Vilnius University Press

Contents lists available at Vilnius University Press

Taikomoji kalbotyra, 16: 42–48
https://www.journals.vu.lt/taikomojikalbotyra

eISSN 2029-8935
DOI: https://doi.org/10.15388/Taikalbot.2021.16.3

Tradicinių aiškinamųjų žodynų antraštyno atnaujinimo būdas

Rūta Petrauskaitė
Vytauto Didžiojo universitetas
ruta.petrauskaite@vdu.lt

Virginijus Dadurkevičius
Vytauto Didžiojo universitetas
virginijus.dadurkevicius@vdu.lt

Anotacija. Straipsnyje pristatomas būdas, leidžiantis atnaujinti tradicinius suskaitmenintus žodynus, pasirėmus žodynuose pateiktų lemų palyginimu su dideliuose tekstynuose vartojama leksika. Tokiam lyginimui galimybę suteikia *Hunspell* platforma, kuri leidžia generuoti visas žodžio formas pagal lemą ir jai priskirtas darybos taisykles. Atlikus *Dabartinės lietuvių kalbos žodyno* 6-osios laidos leidimo skaitmeninės versijos lemų analizę ir palyginus ją su *Jungtinio lietuvių kalbos tekstyno* leksiniais duomenimis, buvo gauti du sąrašai: žodyne pristatytos, bet tekstyne neaptiktos leksikos, ir žodyne nepateiktos, bet tekstyne dažnai vartojamos leksikos. Pastarasis sąrašas pristatomas detaliai tam, kad išryškėtų, kokia leksika turėtų būti papildomi nauji žodyno leidimai.
Raktažodžiai: *tekstynų lingvistika; jungtinis lietuvių kalbos tekstynas; tradicinė leksikografija; dažninis sąrašas; „Hunspell“ platforma; žodynų atnaujinimas*

A method to update traditional explanatory dictionaries

Summary. In the paper the method is presented how to update traditional digitalised dictionaries based on comparison of the dictionary lemmas and a big corpus. *Hunspell* platform is used for generation of all the word forms from the dictionary lemmas. 6th edition of *The Dictionary of Modern Lithuanian* was chosen for its comparison with the lexical data from *The Joint Corpus of Lithuanian.* The outcome of the comparison was two lists of non-overlapping lexis: the list of the dictionary lemmas unused in the present-day Lithuanian and the list of the dictionary gaps, i.e., frequently used words and word forms ignored by the dictionary. The latter is discussed in greater detail to give lexicographers a clue for updates.
Keywords: *corpus linguistics; a joint corpus of Lithuanian; traditional lexicography; frequency list; Hunspell platform; dictionary updates.*

1. Įvadas

Tekstynai kaip pagrindinis duomenų šaltinis pradėti taikyti leksikografijai ne nuo savo atsiradimo, bet šiek tiek vėliau, apie 1980 metus. Tačiau nuo tada iki šiol tekstynai ir iš jų gauti išvestiniai duomenys (dažniniai žodžių formų ar lemų sąrašai, konkordansai, raktažodžiai, kolokacijos ir t. t.) plačiai naudojami leksikografų. Tekstynų lingvistika, kalbos technologijos yra neatskiriamos nuo moderniosios leksikografijos – visi žodynų rengimo etapai yra vienaip ar kitaip su jomis susiję. Vertėtų trumpai apžvelgti tekstynų taikymo būdus žodynams rengti.

Pats paprasčiausias, bet drauge ir populiariausias tekstynų naudojimas yra jų, kaip autentiškų, leksikografų nesugalvotų, tačiau realią komunikacijos situaciją atspindinčių pavyzdžių šaltinio. Pavyzdžiai paprastai atrenkami iš neanotuotų tekstynų antraštinių žodžių reikšmėms, įprastiniams junglumo partneriams, gramatinėms kategorijoms, kitiems vartosenos ypatumams iliustruoti dažnai net nesiekiant išanalizuoti, suklasifikuoti ir apibendrinti visų vartosenos pavyzdžių, t. y. visų konkordanso eilučių. Kitaip sakant, iš daugybės vartosenos pavyzdžių atsirenkami tik tinkamiausi ir reikalingiausi. Tokia tekstynais paremta (angl. *corpus-based*) leksikografija iš esmės priklauso tradiciniam žodynų rengimo būdui. Tekstynų nulemtoje (angl. *corpus-driven*) leksikografijoje tekstynai pritaikomi kur kas įvariau ir produktyviau. Antraštinių žodžių semantinei struktūrai nustatyti ir atskirų reikšmių vartosenai iliustruoti rankiniu būdu ar automatiškai analizuojami konkordansai, iš kurių išgaunami ne tik dažniausios vartosenos pavyzdžiai, bet ir skirtingas žodžių vartosenos skirtingomis reikšmėmis dažnumas (tradiciniuose žodynuose, beje, visai neatsispindintis), funkcinių stilių ir specifinių šaltinių įvairovė, įprastinės gramatinės formos. Visam tam darbui pakanka neanotuotų tekstynų.

Dar daugiau galimybių leksikografams suteikia morfologiškai ir sintaksiškai anotuoti tekstynai. Būtent iš jų išgaunami gramatiniai vartosenos modeliai, atspindintys gramatinių formų ir kategorijų paradigmą, paremtą statistine jų analize. Kiekvienam leksikografus dominančiam žodžiui sudaromas specialus jo vartosenos modelis. Bet net ir tokiu atveju tekstynai dar nėra panaudojami tiek, kiek būtų galima. Geriausiai tekstynai panaudojami leksikografijai tuomet, kai žodynai nuo pat pradžių imami sudaryti tekstynų pagrindu. Adamas Kilgarriffas (2012) šį procesą apibūdino taip: pirmiausiai sudaromas būsimo žodyno tikslus atitinkantis tekstynas, tada jis apdorojamas šiuolaikiškiausiomis automatinėmis kalbos analizės priemonėmis (pasitelkiant lemuoklį, parserį, kolokacijų išgavimo bei jų klasifikavimo įrankius ir t. t.). Iš gautos dažninio žodžių sąrašo viršutinės dalies (priklausomai nuo rengiamo žodyno apimties, imant nuo trečdalio iki pusės sąrašo lemų) sudaromas antraštinių žodžių sąrašas. Vėliau atliekama kiekvieno žodžio analizė remiantis jo konkordansais ir iš tekstyno išgauta statistika (pvz., su *Sketch Engine* žr. Kilgarriff *et al.* 2004), junglumo partneriais, gramatinės vartosenos modeliais ir pan. Parengus žodyninio straipsnio metmenis, atspindinčius leksemos semantinę struktūrą, prie tekstyno grįžtama automatiškai pasirinkti geriausių pavyzdžių (Kilgarriff *et al.* 2008).

Taigi šiandien niekas neabejoja, kad vadinamoji tekstynų revoliucija įgalino leksikografus geriau atspindėti kalbos vartoseną. Be to, ji leido greičiau parengti kur kas geresnių žodynų. Tai akivaizdu naujai ir ištisai tekstynų pagrindu rengiamų žodynų (angl. *born digital*) atveju. Tačiau taip rengiami toli gražu ne visi žodynai. Dalis jų, rengtų tradiciniais metodais, bet vėliau suskaitmenintų, taip pat galėtų ir turėtų būti atnaujinti. Ypač svarbu būtų atnaujinti senų žodynų antraštinių žodžių sąrašus, kurie sensta akivaizdžiausiai. Būtent apie tai ir yra pristatomas žodyno ir tekstyno palyginimo metodas, leidžiantis nustatyti žodyne pateikiamą, bet šiuolaikinėje kalboje nevartojamą leksiką, taip pat žodyno antraštinių žodžių spragas, t. y. trūkstamą naujausią leksiką.

Mūsų tyrimas yra skirtas palyginti *Dabartinės lietuvių kalbos žodyno* 6-osios laidos variantą, vieninteli variantą skaitmeniniu formatu prieinamą vartotojams (toliau – DLKŽ6 arba Žodynas), su gausiais

tekstiniais ištekliais, vadinamuoju jungtiniu tekstynu (toliau – JT arba Tekstynas), siekiant įvertinti, kiek DLKŽ6 antraštinių žodžių sąrašas sutampa su skaitmeniniuose ištekliuose vartojama leksika (plačiau žr. Dadurkevičius, Petrauskaitė 2020). Kitas svarbus išteklius – atvirojo kodo *Hunspell* platformos pagrindu formalizuota lietuvių kalbos morfologija ir visa DLKŽ6 leksinė informacija, kuri buvo pritaikyta palyginti šio žodyno ir JT leksiką (plačiau žr. Dadurkevičius 2017). Šiame straipsnyje susitelkiama į DLKŽ6 neįtrauktų žodžių, vadinamųjų spragų, sąrašą, kuris buvo išanalizuotas vartosenos dažnio, leksiniu semantiniu ir gramatiniu aspektais, parodant dažniausias žodžių formas, kurių tiesioginių lemų nėra DLKŽ6. Trumpai pakomentuojamas ir priešingas, vadinamųjų perteklinių DLKŽ6 žodžių sąrašas, kuris buvo sudarytas kruopščiai patikrinus visas galimas perteklinio žodžio formas jungtiniame tekstyne. Neaptikus nė vienos iš jų, tokia lema buvo įtraukta į perteklinių žodžių sąrašą.

Svarbu pažymėti, kad visas analizės procesas pristatytas tik kaip galimas metodas žodynų ir tekstynų lyginamajai analizei, bet ne anksčiau rengtų žodynų kritikai. Pasirėmus lyginamąja analize, galima sėkmingai ir rentabiliai atnaujinti suskaitmenintus, tradiciniais būdais rengtus žodynus. Suprantama, kad žodynų naujinimas neužkerta kelio rengti žodynus tekstynų pagrindu nuo pat pradžių. Be to, pristatoma metodika yra nuo kalbos nepriklausoma, taigi ją galima taikyti ir kitų kalbų tradiciniams žodynams atnaujinti.

2. Žodyno ir Tekstyno lyginimo būdai

Tyrimas pagrįstas nekontekstinės analizės prielaida, t. y. kiekvienas JT žodis buvo nagrinėjamas atskirai, neatsižvelgiant į jam artimiausių kitų žodžių kaimynystę. Ši prielaida leido vietoje ištisinio teksto masyvų naudoti dažninius žodžių sąrašus, kuriuose pateikiama žodžio forma ir jos dažnis, o būtinųjų skaičiavimų laiką sutrumpinti nuo kelių metų iki kelių dienų. Iš viso JT buvo rasta 5 mln. skirtingų žodžių formų: nuo milijonus kartų pasikartojančių jungtukų bei prielinksnių iki tik kartą pavartotų retų žodžių formų, tokių kaip *darninančią*, *pusantrametis*, *žuvelyčių*, nuo dažnų taisyklingų bendrinės kalbos žodžių iki retų vardų, rašybos klaidų, atsitiktinai patekusių kitų kalbų žodžių ir pan.

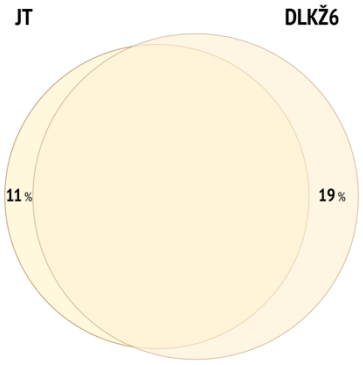
DLKŽ6 ir JT palyginimas nėra paprastas uždavinys, nes žodyne yra nurodomos antraštinės žodžio formos – lemos, o tekstyną sudaro konkrečios žodžių formos. Kadangi lietuvių kalba pasižymi didele žodžių darybos ir kaitybos formų įvairove, tai, pavyzdžiui, iš daiktavardžio lemos galima generuoti keliolika skirtingų žodžio formų, o iš veiksmažodžio – net daugiau nei pusę tūkstančio formų. Siekiant automatizuoti visą šį sulyginimo procesą, buvo pasitelkta universali *Hunspell* platforma[1], leidžianti sujungti lemų sąrašą ir atskirai aprašytas darybos taisykles. Šiam tyrimui buvo naudojamos pagal Dabartinės lietuvių kalbos gramatiką (2006) anksčiau pristatytos darybos taisyklės (Dadurkevičius 2017), o lemų sąrašas buvo sudarytas iš visų DLKŽ6 tiesiogiai užrašytų ir numanomų lemų, prie kurių priskiriamos veiksmažodžio sangrąžinės formos, veiksmažodžių, būdvardžių, prieveiksmių formos su priešdėliais *ne-, nebe-, te-, tebe-*. To rezultatas – du 200 000 lemų ir 5 000 darybos taisyklių failai, taip pat galimybė programiškai iš DLKŽ6 lemų generuoti virš 50 mln. teoriškai galimų jų formų.

Pirmiausiai buvo sudarytas dažninis JT žodžių formų sąrašas, sužymint, kokios formos *Hunspell* platformoje neatpažįstamos kaip teisingos, kitaip sakant, tokios formos, kurioms nepavyko surasti jokios lemos ir jokios ją sukūrusios darybos taisyklės. Šis sąrašas gali būti pravartus vertinant žodyno spragas. Kitas, DLKŽ6 lemų vartojamumo sąrašas, buvo sudarytas nustatant, kiek kartų atskira žodyno

[1] *Hunspell* platforma – https://hunspell.github.io/.

lema galėjo būti panaudota interpretuojant visas JT žodžių formas: nulis kartų reiškia, kad tokia žodyno lema jokia teoriškai galima savo forma nėra aptikta JT.

Abu šie sąrašai kartu su dviem *Hunspell* platformos failais yra pateikti CLARIN-LT atvirosios prieigos resursų saugykloje (Dadurkevičius 2020), o 1,3 mlrd. žodžių JT sandara, DLKŽ6 neatpažintų JT žodžių formų (maždaug kas dešimtas JT žodis) ir nenaudojamų lemų (maždaug penktadalis DLKŽ6 lemų) pagrindiniai statistiniai duomenys yra pristatyti tarptautinėje Baltijos šalių kalbų technologijų konferencijoje (Dadurkevičius, Petrauskaitė 2020). 1 pav. pateikiama šiuos duomenis apibendrinanti diagrama, parodanti sutampančias ir besiskiriančias Tekstyno ir Žodyno dalis.



1 pav. Asociatyvus JT ir DLKŽ6 palyginimas

3. Žodyno ir Tekstyno leksikos palyginimo rezultatai

Įdėmesnis žvilgsnis į JT neaptiktų DLTŽ6 lemų sąrašą, gautą iš DLKŽ6 lemų vartojamumo sąrašo atrinkus lemas, pažymėtas nuliu, atskleidžia kai kuriuos šios leksikos ypatumus, paaiškinančius, kodėl jos neaptinkamos šiandienėje kalboje[2]. Sąrašą sudaro 16 272 lemos, pateiktos abėcėlės tvarka su kalbos dalies žymomis. Visos perteklinės lemos tiriamame Žodyne yra antraštiniai žodžiai. Kitu atveju, jei tai iš veiksmažodžių ar būdvardžių padaryti daiktavardžiai su priesagomis *-amas, -imas, -umas*, tai jų lemos pateikiamos žodyninio straipsnio pabaigoje. Manytina, kad būtent dėl to, kad tokie daiktavardžių abstraktai yra dirbtiniai, vartosenos nepatvirtinti dariniai, jie jokia forma nebuvo atrasti JT. Taip pat svarbu pasakyti, kad didelę perteklinės leksikos sąrašo dalį sudaro po kelis tos pačios šaknies vedinius: *abuojėjimas, abuojėti, abuojybė* arba darybinius sinonimus: *akėčvilkos, akėtvilkos*. Dėl to daugelio jų darybinė reikšmė yra aiški: *alkūniavimas, alkūniuoti, alkūniuotis.* Leksinę reikšmę lemia žodžio šaknis, jei ji yra neatpažįstama, nesuprantami lieka ir jos dariniai.

JT neaptiktų DLTŽ6 lemų sąrašas buvo peržiūrėtas ir įvertintas remiantis gimtakalbių autorių kalbos jausmu, siekiant nustatyti šiandieniam kalbos vartotojui sunkiai suprantamą ar visai nesuprantamą leksiką. Akivaizdu, kad toks atrankos metodas neapsaugo nuo subjektyvumo, nes skirtingi žmonės pateiktų kiek kitokias lemas bei skirtingą jų skaičių. Toliau pateikiamais pavyzdžiais, išrinktais tik iš A (962 vienetai) raidės sąrašo, siekiama ne atlikti išsamią Žodyno perteklinės leksikos analizę – ne toks yra straipsnio tikslas – o tik pailiustruoti atrankos rezultatą:

> *alūnuoti,* ***amžindie, anuotis****, antprusnis,* ***antrur****, apdavai, apdrykti, apdvokėlis,* ***apgrisimas****, apgūžti,* ***apgvaibti,*** *apibubyti, apydžia,* ***apyžlėja, apliaukoti, apmaudimas, apsialsinti, apsičiurinti, apsiviežti, apskalba, apvažumas****, apžindis,* ***apžiūrumas, apžosti****, ardinėti, arkymasis, ašmainis, atasprąstis, atgajumas, atgaleivis, atgamumas, atgandus, atgūlis, atkragus,* ***atlagenti, atmainumas, atminkai, atminkalai, atsigodėjimas****, atskaras, atsklembti,* ***atšlainis,*** *atšolys, atspiromis,* ***atstapas,*** *atvėtos,* ***atžūlėlis, atžūliai, augila****.*[3]

Siekiant įvertinti galimas DLKŽ6 spragas, anksčiau aptartas dažninis JT žodžių formų sąrašas buvo filtruotas paliekant tik neatpažintas formas ir neįtraukiant akivaizdžių rašybos klaidų, kitų kalbų žodžių,

[2] Didelė JT apimtis leidžia daryti prielaidą, kad jis yra reprezentatyvus ir atspindi šiandienę lietuvių kalbą.

[3] Paryškintos lemos jau nebuvo įtrauktos į DLKŽ7, http://lkiis.lki.lt/dabartinis.

tikrinių vardų ir pavardžių. Jame pateikti nekaitomi žodžiai ir kaitomų žodžių formos (toliau – žodžių formos), išdėstytos mažėjančio dažnio tvarka. Visą sąrašą sudaro 254 726 vienetai. Dėl didelės jo aprėpties toliau buvo dirbama tik su viršutine sąrašo dalimi, prasidedančia nuo 100 pavartojimo atvejų. Šioje dalyje rasta 21 388 žodžių formos (8,4 proc. nuo viso sąrašo). Ji baigiasi sąrašo viršuje esančiu dažniausiu žodžiu *šiek* (tiksliau – samplaikos *šiek tiek* dalimi), pavartotu 175 933 kartus. Dar įdėmiau sąrašas buvo peržiūrėtas pasitelkus tik aukščiausiai esančią jo dalį, apimančią žodžių formas, pavartotas dažniau nei 2 000 kartų (1 336 vienetų, 0,5 proc. nuo viso sąrašo). Pažymėtina, kad skirtingos kai kurių žodžių formos užima skirtingas vietas dažniniame sąraše, pavyzdžiui, dažnas deminutyvas *krepšelis*, skirtingomis formomis pačių dažniausių formų sąraše pasirodo 5 vietose: *krepšelio* 13 084*, krepšelį* 4 361*, krepšelis* 4 214*, krepšeliui* 3 389*, krepšelių* 2 580*.* Čia pristatant Tekstyne itin dažnai vartojamus, tačiau Žodyne nepateiktus žodžius, pasitelkiama pagrindinė jų forma, o suminis atskirų formų pavartojimo dažnis nenurodomas. Tačiau informacija apie tai, kokiomis gramatinėmis formomis yra vartojami žodžiai ir koks yra atskirų formų vartojimo dažnumas, yra labai svarbi. Viena iš svarbiausių tekstynų lingvistikos nuostatų, vadinamoji leksikos ir gramatikos vienovė, yra apie tai, kad kiekvienas žodis turi savo gramatiką, t. y. tik jo vartosenai būdingas gramatines formas (Sinclair 2000), taigi jo leksinis gramatinis vartosenos modelis apima toli gražu ne visą galimų gramatinių formų paradigmą. Net ir pavartotų formų dažnumas smarkiai skiriasi. Be to, atskiros kaitybinės žodžių formos (juo labiau darybinės) siejasi su lemos semantine struktūra, su atskiromis daugiareikšmio žodžio reikšmėmis.

DLKŽ6 spragų sąrašas, net ir pati viršutinė jo dalis, pernelyg ilgas nuoseklesnei analizei, ją turėtų atlikti leksikografai, rengiantys naujausius DLKŽ leidimus ar kitus aiškinamuosius lietuvių kalbos žodynus. Čia norima tik paminėti kai kuriuos šio sąrašo ypatumus, prieš tai sugrupavus žodžius į gramatines, darybines ir leksines semantines grupes.

Labiausiai į akis krintanti grupė yra tarptautiniai žodžiai ar skoliniai, anksčiau ar visai neseniai atėję į lietuvių kalbą. Nemenka jų dalis buvo įtraukta į tarptautinių žodžių žodynus, kai kurie – į terminų žodynus, paliekant DLKŽ savai leksikai. Tačiau didžiumos tarptautinių žodžių paplitimas įvairiuose šiandienės kalbos diskursuose, jų vartojimo dažnumas verčia abejoti tokio priverstinio jų atskyrimo racionalumu. Tai ypač pasakytina apie tą tarptautinę adaptuotą leksiką, kuri neturi lietuviškų atitikmenų, plg.: *bliuzas, choreografas, civilis, ekspremjeras, infrastruktūra, inovacija, koma, klasifikatorius, klipas, kremavimas, lageris, mafija, mikroautobusas, politologas, populizmas, reitingas, signataras.* Sąraše dominuoja daiktavardžiai, tačiau esama ir būdvardžių bei prieveiksmių: *alternatyvus, habilituotas, identifikacinis, juvelyrinis, korumpuotas, plastikinis; emociškai, morališkai, tradiciškai.* Yra ir neadaptuotų žodžių: *amplue, ego, liuks, šou*. Pasitaiko darybos požiūriu mišrių darinių su tarptautiniais formantais: *biodegalai, biokuras, bioetika, mikroautobusas, viceministras, vicemeras*.

Kita dažna, nors ir negausi, grupė yra santrumpos, nepatekusios į DLKŽ6 santrumpų sąrašus: *cm, ha, kg, km* ir t. t. Taip pat ir itin dažnai vartojamos dalelytės *juolab, visgi, negi, vėlgi.* Tačiau didžiausią į DLKŽ6 neįtrauktų žodžių grupę sudaro žodžiai, kurių ignoravimas gali būti paaiškintas tik bendrosiomis žodyno sąrangos nuostatomis, aprašytomis įvadiniuose straipsniuose. Jais remiantis galima manyti, kad dalis analizuojamo sąrašo žodžių nepateko į žodyną tik kaip specifinė kalbos dalis, pavyzdžiui, *tradiciškai*, nes žodyne pateikiamas pamatinis daiktavardis *tradicija.* Toliau DLKŽ6 neaptikti žodžiai pateikti suklasifikuoti kalbos dalimis. Kaip jau minėta, jie pateikiami pagrindine forma (lema) nenurodant atskirų gramatinių formų ir jų dažnio abėcėlės tvarka. Sąraše dominuoja daiktavardžiai:

> ***alyvuogė,*** *analitikas, anapilis,* ***animacija, aplinkosauga****, apšvieta, atidėjinys, bankininkystė, dokumentika, dvynukai (yra dvynys), garantas, gimstamumas, išieškotojas, išminuotojas, įžaidėjas, kapavietė, lageris, lyderystė, lieknėjimas, nominacija,* ***nuotekos,*** *pagrobėjas, paminklosauga, pareigybė, perdirbėjas, popmuzika,* ***programišius, rinkodara****, sovietmetis,* ***spausdintuvas, spinduliuotė****, stebė-*

*sena, šauktinis, šiandiena, taikdarys, teisėkūra, teisėsauga, teisėsauginikas, tinklalaidė**, tinklaraštis, ugdymas**, vandentvarka, žyma.*[4]

Į akis krinta visa kategorija žodžių ir jų formų, sistemiškai neatspindėta DLKŽ6. Tai sangrąžiniai veiksmažodiniai daiktavardžiai, t. y. išvestiniai žodžiai:

aiškinimasis, apsikeitimas, atsipalaidavimas, atsisakymas, atsistatydinimas, gelbėjimasis, įsidarbinimas, įsipareigojimas, įsiskolinimas, įsitikinimas, išsilavinimas, išsimokslinimas, išsiskyrimas, išsivystymas, jungimasis, nusidėvėjimas, pasipiktinimas, pasivaikščiojimas, pasirengimas, persirengimas, prisijungimas, susijaudinimas, susikaupimas, susižavėjimas.

Norint juos susirasti žodyne, reikia žengti du ar tris žingsnius atgal link pamatinio žodžio, pavyzdžiui, ieškoti tokia eilės tvarka: *aiškinimasis – aiškinimas – aiškintis – aiškinti*. Tik pastaroji veiksmažodinė forma *aiškinti* yra pateikta kaip lema, matyt, darant prielaidą, kad visi kiti darybos formatai kalbos vartotojo pasidaromi pagal bendras darybos taisykles. Tai, be abejo, tinka gimtakalbiams, tačiau žodynas skirtas ir užsieniečiams, kurie tik mokosi lietuvių kalbos ir kuriems toks paieškos kelias akivaizdžiai per ilgas. Be to, dariniai vartosenoje įgyja ir naujų reikšmių bei konotacijų, ypač vartojami stabiliuose žodžių junginiuose, kurios niekaip negali būti išvedamos iš pamatinio žodžio reikšmės, plg., *santykių aiškinimasis*. Arba *programišius,* kuris tikrai nereiškia ‚programavimo aistruolis', nes šiandienėje vartosenoje atsiradusi neigiama konotacija lemia jo reikšmę, tad dabar šis žodis vartojamas nelegaliam įsibrovėliui, turinčiam piktų kėslų, apibūdinti. Panašiai ir *išsilavinimas – išsilavinti – išlavinti*. Tik nesangrąžinio veiksmažodžio *išlavinti* žodyniniame straipsnyje randama sangrąžinė forma *išsilavinti* ir daiktavardis *išsilavinimas*, šiandien ypač dažnai vartojamas kasdienėje kalboje, o ne Žodyno lema *išlavinti*. Skaitmeninio Žodyno varianto paieškos sistema nesunkiai galėtų padėti atrasti šiuos vedinius antraštinio žodžio straipsnyje, tačiau atrandama tik antraštinė lema. Tas pat pasakytina ir apie kitų veiksmažodinių abstraktų, pavyzdžiui, *privatizavimas, gimstamumas, likvidumas, sertifikavimas, suderinamumas* paiešką, nes neįmanoma atrasti *gimstamumas*, tik *gimti,* o šio iš dalyvio padaryto daiktavardžio reikšmė nuo veiksmažodžio nutolusi. Bet kuriuo atveju svarbu pažymėti, kad būtent daiktavardis yra itin dažnas dėl šių dienų Lietuvos aktualijų, todėl leksikografams jo ignoruoti nederėtų.

Nemažą DLKŽ6 nepateiktų lemų sąrašo dalį sudaro priešdėlio *ne-* vediniai, dažnai vartojami JT:

neaiškumas, neatitikimas, neatsargumas, neatvykimas, nebuvimas, nedarbingumas, nediskriminavimas, neištikimybė, nekantrumas, nemalonumas, nepagarba, nepakankamumas, nepasitenkinimas, nepasitikėjimas, nepatogumas, nepritarimas, nestabilumas, nesugebėjimas, nesutarimas, nešališkumas, netikslumas, nevaisingumas, neveikimas, nevykdymas, nežinojimas.

Jų dažnumas gali rodyti savarankišką reikšmę, tam tikrą neiginių leksikalizaciją, ypač jei jie vartojami teisės, medicinos ir kituose specialiuose diskursuose. Norint nustatyti leksikalizacijos atvejus, būtina išsamiai tirti ir aprašyti jų vartoseną, nedarant prielaidos, kad priešdėlio *ne-* vediniai tėra tik pamatinio žodžio antonimai.

DLKŽ6 neatrasti veiksmažodžiai taip pat yra vediniai – priešdėliniai ir / ar sangrąžiniai. Pastarųjų atveju žodyno sudarytojų teigiama, jog neatsiradus naujai leksinei reikšmei, sangrąžiniai veiksmažodžiai pateikiami prie nesangrąžinių formų. Tačiau kai kada leksikalizacija yra akivaizdi, išryškėjanti iš šiandienės vartosenos, tačiau klaidingai atspindėta DLKŽ6, pavyzdžiui, *pasimylėti*, pateikiamą prie *mylėti* su šiuo pavyzdžiu: *Jie broliškai mylisi*[5]. Kiti šio tipo veiksmažodžiai:

---

[4] Paryškintos lemos jau buvo įtrauktos į DLKŽ7 http://lkiis.lki.lt/dabartinis.

[5] Tas pats pavyzdys paliktas ir DLKŽ7.

aiškintis, apskaityti, atsistatydinti, dalintis, inicijuoti, išanalizuoti, išpildyti, išpopuliarėti, įvardinti (yra įvardyti), kviestis, nesureikšminti, nusifotografuoti, nuvilnyti, padiskutuoti, pakomentuoti, pakoreguoti (yra koreguoti), parklupdyti, pasidalinti, pasiglemžti, pasikonsultuoti, pasižvalgyti, paviešinti, sukonkretinti (yra konkretinti), sulieknėti, sureaguoti, susiformuoti, susikoncentruoti, susimokėti, susisprogdinti, užblokuoti, užsitikrinti.

Žodyne taip pat nerasta kai kurių itin dažnų prieveiksmių:

genetiškai, piktybiškai, sąlyginai, santykinai,

ir būdvardžių:

alternatyvus, aukštakulnis, autorinis, daugiabutis, daugiamandatis, draudiminis, energetinis, globalus, ikiteisminis, iniciatyvinis, intelektinis, internetinis, kabelinis, kamieninis, kvalifikacinis, likutinis, mažaaukštis, mėgėjiškas, mobilus, nuotolinis, palydovinis, poįstatyminis, prioritetinis, projektinis, rinkiminis, sertifikuotas, telefoninis, toksiškas, vakarietiškas, vienmandatis.

Tikėtina, dėl tų pačių nuo seno taikomų šio Žodyno sandaros principų, kuriuos skaitmeninio žodyno statusas leistų peržiūrėti ir iš esmės atnaujinti.

4. Apibendrinamosios išvados

Šiuolaikinės kalbos technologijos, paremtos tekstynais ir jų analizės priemonėmis, leidžia rengti žodynus visai kitaip, ypač jei su jomis siejamas visas žodyno rengimo ciklas. Tačiau ir anksčiau rengtus tradicinius žodynus galima atnaujinti pasitelkus itin didelius tekstynus ir specialias platformas, leidžiančias palyginti suskaitmenintus tradicinius žodynus ir šiuolaikinę kalbą. Lyginimo rezultatai gali padėti atnaujinti antraštinių žodžių sąrašus, patobulinti paieškos sistemas. Be to, tekstynai ir išsamiai juose tiriama leksinių vienetų vartosena leidžia atpažinti leksikalizuotus kalbos vienetus ir juos tinkamai aprašyti. Svarbiausia yra sistemiškai, o ne epizodiškai, remtis vartosena, ypač naujausiais tekstais. Čia pristatomas metodas taikytas *Dabartinės lietuvių kalbos žodynui*, tačiau jis gali būti pravartus ir kitiems lietuvių kalbos žodynams. Šis metodas, paremtas universalia *Hunspell* platforma, gali būti taikomas ir kitų kalbų tradicinių žodynų antraštinių žodžių sąrašui atnaujinti.

Šaltiniai

Dabartinės lietuvių kalbos žodynas. Keinys, S. 6-asis (3-asis elektroninis) leidimas. Vilnius: Lietuvių kalbos institutas, 2006.

Literatūros sąrašas

Ambrazas, V. (red.) 2006. *Dabartinės lietuvių kalbos gramatika*. Mokslo ir enciklopedijų leidybos institutas.

Dadurkevičius, V. 2017. Lietuvių kalbos morfologija atvirojo kodo “Hunspell” platformoje. *Bendrinė kalba* 90, 1-15.

Dadurkevičius, V., Petrauskaitė R. 2020. Corpus based methods for assessment of the traditional dictionaries. In *Human Language Technologies – The Baltic Perspective. Proceedings of the Ninth International Conference Baltic HLT 2020*, A. Utka et al. (eds.). Amsterdam, Berlin, Washington, DC: IOS Press. 123–126.

Dadurkevičius V., 2020, *Assessment Data of the Dictionary of Modern Lithuanian versus Joint Corpora,* CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/36.

Kilgarriff, A., Rychlý, P., Smrz, P., Tugwell, D. 2004. The Sketch Engine In *Proceedings of the Eleventh Euralex Congress*, G. Williams, S. Vessier (eds). Lorient, France: UBS, 105-116.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., Rychlý, P. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In *Proceedings of the XIII Euralex Congress*, E. Bernal, J. DeCesaris (eds). Barcelona: Universitat Pompeu Fabra, 425-431.

Kilgarriff, A. 2012. *Using corpora [and the web] as data sources for dictionaries*. Prieiga internetu https://www.sketchengine.eu/wp-content/uploads/Using_corpora_2012.pdf (žiūrėta 2021-03-15).

Sinclair, J. 2000. Lexical Grammar. *Darbai ir dienos* 24, 191–203.