti samplaikinių formų atpažinimu, žodžio kirčiavimo/ nekirčiavimo statistiniu dažniu, gramatikos taisyklėmis bei gretimų žodžių kirčiavimu. Kiekvienam metodui apibrėžtos žodžių klasės, kurioms jis geriausiai tinka. Paaiškinta, kaip visus metodus sujungti į vieną algoritmą. Kuriant algoritmą buvo stengiamasi minimizuoti pirmosios ir antrosios rūšies klaidų sumą. Taikant šį algoritmą testavimo duomenims gauta 4,1% klaidų tarp visų žodžių, o klaidų ir nekirčiuotų žodžių santykis yra 18,8%.

Literatūra

1. Ambrazas, V. (red.) (1996) Dabartinės lietuvių kalbos gramatika (DLKG), Mokslo ir enciklopedijų leidykla, Vilnius.
2. Keinys, S. (vyr. red.) (1993) Dabartinės lietuvių kalbos žodynas (DLKŽ) Mokslo ir enciklopedijų leidykla, Vilnius.
3. Kasparaitis, P. (2000) Automatic Stressing of the Lithuanian Text on the Basis of a Dictionary, *Informatica* 11 (1), pp19-40.
4. Kasparaitis, P. (2001) Automatic Stressing of the Lithuanian Nouns and Adjectives on the Basis of Rules, *Informatica* 12 (2), pp315-336.
5. Kazlauskienė, A., Norkevičius, G., Raškinis, G. (2004) Automatizuotas lietuvių kalbos veiksmažodžių kirčiavimas: problemos ir jų sprendimas. Baltų ir kitų kalbų fonetikos ir akcentologijos problemos, pp166-173.
6. Mauricaitė, V. (1985) Kai kurių frazės faktorių įtaka žodžių šlijimui, Kalbotyra 36 (1), pp38-43.
7. Mauricaitė, V. (1986) Kiek žodžių, tiek kirčių? Lietuvių kalba mokykloje 2, Šviesa, Kaunas, pp95-103.
8. Mauricaitė, V. (1987) Samplaikinių formų dėmenų akcentinis šlijimas, Mūsų kalba 2, pp3-6.
9. Mauricaitė, V. (1994) Akcentinis šlijimas frazės viduje (statistinis tyrimas), Kalbotyra 43 (1), pp61-64.
10. Norkevičius, G., Kazlauskienė, A., Raškinis, G. (2004) Bendrinės lietuvių kalbos daiktavardžių ir būdvardžių kirčiavimo struktūrinis modelis, algoritmas ir realizacija, Kalbų studijos 6, Technologija, Kaunas, pp72-76.
11. Pakerys, A., Pupkis, A. (1976) Lietuvių kalbos bendrinė tartis. Plokštelių tekstai, Vilnius.
12. Rimkutė, E., Jarašiūnaitė, G., Homola, P. (2005) Morfologinių samplaikų atpažinimas ir klasifikavimas, Lituanistica 62 (2), pp58-75.
13. Stundžia, B. (1991) Kirtis tekste. Žodžiai ir prasmės 1, Mokslas, Vilnius, pp86-92.

Tomas Anbinderis, Pijus Kasparaitis

Algorithms for Detecting Clitics in the Lithuanian Text

Summary

The task of automatic stressing of the Lithuanian text is analyzed in the paper. Stressed text can be used when teaching the Lithuanian language, in the text-to-speech synthesis systems, etc. In spoken language, some words are left unstressed (called clitics) and join the stressed ones. In linguistic papers it is possible to find only common tendencies of clisis however algorithms of clitics‘ search in Lithuanian language text necessary in human language technologies is a completely not researched field. Factors influencing clisis are reviewed and methods for detecting clitics are offered in the present paper. The methods are based on 1) recognizing combinational forms, 2) the statistical frequency of word being stressed/ unstressed, 3) some grammatical rules, 4) stressing of adjacent words. The second method is very simple and quite reliable but better results were achieved when using the third and the fourth methods for some classes of words. Words’ classes are defined as well as which method suits best. It is explained, how to join all the methods into one algorithm. We attempt to minimize the sum of mistakes of 1st type and 2nd type when creating this algorithm. By applying this algorithm to the testing data 4.1% mistakes are received among all the words, and the ratio of mistakes and unstressed words is 18.8%.

Straipsnis įteiktas 2006 08
Parengtas publikuoti 2007 04

Autoriai

Tomas Anbinderis, studijuoja Vilniaus universiteto Informatikos doktorantūroje ir dirba Vilniaus universiteto Matematikos ir informatikos fakulteto Kompiuterijos katedroje asistentu.
Mokslinių tyrimų sritis apima balso sintezę iš teksto ir skaitmeninį vaizdų apdorojimą.
Adresas: Vilniaus universitetas Matematikos ir informatikos fakultetas Kompiuterijos katedra, Naugarduko 24, 03225 Vilnius, Lietuva.
El. paštas: Tomas.Anbinderis@mif.vu.lt

Pijus Kasparaitis, dr. (fiziniai mokslai), Vilniaus universiteto Matematikos ir informatikos fakulteto Kompiuterijos katedros lektorius. 2001 m. apgynė daktaro disertaciją „Lietuvių kalbos kompiuterinė sintezė". Per pastaruosius 5 metus paskelbė 10 publikacijų, tris iš jų recenzuojamuose žurnaluose.
Mokslinių tyrimų sritis apima balso sintezę iš teksto ir kitas kompiuterinės lingvistikos sritis.
Adresas: Vilniaus universitetas Matematikos ir informatikos fakultetas Kompiuterijos katedra, Naugarduko 24, 03225 Vilnius, Lietuva.
El. paštas: pkasparaitis@yahoo.com