Senųjų raštų rašybos keitimas paieškos sistemai

Mindaugas Šinkūnas
Lietuvių kalbos institutas
Raštijos paveldo tyrimų centras
mindaugas.sinkunas@lki.lt

LINGVISTINEI ANALIZEI reikia skaitmeninių tekstų, tinkamų programiniam apdorojimui. Lietuvių kalbos instituto duomenų bazei senieji raštai skaitmeninami laikantis dokumentinio perrašo principų, nekeičiant originalo rašybos. Senoji rašyba dažnai yra variantiška, nenusistovėjusi ir gerokai skiriasi nuo dabartinės, tai trukdo pritaikyti technologijas, kuriamas dabartinei lietuvių kalbai tirti. Straipsnyje aprašomas empirinėmis taisyklėmis paremtas būdas iš žodžių formų senąja rašyba automatiškai sugeneruoti formas dabartine rašyba perraše išlaikant originalios rašybos ypatybes. Sugeneruoti atitikmenys naudojami paieškos sistemoje.

REIKŠMINIAI ŽODŽIAI: tekstynų lingvistika, rašybos, ortografijos normalizavimas, normalizacija, reguliacija, transliteracija.

Lingvistinei analizei reikia skaitmeninių tekstų, tinkamų programiniam apdorojimui. Juos rengiant remiamasi dokumentinio leidimo principais – išlaikoma originalo makrostruktūra, didžiųjų ir mažųjų raidžių rašyba, skyryba, specifinės jau nebevartojamos raidės, ligatūros, diakritiniai ženklai ir pan. (Aleknavičienė 2008, 321). Perrašas papildomas įvairiomis metainformacinėmis žymomis apie autorystę, geografiją, kalbą ir pan. Tokiais principais parengti skaitmeniniai perrašai patikimai perteikia originalą, kartu atspindi kalbos ir rašybos įvairovę. Vie-

ni autoriai rašė vartodami ryškias savo gimtosios ar pasirinktos tarmės ypatybes, kiti laikėsi daugiau ar mažiau nusistovėjusių kalbos ar rašybos standartų, kai kurie tobulino rašybą, susikūrė specifinių grafemų (raidžių ar jų junginių). Senoji rašyba ne tik gerokai įvairuoja, bet ir skiriasi nuo dabartinės, todėl kalbos tyrimams sunku pritaikyti technologijas, kurtas dabartinei lietuvių kalbai. Rašyba dažnai trukdo naudotis programiniais kalbos analizės įrankiais, automatizuoti morfologinę analizę, sieti žodžius su kitais tekstynais ar kalbos duomenų rinkiniais.

Minėta problema iškyla ir organizuojant paieškos sistemą. Pavyzdžiui, užklausos langelyje įvedę formą *žodis*, turėtume rasti ir kirčiuotas formas, ir dialektines formas su <a>, ir formas su įvairiai užrašyta [ž], tokias kaip: Martyno Mažvydo *ʒadis*, *S3adis*, *βadis*, *βodis*; Baltramiejaus Vilento *βodis*, *Sʒodis*; Jono Jaknavičiaus *ʒodis*, *3odis*; Danieliaus Kleino *ʒodis* ar Jokūbo Brodovskio *ʒódis* ir t. t. Kad visa tai veiktų, reikia originalią formą susieti su paieškoje įvesta forma *žodis*. Reikalinga formos perraša dabartine rašyba vakarų aukštaičių tarmės pagrindu, taip pat turi būti išlaikyta informacija, kurią teikia senoji rašyba ir užfiksuota kalba. Vadinasi, turi būti sukurtas ne vienas, bet keli atitikmenys.

Žodžių ar jų fragmentų paieška turi būti kuo įvairiapusiškesnė, paieška turi veikti pagal originalią kalbos formą senąja rašyba (*szadis*), pagal senąją formą sumoderninta rašyba (*žadis*), pagal fonetiškai sumodernintą formą senąja rašyba (*szodis*), pagal sumodernintą formą ir rašybą (*žodis*). Reikia, kad Mažvydo forma *βadis* būtų susieta su jos atitikmenimis *szadis*, *žadis*, *szodis* ir *žodis*.

Tokie rašybos keitimai įprasti istoriniams tekstynams ir priklausomai nuo siekiamų tikslų vadinami normalizacija, standartizacija, modernizacija, reguliacija, kanonizacija ar transliteracija.

Normalizacijos metodų yra įvairių, skiriasi jų efektyvumas ir paruošiamųjų darbų apimtys. Gali būti remiamasi dabartinės kalbos žodynais ar rašybos kontrolės, tikrinimo programomis

(Baron, Rayson 2008), pritaikomi statistiniai mašininio vertimo metodai (Ljubešić, Zupan, Fišer, Erjavec 2016), geriausių rezultatų pasiekiama hibridiniais metodais: naudojant žodynus ir taikant neuroninio mašininio vertimo įrankius (Tang, Cap, Pettersson, Nivre 2018). Transliteracijos taisykles įmanoma sugeneruoti ir automatiškai iš tekstų, turinčių perrašus normalizuota rašyba (Bollmann, Petran, Dipper 2011). Šiems metodams reikalingi jau paruošti pavyzdiniai duomenų rinkiniai, analogai. Lietuvių kalbos anotuotų senųjų šaltinių parengta dar labai nedaug (6, žr. <http://titus.uni-frankfurt.de/sliekkas>), per mažai, kad juos galima būtų panaudoti taisyklėms generuoti. Dabartinės kalbos tekstynus naudoti normalizacijai sudėtinga, nes istorinė rašyba gerokai skiriasi nuo dabartinės.

Apie galimybę pagreitinti senųjų raštų anotavimą pradėta mąstyti rengiant senųjų raštų indeksus (Ambrazas, Aleknavičienė, Zinkevičius 1998) ir morfologiškai anotuojant šaltinius senosios lietuvių kalbos korpuse „SLieKKas“ (Gelumbeckaitė, Šinkūnas, Zinkevičius 2012a; 2012b; Mittmann 2013). Morfologijos analizei buvo iš dalies pritaikytas dabartinės bendrinės lietuvių kalbos anotatorius ir sintezatorius „Lemuoklis“ (Zinkevičius 2000), reikėjo sukurti kelią, kuriuo iki jo atkeliautų žodis senąja rašyba. Buvo atliekama pirminė normalizacija ir sukuriamas vienas formos atitikmuo, vėliau taikytos raidžių keitimo, įterpimo ar šalinimo taisyklės ir tikėtasi, kad viena ar kelios pagal jas modifikuotos formos atitiks formą dabartine rašyba, kurią jau galės perskaityti ir morfologiškai interpretuoti „Lemuoklis“. Toks metodas nereikalavo didesnių parengiamųjų darbų ir veikė perkeliant duomenis į anotavimo programą „Toolbox“. Anotatorius dirbdamas turėjo parinkti tinkamą formos atitikmenį ir jį lydinčią morfologinę informaciją. Žinoma, transliteracija tinkamai veikdavo ne visada, pavyzdžiui, formai *apſifiunkinaiuſi* pagaminti 14 335 variantai, bet automatinė anotacija neįvyko. Sėkmingais anotavimo atvejais žodžio forma buvo susiejama su vienu transliteruotu jos atitikmeniu.

Lietuvių kalbos instituto senųjų raštų tekstyną sudaro maždaug 6 milijonai žodžių formų (iš jų apie milijonas – ne lietuvių kalba), tekstynas vis papildomas naujais šaltiniais. Nesant žodžių sąsajų su dabartine rašyba, naudotis įvairių autorių ir skirtingų laikotarpių kalbos duomenimis tampa sudėtinga. Paprasčiausia rašybą suniveliuoti būtų žodžius perrašinėjant rankomis, tačiau dėl didelės tekstyno apimties toks būdas reikalautų ilgo ir atidaus darbo. Automatizuotas mechanizmas, savarankiškai atliekantis perrašinėjimą, galėtų visa tai gerokai pagreitinti. Tam reikia sukurti normalizacijos taisykles, pritaikytas skirtingoms rašybos sistemoms.

Kuriant taisykles svarbu ištirti lietuvių rašybos istoriją ir atsižvelgti į jos raidą. Mažojoje Lietuvoje skirtini bent du ortografijos raidos etapai: pirmajai grupei priklauso raštai, pasirodę nuo 1547 iki 1640 metų, antrajai – nuo 1640 metų. Beveik kiekvieno pirmosios grupės autoriaus vartojamoje rašyboje ir kalboje galima rasti individualių bruožų (pavyzdžiui, reikalingos specialios taisyklės Mažvydo *Katekizmui*, bet jos netaikytinos kitiems jo kūriniams). Didžiosios Lietuvos XVI–XVIII a. raštus pirmiausia reikia skirstyti pagal tarminį jų pagrindą ir rašomosios kalbos variantą, vėliau ieškoti smulkesnių grupių, kurioms būdingos tos pačios rašybos ypatybės.

Transliteracija vyksta keliais etapais: raidžių seka skaitoma iš kairės į dešinę, pagal pritaikytą taisyklę grafema ar jų grupė modifikuojama, variantai kaupiami, priėjus žodžio pabaigą sugeneruojami formos atitikmenys. Pavyzdžiui, Mažvydo forma *ßadis* transliteruoti pradedama nuo ligatūros <ß>, kuri yra grafemos <sz> realizacija. Ji gali žymėti [š] arba [ž], o konkreti reikšmė priklauso nuo šaltinio geografijos ir parašymo laiko: <sz> reikšme [ž] vartojama Mažosios Lietuvos raštuose, sukurtuose iki 1640 metų, kituose raštuose <sz> žymi [š]. Kadangi Mažvydo raštai patenka į pirmąją raštų grupę, <ß> formoje ***ß****adis* keičiama variantu <ž>. Mažvydas priklauso grupei autorių, kurie netiksliai žymi priebalsių minkštumą, todėl įjungiama taisyklė,

numatanti, jog priebalsė prieš užpakalinės eilės balsę turi būti papildyta minkštumo ženklu <i>, sugeneruojami dar du variantai <szi> ir <ži>. Toliau apdorojama antroji raidė <a>. Mažvydo raštuose, ypač *Katekizme*, <a> gali žymėti [ā], kuri atitinka dabartinės kalbos [o], todėl formos *ßadis* <a> papildoma variantu <o>. Raidė <d> nekeičiama. Galūnė <is> gauna variantus <is>, <ys> ir <įs> (nosinė raidė atsiranda dėl taisyklės, skirtos esamojo laiko veikiamojo dalyvio galūnei, pvz., *žydįs*). Po šių operacijų atmintyje sukaupti variantai parodyti 1 lentelėje. Sudėliojus visas galimas jų kombinacijas sugeneruojami 24 Mažvydo formos atitikmenys ir pridedama originali forma: *ßadis*; *sziodįs*; *sziodis*; *sziodys*; *sziadįs*; *sziadis*; *sziadys*; *žiodįs*; *žiodis*; *žiodys*; *žiadįs*; *žiadis*; *žiadys*; *szodįs*; *szodis*; *szodys*; *szadįs*; *szadis*; *szadys*; *žodįs*; *žodis*; *žodys*; *žadįs*; *žadis*; *žadys*.

1 LENTELĖ. Mažvydo formos *ßadis* variantų generavimas

ß	a	d	i	s
sz	a	d	i	s
ž	o		y	
szi			į	
ži				

Automatinė taisyklė, keičianti raidę <a> formoje *ßadis*, veiks ir visose kitose formose, kurios turi <a>, todėl bus prikurta ir daug neegzistuojančių ir nepanaudojamų formų. Sistema turėtų veikti kuo tiksliau ir taupiau, generuoti kuo mažiau variantų, todėl tikslinga riboti taisyklių taikymą atsižvelgiant į šaltinio specifines kalbos ir rašybos ypatybes. Pavyzdžiui, raidę <a> į <o> verta keisti Mažvydo *Katekizme* visose pozicijose, bet *Formoje krikštymo* ją tikslinga apdoroti tik žodžio galūnėje; Jono Bretkūno Evangelijos pagal Luką vertime <a> modifikuotina į <o> ir šaknyje, bet kitose Biblijos dalyse – tik galūnėje. Tokiu

būdu išlaikoma originali forma, sukuriama fonetiškai modernizuota forma ir sąlyginai veikia ekonomijos principas.

Siekiant sumažinti atitikmenų skaičių, dalį rašybos ypatybių galima ignoruoti performuojant paieškos užklausą. Pavyzdžiui, ieškomos raidės <y> ir <į> traktuojamos kaip <i>, tada paieškai reikalingų Mažvydo formos *ßadis* atitikmenų sumažėja iki 9: *ßadis*; *sziodis*; *sziadis*; *žiodis*; *žiadis*; *szodis*; *szadis*; *žodis*; *žadis*. Atitikmenų skaičius šaltiniuose po balsių sutapatinimo ir panašių optimizacijų nurodytas 2 lentelės III skiltyje.

Variantų generavimo taisyklės automatiškai išrikiuojamos pagal modifikuojamų raidžių kiekį, turi suveikti ilgiausią junginį apimančioji taisyklė. Jei taisyklės dubliuojasi, veikia ta, kuri yra detalesnė, pavyzdžiui, priebalsio prieš balsį taisyklė nesuveikia, bet užleidžia vietą priebalsio prieš priešakinį balsį taisyklei, o ši – nusileidžia priebalsio prieš *e*-tipo balsius taisyklei. Iš esmės, taisyklės formuojamos panašiai kaip istorinės fonetikos dėsniai, tik operuojama ne garsais, o rašmenimis.

Formuluojant taisykles patogu operuoti grafemų grupėmis, t. y. kintamaisiais, kurie leidžia sumažinti analogiškų taisyklių skaičių. Pavyzdžiui, apibrėžiama, kad klasė „Ž“ apima bet kurią iš raidžių <ʒ́, ʒ̀, ź, Ź, ʒ̀, ʒ̀, ż, Ż, ʒ̇, ʒ̇, ʒ̀, ż, Ż, ʒ̃, ʒ̃, ʒ̄, ʒ̄, z̄, Z̄, ž̌, ž̌, ž, Ž, ẑ, Ẑ, ᴢ, Z>. Iš kelių poklasių gali būti sudaromos platesnės klasės, pavyzdžiui, pučiamieji priebalsiai „ConsS“ apibrėžiami kaip „S, Š, Z, Ž“.

Viena iš dažniausiai pritaikomų taisyklių yra normalizuojanti priebalsių palatalizacijos žymėjimą. Daugumoje Mažosios Lietuvos raštų priebalsių minkštumas žymimas nenuosekliai. Tyrėjai nesutaria, ar tai yra vien rašybos ypatybė, ar ji atspindi ir fonetinius pokyčius. Suformuluota taisyklė veikia priebalses prieš užpakalinės eilės balsius (pvz., ***K*a*r*a*l*au*cʒ*ui*), šioje pozicijoje <i> pridedama Prūsijos raštuose. Didžiosios Lietuvos autoriams ši taisyklė nereikalinga, nes jie priebalsių minkštumą žymėjo ganėtinai nuosekliai (pvz., *Wil**ni**uie*), išskyrus priebalsį [l]. LDK raštuose <i> turi būti pridedama po <l> (pvz., *Kara-*

***l**aus*) ir pakeičiamas veliarinio priebalsio žymėjimas perbraukta <ł> (pvz., *pas**ł**apcziomis*). Kai kuriuose raštuose yra pozicijų su pertekliniu palatalizacijos žymėjimu (pvz.: ***gi**erti* ‘gerti’, ***ki**elefi* ‘keliasi’), jose <i> naikinama po gomurinių ir lūpinių priebalsių prieš priešakinės eilės [e] tipo balsius.

Priebalsių asimiliacijos pozicijose kai kurie autoriai renkasi fonetinę rašybą (pvz.: *el**k**tifi* ‘elgtis’, *vʒdin**k**s* ‘uždengs’, *ab**d**ingie* ‘apdengė’, *wei**ſ**deck* ‘žiūrėk’), ją įmanoma normalizuoti taisyklėmis, kuriose skardieji ir duslieji priebalsiai aprašomi klasėmis, arba turi būti sukurtos atskiros taisyklės tokiems atvejams kaip, pavyzdžiui, *wen**gk***.

Labai nevienodai senuosiuose raštuose žymimas priebalsis [j] (pvz.: *y**i**o**i**ima*, *b**i**j**o**i**o**fi*, *klo**y**a*, *giwato**i**h**e***). Mažosios Lietuvos raštuose raidė <j> įsitvirtino nuo Kleino laikų, LDK katalikų raštuose vietoje jos rašyta <i>, reformatai vartojo ir <i>, ir <j> (plačiau žr. Šinkūnas 2015, 228). Kitais atvejais (pvz.: ***gh**is*, ***g**is* ‘jis’) taisyklių tiksliai poziciškai apibrėžti beveik neįmanoma arba jos generuoja daugybę nereikalingų atitikmenų. Formuluojant šias taisykles reikia numatyti, ar jos aktualios vienam šaltiniui, ar jų grupei, ar visiems šaltiniams be išimties. Pavyzdžiui, nuo Mažvydo iki Simono Vaišnoro (gal dar Jono Rėzos) verta įterpti <j> tarp balsių tokiose formose kaip *Mar**i**o**s*** ‘Marijos’. Tokių formų pasitaiko ir vėlesniuose raštuose, bet jos gerokai retesnės, todėl taisyklė juose neįjungiama ir išvengiama daugybės variantų ten, kur jie nepageidaujami (pvz., *mar**i**o**s***). Kas kita su tokiomis formomis kaip *prar**ya*** – taisyklės riboti nėra prasmės, nes raštuose, kuriuose taip nerašoma, ji tiesiog nesuveiks.

Paprastesnės taisyklės reguliuoja priebalsių ar balsių sudvejinimą (geminaciją, pvz.: ***app**iauti* ‘apipjauti’, ***Suu**naus*), grafemas <h> ir <ch, gh, ph, th> (pvz.: ***H**VKINIKV* ‘ūkininkų’, ***Ch**rikſtito**gh**is* ‘krikštytojis’, ***Ph**araona* ‘faraoną’, ***th**iewe* ‘tėve’), grafemų <x, ck> vartojimą (pvz.: *an**x**ti*, *Dauk**x**fink* ‘daugsink’; *aplin**ck**ui***). Kai kuriais atvejais reikia numatyti taisyklių išimtis, pavyzdžiui, degeminacijos taisyklė turi veikti kitaip priešdėliuo-

se *ap-*, *at-*, jei šaknis prasideda tuo pačiu priebalsiu (***app****iauti* ‘apipjauti’, ***att****olink* ‘atitolink’).

Pučiamųjų priebalsių [š], [ž] žymėjimas skirtingo laikotarpio raštuose keičiamas atsižvelgiant į šaltinio geografiją ir chronologiją (pvz.: Mažvydo ***ſch****iſcha* ‘šičia’, *neu****sß****mirſti*, ***ß****mones*, *krikſtidame*, ***Sʒ****idump*; Kleino, Daukšos ir kitų autorių *du****ß****es*; Kleino *iſʒwydus*’, ***Sʒ****irdis*; Mažvydo ir kt. ***ʒ****mones*, Daukšos ir kt. ***ʒ****ódʒus*, ***Ʒ****ódʒius*). <sch> keičiama į <š> visuose raštuose, <sz> LDK raštuose keičiama į <š>, Mažosios Lietuvos raštuose iki 1640 m. keičiama į <ž>, vėlesniuose – į <š>, šalia <z> sukuriamas variantas su <ž> (plačiau apie šių priebalsių rašybos istoriją žr. Šinkūnas 2014). Mažosios Lietuvos raštų <ś> keičiama į <š>, kitų regionų autoriams taisykles reikia reguliuoti kiekvienam atskirai (plg. Sirvydo ***ś****irdi* ‘širdį’ ir vad. Morkūno postilės *pa****ś****iutinimu* ‘pasiutinimu’).

Afrikatų [č], [dž] rašybos įvairovė taip pat reikalauja gausoko taisyklių pluošto (pvz.: *tre****tc ʒ****e*, *atpen****t ʒ***, *atpen****c ʒ***, *Buc****ʒ****ůju*, *kenćiu*, *iau****ci****as* ‘jautis’; *ßo****dc ʒ****iuu*, *ßa****dß****ius*, *Ʒo****dʒ****us*, *ʒo****dʒ****iû*).

Balsių normalizacijos taisyklės apima nenuoseklų ilgumo žymėjimą (pvz.: *baʒ****ni****cʒe*, *d****u****ßes*), grafemų <æ>, <ea>, <e̊> vartojimą (pvz.: *fædetumbit*, *læpus*, *up****ea****lis*, *ge̊rai*; plg. *neabejoiánti*), specifinę Daukšos <ę> (pvz., *gę****r****as*), kirčiuotos balsės verčiamos nekirčiuotomis. Taip pat reikia išspręsti pasitaikantį [a] priešakėjimą po minkštojo priebalsio (pvz.: *patogi****e****i*, *patog****e****y*, *baʒnicʒ****e***, *prießtarauj****e****ncʒus*; *baʒnicʒ****e***), stengiamasi teisingai išskleisti abreviatūras (pvz.: *id****ā****t* ‘idant’, *i****ā*** ‘jam’, *Sʒwēcʒáuſojo, ſkundʒ****ē*** ‘skundžiam’, *danga****9*** ‘dangaus’, *Na****m̃****ůſe*, *moki****ñ****a*, *priwerſt****ū****bei*, *brang****ū****mu*, *ad****ū****t* ‘adunt’, *d****ū****gui* ‘danguj’, lot. *n****ō*** ‘non’, *c****õ****munibus*), sutvarkyti XVI amžiaus raštuose dažną [u] žymėjimą raide <v> ir [v] žymėjimą raide [u] (***v****ßgimima*, *gi****u****enima*).

Daugumoje raštų labai nenuosekliai žymimi nosiniai balsiai, norint juos sužymėti tenka generuoti labai daug atitikmenų. Konteksto, morfologijos ar sintaksės analizė nėra atliekama, tad neįmanoma nustatyti, ar nosinė raidė yra reikalinga tokiose for-

mose kaip *důn**a***, *linxſmib**e***, *i**e***, *kur**ij*** ir tik retais atvejais padeda akcentografijos taisyklės (pavyzdžiui, Mažojoje Lietuvoje nuo XVII a. pirmojo ketvirčio *-û* žymi daugiskaitos kilmininko formas, pvz., *j**û*** ‘jų’). Mažiau problemų kelia nosinių balsių rašyba priešdėliuose ar šaknyse, jose nosinės raidės dėliojamos pagal sudarytą šaknų sąrašą. Į jį nepatekusios nosinės raidės laikomos spaudos klaidomis (pvz.: *ſatiew**a**nems*, *Linxſmik**ę**tefe*).

Atskira taisyklių grupė reguliuoja dvibalsių žymėjimą prieš sprogstamuosius (išskyrus lūpinius) priebalsius (pvz.: *βab**a**ga*, *atl**a**kijk*, *ſa**a**draugiste*, *ſa**a**tiewanems*; *Vſchβ**ę**gie*, *p**ę**kis*, *świ**ę**tes*, *β**ę**dien*; *Pabudd**i**k*, *uʒrak**i**dino*, *ſudégi**s***). Kartais šios taisyklės aktualios tik šaltinio fragmentui, pavyzdžiui, tokios formos kaip *g**i**ti* vietoje *ginti* vartojamos tik vienoje 1728 metų psalmyno dalyje. Dvibalsių rašyba tam tikrais atvejais taip pat reikalauja specialaus dėmesio (pvz.: *miel**aw**ſis*, *d**ů**na*, *d**ů**na*).

Dialektines fonetines ypatybes reguliuoja taisyklės, apimančios šių balsių atitikmenis: [o] (pvz.: *βadʒ**ia*** ‘žodžio’, *raſcht**a*** ‘rašto’), [ė] (pvz., *th**ie**we* ‘tėve’), [e] (pvz.: *ʒ**ie**mes*, *ʒ**iá**mes* ‘žemės’ ar ‘žemes’), [ę] (pvz., *g**ią**sta* ‘gęsta’), [ie] (pvz.: *w**e**nas* ‘vienas’, *Gėdó-kit* ‘giedokite’), [uo] (pvz.: *d**o**na* ‘duona’, *ap**ó**kai* ‘apuokai’). Didžiosios Lietuvos tekstuose [l] kietinamas prieš *e*-tipo vokalizmą (pvz., *ap**ł**ay**s**ti* ‘apleisti’), [an, am, en, em, ą, ę] atitikmenys rytų aukštaičių tarmėje yra <un, um, in, im, u, i> (pvz.: *d**un**gun* ‘dangun’, *k**in**ćia* ‘kenčia’, *r**un**k**u*** ‘ranką’, *iβei**i*** ‘išėję’, *dar**i**s* ‘daręs’ ir pan.). Šios taisyklės taikomos tik ten, kur jos būtinos, pavyzdžiui, formoms *dar**is***, *gal**is*** sukuriami variantai *-ys* ir *-įs*, o rytų aukštaičių tarme rašytuose tekstuose – dar papildomas variantas *-ęs*.

Morfologinių elementų keisti nesiekiama, bet dalis taisyklių formuluojamos morfemų lygmenyje, jų prireikia siekiant sutvarkyti priešdėlių, dalies priesagų (pvz., suvienodinamos *-ininkas* ir *-inykas*, pvz., *muyt**inik**as*) ir galūnių rašybą (pvz., įvardžiuotinių *ſufiej**ę**je*** ‘susiėjęjie’, *pirm**amjam***, *teiſ**ámujam***, *ſchwent**aſes*** ‘šventosios’ ar ‘šventąsias’) bei atskirti dalelytę *gi* žodžio pabaigoje (*y**gi***, *kaſ**gi***).

Žodžiai ne lietuvių kalba transliteruojami pagal atskiras taisykles.

Taisyklės rašomos tik dėsningoms rašybos ypatybėms. Dalį nedėsningų formų mėginta taisyti neautomatizuotai. Atsitiktinė rašyba, spaudos klaidos arba anachronistinė rašyba normalizuojama rengiant ar skaitant tekstus, šalia originalios formos specialiose žymose įrašant pataisytąją. Pagal šias žymas automatiškai sudaromi atitikmenų sąrašai, jie pridedami prie kitų taisyklių ir įgyja prioritetą taisyklių eilėje. Tada šios formos transliteruojamos kartu su „ištaisytais" jų atitikmenimis pagal bendrąsias taisykles. Taisymų apimtis nurodo pasirinktos žymos tipas: klaidos gali būti taisomos konkrečioje šaltinio vietoje, vieno šaltinio ribose arba visiems apdorojamiems tekstams (jei klaida aprašyta viename šaltinyje, taisymas suveiks ir tame, kuriame ji nebuvo aprašyta).

Taisyklės senųjų raštų rašybai keisti buvo suformuluotos empiriškai. Šiuo metu jos parašytos ir pritaikytos Mažvydo (*Katekizmas*, toliau – MžK; *Giesmė Ambraziejaus* – MžGA, *Forma krikštymo* – MžF), Vilento (*Evangelijos bei Epistolos* – VEE), Jaknavičiaus (*Evangelijos* – JE) ir Kleino raštų (*Grammatica* – KlGr, *Naujos giesmių knygos* – KlNG, *Maldų knygelės* – KlM), 1727 m. Naujojo Testamento (NT), 1728 m. psalmyno (Ps), 1735 m. Biblijos (B) ir 1816 m. Juozapo Arnulfo Giedraičio Naujojo Testamento (GNI) rašybai.

Šaltinių apimtys parodytos 2 lentelėje. Nors *Giesmė šv. Ambraziejaus* yra trumpiausias šaltinis (I skiltyje nurodyta, jog jį sudaro 517 žodžių), jam sugeneruota net 15 290 atitikmenų (II skiltis), iš jų paieškai naudojami 3 257 atitikmenys (III), vidutiniškai tai sudaro 6,3 atitikmens vienam žodžiui (IV). Mažiausiai atitikmenų sugeneruota Giedraičio Naujajam Testamentui bei nelietuviškiems žodžiams Kleino gramatikoje. XVI amžiaus raštams atitikmenų reikia gerokai daugiau dėl nenusistovėjusios rašybos, o Jaknavičiaus perikopėms ir ypač Mažvydo raštams – dėl dialektinių ypatybių.

2 LENTELĖ. Normalizacijos mastai ir efektyvumas: I – žodžių formų skaičius šaltinyje, II – sugeneruota jų atitikmenų, III – paieškai naudojami atitikmenys, IV – vidutinis atitikmenų skaičius vienai formai (III/I), V – tinkamai sugeneruotų atitikmenų dalis po pirminės normalizacijos (%), VI – tinkamai taisyklėmis sugeneruota atitikmenų (%)

Metai	Autorius	Šaltinis	I	II	III	IV	V	VI
1547	Mažvydas	MžK	6 873	120 379	40 258	5,86	29,0	82,5
1549	Mažvydas	MžGA	517	15 290	3 257	6,30	33,0	83,0
1559	Mažvydas	MžF	2 907	51 376	14 297	4,92	37,5	87,0
1579	Vilentas	VEE	32 504	458 338	130 312	4,01	37,0	88,5
1647	Jaknavičius	JE	15 570	163 084	68 292	4,39	38,5	89,5
1653	Kleinas	KlGr	31 869	128 096	61 950	1,94	80,0	96,0
1666	Kleinas	KlNG, KlM	81 302	672 896	236 166	2,90	51,0	91,5
1727	–	NT	144 966	1 283 533	430 067	2,97	53,5	93,0
1728	–	Ps	34 695	347 589	108 649	3,13	58,5	94,5
1735	–	B	715 050	6 724 484	2 246 573	3,14	56,0	95,5
1816	Giedraitis	GNI	135 831	819 264	232 884	1,71	54,0	97,0
Iš viso:			1 202 084	12 841 128	3 572 746	2,97	47,9	90,6

Panašus darbas buvo atliktas su švedų istoriniu tekstynu, eksperimentuojant, kaip rašybos normalizacija veikia su 29 taisyklėmis (Pettersson, Megyesi, Nivre 2012, 335). Lietuvių seniesiems raštams taisyklių suformuluota daugiau, jų prireikė nuo 74 iki 495: MžGA – 74 (206), MžF – 116 (251), MžK – 126 (383), KlGr – 194 (403), JE – 237 (404), Ps – 251 (444), VEE – 288 (463), KlNG-KlM – 367 (537), GNI – 359 (502), NT – 377 (550), B – 495 (617), čia suskaičiuotos taisyklės ar jų dalys, panaudotos daugiau negu 9 kartus, skliaustuose nurodytos visos suveikusios taisyklės. Panaudotų taisyklių kiekis priklauso nuo šaltinio apimties ir jo tarminio pagrindo.

Atliktą rašybos normalizaciją mėginta įvertinti pagal efektyvumą. Sėkmingu laikomas atvejis, jei taisyklėmis modifikuota forma įgavo reikiamą dabartinę rašybą bendrinės kalbos fonetiniu pagrindu. Buvo perrašyta 200 žodžių formų iš kiekvieno šaltinio (imant po 100 nuo pirmojo ketvirčio ir šaltinio vidurio), tada tikrinta, ar žmogaus perrašytoji forma yra tarp mašinos sugeneruotų atitikmenų. Nors tikrinamų duomenų imtis yra ganėtinai maža, manytina, jog ji parodo pasiektas transliteravimo galimybes (žr. 2 lentelę).

Dalis žodžių nereikalauja didesnių transformacijų, pageidaujama forma atsiranda didžiąsias raides pavertus mažosiomis ar pakeitus kai kurias priebalses ar balsius su diakritikais. Daugiau negu pusė tokių formų yra Kleino raštuose ir XVIII–XIX a. Biblijos leidimuose (2 lentelės V skiltis), o lotyniškoje Kleino gramatikoje jų ypač daug dėl neproblemiškos lotynų kalbos ortografijos.

2 lentelės VI skiltyje apskaičiuotas sėkmingų formų skaičius po galutinės normalizacijos. Jis svyruoja tarp 82 ir 97 proc. Didžiausias tikslumas pasiektas formose iš Giedraičio Naujojo Testamento ir lotyniškos Kleino gramatikos. Panašus rezultatas ir XVIII amžiaus Biblijos leidimuose. Ganėtinai gerai veikia ir žodžių rytų aukštaičių tarme išspausdintose Jaknavičiaus *Evangelijose* transliteracija. Mažiausiai teisingų atitikmenų sugeneruota pirmojoje spausdintoje lietuviškoje knygoje.

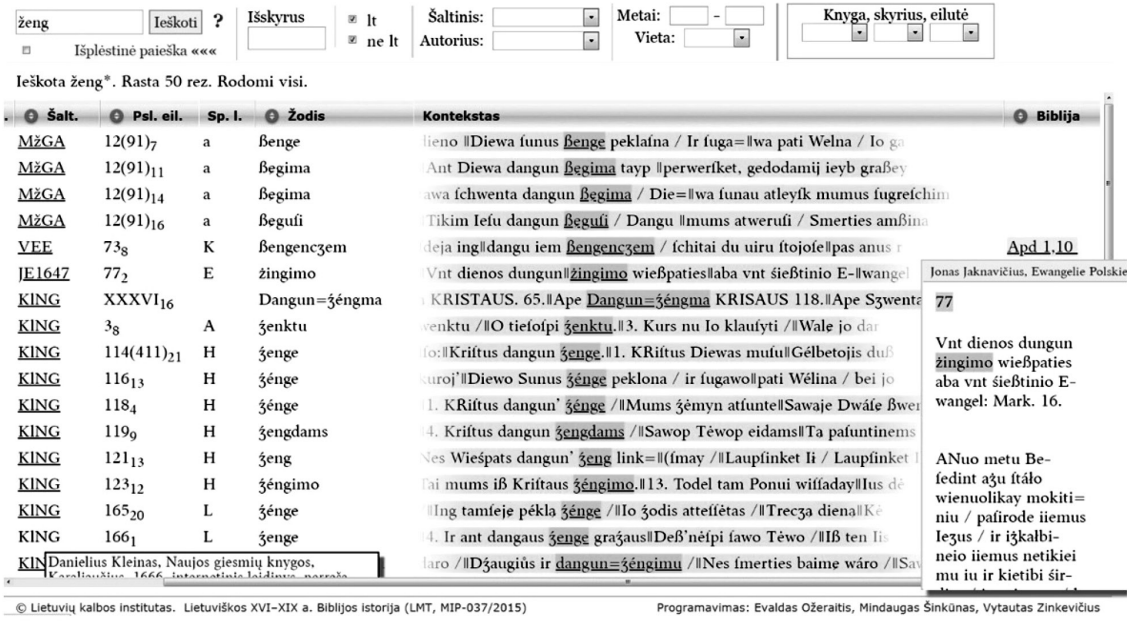
Sugeneruotos formos naudojamos internetiniame puslapyje http://sr.lki.lt. Svetainė veikia kaip žodžių formų ar jų dalių paieškos sistema, šaltinių peržiūros įrankis arba interlinearinis senųjų Biblijos vertimų leidimas. Galima dviejų tipų paieška: universalioji ir tikslioji. Jei pasirenkama ieškoti „Žodis originalia rašyba“ (pažymima varnelė po užklausa), ieškoma tik tiksliai originalų užrašymą atitinkančių raidžių, pavyzdžiui, „ʒeng“ parodys formas, prasidedančias šiomis raidėmis. Tokius rezultatus galima modifikuoti užpildant laukelį „Išskyrus“, pavyzdžiui, jame įrašius „ʒengim“ iš rezultatų bus eliminuotos formos *ʒengimas*, *ʒengimo* ir pan.

Numatytasis paieškos režimas yra universalus, formų juo galima ieškoti originalo („ßeng“), originalia modifikuota („szeng“) arba dabartine („ženg“) rašyba. Pagal užklausą „ženg“ surandamos formos: *ʒeng*, *ßenge* (su *sz*), *ne žeńgia*, *Ʒengimą* (su *Z*), *ßegima* (su *ę*), *ʒénge* (su *én*), *žingimo* (su rytų aukštaičių *in*), *ʒenktu* (su *k*) ir kt. (žr. 1 pav.).

Sudėtingesnės paieškos sąlygos formuojamos pasitelkiant ženklus „|“ (pridėjus jį užklausos gale ieškoma tik formų, visiškai atitinkančių įvestą tekstą), „*“ (bet koks kiekis bet kokių raidžių šioje teksto vietoje), „_“ (viena bet kokia raidė). Įvedus laužtinių skliaustų porą ieškoma visų juose nurodytų raidžių ar jų diapazono, pavyzdžiui, užklausa „[gž]yvat“ leis surasti formas *giwatas*, *gywata*, *ʒiwata*, *ʒywats*, *ßiwata* ir pan.

Rezultatus ribojantys kriterijai nustatomi pasirinkus išplėstinę paiešką: kalba, šaltinis, autorius, laikotarpis (metai nuo–iki), vieta, Biblijos knyga, skyrius, eilutė.

Rezultatuose pateikiamas rastų formų skaičius, šaltinio pavadinimas (santrumpa su iššokančiu aprašu), autorius, metai, metrika (puslapis, eilutė, spaudos lankas), žodžio forma, kontekstas ir Biblijos eilutė (jei žodis jai priklauso). Rezultatus galima rikiuoti abėcėlės tvarka pagal daugumą iš šių kriterijų (pagal kelis kriterijus vienu metu surikiuojama nuspaudus „Shift“). Platesnis kontekstas matomas atveriant šaltinio tekstą naujame lange (lei-



1 PAV. Senųjų raštų paieškos svetainė (http://sr.lki.lt). Užklausa *ženg** grąžina rezultatus ***ßenge***, ***ßegima***, ***žingimo***, ***ʒénktu***, *dangun=**ʒéngimu*** ir kt.; parodoma formos pavartojimo šaltinyje metrika, kontekstas, Biblijos eilutė; dominanti forma šaltinyje gali būti atverčiama atskirame lange (dešinėje)

džiami keli atviri langai vienu metu). Žodis iš Biblijos yra susietas su atitinkama Biblijos eilute, pavartota kituose šaltiniuose.

Svetainė yra sumanyta ir kaip interlinearinis Biblijos vertimų leidimas. Biblijos skyrius ar eilutė (-ės) atverčiamos palikus tuščią žodžio paieškos eilutę. Biblinės eilutės išrikiuojamos pagal šaltinio chronologiją. Gali būti rodomos ir visos vieno Biblijos skyriaus ar knygos eilutės. Rezultatus taip pat galima riboti pagal laikotarpį ar kurį kitą kriterijų.

Išvados

Lietuvių kalbos senųjų raštų tekstyne kaupiami šaltiniai, perrašyti pagal dokumentinio leidimo principus, išlaikomos originalo rašybos ar ja užkoduotos kalbos ypatybės. Skirtingais raštijos raidos etapais vartotos kelios ortografijos sistemos neleidžia taikyti kalbos analizės technologijų, paremtų dabartine rašyba. Rašybos skirtumus apeiti leidžia istorine rašyba parašytų žodžių sąsaja su ekvivalentais normalizuota rašyba.

Rašybos normalizacijos tikslai yra suniveliuoti skirtingus grafinius grafemos realizacijos būdus nepakeičiant ortografijos ir perrašyti dabartinės lietuvių kalbos abėcėle. Kartu vykdoma ir fonetikos normalizacija, kuri apima dialektinių fonetinių ypatybių panaikinimą bei fonemų realizavimo asimiliacinėse pozicijose sureguliavimą. Šie principai leido sukurti universalų paieškos mechanizmą, kuriame užklausos gali būti formuluojamos keliomis ortografinėmis sistemomis.

Didelė tekstyno apimtis ir riboti ištekliai verčia ieškoti būdų, kaip atitikmenų gamybą automatizuoti. Sukurtas taisyklių rinkinys remiasi empiriniais rašybos istorijos tyrimais. Taisyklės hierarchiškai rikiuojamos pagal apdorojamų ženklų ilgį, jų taikymas ribojamas pagal metaduomenis atsižvelgiant į konkretaus šaltinio rašybos ypatybes. Pasiektas 82–97 proc. teisingai sugeneruojamų atitikmenų tikslumas.

Taisyklėmis paremtos normalizacijos privalumas yra darbo automatizavimas ir perrašo nuoseklumas, trūkumai – sugeneruoja-

mas ne vienas, bet keli atitikmenys, o dėl daugiareikšmių taisyklių sukuriama ir kalboje neegzistuojančių formų. Atitikmenų, naudojamų paieškos sistemoje, gausa sumažinta pagal neegzistuojančius lietuvių kalboje raidžių junginius ir siaurinant užklausoje vartojamą abėcėlę. Tolesnė sugeneruotų atitikmenų atranka galėtų vykti pagal žodynus ir pritaikant dabartinei lietuvių kalbai sukurtus automatinės morfologijos, o vėliau ir sintaksės, analizės įrankius.

Šaltiniai

B – *Biblija*, Karaliaučius, 1735; Biblioteka Jagiellońska, sign.: SD 52067 I. Dokumentinis perrašas, parengė Mindaugas Šinkūnas, Ernesta Kazakėnaitė, Jurgita Venckienė, Birutė Triškaitė (2018).

GNI – Juozapas Arnulfas Giedraitis, *Naujas įstatymas*, Vilnius, 1816. Dokumentinis perrašas, parengė Mindaugas Šinkūnas, Jurgita Venckienė (2018).

JE – Jonas Jaknavičius, *Ewangelie polskie i litewskie*, Vilnius, 1647; Kauno technologijos universiteto biblioteka, sign.: C 72772. Dokumentinis perrašas, parengė Milda Lučinskienė (1998).

KlGr – Danielius Kleinas, *Grammatica Litvanica*, Karaliaučius, 1653; Vilniaus universiteto biblioteka, sign.: Lr 484. Dokumentinis perrašas, parengė Mindaugas Šinkūnas (2011).

KlM – Danielius Kleinas, *Naujos maldų knygelės*, Karaliaučius, 1666; Biblioteka Uniwersytecka w Toruniu, sign.: Pol.7.II.4335. Dokumentinis perrašas, parengė Vaidotas Rimša (2003), Mindaugas Šinkūnas (2006).

KlNG – Danielius Kleinas, *Naujos giesmių knygos*, Karaliaučius, 1666; Biblioteka Uniwersytecka w Toruniu, sign.: Pol.7.II.4334. Dokumentinis perrašas, parengė Vaidotas Rimša (2003), Mindaugas Šinkūnas (2006).

MžF – Martynas Mažvydas, *Forma krikštymo*, Karaliaučius, 1559; Biblioteka Uniwersytecka w Toruniu, sign.: Pol.6.II.5. Dokumentinis perrašas, parengė Mindaugas Šinkūnas (2014).

MžGA – Martynas Mažvydas, *Giesmė šv. Ambražiejaus*, Karaliaučius, 1549; Biblioteka Kórnicka PAN, sign.: Cim.O.279. Dokumentinis perrašas, parengė Mindaugas Šinkūnas (2014).

MžK – Martynas Mažvydas, *Katekizmas*, Karaliaučius, 1547; Vilniaus universiteto biblioteka, sign.: Lr 5650. Dokumentinis perrašas, parengė Diego Ardoino, Mindaugas Šinkūnas (2014).

NT – *Naujas Testamentas*, Karaliaučius, 1727; Lietuvių literatūros ir tautosakos instituto biblioteka, sign.: B1901. Dokumentinis perrašas, parengė Mindaugas Šinkūnas, Jurgita Venckienė, Birutė Triškaitė (2016).

Ps – *Psalteras Dovydo*, Karaliaučius, 1728; Lietuvių literatūros ir tautosakos instituto biblioteka, sign.: B1901. Dokumentinis perrašas, parengė Mindaugas Šinkūnas (2016).

VEE – Baltramiejus Vilentas, *Evangelijos bei Epistolos*, Karaliaučius, 1579; Vilniaus universiteto biblioteka, sign.: Lr 1387, Geheimes Staatsarchiv Preußischer Kulturbesitz, sign.: XX StUB Kgb. Nr. 54. Dokumentinis perrašas, parengė Ona Aleknavičienė (2016).

Literatūra

Aleknavičienė Ona 2008, Senoji lietuvių raštija internete, *Archivum Lithuanicum* 10, 297–350.

Ambrazas Saulius, Ona Aleknavičienė, Vytautas Zinkevičius 1998, Istorinis lietuvių kalbos žodynas ir senųjų raštų kompiuterizavimas, *Lietuvių kalbotyros klausimai* 39, 192–210.

Baron Alistair, Paul Rayson 2008, VARD 2: A Tool for Dealing with Spelling Variation in Historical Corpora, *Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, UK, May 2008.

Bollmann Marcel, Florian Petran, Stefanie Dipper 2011, Rule-Based Normalization of Historical Texts, *Proceedings of Language Technologies for Digital Humanities and Cultural Heritage Workshop*, Hissar, Bulgaria, September 2011, 34–42.

Gelumbeckaitė Jolanta, Mindaugas Šinkūnas, Vytautas Zinkevičius 2012a, Senosios lietuvių kalbos tekstynas (SLIEKKAS) – nauja diachroninio tekstyno samprata, *Darbai ir dienos* 58, 257–281.

Gelumbeckaitė Jolanta, Mindaugas Šinkūnas, Vytautas Zinkevičius 2012b, Old Lithuanian Reference Corpus (SLIEKKAS) and Automated Grammatical Annotation, *Journal for Language Technology and Computational Linguistics* 27 (2), 83–96.

Ljubešić Nikola, Katja Zupan, Darja Fišer, Tomaž Erjavec 2016, Normalising Slovene Data: Historical Texts vs. User-Generated Content, *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 146–155.

Mittmann Roland 2013, Old German and Old Lithuanian: the Creation of Two Deeply-Annotated Historical Text Corpora, Коллектив авторов (Ответственные редакторы: Виктор П. Захаров, Ольга А. Митрофанова, Мария В. Хохлова), *Труды международной научной конференции «Корпусная лингвистика – 2013» / Proceedings of the international conference „Corpus linguistics – 2013“*, Санкт-Петербург: Санкт-Петербургский государственный университет, Филологический факультет / St. Petersburg: St. Petersburg State University, Philological Faculty, 103–111.

Pettersson Eva, Beáta Megyesi, Joakim Nivre 2012, Rule-Based Normalisation of Historical Text – A Diachronic Study, *Proceedings of the First international Workshop on Language Technology for Historical Text(s). KONVENS*, Vienna, Austria, September 2012, 333–341.

Šinkūnas Mindaugas 2014, Mažosios Lietuvos raštų ortografijos reforma XVII amžiuje. I. Pučiamųjų priebalsių ir afrikatų žymėjimas, *Archivum Lithuanicum* 16, 2014, 9–58.

Šinkūnas Mindaugas 2015, Lietuvių ortografijos skirtumai katalikų ir evangelikų reformatų raštuose (XVI–XVII a.): priebalsio /j/ ir diftongų /ai/, /ei/ rašyba, *Baltistica* 50 (2), 197–244.

Tang Gongbo, Fabienne Cap, Eva Pettersson, Joakim Nivre 2018, An Evaluation of Neural Machine Translation Models on Historical Spelling Normalization, *Proceedings of COLING 2018*, arXiv:1806.05210v2

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

THE NORMALIZATION OF OLD LITHUANIAN ORTHOGRAPHY FOR USAGE IN A SEARCH ENGINE

Summary

The Lithuanian historical corpus consists of machine-readable texts, transcribed according to the principles of documentary edition; the original spelling and the language features it encodes are preserved. Several orthographic systems were used during the various stages of the history of Lithuanian language, and some of them differ from the modern one to a relatively great extent. The historical orthography does not allow the use of language analysis tools, which were developed on the basis of the modern spelling. A link is therefore needed that would connect the historical orthography to the modern orthography used today.

In normalizing spelling, various challenges must be dealt with: the same grapheme must be differently realized without changing the orthography and by rewriting the form in the modern Lithuanian alphabet. At the same time, the normalization of phonetics has to be carried out, which includes the elimination of dialectal phonetic features and the representation of phonemes in the assimilated position. These principles can be used in constructing a universal search engine, in which queries can be processed across different orthographic systems (http://sr.lki.lt).

The size of the corpus and the available limited resources stimulate the search for an automated way of normalizing orthography. A set of rules was developed based on the empirical research on the history of orthography; these rules were then arranged hierarchically in accordance with the length of the sequence of processed characters, their implementation being limited to using the metadata according to the spelling features of a particular source. A 82–97% accuracy level of correct normalization was achieved.

The advantage of a rules-based transliteration is the consistency of changes; the disadvantage can be seen in generating not a single but several equivalents of the word, and the ambiguous rules in certain cases generate many tokens that do not exist in the natural language. The number of generated forms being fed to the search engine was reduced based on non-existent letter sequences and by narrowing the query alphabet. A further selection of the correct forms could be done using dictionaries or tools for analyzing the morphology and syntax of modern Lithuanian.