

BENDRINĖS LIETUVIŲ KALBOS DAIKTAVARDŽIŲ IR BŪDVARDŽIŲ KIRČIAVIMO STRUKTŪRINIS MODELIS, ALGORITMAS IR REALIZACIJA

Giedrius Norkevičius, Asta Kazlauskienė, Gailius Raškinis

Vytauto Didžiojo universitetas, K. Donelaičio g. 58, 3000 Kaunas

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 automatinių lietuvių kalbos kirčiavimo algoritmus. Aptariamos pagrindinės problemos, kurios išskilo įprastas lingvistines taisykles pritaikant kuriamuose algoritmuose. Nagrinėjama sąveika su morfologinės ir leksinės informacijos duomenų bazėmis. Pateikiamas preliminarus automatinių kirčiavimo tikslumo įvertinimas, gautas kirčiuojant daugiau nei 24 mln. daiktavardžių beveik 8 mln. būdvardžių, esančių VDU tekstyne, ir kirčiavimo tikslumo gerinimo būdai. Analizuojamos galimybės, leisiančios 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, automatinių 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 skiemuo. Šią sistemą turi latvių, čekų, slovakų, islandų, estų, suomių, vengrų kalbos
2. Pastoviai kirčiuojamas paskutinis skiemuo. 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 skiemuo. Priešpaskutinio skiemens kirtį turį 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 skiemuo, o kai visi žodžio skiemenys trumpi, pirmas skiemuo. 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 vienodai, 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 skiemuo, be to, žodžio paradigmoje kirtis nėra pastovus: gali šokinėti iš vieno skiemens į kitą. Todėl lietuvių kalbos kirčiavimas labai sudėtingas. Vis dėlto yra bandymų spręsti lietuvių kalbos automatinių kirčiavimo problemą. Gana plačiai daiktavardžių ir būdvardžių kirčiavimą išnagrinėjo P. Kasparaitis [1]. 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

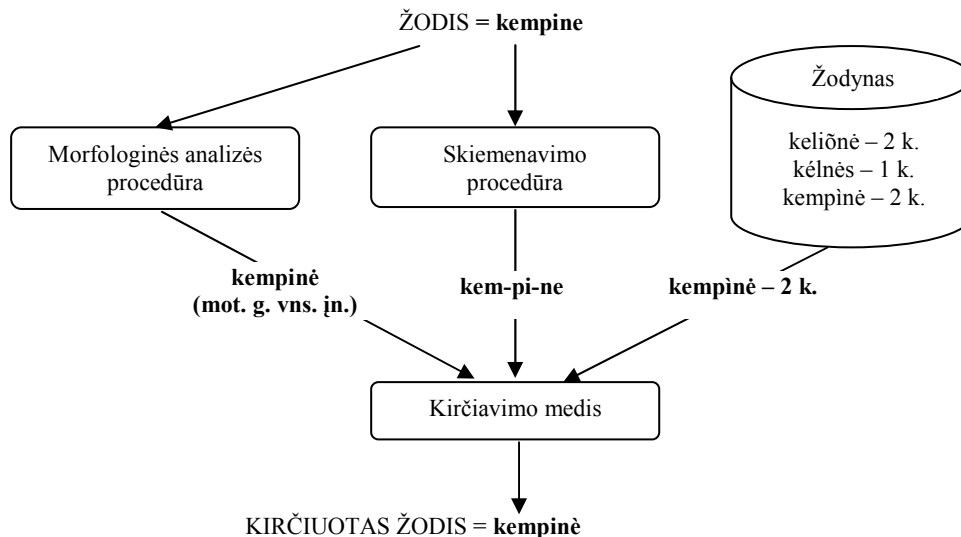
Algoritmiškai paprasčiausias automatinių kirčiavimo būdas – sudaryti visų lietuvių kalboje vartojamų žodžių sąrašą, įskaitant visas jų formas (linksnius, skaičius, giminę), ir kiekvienam sarašo žodžiui nurodyti reikiamą kirčio ženklą ir vietą. Kirčiuojant tekstą, kiekvienas žodis visų pirma būtų ieškomas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

daiktavardžio 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 [3]:

Pirmos kirčiuotės daugiaskiemieniai 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: a) žodžio morfologines savybes ir jo pagrindinę formą (pabraukta), b) reikiamo žodžio skiemenų kiekį (paryškinta), c) 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 suteiktų visų trijų tipų žinias:

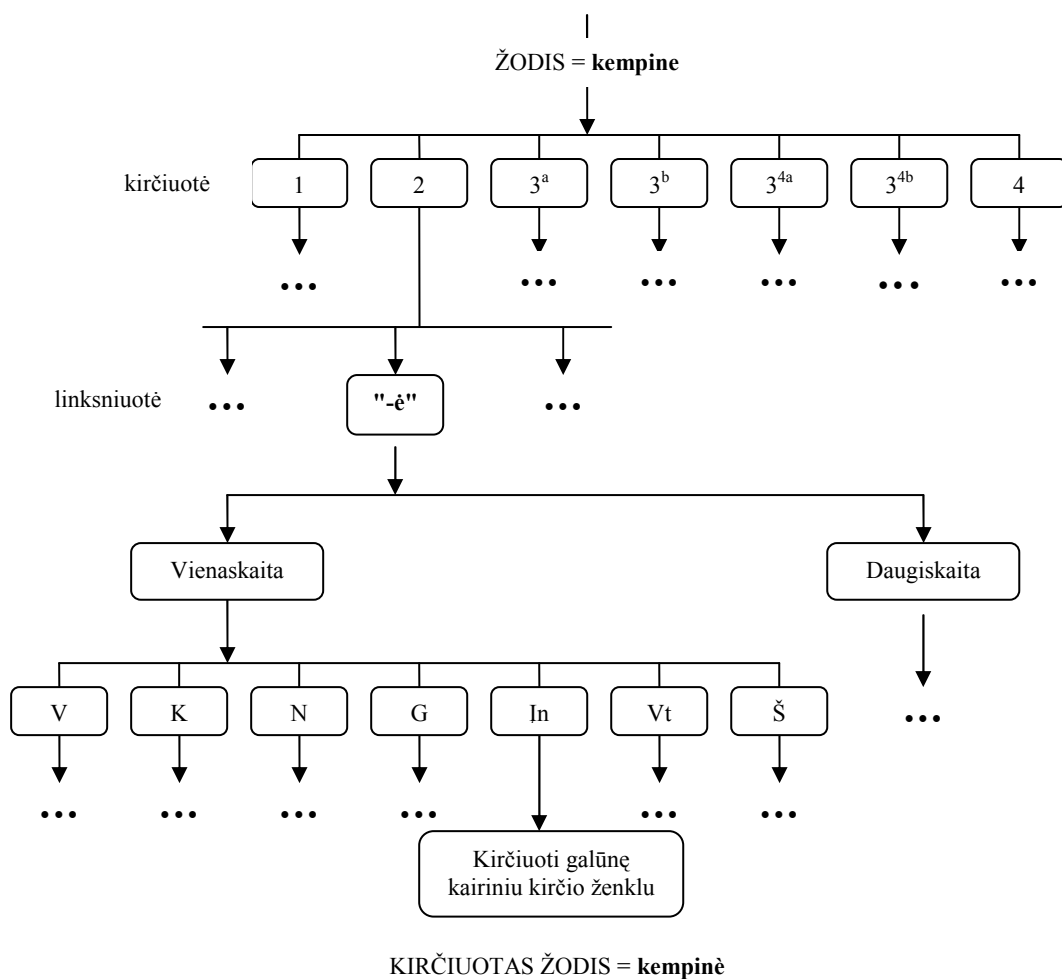


1 pav. Daiktavardžių ir būdvardžių kirčiavimo algoritmo schema

Morfologinės analizės procedūra [2] nustato nagrinėjamo žodžio giminę, skaičių, linksnį ir pagrindinę formą. Skiemenavimo procedūra suskaido nagrinėjamą žodį skiemenimis 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:

- a) medis yra lengvai sudaroma ir modifikuojama žinių struktūra;
- b) medyje aiškus operacijų eiliškumas, todėl jis lengvai paverčiamas algoritmu.



2 pav. Daiktavardžių ir būdvardžių kirčiavimo medžio iliustracija

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

1. 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. 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.
2. 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. Algoritiškai 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*, neišku, ar kokio kirčio ženklo reikia (nes nežinoma balsio kiekybė: padėtinio ilgumo ar trumpasis balsis): kairinio ar riestinio (pvz., *kàsą*).
3. Negalima sudaryti elementarių taisyklių priešdėliniams, sudurtiniams daiktavardžiams, nes jų kirčiavimas nėra absoliučiai vienodas, dėsningumai akivaizdūs, 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 [5]. 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

šiems žodžiams taisyklingai sukirčiuoti trūko reikiamos informacijos. Kirčiavimo tikslumas buvo vertinamas pagal du kriterijus:

1. Žodžių procentą, kurį algoritmas galėjo kirčiuoti.
2. 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 (%)	
vyt. 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

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	
daugiskaita	V		320 190	48 407	85	
	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	

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 automatiniame kirčiavime 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:

a) pastoviai arba pirmąją kirčiuote kirčiuojami tie vediniai, kurie padaryti iš daugiaskiemenių būtojo kartinio laiko veiksmažodžių, t. y. šie žodžiai išlaiko saugomų daugiaskiemenių mišriųjų veiksmažodžių žodyno [4] 2 formos kirčio vietą ir priegaidę.

b) pagal antrąją kirčiuotę kirčiuojami tie vediniai, kurie padaryti iš dviskiemenių būtojo kartinio laiko veiksmažodžių.

Kirčiuodami minėtus vedinius ne pagal žodyną, o pagal struktūrizuotas/algoritmuotas 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

1. 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.
2. 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%.
3. Tobulinant sukurtą algoritmą, pirmiausiai reikia atkreipti dėmesį į turimą žodyną: a) išanalizuoti priesaginių daiktavardžių ir būdvardžių kirčiavimo dėsningumus 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 įvardžiuotinių formų kirčiavimą.

Literatūros sąrašas

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

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%.