

KOMPIUTERINĖ LINGVISTIKA/ COMPUTATIONAL LINGUISTICS

Bendrinės lietuvių kalbos daiktavardžių ir būdvardžių kirčiavimo struktūrinis modelis, algoritmas ir realizacija

Giedrius Norkevičius, Asta Kazlauskienė, Gailius Raškinis

Anotacija. Lietuvių kalbos vardažodžių (ir ne tik jų) kirčiavimas išsamiai išanalizuotas, neblogai kodifikuotas ir aprašytas daugelio lietuvių kalbininkų, tam skirta ne viena įvairaus pobūdžio knyga ar vadovėlis. Vis dėlto kalbos technologijų darbuose pasigendama išsamaus struktūrinio lietuvių kalbos kirčiavimo modelio, be kurio neįmanomi kokybiški šnekos sintezės programiniai produktai, negalima parengti modernių kompiuterinių mokomųjų kirčiavimo programų, kalbos vartotojai neturi šiuolaikinių galimybių greitai pasitikrinti, ar taisyklingai kirčiuoja konkrečius žodžius.

Pranešime pristatoma tik viena mūsų kuriamos kompiuterinės programos dalis – veikiantis daiktavardžių ir būdvardžių kirčiavimo algoritmas ir svarbiausia jo sudedamoji dalis – formalizuota kirčiavimo taisyklių medžių struktūra. Aptariamos pagrindinės problemos, kurios iškilo įprastas lingvistines taisykles pritaikant kuriamuose algoritmuose. Nagrinėjama sąveika su morfologinės ir leksinės informacijos duomenų bazėmis. Pateikiami preliminarūs programos testavimo rezultatai ir jų gerinimo būdai. Analizuojamas galimybės, leidiančios patobulinti kai kurias programos dalis.

Darbe pristatomas veikiantis daiktavardžių ir būdvardžių kirčiavimo algoritmas ir svarbiausia jo sudedamoji dalis – formalizuota medžio pavidalo kirčiavimo taisyklių struktūra. Pateikiami struktūrinio modelio sudarymo principai, pagrindžiamas jo tinkamumas kurti automatinio lietuvių kalbos kirčiavimo algoritmus. Aptariamos pagrindinės problemos, kurios iškilo įprastas lingvistines taisykles pritaikant kuriamuose algoritmuose. Nagrinėjama sąveika su morfologinės ir leksinės informacijos duomenų bazėmis. Pateikiamas preliminarus automatinio kirčiavimo tikslumo įvertinimas, gautas kirčiuojant daugiau nei 24 mln. daiktavardžių beveik 8 mln. būdvardžių, esančių VDU tekstyne, ir kirčiavimo tikslumo gerinimo būdai. Analizuojamas galimybės, leidiančios patobulinti kai kurias algoritmo dalis.

Įvadas

Lietuvių kalbos daiktavardžių ir būdvardžių kirčiavimas išsamiai išanalizuotas, neblogai kodifikuotas ir aprašytas daugelio lietuvių kalbininkų, tam skirta ne viena įvairaus pobūdžio knyga ar vadovėlis. Vis dėlto kalbos technologijų darbuose pasigendama išsamaus struktūrinio lietuvių kalbos kirčiavimo modelio, be kurio neįmanomi kokybiški šnekos sintezės programiniai produktai, negalima parengti modernių kompiuterinių mokomųjų kirčiavimo programų, kalbos vartotojai neturi šiuolaikinių galimybių greitai pasitikrinti, ar taisyklingai kirčiuoja konkrečius žodžius. Šiame darbe nagrinėjame tik vieną mūsų kuriamos kompiuterinės programos, skirtos automatiniam lietuvių kalbos žodžių kirčiavimui, dalį – daiktavardžių bei būdvardžių automatinį kirčiavimą, automatinio kirčiavimo struktūrinio modelio sudarymo bei lingvistinių taisyklių pritaikymo kompiuteriui problemas.

Analogiški darbai

Dauguma kitų kalbų turi fiksuotą kirtį, t.y. kirčio vietą galima nusakyti griežtomis taisyklėmis. Dažniausiai tai būna visai paprasti teiginiai, nurodą kirčio nutolimą nuo žodžio

pradžios ar pabaigos. Pagal nuotolį skiriami trys fiksuoto kirčio modeliai:

1. Pastoviai kirčiuojamas pirmasis žodžio skiemo. Šią sistemą turi latvių, čekų, slovakių, islandų, estų, suomių, vengrų kalbos
2. Pastoviai kirčiuojamas paskutinis skiemo. Šios rūšies kirčiavimas būdingas daugumai tiurkų kalbų, taip pat persų (ir tadžikų) kalbai. Panašiai kirčiuojama ir prancūzų kalboje, tik kirtį gauna ne žodžiai, o tam tikros reikšminės jų grupės.
3. Pastoviai kirčiuojamas priešpaskutinis skiemo. Priešpaskutinio skiemens kirtį turi lenkų kalba.

Galimi ir sudėtingesni fiksuoto kirčio modeliai, kai kirčio vieta priklauso ne tik nuo žodžio ribų, bet ir nuo balsių bei skiemenų kiekybės. Pavyzdžiui, mongolų kalboje kirtį gauna pirmasis ilgas žodžio skiemo, o kai visi žodžio skiemenys trumpi, pirmas skiemo. Fiksuotą kirtį turinčiose kalbose automatinis kirčiavimas nesukelia ypatingų problemų.

Lietuvių kalba, kaip ir rusų, bulgarų, italų, ispanų, anglų turi laisvą kirtį. Kai kuriose laisvą kirtį turinčiose kalbose daugelis vienodas galūnes turinčių žodžių kirčiuojami vieno-

dai, pavyzdžiui taip yra italų kalboje. Daugelyje laisvojo kirčio kalbų žodžio paradigmoje kirčio vieta nesikeičia – jos turi vadinamąjį laisvąjį pastovų kirtį. Šiuo atveju automatinis kirčiavimas nėra sudėtingas – informacija apie kirtį glūdi pačiame žodyje. Lietuvių kalboje gali būti kirčiuojamas bet kuris skienu, be to, žodžio paradigmoje kirtis nėra pastovus: gali šokinėti iš vieno skienens į kitą. Todėl lietuvių kalbos kirčiavimas labai sudėtingas. Vis dėlto yra bandymų spręsti lietuvių kalbos automatinio kirčiavimo problemą. Gana plačiai daiktavardžių ir būdvardžių kirčiavimą išnagrinėjo P. Kasparaitis (2001). Autoriaus pateikta metodika pasižymi tuo, kad yra nagrinėjama ne morfologinė informacija, o žodžio sandaros ypatumai, todėl kirčiavimo taisyklės paremtos žodžio sandaros ypatumais, o tai lemia sudėtingą jų suvokimą, modifikavimą bei patikrinimą, kaip atitinka tradicines kirčiavimo taisykles, kurios remiasi morfologine informacija.

Automatinio kirčiavimo algoritmas

Algoritiškai paprasčiausias automatinio kirčiavimo būdas – sudaryti visų lietuvių kalboje vartojamų žodžių sąrašą, įskaitant visas jų formas (linksnius, skaičius, giminę), ir kiekvienam sąrašo žodžiui nurodyti reikiamą kirčio ženklą ir vietą. Kirčiuojant tekstą, kiekvienas žodis visų pirma būtų ieškomas šiame sąraše, o, jį radus, būtų tiesiog išspausdinamas taisyklingas jo kirčiavimas. Deja, toks kirčiavimo modelis būtų neefektyvus. Vien tik atsižvelgus daiktavar-

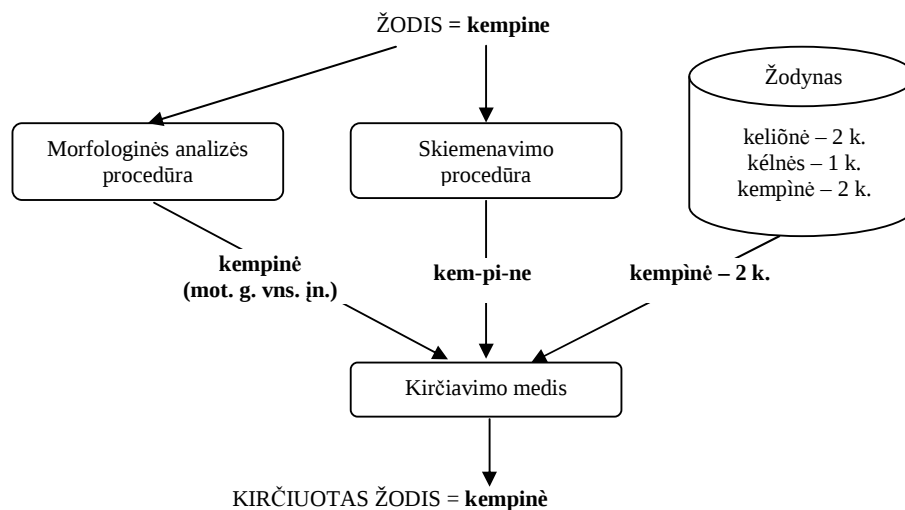
džio kaitymą skaičiais ir linksniais, duomenų bazėje reikėtų saugoti 13 skirtingų to paties žodžio formų. Todėl automatiniam kirčiavimui verta pasinaudoti akcentologų sukurtomis lietuvių kalbos kirčiavimo taisyklėmis.

Panagrinėkime tokią daiktavardžių kirčiavimo taisyklę, kuri pateikiama lietuvių kalbos vadovėlyje (1989):

Pirmos kirčiuotės daugiaskiemeniai daiktavardžiai turi pastovų kirtį visuose linksniuose

Nors ši taisyklė yra viena iš paprasčiausių, ji akivaizdžiai rodo, kad, norėdami automatiškai sukirčiuoti tam tikrą daiktavardžių tipą, visų pirma privalome turėti algoritminiam naudojimui tinkamu būdu užrašytas specifines kirčiavimo žinias. Šiuo atveju turime mokėti nustatyti:

- žodžio morfologines savybes ir jo pagrindinę formą (pabrakta),
- reikiamo žodžio skienų kiekį (paryškinta),
- kirčiuotę bei žinoti kaip kirčiuojama pagrindinė forma (kursyvas). Ši trejopa informacija apie kirčiuojamą žodį reikalinga ir kitų daiktavardžio ir būdvardžio formų kirčiavimui. Taigi kuriamas daiktavardžių ir būdvardžių automatinio kirčiavimo algoritmas (1 pav.) turi apimti procedūras, kurios suitektų visų trijų tipų žinias:



1 pav. Daiktavardžių ir būdvardžių kirčiavimo algoritmo schema

Morfologinės analizės procedūra (2000) nustato nagrinėjamo žodžio giminę, skaičių, linksnį ir pagrindinę formą. Skiemenavimo procedūra suskaido nagrinėjamą žodį skienimis ir nustato jų skaičių. Žodyne saugoma informacija apie kiekvieno daiktavardžio ir būdvardžio pagrindinės formos kirčiavimą: kirčio vietą, tipą ir kirčiuotę. Visa ši informacija vėliau pateikiama kirčiavimo procedūrai, kuri turi medžio pavidalą ir yra vadinama kirčiavimo medžiu (2 pav.).

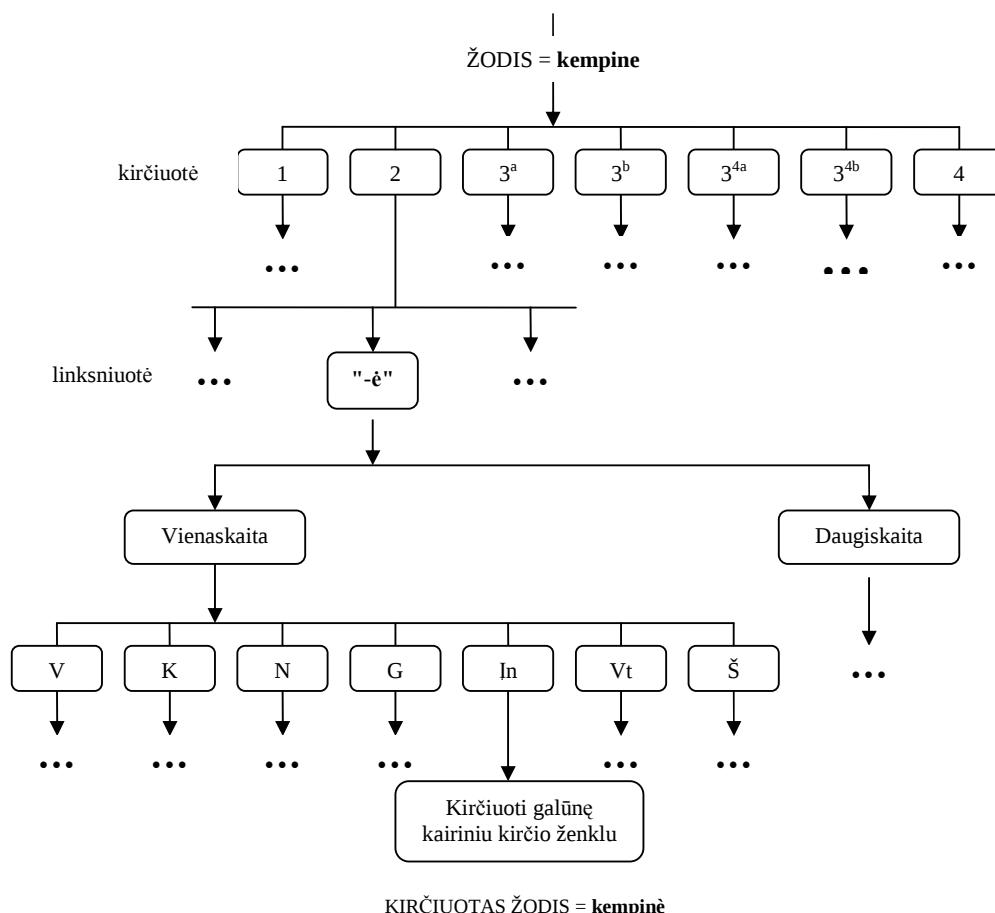
Lietuvių kalbos daiktavardžių ir būdvardžių kirčiavimo taisyklių struktūrizavimui medžio struktūra pasirinkta neatsitiktinai. Pasirinkimą lemia tai, kad:

- medis yra lengvai sudaroma ir modifikuojama žinių struktūra;

- medyje aiškus operacijų eiliškumas, todėl jis lengvai paverčiamas algoritmu.

Analizuojant turimus duomenis bei vadovėliuose aprašytas kirčiavimo taisykles, susidurta su keliomis problemomis:

- Visi daiktavardžiai pagal kirčiavimo panašumą akcentologų yra skirstomi į keturias grupes – kirčiuotes (jeigu skaičiuotume ir 3 kirčiuotės variantus 3^a , 3^b , 3^{4a} , 3^{4b} , į 7 grupes). Tokio skirstymo visiškai užtenka žmogui, bet jo nepakanka kompiuteriui.



2 pav. Daiktavardžių ir būdvardžių kirčiavimo medžio iliustracija

Taip yra todėl, kad kiekviena kirčiuotė turi bent po penkis skirtingos linksniuotės pavyzdžius. Vadinasi, kiekvienas kirčiuotės linksnis gali turėti bent po penkias skirtingas galūnes. Jeigu kirčiuojama galūnė, tada kompiuteris turi turėti tos galūnės kirčiavimo modelį / algoritmą.

- Žinodami kirčiuotę ir galūnės kirčiavimo modelį, galime nustatyti, kaip kirčiuojamas konkretus daiktavardis. Jeigu pagrindinė forma kirčiuojama taip pat galūnėje (pvz., *kasà*), neturime informacijos, koks kirtis būdingas šakniam. Ši problema aktuali antrosios, trečiosios bei ketvirtosios kirčiuotės žodžiams. Algoritmiškai problemą galima išspręsti, remiantis tokiu dėsniu: jeigu kirčiuotinoje šaknyje turime mišrųjį dvigarsį, dvibalsį ar prigimtinio ilgumo balsius *a, e, é, i, u, ū, y*, tai antrosios ir ketvirtosios kirčiuotės žodžiai kirčiuotini tik tvirtagališkai (riestiniu kirčio ženklu), o trečiosios kirčiuotės – tvirtapradiškai (dvigarsiai su pirmuoju dėmeniui *i, u* kairiniu kirčiu, kiti dešininiai). Jeigu antroje ir ketvirtoje kirčiuotėje kirtis nušoka į kirčiuotą šaknies balsį *a* arba *e*, neaišku, ar kokio kirčio ženklo reikia (nes nežinoma balsio kiekybė: padėtinio ilgumo ar trumpasis balsis): kairinio ar riestinio (pvz., *kàsq*).
- Negalima sudaryti elementarių taisyklių priešdėliams, sudurtiniams daiktavardžiams, nes jų kirčiavimas nėra absoliučiai vienodas, dėsningumai akivaiz-

dūs, tačiau algoritmizuojant reikėtų labai ilgo išimčių sąrašo.

Daiktavardžių ir būdvardžių kirčiavimo tikslumas

Sukurto algoritmo kirčiavimo tikslumas buvo vertinamas pasitelkiant 85 mln. žodžių apimties VDU lietuvių kalbos tekstyną (2000). Šiame tekстыne morfologinės analizės algoritmas atpažino apie 24 mln. daiktavardžių ir 7,9 mln. būdvardžių, kurie ir buvo pateikti kirčiavimo algoritmui. Algoritmas galėjo nekirčiuoti kai kurių žodžių tuo atveju, jei šioms žodžiams taisyklingai sukirčiuoti trūko reikiamos informacijos. Kirčiavimo tikslumas buvo vertinamas pagal du kriterijus:

- Žodžių procentą, kurį algoritmas galėjo kirčiuoti.
- Taisyklingai sukirčiuotų žodžių procentą.

Žodžių, besiskiriančių gimine, skaičiumi ar linksniu, kirčiavimo tikslumas buvo skaičiuojamas atskirai. Kirčiavimo tikslumą vertino ekspertas: jis peržiūrėjo 100 dažniausių kiekvienos daiktavardžio ir būdvardžio formos kompiuteriu sukirčiuotų žodžių. Tikslumo įvertinimai pateikti 1 ir 2 lentelėse.

1 lentelė. Daiktavardžių automatinio kirčiavimo tikslumas

Daiktavardžio forma			Rasta tekste (žodžiai)	Negalėjo kirčiuoti (žodžiai)	Galėjo kirčiuoti (%)	Teisingai kirčiavo (%)
vyr. g.	Vienaskaita	V	2 156 455	370 074	83	99
		K	3 354 490	577 476	83	100
		N	329 478	34 047	90	97
		G	1 346 184	139 321	90	95
		In	678 698	45 387	93	98
		Vt	532 495	133 414	75	99
		S	261 599	50 038	81	94
	Daugiskaita	V	1 260 710	123 715	90	96
		K	1 947 068	171 577	91	93
		N	250 048	9 727	96	96
		G	859 640	68 159	92	91
		In	406 913	18 767	95	93
		Vt	152 295	18 537	88	93
		mot.g.	Vienaskaita	V	1 372 248	354 004
K	2 293 029			650 784	72	88
N	460 803			201 361	56	84
G	1 212 933			160 528	87	85
In	900 049			223 323	75	99
Vt	615 110			167 332	73	92
S	6 741			1 331	80	75
daugiskaita	V		1 384 943	70 152	95	91
	K		1 214 491	270 465	78	94
	N		95 053	4 859	95	92
	G		608 517	92 224	85	93
	In		163 304	10 128	94	89
	Vt		150 987	3 821	97	80
	Iš viso		24 014 281	3 970 551	83,47	92,54

Iš 1 lentelės matyti, kad algoritmas galėjo kirčiuoti 83,47% lietuvių kalbos tekstuose rastų daiktavardžių ir 92,54% jų sukirčiavo taisyklingai. Dažniausiai negalėjo kirčiuoti dėl to, kad šiuo metu automatiniam kirčiavimui naudojamas daiktavardžių žodynas yra ilgas ir neišsamus (beveik 45 tūkst. daiktavardžių). Preliminari žodyno analizė parodė, kad jį būtina ir galima gerokai patobulinti. Pavyzdžiui, žodyne trūksta daugelio dažnai vartojamų asmenvardžių ir vietovardžių, be to, žodyne kaip antraštinė forma teikiamas tik vyriškosios giminės daiktavardis (dažniausiai veikėjų pavadinimai), nėra moteriškosios giminės. Kita vertus, veiksmažodinių abstraktų su priesagomis -imas, -ymas dabartiniame žodyne yra beveik 17 tūkst. Šiuos vedinius vertėtų pašalinti iš žodyno, nes jų kirčiavimas nesudėtingas ir jį galima algoritmizuoti:

- pastoviai arba pirmą kirčiuotą kirčiuojami tie vediniai, kurie padaryti iš daugiaskiemenių būtojo kartinio laiko veiksmažodžių, t.y. šie žodžiai išlaiko saugomų daugiaskiemenių mišriųjų veiksmažodžių žodyno (2004) 2 formos kirčio vietą ir priegaide.
- pagal antrą kirčiuotą kirčiuojami tie vediniai, kurie padaryti iš dviskiemenių būtojo kartinio laiko veiksmažodžių.

2 lentelė. Būdvardžių automatinio kirčiavimo tikslumas

Būdvardžio forma			Rasta tekste (žodžiai)	Negalėjo kirčiuoti (žodžiai)	Galėjo kirčiuoti (%)	
vyr. g.	vienaskaita	V	682 655	119 438	83	
		K	530 417	109 572	79	
		N	63 632	6 473	90	
		G	607 127	92 332	85	
		In	185 552	36 316	80	
		Vt	55 990	7 370	87	
		S	14 926	3 911	74	
	daugiskaita	V	722 839	385 231	47	
		K	425 804	139 080	67	
		N	50 624	12 468	75	
		G	233 169	43 072	82	
		In	107 310	19 019	82	
		Vt	33 867	7 532	78	
		mot.g.	Vienaskaita	V	565 840	95 119
K	320 190			48 407	85	
N	1 009 938			83 212	92	
G	303 331			40 627	87	
In	313 361			52 805	83	
Vt	93 341			17 851	81	
S	320 190			48 407	85	
daugiskaita	K		424 224	137 500	68	
	N		26 275	3 364	87	
	G		281 402	41 046	85	
	In		53 581	7 043	87	
	Vt		39 022	5 613	86	
	bevardė. g.		419 937	49 351	88	
	Iš viso		7 884 544	1 612 159	79.55	

Kirčiuodami minėtus vedinius ne pagal žodyną, o pagal struktūrizuotas / algoritmizuotas taisykles, sutrumpintume naudojamo žodyno apimtį beveik trečdaliu, be to, remiantis taisyklėmis, būtų galima kirčiuoti naujausius vedinius. Tai labai svarbu, nes šis darybos tipas labai produktyvus: tokius vedinius galima padaryti iš bet kurio veiksmažodžio, dabar jų pasidaroma ir iš įvairių tarptautinių žodžių. Šių vedinių tekste tikrai nemažai pasitaiko, nes pastebimas akivaizdus polinkis kalbėti abstrakčiau.

Išvados

- Darbo metu sukurtas daiktavardžių ir būdvardžių kirčiavimo algoritmas, kuris remiasi kirčiuojamų žodžių morfologine ir skiemenine analize, kirčiavimo žodynu bei į medžio pavidalą pertvarkytomis lietuvių kalbos kirčiavimo taisyklėmis.
- Sukurtų algoritmų veikimo tikslumas buvo patikrintas, kirčiuojant 24 mln. daiktavardžių ir 7,9 mln. būdvardžių. Algoritmas turėjo pakankamai informacijos ir galėjo kirčiuoti 83,47% daiktavardžių ir 79,55% būdvardžių, o kirčiavimo tikslumas viršijo 92%.
- Tobulinant sukurtą algoritmą, pirmiausiai reikia atkreipti dėmesį į turimą žodyną:

- a) išanalizuoti priesaginių daiktavardžių ir būdvardžių kirčiavimo dėsningumą ir įmanomais atvejais priesaginius vedinius kirčiuoti remiantis taisyklėmis, o ne žodynu;
- b) į žodyną įtraukti ir dažniausius vietovardžius bei asmenvardžius;
- c) papildyti žodyną moteriškosios giminės daiktavardžiais ir būdvardžiais. Be to, būtina patikslinti būdvardžio įvardžiutinių formų kirčiavimą.

Literatūra

1. Kasparaitis, P. (2001). Lietuvių kalbos kompiuterinės sintezė, daktaro disertacija.
2. Zinkevičius, V. (2000). Lemuoklis – morfologinei analizei. Darbai ir dienos, 24, Vytauto Didžiojo universitetas, p. 245-274.
3. Kniūkšta, P., Lyberis, A. (1989). Mokomasis lietuvių kalbos ir kirčiavimo žodynas, Šviesa.
4. Kazlauskienė, A., Norkevičius, G., Raškinis, G. (2004). Automatizuotas lietuvių kalbos veiksmažodžių kirčiavimas: problemos ir jų sprendimo būdai, prešimas pateiktas konferencijai „Baltų ir kitų kalbų fonetikos ir akcentologijos problemos“, VPU.
5. Marcinkevičienė, R. (2000). Tekstynų lingvistika (teorija ir praktika) Darbai ir dienos, 24, Vytauto Didžiojo universitetas, p. 7-63.

Giedrius Norkevičius, Asta Kazlauskienė, Gailius Raškinis

Accentuation of Lithuanian Nouns and Adjectives: Structural Model, Algorithm and Implementation

Summary

This paper describes an algorithm of automatic accentuation of Lithuanian nouns and adjectives. The algorithm incorporates the procedure of morphological analysis, and the procedure of word segmentation into syllables. It uses accentuation thesaurus, and incorporates tree-structured accentuation knowledge. Paper shows the feasibility of such tree-structured model for creating practical automatic accentuation tool. The algorithm was tested on 24 and 7,9 millions of nouns and adjectives respectively taken from VMU text corpus. It showed accuracy levels superior to 80%.

Straipsnis įteiktas 2002 05
Parengtas publikuoti 2002 11

Autoriai

Gailius Raškinis, dr. (fiziniai mokslai), Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro mokslo darbuotojas ir Informatikos fakulteto Taikomosios informatikos katedros dėstytojas. Autoriau disertacinio darbo pagrindu sukurta muzikos atpažinimo sistema yra naudojama 25 pasaulio šalyse. Per pastaruosius 4 metus autorius dalyvavo 4 tarptautiniuose ir respublikiniuose mokslo projektuose, paskelbė daugiau nei 10 publikacijų.

Mokslinių tyrimų sritis apima mašininio mokymo metodų kūrimą bei taikymą kalbos technologijų uždaviniams spręsti.

Adresas: Vytauto Didžiojo universitetas, Kompiuterinės lingvistikos centras, K. Donelaičio g. 52, Kaunas LT-44244, Lietuva.

El. paštas: g.raskinis@if.vdu.lt

Giedrius Norkevičius, studijuoja Vytauto Didžiojo universiteto Informatikos doktorantūroje ir dirba asistentu Informatikos fakulteto Taikomosios informatikos katedroje. Autorius yra paskelbęs publikacijų mokslinėje spaudoje.

Mokslinių tyrimų sritis susijusi su įvairių lietuvių šnekos aspektų kompiuteriniu modeliavimu: automatinio lietuvių kalbos teksto kirčiavimu, kompiuteriniais lietuvių šnekos intonacijos ir garsų trukmės modeliais.

Adresas: Vytauto Didžiojo universitetas, Informatikos fakultetas, Taikomosios informatikos katedra, Vileikos g. 8, Kaunas LT-44404, Lietuva.

El. paštas: Giedrius_Norkevicius@fc.vdu.lt

Asta Kazlauskienė, dr. (humanitariniai mokslai) ir Vytauto Didžiojo universiteto Lietuvių kalbos katedros docentė. Per pastaruosius 4 metus paskelbė daugiau nei 10 publikacijų.

Mokslinių tyrimų sritis: lietuvių kalbos ir tarmių fonetika ir akcentologija, taikomoji kalbotyra.

Adresas: Vytauto Didžiojo universitetas, Lietuvių kalbos katedra, K. Donelaičio g. 52, Kaunas LT-44244, Lietuva.

El. paštas: Asta_Kazlauskiene@fc.vdu.lt

