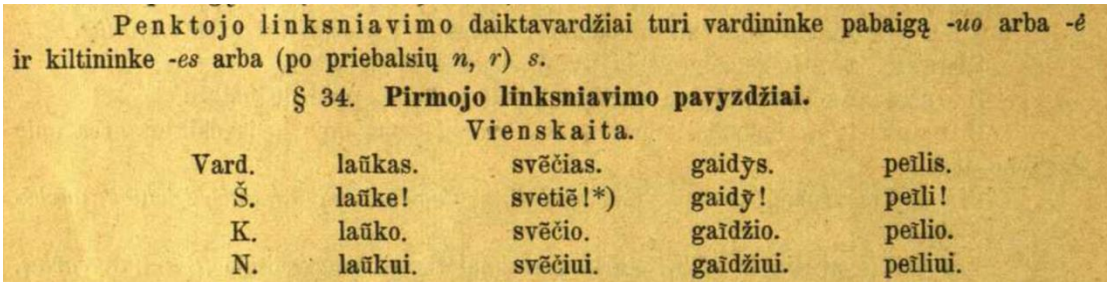
VIRGINIJUS DADURKEVIČIUS
Vilniaus universitetas

LIETUVIŲ KALBOS MORFOLOGIJA ATVIROJO KODO *HUNSPELL* PLATFORMOJE[1]

ESMINIAI ŽODŽIAI: kompiuterinė lingvistika, lietuvių kalba, gramatika, morfologija, morfologinė analizė, rašybos tikrinimas, teksto indeksavimas, paieška, atvirasis kodas, *Hunspell*.

ĮVADAS

Prieš daugiau kaip 100 metų Jono Jablonskio pastangomis (1901, 1918, 1922) lietuvių kalba pradėta norminti ir pritaikyta būti pagrindine valstybę vienijančia priemone. Jau pirmojoje Jono Jablonskio *Lietuviškos kalbos gramatikoje* (1901) lietuvių kalbos fonetika, morfologija ir sintaksė buvo formalizuotos: apibrėžtos pagrindinės sąvokos, pateiktos žodžių kaitybos ir sakinio sandaros taisyklės (žr 1 pav.).



Penktojo linksniavimo daiktavardžiai turi vardininke pabaigą *-uo* arba *-ė* ir kiltininke *-es* arba (po priebalsių *n*, *r*) *s*.

§ 34. **Pirmojo linksniavimo pavyzdžiai.**

Vienskaita.

Vard.	laũkas.	svẽčias.	gaidỹs.	peĩlis.
Š.	laũke!	svetiẽ!*)	gaidỹ!	peĩli!
K.	laũko.	svẽčio.	gaĩdžio.	peĩlio.
N.	laũkui.	svẽčiui.	gaĩdžiui.	peĩliui.

1 PAV. Ištrauka iš Jono Jablonskio *Lietuviškos kalbos gramatikos* 1901[2]

Neįkainojamas Jono Jablonskio indėlis norminant lietuvių kalbos leksiką, aiškinant semantinius žodžių ypatumus.

Ilgą laiką to visiškai pakako mokiniams ir mokytojams, rašytojams ir skaitytojams, valdininkams ir mokslininkams. Atsiradus civilizaciją keičiančiai naujovei – kompiuteriams – klasikinei gramatikai ir leksikai iškilo būtinybė „apsivilkti naują drabužį“ ir tapti visaverte naujųjų

[1] Straipsnis parengtas pagal pranešimą *Lietuvių kalbos gramatika atvirojo kodo pasaulyje*, skaitytą 24-ojoje tarptautinėje Jono Jablonskio konferencijoje *Skaitmeniniai kalbos ištekliai, jų plėtros kryptys ir panaudos galimybės* (2017 09 29); ją organizavo Lietuvių kalbos instituto Bendrinės kalbos tyrimų centras ir Vilniaus universiteto Lietuvių kalbos katedra.

[2] Paveikslas iš http://www.epaveldas.lt/vbspi/biRecord.do?biRecordId=25383.

technologijų dalimi. Pirmiausia kalbos dalykų kompiuteriuose prireikė koduojant tekstą (keičiant raides į skaičius), automatiškai koreguojant įvedamo teksto klaidas, skiemenuojant abipusės lygiuotės skilčių tekstą. Kiek vėliau – informacijos paieškai, šnekos analizei bei sintezei, mašininiam vertimui iš vienos kalbos į kitą, automatiniam turinio suvokimui, teksto „jausminei“ analizei ir t. t. Kalbos dalykai privalėjo būti smulkmeniškai formalizuoti ir užrašyti kitokiu, kompiuteriams suvokiamu būdu. Svarbiausias iš šių dalykų – morfologija, nes tai yra bet kokių nuodugnesnių kompiuterinės lingvistikos darbų (sintaksė, semantinė analizė ir pan.) pagrindas.

Pirmą tokį darbą amžių sandūroje atliko eksperimentinės gamyklos „Bitas“ programuotojas Vytautas Zinkevičius (1996, 2000). Beveik du dešimtmečius tai buvo nepamainoma ir vienintelė priemonė šiuolaikiškam lietuvių kalbos tyrimui ir kompiuteriniams taikymams. Ši sistema buvo sėkmingai naudojama rašybos klaidoms aptikti tekstų redagavimo programose, taip pat morfologinei analizei bei sintezei atlikti moksliniuose ir taikomuosiuose darbuose Lietuvių kalbos institute, Vytauto Didžiojo (ten jo darbai gavo „Lemuoklio“ vardą) ir Vilniaus universitetuose. Buvo netgi sukurtas šios sistemos variantas, pritaikytas senajai lietuvių kalbai analizuoti (Gelumbeckaitė ir kt. 2012). Tačiau plečiantis kompiuterinės lingvistikos poreikiams pradėjo ryškėti šios sistemos trūkumai: nestandartinės, neaprašytos duomenų struktūros; uždaras kodas; nevisavertis tikrinių vardų, iliatyvo, dviskaitos, pirminių veiksmažodžių būsimojo laiko, sutrumpėjusių bei retų formų realizavimas. Ypač šios sistemos tolimesniam plėtojimui trukdė uždaras kodas – duomenis ir algoritmą praktiškai galėjo keisti tik pats autorius.

Siekiant išvengti anksčiau aprašytų trūkumų ir sukurti naujos kartos kompiuterinę lietuvių kalbos morfologiją, buvo iškelti šie reikalavimai:

1) naudoti ir kurti tik atvirąjį kodą;
2) tik duomenys, bet ne jų forma ir interpretavimas, gali turėti specifinių lietuvių kalbos savybių;
3) visas programinis kodas turi būti universalus, tikti bet kuriai kitai kalbai;
4) pasinaudoti kitoms pasaulio kalboms sėkmingai pritaikytais sprendimais.

Minėtus reikalavimus gerai atitinka *Hunspell* platforma[3]. Be to, tokiu būdu aprašius kalbą, vėliau nesunkiai galima tikrinti rašybos klaidas daugelyje (per 50) taikomųjų programų (*OpenOffice*, *LibreOffice*, *Firefox*, *Chrome*, *Safari*, *InDesign* ir t. t.), atlikti žodžių morfologinę analizę bei sintezę, vykdyti intelektinę tekstinės informacijos paiešką ir pan.

[3] Prieiga internete: https://en.wikipedia.org/wiki/Hunspell.

Šiame straipsnyje apžvelgiamas naujas lietuvių kalbos morfologijos realizavimas *Hunspell* platformos pagrindu. Aptariamos su tuo susijusios problemos ir jų sprendimo būdai, pateikiami sėkmingo testavimo ir praktinio taikymo pavyzdžiai.

1. LIETUVIŲ KALBOS APRAŠYMO *HUNSPELL* FORMATU YPATYBĖS

1.1. *Hunspell* platformos atsiradimo prielaidos ir bendrieji principai

Atvirojo kodo *Hunspell* platforma išsiplėtojo iš paprasčiausių rašybos tikrinimo priemonių *Ispell*[4], *Aspell*[5], *MySpell*[6] ir pan. Pamažu darant jas sudėtingesnes ir pritaikant vis įvairesnėms kalboms (ypač agliutinacinėms – suomių, vengrų, turkų ir pan.), tapo įmanoma spręsti ne tik rašybos tikrinimo, bet ir morfologinės analizės bei sintezės uždavinius. Šios pasaulyje itin populiarios platformos autoriai – daugiausia vengrų kilmės mokslininkai ir inžinieriai-programuotojai (Trón ir kt. 2005). Nenuostabu, kad jos pavadinimo pirmas dėmuo „*Hun-*“ yra pagal Vengrijos pavadinimą anglų kalba.

Kalbos morfologiją *Hunspell* platformoje aprašo du tekstiniai failai: plėtiniu **.aff** užrašomos kaitybos taisyklės, o plėtiniu **.dic** užrašomas žodžių formų žodynas su pradine morfologine informacija ir nuorodomis į vieną ar kelias kaitybos taisyklių grupes, kurios papildomai gali būti taikomos. Atskirą taisyklę galima įsivaizduoti kaip instrukciją, kurioje nurodoma:

1) Ką į ką galima keisti žodyje iš formų žodyno. Pvz., taisyklėje

```
SFX 85 čias ty [^š]čias  is:Masc_Sg_Voc
```

„`SFX`“ žymi, kad keičiama žodžio pabaiga, o keisti galima „`-čias`“ į „`-ty`“.

„`PFX`“ žymėtų, kad keičiama žodžio pradžia. Žodžio vidurio keitimai nėra numatyti.

2) Kaitybos taisyklių grupės, kuriai priskiriama ši taisyklė, numeris. Grupę gali sudaryti viena arba kelios taisyklės, kurios visada taikomos kartu. Ankstesniame pavyzdyje taisyklė priskiriama grupei, kurios numeris yra „`85`“.

3) Sąlygos, kurioms esant, keitimas yra galimas (taikomas tam tikras reguliariųjų išraiškų formalizmas). Mūsų pavyzdyje keitimas galimas tik su sąlyga, kad žodžio iš formų žodyno pabaiga nėra „`-ščias`“ (pvz., „vaikiščias“).

4) Morfologinė informacija, susijusi su šiuo keitimu. Mūsų pavyzdyje su tokiu žodžio pabaigos keitimu yra siejamas morfologinis pažymėjimas „`Masc_Sg_Voc`“ (vyriškoji giminė, vienaskaita, šauksmininko linksnis, pvz., „svety“).

[4] Prieiga internete: https://www.cs.hmc.edu/~geoff/ispell.html.
[5] Prieiga internete: http://aspell.net.
[6] Prieiga internete: https://code.google.com/archive/a/apache-extras.org/p/ooo-myspell.

Išsamiau su *Hunspell* kaitybos taisyklių sudarymo ir morfologinės informacijos nurodymo principais galima susipažinti šios platformos svetainėje[7], o įvairų kalbų morfologijos aprašymo šioje platformoje galimybės gana gerai išnagrinėtos Tommi A. Pirineno disertacijoje (2014).

1.2. Specifiniai reikalavimai lietuviškajam realizavimui

Kad toks kalbos formalizavimas tiktų morfologinei analizei, sintezei, tekstinės informacijai indeksuoti ir paieškai, reikia tenkinti šiuos du reikalavimus:

1) į žodžių formų žodyną gali būti įrašomos tik lemos (antraštinės formos);
2) priešdėlinė, priesaginė ir sangrąžinė daryba turi atsispindėti ne **.aff**, o **.dic** faile.

Pirmąjį reikalavimą lietuvių kalboje nėra lengva įgyvendinti dėl didelės veiksmažodžių kaitybos įvairovės (apie 170 variantų), bet spręsti šią problemą padėjo *Hunspell* platformos galimybė iš vienos kaitybos taisyklės kviestis kitą. Nors tokių kvietimų „gylis“ gali būti lygus tik vienetui, tai gerokai sumažina kaitybos taisyklių failo dydį – nuo kelių milijonų iki keliolikos tūkstančių eilučių.

Antrojo reikalavimo tenkinimas kiek išplečia **.dic** failo dydį, nes, pvz., tos pačios šaknies „imti“, „paimti“, „suimti“ ir „nesiimti“ priešdėliniai–priesaginiai–sangrąžiniai vediniai bus įrašomi atskiromis eilutėmis. Ir tai yra ne vienintelė šio reikalavimo neigiama pasekmė – galima lengvai apsirikti ir neįtraukti kokio nors rečiau vartojamo veiksmažodžio ar būdvardžio darinio. Tiesa, ši problema yra gana išsamiai sprendžiama lemas atrenkant visų pirma ne iš išleistų žodynų, bet iš sukauptų tekstynų.

1.3. Lietuviškųjų .dic ir .aff failų fragmentai (pavyzdžiai)

Keli lietuvių kalbos **.dic** failo (žodyno) fragmentai su paryškintais Jono Jablonskio pavyzdinių žodžių atitikmenimis parodyti 2 paveiksle.

[7] Prieiga internete: https://github.com/hunspell/hunspell.

```
laukakmenis/118,122,132,137,141,145,157,159,9999 po:noun
laukan po:adverb
laukas/1,3,7,10,12,14,20,21,30,9999 po:noun
Laukavičienė/738,740,745,9999 po:noun_family_name
...
svečias/1467,1767,1775,1781,1784,1841,9999 po:adjective
svečias/71,73,85,93,98,100,106,116,9999 po:noun
svečiavimasis/1308,1310,1314,1315,1317,1319,9999 po:noun_reflexive
Svečiulienė/738,740,745,9999 po:noun_family_name
...
Gaidulis/118,122,132,141,145,157,9999 po:noun_family_name
Gaidulytė/738,740,745,749,754,760,9999 po:noun_family_name
gaidys/264,268,278,280,284,288,300,302,9999 po:noun
Gaidys/264,268,278,284,288,300,9999 po:noun_family_name
...
peilinis/2036,9999 po:adjective
peilis/118,122,132,137,141,145,157,159,9999 po:noun
peiliukas/1,3,8,10,12,14,20,21,30,9999 po:noun
Peipus/444,446,452,9999 po:noun_geographic_name
```

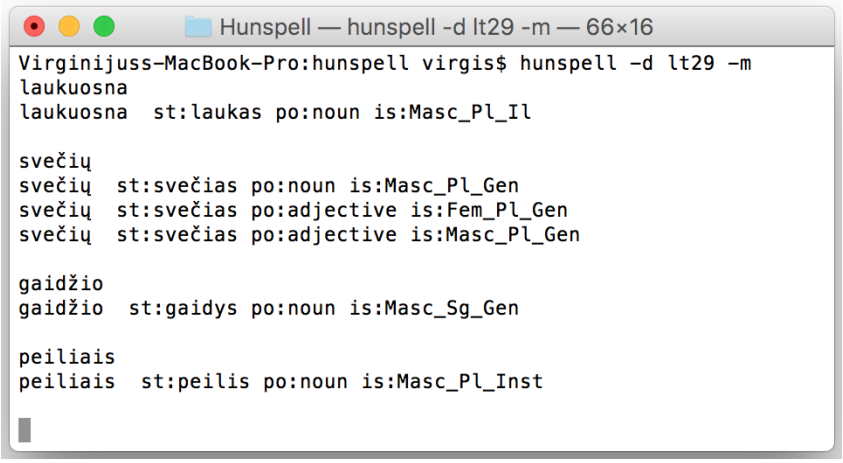
2 PAV. Žodžių formų **.dic** failo fragmentas. Skaičiai po pasvirojo brūkšnio žymi nuorodas į adresuojamas kaitybos taisyklių grupes. Raudonai paryškintos nuorodos į taisyklių grupes, kurios parodytos 3 PAV.

3 pav. pateikiami keli lietuvių kalbos kaitybos taisyklių **.aff** failo fragmentai:

```
SFX 21 Y 1
SFX 21 as uosna .   is:Masc_Pl_Il
...
SFX 98 Y 2
SFX 98 as ai .      is:Masc_Pl_Nom
SFX 98 as ų .       is:Masc_Pl_Gen
...
SFX 264 Y 6
SFX 264 is is .     is:Masc_Sg_Nom
SFX 264 ys ys .     is:Masc_Sg_Nom
SFX 264 is žio .    is:Masc_Sg_Gen
SFX 264 ys žio .    is:Masc_Sg_Gen
SFX 264 ai ai .     is:Masc_Pl_Nom
SFX 264 ai ų .      is:Masc_Pl_Gen
...
SFX 145 Y 12
SFX 145 is iams .   is:Masc_Pl_Dat
SFX 145 ys iams .   is:Masc_Pl_Dat
SFX 145 is iam .    is:Masc_Pl_Dat_short
SFX 145 ys iam .    is:Masc_Pl_Dat_short
SFX 145 is ius .    is:Masc_Pl_Acc
SFX 145 ys ius .    is:Masc_Pl_Acc
SFX 145 is iais .   is:Masc_Pl_Inst
SFX 145 ys iais .   is:Masc_Pl_Inst
SFX 145 is iuose .  is:Masc_Pl_Loc
SFX 145 ys iuose .  is:Masc_Pl_Loc
SFX 145 is iuos .   is:Masc_Pl_Loc_short
SFX 145 ys iuos .   is:Masc_Pl_Loc_short
```

3 PAV. Kaitybos taisyklių **.aff** failo fragmentai. Paryškintos kelios taisyklės, kurios toliau naudojamos morfologinei analizei iliustruoti.

Panaudoję pavyzdžiuose pateiktą kalbos aprašą ir standartines (nuo kalbos nepriklausomas) *Hunspell* analizės priemones, Jono Jablonskio pavyzdiniams žodžiams gautume tokį atsakymą (žr. 4 pav.).



4 PAV. Morfologinės žodžių analizės pavyzdys

2. DARBO EIGA IR REZULTATAI

Kai jau sudaromas visavertis kalbos aprašas *Hunspell* formatu (t. y. sukuriami atitinkami **.dic** ir **.aff** failai), atsiveria praktiškai neribotos galimybės praktiniams taikymams – pradedant rašybos ar gramatikos tikrinimu ir baigiant išmaniosiomis informacijos paieškos sistemomis. Dėl gana sudėtingos lietuvių kalbos morfologijos (daug taisyklių ir jų grupių) ir būtinybės lemų sąrašą derinti su dideliais (daugiau kaip 1 mlrd. žodžių) tekstynais, atspindinčiais realią dabartinės rašytinės lietuvių kalbos būklę, prireikė maždaug 5 metų nuo pirmųjų bandymų iki sklandžiai veikiančio morfologijos varianto sukūrimo. Anksčiau šiame straipsnyje pateikti pavyzdžiai yra labiau iliustratyvūs, skirti tik idėjai suprasti; tikrovėje taisyklės ir jų grupės yra daug sudėtingesnės, ypač tos, kurios susijusios su veiksmažodžiams.

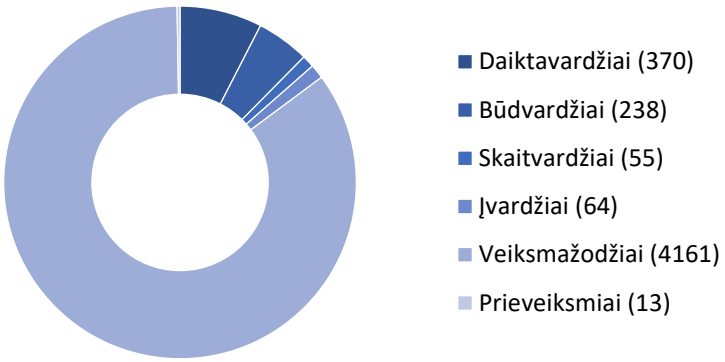
Siekiant pagreitinti morfologijos sukūrimo laiką ir išvengti klaidų surašant tūkstančius taisyklių ir jas priskiriant lemoms, iš pradžių buvo sukurtas pirminis, abstraktesnis nei *Hunspell* specifikacijose reikalaujamas morfologijos aprašas (taisyklės – *MS Excel* lentelėje, lemos ir jų klasifikavimas – *MS Access* duomenų bazėje). Taip pat buvo sukurtos programinės priemonės šioms pirminėms, žmogui lengviau skaitomoms ir paprasčiau suvokiamoms struktūroms automatiškai pertvarkyti į **.aff** ir **.dic** failus arba kitą, taikymo specifikos nusakytą formą. Naujausia lietuviškųjų **.aff** ir **.dic** failų versija pateikiama autoriaus *GitHub* svetainėje[8].

Viso morfologijos formalizavimo proceso metu buvo naudojami šie pagrindiniai šaltiniai:

[8] Prieiga internete: https://github.com/dadurka/hunspell_morphology_lt.

1) *Dabartinės lietuvių kalbos gramatika* (DLKG 2006);
2) Vytauto Didžiojo universiteto (toliau – VDU) Dabartinės lietuvių kalbos tekstynas[9] – ~140 mln. žodžių;
3) Vilniaus universiteto mašininio vertimo tekstynas[10] – ~800 mln. žodžių, ~4 mln. unikalių;
4) Lietuvos Respublikos Seimo (toliau – LRS) dokumentai[11] – ~400 mln. žodžių, ~1 mln. unikalių;
5) *Dabartinės lietuvių kalbos žodynas* (DŽ 2006) – ~60 tūkst. lemų;
6) *Lietuvių kalbos žodynas*[12] – ~500 tūkst. straipsnių;
7) *Tarptautinių žodžių žodynas* (VTŽŽe) – ~22 tūkst. lemų;
8) *Lietuvių pavardžių žodynas* (LPŽ) (~80 tūkst. pavardžių).

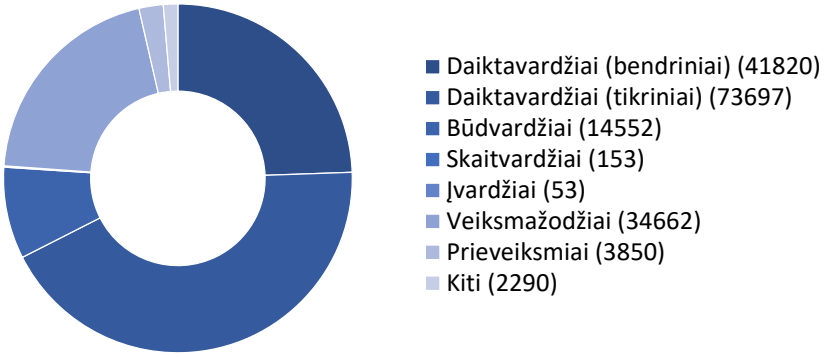
Iš viso buvo sudaryta maždaug 5000 adresuojamų taisyklių grupių (18 000 individualių taisyklių). Kaip taisyklių grupės pasiskirsčiusios pagal kalbos dalis, parodyta 5 paveiksle.



5 PAV. Kaitybos taisyklių grupių pasiskirstymas pagal kalbos dalis

.dic failas (žodynas) buvo sudarytas iš 171 000 lemų. Kaip lemos pasiskirsčiusios pagal kalbos dalis, pavaizduota 6 paveiksle.

[9] Prieiga internete: http://tekstynas.vdu.lt/tekstynas/.
[10] Prieiga internete: https://www.versti.eu/.
[11] Prieiga internete: http://www.lrs.lt/.
[12] Prieiga internete: http://www.lkz.lt/.



6 PAV. Lemų pasiskirstymas pagal kalbos dalis

Vykdant atranką buvo apsiribojama tiktai taisyklinga šiuolaikine lietuvių kalba ir vengiama klaidingų, nelietuviškų, neteiktinų, įžeidžiančių, nebevartojamų žodžių. Tikriniai daiktavardžiai buvo papildomai skirstomi į pavardes, vardus, geografinius pavadinimus ir kitus atvejus. Tokia atranka palengvina sukurtos morfologijos taikymą rašybai tikrinti ir leidžia analizės rezultatus lengviau taikyti vėlesniuose teksto analizės etapuose (sintaksė, įvardintų esybių atpažinimas, indeksavimas ir paieška).

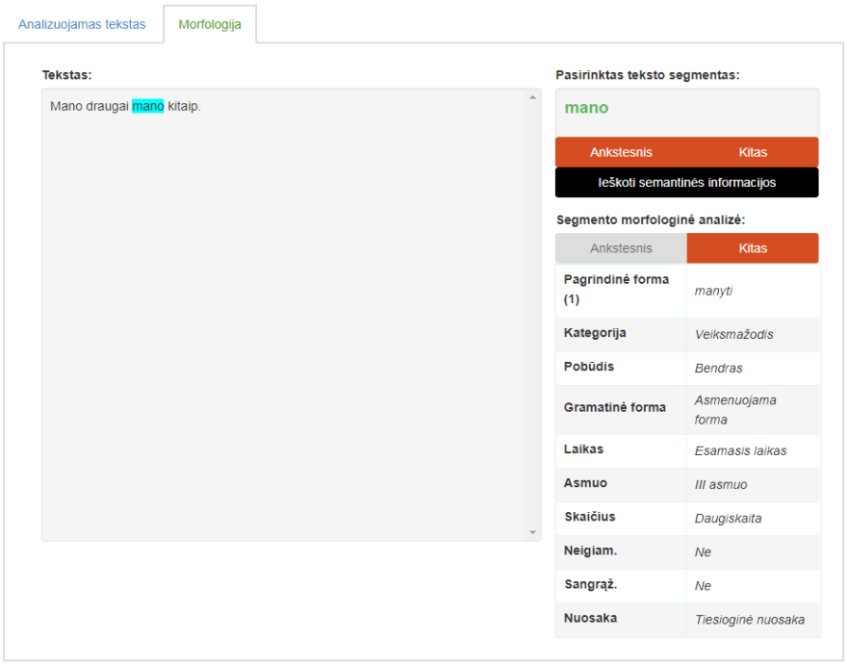
Naujai sukurto morfologinio analizatoriaus tikslumas yra apie 98 proc., t. y. vidutiniškai 98 iš 100 žodžių morfologinis analizatorius interpretuoja teisingai (Kapočiūtė-Dzikienė ir kt. 2017). Likusius 2 proc. neatpažintų žodžių dažniausiai sudaro rašybos klaidos, kitų kalbų žodžiai (vardai, pavardės, frazių citatos ir pan.), neteiktini, nenorminiai žodžiai ir taisyklingi žodžiai, kurie nebuvo įtraukti į lemų sąrašą.

3. PRAKTINIS TAIKYMAS

3.1. VDU Sintaksinės-semantinės analizės sistema

Naujai sukurta lietuvių kalbos morfologija *Hunspell* platformoje 2011–2015 metais buvo pritaikyta VDU Sintaksinės-semantinės analizės sistemoje[13] rašybai ir gramatikai tikrinti, morfologinei ir sintaksinei teksto analizei atlikti (žr. 7 pav.). Morfologinė analizė buvo papildyta mišriu statistiniu-taisykliniu vienareikšminimu (paslėptas Markovo modelis, pagal auksinį standartą aproksimuotos trigramų tikimybės, Viterbio algoritmas, kelios išimtys). Plačiau apie vienareikšminimą morfologinėje analizėje rašoma Jurgitos Kapočiūtės-Dzikienės, Erikos Rimkutės ir Loïc Boizou publikacijoje (2017).

[13] Prieiga internete: http://semantika.lt/SyntaticAndSemanticAnalysis/Analysis.

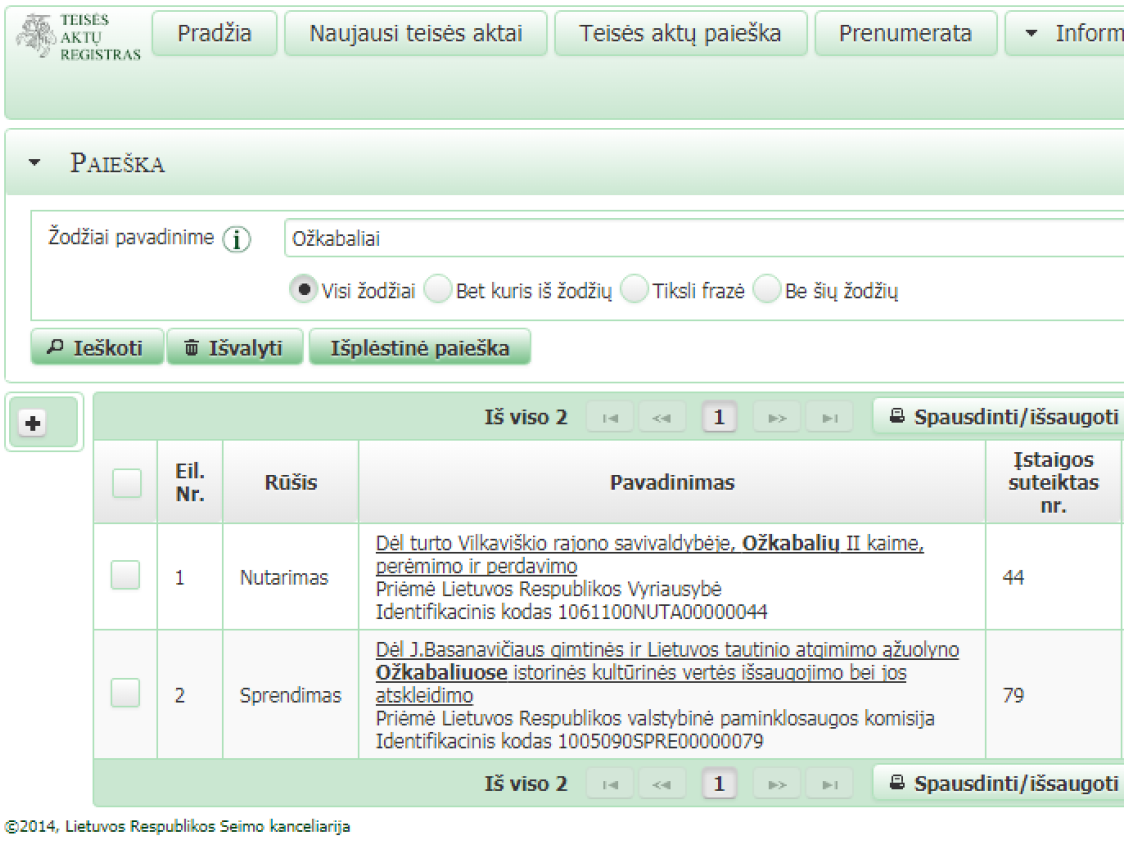


7 PAV. *Hunspell* lietuvių kalbos morfologijos taikymas VDU Sintaksinės-semantinės analizės sistemoje

3.2. LRS teisės aktų registro paieškos sistema

2011–2013 metais lygiagrečiai su naujos lietuvių kalbos morfologijos kūrimu vyko jos pritaikymas Teisės aktų registro paieškos sistemoje LRS kanceliarijoje[14]. Paprastai tokiose sistemose dar prieš vykdant pirmąją paiešką visi dokumentai yra indeksuojami lemuojant, t. y. visi dokumentų žodžiai keičiami į jų lemas ir įsidėmima, kur kokia lema buvo. Prieš vykdant paieškos užklausą jos tekstas irgi lemuojamas tokiu pačiu būdu, todėl vėliau randama ne tik užklausoje užrašyta žodžio forma, bet ir bet kuri kita to žodžio forma, esanti tarp ieškomų dokumentų. Kadangi šioje sistemoje nebuvo galima tiesiogiai panaudoti .**dic**, .**aff** duomenų, sistemai reikalingas originalus formatas (tekstinis failas, kurio kiekviena eilutė yra [*lema*] [*kalbos dalis*] → [*formų sąrašas*]) buvo generuojamas iš anksčiau minėto pirminio morfologijos aprašo. Nors iš šio aprašo galima generuoti apie 15 mln. teoriškai galimų žodžių formų, dėl paieškos sistemos vidinių ribojimų iš jų buvo panaudotos tik apie 1 mln. pačių dažniausiųjų. Toks sprendimas pasiteisino, paieška veikia gana sklandžiai ir naudojama iki šiol (žr. 8 pav.).

[14] Prieiga internete: https://www.e-tar.l.t

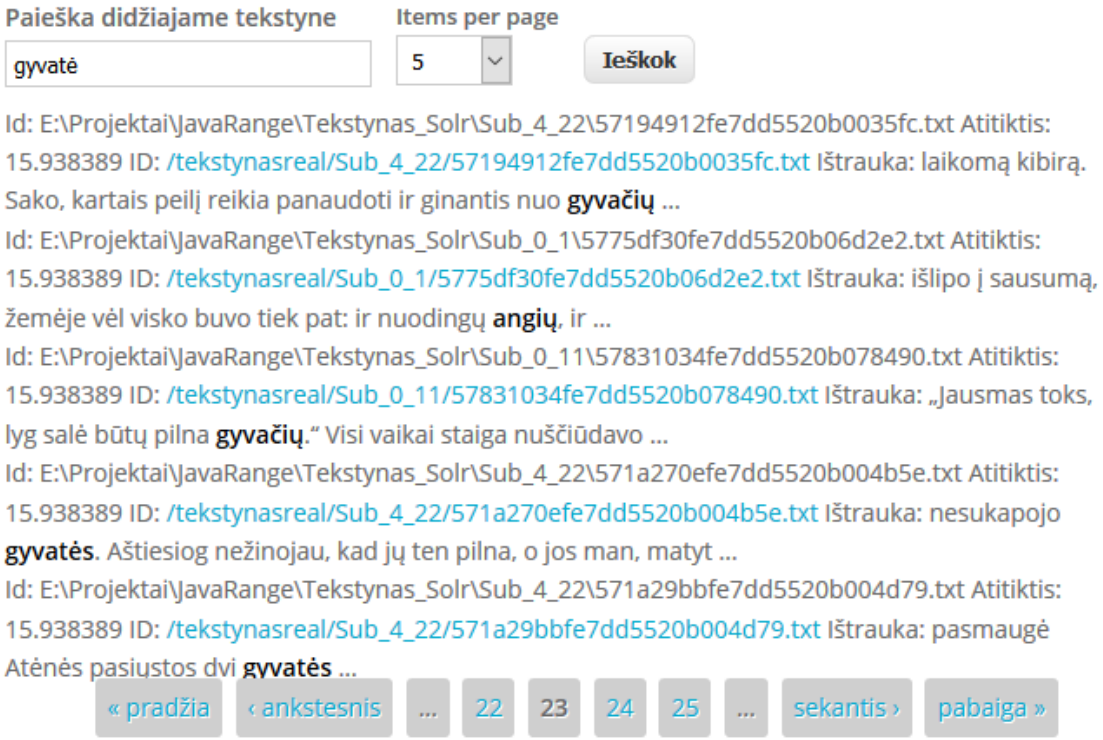


8 PAV. Lietuvių kalbos morfologijos taikymas LRS Teisės aktų registro paieškoje

4.3. Dokumentų indeksavimo ir paieškos atvirojo kodo sistema *Solr/Lucene*[15]

Siekiant plačiau panaudoti atvirojo kodo privalumus ir patirtį, sukauptą realizuojant lietuvių kalbos morfologiją Teisės aktų registre, buvo sukurta lietuviškų dokumentų indeksavimo ir paieškos sistema *Solr/Lucene* pagrindu. Tiesioginis **.dic**, **.aff** duomenų taikymas šioje sistemoje yra galimas, bet labai neefektyvus, todėl buvo pakeistas baigtiniu būsenų keitikliu (FST – *Finite State Transducer*), sudarytu iš anksčiau minėto pirminio morfologijos aprašo ir generuojančiu visas pagal šį aprašą teoriškai galimas žodžių formas. Toks sprendimas leidžia ne tik tiksliai indeksuoti ir surasti visas norimas žodžių formas, bet ir pagreitinti indeksavimą šimtus kartų, o paieškos rezultatus gauti panaudojant įvairius semantinius ryšius, atsižvelgiant į papildomą taksonominę-semantinę informaciją. Pvz., jei užklausoje nurodomas žodis „gyvatė“, tai sistema papildomai ieško „angis“, „barškuolė“ ir pan. (žr. 9 pav.), o užklausoje nurodžius žodį „angis“ – tarp paieškos rezultatų jau nebus nei „gyvačių“, nei „barškuolių“ (Ralys, Dadurkevičius 2017).

[15] Prieiga internete: http://lucene.apache.org/solr/.



9 PAV. Lietuvių kalbos morfologijos taikymo dokumentų indeksavimo ir paieškos atvirojo kodo sistemoje *Solr/Lucene* pavyzdys

4. PERSPEKTYVOS

Norint plačiau taikyti lietuvių kalbos morfologijos *Hunspell* platformoje realizavimą, reikėtų plėtoti dvi jos kryptis: norminės kalbos variantą ir pilnąjį variantą (norminė ir nenorminė kalba). Pirmasis variantas leistų ir toliau atlikti rašybos tikrinimo užduotis ir duoti rekomendacijas dėl kalbos normų pažeidimų taisymo, o antrasis būtų tinkamas taikomiesiems gyvosios kalbos analizės uždaviniams spręsti.

Ne mažiau svarbi ateities užduotis yra nuolat papildyti morfologijos žodyną neįtrauktais ir naujai atsirandančiais žodžiais, taisyti išryškėjusias kaitybos taisyklių bei žodžių klasifikavimo klaidas ir bent kartą per metus publikuoti naujausias **.dic**, **.aff** versijas.

Panaudojus jau sukauptą patirtį nesunku būtų parengti ir formalų lietuviškų tarmių (ypač žemaičių) ar senosios lietuviškų raštų kalbos aprašą.

ŠALTINIAI

DLKG 2006 – *Dabartinės lietuvių kalbos gramatika*. Red. Vytautas Ambrazas. 4-oji pataisyta laida, Vilnius: Mokslo ir enciklopedijų leidybos institutas.

DŽ 2006 – *Dabartinės lietuvių kalbos žodynas* (elektroninis). Vyr. red. Stasys Keinys. 6-as leid. Vilnius: Lietuvių kalbos institutas.

LKŽe – *Lietuvių kalbos žodynas*, elektroninis variantas. Red. kolegija: Gertrūda Naktinienė (vyr. red.), Jonas Paulauskas, Ritutė Petrokienė, Vytautas Vitkauskas, Jolanta Zabarskaitė, Vilnius: Lietuvių kalbos institutas, 2005. Atnaujinta versija, 2008. Prieiga internete: www.lkz.lt.

LPŽ – Aleksandras Vanagas (red.), *Lietuvių pavardžių žodynas* 1–2, Vilnius: Mokslas, 1985–1989.

VTŽŽe 2003 – Vaitkevičiūtė V. *Tarptautinių žodžių žodynas* (elektroninis). Vilnius: Fotonija.

LITERATŪRA

Gelumbeckaitė ir kt. 2012: Gelumbeckaitė J., Šinkūnas M., Zinkevičius V. Old Lithuanian Reference Corpus (SLIEKKAS) and Automated Grammatical Annotation. - *Journal for Language Technology and Computational Linguistics (JLCL), 2012. - 27.2. Altüberlieferte Sprachen als Gegendstand der Texttechnologie. Ancient Languages as the Object of Text Technology*, ed. by Armin Hoenen, Thomas Jügel, Gesellschaft für Sprachtechnologie und Computerlinguistik (GSCL), 83–96 p.

Jablonskis J. 1901: *Lietuviškos kalbos gramatika*. Rašytojams ir skaitytojams vadovėlis. Parašė Petras Kriaušaitis. Tilžeje 1901. Spauzdinta pas Otto v. Mauderode. 88 p.

Jablonskis J. 1918: *Lietuvių kalbos gramatika*: etimologija : pirmosioms mūsų aukštesniosioms mokslo įstaigoms / P. Kriaušaičio ir Rygiškių Jono. Vilnius : Lietuvių mokslo draugija. 237 p.

Jablonskis J. 1922: *Lietuvių kalbos gramatika*: etimologija : vidurinėms mokslo įstaigoms / Rygiškių Jonas. 2-asis leid. Kaunas; Vilnius: „Švyturio“ b-vė, (Tilžė : Otto v. Mauderodės sp.). 280 p.

Kapočiūtė-Dzikienė J. ir kt. 2017: Jurgita Kapočiūtė-Dzikienė, Erika Rimkutė, Loïc Boizou Comparison of Lithuanian Morphological Analyzers. – *Text, Speech, and Dialogue*. 20th International Conference, TSD 2017, Prague, August 27–31, 2017, Proceedings, 47–56.

Ralys D., Dadurkevičius V. 2017: *Informacijos paieškos sistemos ir lietuvių kalba*, [žiūrėta 2017-10-30]. Prieiga internete: https://www.slideshare.net/DanieliusRalys/informacijos-paiekos-sistemos-ir-lietuvi-kalba-81373517.

Tommi A Pirinen 2014: *Weighted Finite-State Methods in Spell-Checking and Correction*, Doctoral dissertation, University of Helsinki.

Trón V. ir kt. 2005: Viktor Trón, András Kornai, György Gyepesi, László Németh, Péter Halácsy, and Dániel Varga. 2005. Hunmorph: open source word analysis. – *Proceedings of the Workshop on Software (Software '05)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 77-85.

Zinkevičius V. 1996: Lietuvių kalbos morfologinių reiškinių kompiuterizavimo lingvistiniai aspektai. – *Lietuvių katalikų mokslų akademijos suvažiavimo darbai*, XVI tomas, 148-154.

Zinkevičius V. 1996: Lietuvių kalbos morfologinių reiškinių kompiuterizavimas. – *Lietuvių katalikų mokslų akademijos suvažiavimo darbai*, XVI tomas, 155-162.

Zinkevičius V. 2000: *Lemuoklis* – morfologinei analizei. – *Darbai ir dienos* 24, 245–273.

Gauta 2017 10 31

Priimta 2017 12 08

VIRGINIJUS DADURKEVIČIUS
Vilniaus universiteto Taikomųjų mokslų institutas
M. K. Čiurlionio g. 29, 03100 Vilnius
virginijus.dadurkevicius@tmi.vu.lt
dadurka@gmail.com

LITHUANIAN MORPHOLOGY IN THE *HUNSPELL* FRAMEWORK

Summary

The paper presents the results of an attempt to build basic Lithuanian language resources using the widespread *Hunspell* platform. The spelling is actually the primary target of this open-source platform but the morphological analysis and synthesis are also possible. Moreover, the ability to efficiently perform lemmatization (stemming) makes this platform the best option for text search engines (e.g. *Solr/Lucene*) and information retrieval. Taggers, grammar checkers and other basic natural language processing tools can also be build using properly built *Hunspell* language resources.

Every *Hunspell* language resource consists of two files: dictionary and affixes (it may be empty). The dictionary contains main forms (lemmas) whereas the affixes contain the morphological rules to generate all possible forms. As a source for the dictionary we have used the Modern Lithuanian Dictionary (6-th edition), Corpus of the Contemporary Lithuanian Language compiled at the Center of Computation Linguistics of Vytautas Magnus University, database of documents of the Lithuanian Parliament, *versti.eu* machine translation corpus of Vilnius University and various public internet sources (totally 1.3 billion tokens). Main criteria for semi-manual compilation of the Lithuanian dictionary of lemmas from these sources was correctness, usability, actuality and approval by language authorities. Deprecated loanwords or extremely rare, exotic, obsolete, jargon, insulting forms were discarded from the list. Resulting dictionary consists of 171 000 lemmas: 42 000 common nouns, 73 000 proper nouns, 15 000 adjectives, 53 pronouns, 153 numerals, 35 000 verbs, 4 000 adverbs and 2 000 others (prepositions, conjunctions, particles, onomatopoeias, interjections, acronyms and abbreviations).

The second component of language resource, the so called "affix file", contains information of various kind: metadata, preferable suggestions for spelling correction, grouping of rules, explicit tags for flexing and non-flexing properties, rules for suffix and affix alteration.

In order to make the *Hunspell* resources suitable for creating basic language tools, e.g. morphological analyzer and synthesizer, some principles should be kept:

1) every flexion paradigm (consisting of one or more rules) should be thoroughly generated from one single lemma in dictionary file (it is not trivial, especially for irregular verbs);

2) every individual alteration case should have its own morphological tag, e.g. '*Masc_Sg_Il*' for masculine + singular + illative;
3) every dictionary item should have references for part of speech and other non-flexing information;
4) avoid prefixation via rules, use dictionary instead – affixed forms may have completely different meanings and using them under single lemma may cause problems for text search engines;
5) do not rely much on calling rules from rules – calling depth can by no more than 1.

The coverage of the contemporary Lithuanian by this implementation of Lithuanian morphology is about 98 percent. The full list of all the theoretically possible forms generated by this resource contains about 17 million entries.

This work clearly shows an efficient way for any language (especially with scarce funding resources) to make basic language tools using a single open source development platform – the *Hunspell*.

KEYWORDS: Lithuanian, grammar, morphology, computational linguistics, Hunspell, speller, tagger, analyser, parsing, disambiguation, indexing, search.

VIRGINIJUS DADURKEVIČIUS

Institute of Applied Research
Vilnius University
M. K. Čiurlionio g. 29, 03100 Vilnius, Lithuania
virginijus.dadurkevicius@tmi.vu.lt
dadurka@gmail.com