

PAIEŠKOS SISTEMOS *GOOGLE* NAUDOJIMO GALIMYBĖS TIRIANT KALBOS REIŠKINIUS

Lietuvių kalbininkai vis dažniau straipsniuose nurodo, kad renkant tam tikrus duomenis – nuo vartosenos pavyzdžių iki skaitinių dažnumo išraiškų – buvo naudotasi kokia nors interneto paieškos sistema. Dažniausiai minima didžiausia *Google* paieškos sistema, tačiau, deja, nepakankamai pritaikyta kalboms tirti. Todėl gana keista, kad paprastai apsiribojama tik nuoroda į *Google* kaip į vieną iš šaltinių, bet nenurodomas paieškos metodas ir nesigilinama į tam tikrus paieškos sistemos ir kalbos ypatumus, kurie gali palengvinti arba apsunkinti paiešką, patikslinti arba iškreipti rezultatus. Paieškos metodas svarbus tiek renkant vartosenos pavyzdžius, tiek tiriant įvairių kalbos reiškinių paplitimą, kur svarbios skaitinės išraiškos.

Šiame straipsnyje aptariami paieškos eilutės sudarymo principai, atsižvelgiant į paieškos sistemos ir lietuvių kalbos ypatumus, lemiančius paieškos sėkmę. Straipsnyje remiamasi 2007–2008 m. atliktu santykinio naujųjų svetimžodžių (toliau – NS) dažnumo tyrimu, kurio pagrindinis tikslas – nustatyti ribas tarp retais, vidutiniškais ir dažnais laikytinų NS. Darbas rėmėsi prielaida, kad atliekant tyrimą nustatyti dažnumo atskaitos taškai vėliau pravers vertinant ir kitų NS paplitimą vartosenoje. Tai galėtų būti naudinga ir kalbos norminimo veikloje, taigi straipsnio tikslas – aprašyti paieškos eilutės sudarymo būdą.

Atlikto tyrimo medžiaga – 2005–2006 m. spaudoje ir dokumentuose bei 2007 m. internete rasti NS¹ – iš viso 201. Sąrašas labai įvairus: jame 33 veiksmazodžiai, 168 vardazodžiai, tarp kurių pasitaikė 25 trūktingos paradigmos (20 vienaskaitinių ir 5 daugiskaitiniai daiktavardžiai), yra ir anglų, ir kitų

¹ Sąrašas sudarytas pagal Valstybinės kalbos norminimo, vartojimo, ugdymo ir sklaidos 2006–2015 m. programą naudojantis Lietuvių kalbos instituto Kalbos kultūros skyriuje vykdomo projekto „Naujųjų lietuvių kalbos svetimžodžių vartosenos tyrimas: 1998–2008“ karnoteka.

kalbų kilmės NS. Pastebėtina, kad kai kurie žodžiai, tirti kaip NS, yra įtraukti į *Tarptautinių žodžių žodyną* (TŽŽ 2001) ir gali būti laikomi tarptautiniais žodžiais.

Internetas yra itin heterogeniška ir nuolat kintanti erdvė. Atliekant tyrimą interneto tinklalapiai nebuvo skirstomi nei pagal temas, nei pagal stilius (nors *Google* paieškos sistemoje įmanoma ir tai), nes norėta susidaryti bendrą NS paplitimo visuotinėje vartosenoje vaizdą. Dėl specifinių internetinio diskurso ypatybių šiame straipsnyje kalbama kiek apibendrintai, pabrėžiamas sąlygiškumas, nemaža dėmesio skiriama galimoms paklaidoms ir daug kur vadovaujamasi kalbine nuojauta.

Pirmoje straipsnio dalyje pateikiami pagrindiniai paieškos eilutės sudarymo principai, tiek bendrieji, tiek būdingi konkrečiam aprašomam tyrimui. Antroje – nurodomi paieškos lauko siaurinimo būdai ir priežastys, kodėl tai reikalinga.

1. PAGRINDINIAI PAIEŠKOS EILUTĖS SUDARYMO PRINCIPAI

Paprastai *Google* paieška atliekama į paieškos laukelį įrašant pamatinę rūpimo žodžio formą. Tokio paieškos būdo iš dalies pakanka renkant vartosenos pavyzdžius, tačiau tikėtina, kad nemaža jų dalis gali likti nepastebėta, o tai dar mažiau paranku tiriant žodžių dažnumą, kur svarbios skaitinės išraiškos.

Problema kyla dėl to, kad *Google* paieškos sistema iš dalies supranta, jog įrašytas žodis gali turėti kitų formų ir įtraukia jas į rezultatus. Tačiau, visų pirma, paieškos algoritmas paprastam naudotojui nėra visiškai aiškus, neaišku, kiek ir kokių žodžio formų pakliūs tarp rezultatų. Antra, paieškos sistema neatpažįsta kalbos dalių. Tarkim, įrašius veiksmažodžio bendratį, tarp rezultatų pasitaiko ir tos pačios šaknies daiktavardžių, pavyzdžiui, įrašius žodį *autorizuoti*, sistema pateiks ir tokių tekstų, kuriuose pavartotas žodis *autorizacija*.

Dėl pastarųjų dviejų priežasčių parankiau atlikti kiek sudėtingesnę, tačiau kalbos tyrimams naudingesnę paiešką. Atliekant įvairius bandymus nustatyta, kad tikslingiausia įrašyti visą žodžio kaitybos paradigmą ar didžiąją jos dalį. Paieškos eilutę riboja tai, kad *Google* paieškos laukelyje daugiausia telpa 32 reikšminiai žodžiai.

Patartina žodžio kaitybės formas išskirti kabutėmis, nes tik taip pažymėtas tekstas *Google* sistemoje suprantamas kaip tiksli citata. Visos žodžio kaitybės formos jungiamos loginiu jungtuku OR (angl. *arba*). Šiuo būdu sudaryta paieškos eilutė atrodo taip:

„akreditavimas“ OR „akreditavimo“ OR „akreditavimui“ OR „akreditavimą“ OR „akreditavimu“ OR „akreditavime“ OR „akreditavimai“ OR „akreditavimų“ OR „akreditavimams“ OR „akreditavimus“ OR „akreditavimais“ OR „akreditavimuose“

Minėtas santykinio NS dažnumo tyrimas buvo atliktas naudojantis *Google* paieškos sistemos filtru „puslapiuose lietuvių kalba“, dėl to jau yra tekę susilaukti kritikos. Motyvuota tuo, kad netenkama maždaug dviejų trečdalių informacijos, t. y. maždaug tiek vartosenos pavyzdžių, nes ne visų interneto svetainių savininkai nurodo kalbinius svetainių metaduomenis. Tačiau santykinio žodžių dažnumo tyrimams to visiškai pakanka. Be to, nesinaudojant šiuo filtru yra itin sunku atsiriboti nuo paieškos sistemos pateikiamų tekstų anglų kalba, jie kliudo renkant vartosenos pavyzdžius. Todėl manytume, kad paieška tinklalapiuose lietuvių kalba yra tinkamesnė, tik svarbu, kad šio apribojimo visur būtų laikomasi nuosekliai.

Skrybės ženklai

Google paieškos sistema vienodai reaguoja į bet kokius skrybės ženklus, net jei jie pateikiami kabutėse ir turėtų būti suprantami kaip tiksli citata. Vadinasi, paieškos laukelyje įrašytas „account‘as“ bus įvertintas kaip du atskiri žodžiai ir suplaktas su „account:as“, „account as“ ar su kitais panašiais atvejais. Ši paieškos sistemos ypatybė svarbi dviem atvejais. Pirma, kai NS sudaro du žodžiai, sistema neleidžia atskirti tokių rašybos variantų, kaip „ex post“ ir „ex-post“, tai galėtų būti svarbu vertinant brūkšnelio vartojimą. Šiais atvejais dažnumo rezultatai suplakami į vieną, tačiau lyginant bendrą žodžių, o ne atskirų jų rašybos variantų dažnumą paieškoje nėra būtina atskirti tokių formų vieną nuo kitos. Antra, kai tiriami NS rašybos variantai su apostrofu, susiduriama su kiek sudėtingesnėmis problemomis, nes kai kurios lietuviškos galūnės sutampa su:

- a) kai kuriais anglų kalbos žodžiais (pvz.: *as* – „kaip“, *is* – „yra“, *us* – „mus“);
- b) kai kuriais lietuvių kalbos žodžiais (jungtuku *o*, prielinksniu *į*, jaustukais *ei*, *ai*);
- c) be diakritinių ženklų („šveplai“) užrašytais lietuvių kalbos žodžiais (*is* – „iš“, *as* – „aš“).

Atliekant paiešką pastaruosius netikslumus kontroliuoti yra labai sunku, o ortografinio NS galūnių sutapimo su lietuviškais žodžiais net neįmanoma. Vis dėlto atvejai, kai po NS iš karto eina kuris nors minėtų lietuviškų žodžių, sutampančių su kaitybos formanto rašytine išraiška, nėra tokie dažni, kad ryškiai paveiktų tyrimų rezultatus.

Paieškos eilutę sudarančių kaitybos formų parinkimas

Vardažodžiai

Tiriant vardažodinių NS ar jų rašybos variantų dažnumą, į paieškos eilutę daugeliu atvejų buvo įrašoma visa linksnio ir skaičiaus kaitybos paradigma. Išimtį sudarė žodžiai su trūktinomis kaitybos paradigmomis (20 vienaskaitinių ir 5 daugiskaitiniai daiktavardžiai) ir daugiareikšmiškumo atvejai (plačiau nagrinėjami antroje straipsnio dalyje). Taip pat dėl ortografinio sutapimo su kitomis kaitybos formomis ir mažo paplitimo neverta atskirai įtraukti šauksmininko. Taigi, taip galima apimti beveik visą vardažodžių kaitybos paradigmą.

Kalbant apie vardažodinių NS kaitymą, dar reikėtų aptarti problemišką giminės klausimą. Vienur moteriškąją giminę galima būtų laikyti NS rašybos variantu, pavyzdžiui, *googlė / gūglė / google'ė / googl'ė; puzzlė / puzzlė / pūzlė; skypė / skaipė*. Šiems NS bendra, kad jie žymi negyvus objektus ir todėl lietuvių kalboje giminėmis nekaitomi, ir tai, kad jų autentiškų formų pabaigoje yra nors ir netiriamas, tačiau rašte žymimas baigmuo *-e*, plg. *google, puzzle, skype*. Tiriant pastebėta, kad moteriškosios giminės formos tokiais atvejais yra gerokai retesnės nei vyriškosios giminės, tai galima aiškinti optimalumo teorijos prielaida, kad „esant prieštaravimui, renkiasi mažiausiai žymėta kategorija“ (Vaičekauskienė 2007: 177).

Būdvardžių ir gyvų objektus žyminčių daiktavardžių vyriškoji ir moteriškoji giminė taip pat nebuvo laikoma atskiromis lemomis, nors ir esama tam tikro reikšmės skirtumo. Loreta Vaičekauskienė, išsamiai nagrinėjusi NS lietuvių kalboje, pažymi, kad „[t]ais atvejais, kai daiktavardžiu galima pavadinti abiejų lyčių asmenis, atrodo, iš pradžių apibendrinama vyriškoji giminė, iš kurios prireikus kartais pasidaroma moteriškosios giminės forma“ (Vaičekauskienė 2007: 175). Vyriškosios ir moteriškosios giminės formos taip pat nebuvo laikomos atskiromis lemomis rengiant *Dažninį dabartinės*

rašomosios lietuvių kalbos žodyną (DDRLKŽ 1998: XIV–XV, XVI). Sudarant paieškos eilutes reikia į tai atsižvelgti, nes dirbant su paieškos sistema *Google* neįmanoma atskirti sutampančių vyriškosios ir moteriškosios giminės kaitybos formų.

Veiksmazodžiai

Veiksmazodžių atvejis sudėtingesnis, nes veiksmazodžio gramatinių kategorijų skaičius kur kas didesnis. Kadangi, kaip jau minėta anksčiau, *Google* paieškos eilutės ilgis yra ribotas, reikia kelti klausimą, kurios veiksmazodžio formos yra reprezentatyviausios dažnumo atžvilgiu ir turėtų būti įtrauktos į paieškos eilutę.

Savaime suprantama, kad, norint apimti didesnę veiksmazodžių distribuciją, į paieškos eilutę reikia įrašyti dažniausiai vartojamas formas. Tam reikia turėti duomenų apie gramatinių kategorijų dažnumą lietuvių kalboje, pagedautina, elektroniniame diskurse (ED). Iki šiol plačiau ED, kaip atskirą kalbos atmainą šalia rašytinės ir sakytinės kalbos, nagrinėjo Asta Ryklienė, bet jos pasirinkta tyrimo metodika neteikia reikalingų duomenų apie gramatinių kategorijų pasiskirstymą ED. Šios autorės darbo tikslas buvo palyginti ED su rašytine ir sakytine kalba (Ryklienė 2001: 3), o kalbos dalių bei žodžių formų kiekybiniai skirtumai rašytinėje kalboje ir ED nagrinėti sugretinus penkiasdešimties dažniausiai šiose kalbos atmainose vartojamų žodžių bei žodžių formų sąvadás (Ryklienė 2001: 9).

Šiame straipsnyje remiamasi Vidos Žilinskienės ir Erikos Rimkutės žodžių formų vartojimo lietuvių kalboje tyrimais. Tiesa, šių autorių darbai skiriasi tuo, kad V. Žilinskienė, kitaip nei E. Rimkutė, tirdama žodžių formų dažnumą lygina du funkcinis stilius – publicistinį ir dalykinį. Laikydamosi nuomonės, kad ED kalba yra artimesnė publicistiniam nei dalykiniam stiliui, šio straipsnio autorė remiasi publicistinio stiliaus gramatinių formų dažnumo statistika. V. Žilinskienės ir E. Rimkutės duomenys labai panašūs – tiesioginė veiksmazodžio nuosaka apima 92,12 proc. (Žilinskienė 2002: 110), 91,5 proc. (Rimkutė 2006: 35) visos veiksmazodžio distribucijos; asmenuojamosios veiksmazodžio formos – 55,48 proc. (Žilinskienė 2002: 110), 50,7 proc. (Rimkutė 2006: 35); bendratis – 15,85 proc. (Žilinskienė 2002: 111). Tiesa, E. Rimkutė (2006: 35) kiek kitokia forma nurodo bendratis paplitimą – jos teigimu, bendratis apima 33,6 proc. neasmenuojamųjų veiksmazodžio formų, kurių bendras paplitimas yra 49,3 proc.

Atsižvelgiant į pastaruosius duomenis, į paieškos eilutę parankiausia įtraukti visas tiesioginės nuosakos asmenuojamąsias veiksmažodžio formas, bendratį ir neveikiamosios rūšies dalyvio vyriškosios giminės daugiskaitos vardininką (palyginti ortografinį sutapimą su bendratimi). Visos 26 gramatinės formos (iš viso 21 reikšminis paieškos žodis, nes sutampa ir visų laikų trečiasis asmuo) sudaro kiek daugiau nei pusę visos veiksmažodžio distribucijos. Paieškos eilutės pavyzdys:

„baninti“ OR „baninu“ OR „banini“ OR „banina“ OR „baniname“ OR „baninate“ OR „baninau“ OR „baninai“ OR „banino“ OR „baninome“ OR „baninote“ OR „baninsiu“ OR „baninsi“ OR „banins“ OR „baninsime“ OR „baninsite“ OR „banindavau“ OR „banindavai“ OR „banindavo“ OR „banindavome“ OR „banindavote“

Kaip jau minėta anksčiau, paieškos sistema *Google* nekreipia dėmesio į skyrybos ženklus ir rašybos variantus su apostrofu laiko dviem atskirais žodžiais. Kadangi *Google* paieškos eilutės ilgis yra ribotas, tiriant veiksmažodžių rašybos variantus su apostrofu į paieškos eilutę galima neįtraukti būtojo dažninio laiko kaitybos formų, nes būtasis dažninis laikas tiek V. Žilinskienės, tiek E. Rimkutės duomenimis yra rečiausias iš visų veiksmažodžio laikų.

Žinoma, esant poreikiui palyginti skirtingų to paties žodžio rašybos variantų dažnumą, galima būtų atlikti platesnę paiešką, t. y. sudaryti keletą paieškos eilučių. Tačiau lyginant žodžių, o ne jų kaitybos formų dažnumą, aptartas metodas, manytina, yra pakankamas. Tik svarbu visur laikytis nuoseklumo ir tuos pačius paieškos kriterijus taikyti visiems veiksmažodžiams.

Iš to, kas pasakyta šiame skyrelyje, aišku, kad paieškoje įmanoma aprėpti daugiau ar mažiau visą vardažodžių kaitybos paradigmą, o veiksmažodžių – tik kiek daugiau nei pusę kaitybos paradigmos, todėl netenkama svaraus lyginamojo pagrindo. Žinoma, būtų galima apskaičiuoti tam tikrą veiksmažodžių ir vardažodžių dažnumo santykio koeficientą, tačiau tam reikia atlikti išsamesnius ED tyrimus. Kol kas lyginti atskirų veiksmažodžių ir vardažodžių dažnumą būtų netikslinga. Taigi, skirtinga vardažodžių ir veiksmažodžių kaitybos formų imtis lemia tai, kad dažnumo tyrimuose parankiau juos skirti į dvi grupes.

Rašybos variantai

Tiriant vartoseną interneto paieškos būdu atliekamas atvirkštinis veiksmas nei tekstynų metodu, nes sudarant paieškos eilutę reikia tarsi iš anksto nuspėti, kokius rašybos variantus gali turėti kiekvienas žodis.

Atliekant kalbos reiškinių paiešką internete, reikėtų vengti anglų kalbos kilmės autentiškos rašybos variantų, o atsitiktinai pasitaikiusius paieškos eilutėse – šalinti. Minėto tyrimo metu pastebėta, kad to nepadarius beveik neįmanoma taip susiaurinti paieškos lauko, kad tarp rezultatų nepatektų anglų kalbos intarpų ar frazių, laikytinų citatomis. Pavyzdžiui, atliekant žodžių *baby* ar *crazy* paiešką gaunama labai daug tekstų, kuriuose minimi dainų pavadinimai: *Baby one more time*, *Crazy* ir pan. Panašus ir *funky* atvejis, nes šiuo NS įvardijamas net tik muzikos stilius, bet ir seniau labai populiarūs lietuvių muzikos grupė. Vis dėlto vertinant santykinę tokių žodžių dažnumą į šiuos rašybos variantus reikia atsižvelgti, nes vartosenoje jie gana dažni.

Tiriami tokie rašybos variantai, kurie vadovaujantis kalbine nuojauta atrodo įmanomi. Prie ortografiškai adaptuojamo svetimžodžio pridedamos darybiškai įmanomos kaitybos galūnės. Kai NS ortografiškai neadaptuojamas, kaitybos galūnės pridedamos tiesiog prie nekaitomos NS formos. Ypatingesni atvejai, kai nekaitoma NS forma baigiasi balse, nes, tiriant NS rašybos variantus be apostrofo, ji numetama (pasirodė, kad *smileas*, *skypeas*, *googleas* ir į juos panašūs rašybos variantai vartosenoje yra reti).

Tiriant iki šiol dar nenormintus kalbos reiškinius dažnai gali kilti klausimas, kur riba tarp rašybos klaidos ir rašybos varianto. Šią ribą lemia dažnumo charakteristika. Pavyzdžiui, minėtame NS tyrime 10 pavartojimo atvejų ribos nesiekiantys NS rašybos variantai neįtraukti skaičiuojant bendrą atskiros NS dažnumą ir laikytini rašybos klaidomis – pavieniais apsirikimais, nerodančiais bendrųjų vartosenos polinkių. Tiesa, šis principas taikytas tik 1000 pavartojimo atvejų ribą viršijusiems žodžiams, retesniems – ne. Vis dėlto kartais gali būti reikšminga ir tai, kad tiriamo kalbos reiškinių internete visai nebuvo rasta arba jis nesiekė 10 pavartojimo atvejų ribos.

Šiais principais sudarytų ir atrinktų rašybos variantų dažnumą žyminčios skaitinės išraiškos sumuojamos ir gaunamas apytikris žodžio dažnumas. Toliau pateikiama lentelė, kurioje paieškos rezultatų apdorojimas matyti aiškiau.

1 LENTELĖ. Duomenų kaupimo būdas

NS rašybos variantas	Paieškos eilutė atskiram rašybos variantui	NS rašybos varianto dažnumas
<i>puzlas</i>	„puzlas“ OR „puzlo“ OR „puzlui“ OR „puzlą“ OR „puzlu“ OR „puzle“ OR „puzlai“ OR „puzlų“ OR „puzlams“ OR „puzlus“ OR „puzlais“ OR „puzluose“	6430
<i>puzlė</i>	„puzlė“ OR „puzlės“ OR „puzlei“ OR „puzlę“ OR „puzlėje“ OR „puzlėms“ OR „puzles“ OR „puzlėmis“ OR „puzlėse“	3210
<i>puzlis</i>	„puzlis“ OR „puzlio“ OR „puzliui“ OR „puzlį“ OR „puzliu“ OR „puzlyje“ OR „puzliai“ OR „puzlių“ OR „puzliams“ OR „puzlius“ OR „puzliais“ OR „puzliuose“	711
<i>puzzlė</i>	„puzzlė OR „puzzlės“ OR „puzzlei“ OR „puzzlę“ OR „puzzlė- je“ OR „puzzlėms“ OR „puzzlėmis“ OR „puzzlėse“	458
<i>puzzlis</i>	„puzzlis“ OR „puzzlio“ OR „puzzliui“ OR „puzzlį“ OR „puzzliu“ OR „puzzlyje“ OR „puzzliai“ OR „puzzlių“ OR „puzzliams“ OR „puzzlius“ OR „puzzliais“ OR „puzzliuose“	315
<i>pūzlė</i>	„pūzlė“ OR „pūzlės“ OR „pūzlei“ OR „pūzlę“ OR „pūzlėje“ OR „pūzlėms“ OR „pūzles“ OR „pūzlėmis“ OR „pūzlėse“	239
<i>puzzlas</i>	„puzzlas“ OR „puzzlo“ OR „puzzlui“ OR „puzzlą“ OR „puzz- lu“ OR „puzzlai“ OR „puzzlų“ OR „puzzlams“ OR „puzzlus“ OR „puzzlais“ OR „puzzluose“	167
<i>pūzlis</i>	„pūzlis“ OR „pūzlio“ OR „pūzliui“ OR „pūzlį“ OR „pūzliu“ OR „pūzlyje“ OR „pūzliai“ OR „pūzlių“ OR „pūzliams“ OR „pūzlius“ OR „pūzliais“ OR „pūzliuose“	101
<i>puzzle'as</i>	„puzzle'as“ OR „puzzle'o“ OR „puzzle'ui“ OR „puzzle'ą“ OR „puzzle'u“ OR „puzzle'e“ OR „puzzle'ai“ OR „puzzle'ų“ OR „puzzle'ams“ OR „puzzle'us“ OR „puzzle'ais“ OR „puzzle'uose“ -are	80
		11 711

Kiekvienam tirtam NS buvo sudaroma panaši lentelė. Pirmoje skiltyje išvardijami visi tirti NS rašybos variantai, surikiuoti pagal dažnumą, nes

straipsnyje atsiribojama nuo svarstymų, kuris NS rašybos variantas galėtų būti laikomas antraštiniu žodžiu (lema). Antroje skiltyje pateikiamos kiekvieno rašybos varianto paieškos eilutės jau susiaurinus paieškos lauką, t. y. pašalinus visas daugiareikšmes formas (žr. 2 skyrių). Trečioje skiltyje nurodomas kiekvieno rašybos varianto dažnumas (šie duomenys tik tarpiniai), kuris vėliau sudedamas ir gaunamas apytikris tiriamo žodžio dažnumas. Tik pastarieji duomenys yra reikšmingi nustatant santykinio NS dažnumo ribas (lentelėje pateikti duomenys rinkti 2008 m. kovo – balandžio mėn.).

2. PAIEŠKOS LAUKO SIAURINIMAS

Dėl rezultatų anglų kalba pertekliaus

Problemą dėl anglišku homografinių (sutampančios rašybos) žodžių ir rašybos variantų su apostrofu iš dalies galima išspręsti naudojantis logine funkcija „išskyrus“, kuri *Google* sistemoje žymima ženklu „-“. Tai reiškia, kad paieška nebus atliekama tinklalapiuose, kuriuose pasitaikys po šio ženklo einantis žodis. Todėl paieškos eilutės pabaigoje nurodžius vieną dažniausių anglų kalbos žodžių ir pažymėjus jį minėtu ženklu, galima labai sumažinti anglų kalbos atvejų skaičių.

Minėtame tyrime visur buvo taikomas apribojimas vienu iš dažniausiai anglų kalboje vartojamų žodžių *are*² (*yra* – esamojo laiko daugiskaitos trečiasis asmuo), tačiau galimas ir kitoks pasirinkimas. Reikia tik atkreipti dėmesį į tai, kad šiam tikslui anglų kalbos žodį reikia rinktis itin atidžiai, nes kai kurie dažniausi anglų kalbos žodžiai gali ortografiškai sutapti ir su kai kuriais lietuvių kalbos žodžiais (angl. *it*, *be*), kaitybos galūnėmis (angl. *is*, *as*, *a*, *I*) ar „šveplės“ rašybos atvejais (angl. *is*, *as*, *I*). Be to, nemaža dalis dažniausių anglų kalbos žodžių (*a*, *the*, *of*) atlieka tarnybinę funkciją ir neretai vartojami įvairiuose pavadinimuose, kurie gali pasitaikyti ir lietuviškuose interneto tinklalapiuose.

Tas pats apribojimas taikytinas ir visais atvejais, kai paieškos eilutę sudaro nekaitoma (autentiškos ar adaptuotos rašybos) NS forma, nepriklausomai nuo kalbos, iš kurios konkretus žodis galėjo pasiekti lietuvių kalbą.

² The 500 Most Commonly Used Words in the English Language: <http://www.world-english.org/english500.htm> (žiūrėta 2008 06 07).

Pastebėta, kad net atliekant ne anglų kilmės NS paiešką ir tik tinklalapiuose lietuvių kalba tarp rezultatų vis tiek pasitaiko daug pavyzdžių anglų kalba. Tai sietina su šios kalbos *lingua franca* statusu ED. Be to, kai kurie tirti NS anglų kalboje jau gali būti virtę tarptautiniais žodžiais arba tiesiog neadaptuojami.

Dėl žodžių ir jų formų daugiareikšmiškumo

Naudojantis *Google* paieškos sistema neišvengiamai tenka susidurti su daugiareikšmiais žodžiais ar jų formomis. Daugiareikšmiškumo³ terminas pasirinktas, visų pirma, norint išvengti homonimų, homografų, homoformų ir pan. terminų painiavos lietuvių kalbotyroje⁴. Antra, aiškinant pastarųjų terminų apibrėžimus, dažnai įtraukiamas fonetinis žodžio planas, visiškai neaktualus atliekant paiešką internete. Trečia, minėtų terminų neužtenka įvardyti kai kuriems daugiareikšmiškumo atvejams, kai, pavyzdžiui, savo rašyba sutampa to paties žodžio, bet skirtingų rašybos variantų kaitybos formos.

Daugiareikšmiškumas gali būti kelių lygmenų: darybinis (tas pats darybos formantas suteikia skirtingas reikšmes), tekstinis, sintaksinis semantinis, leksinis semantinis (smulkiau skirstomas į homonimus, homografus, homofonus ir polisemiją) ir morfologinis (Rimkutė 2006: 7, 40). Šiame straipsnyje yra aktualus morfologinis daugiareikšmiškumas (toliau – MD) ir iš dalies – leksinis semantinis.

Morfologiškai daugiareikšmės to paties žodžio formos

To paties žodžio formų morfologinis daugiareikšmiškumas yra sistemingas reiškinys, nes sietinas su lietuvių kalbos morfologinės sistemos ypatybėmis ir linksnių sinkretizmu. Šiame straipsnyje išvardijami tik tie morfologinio daugiareikšmiškumo atvejai, su kuriais buvo susidurta atliekant tyrimą, ir nurodomi būdai, kaip tais atvejais reikėtų elgtis *Google* sistemoje.

³ *Daugiareikšmiškumo* terminą vartoja ir šį reiškinį savo darbuose išsamiai nagrinėja E. Rimkutė (Rimkutė 2002, 2003; Rimkutė, Grybinaitė 2004).

⁴ Plg. homonimų, homografų, homofonų, homoformų apibrėžimus bei vartojimo ypatumus šiuose šaltiniuose: Jakaitienė 1980: 42–48; 2005: 53–55, 161–168; Gaivenis, Keinys 1990: 79–80; LKE 1999: 247–249.

Ortografiškai sutampančios to paties rašybos varianto kaitybos formos:

(i)o kamieno vardažodžių vns. vard. ir įnag., pvz.:

akreditacija, alibidemija, aromaterapija, autorizacija, blogosfera / blogsfera, digitalizacija, harisa / harissa, konvergencija, looseriška / lūzeriška, medija / media, milonga, salsa, spamintoja;

(i)o kamieno vardažodžių vns. kilm. ir dgs. vard., pvz.:

akreditacijos, autorizacijos, blogosferos / blogsferos, harisos / harissos, looseriškos / lūzeriškos, medijos / medias, spamintojos;

ė kamieno vardažodžių vns. kilm. ir dgs. vard., pvz.:

baikerės, blogerės / bloger'ės / blog'erės, gliancinės, golbolininkės, karaokės, lūzerės / looserės / looser'ės / loser'ės / loserės, menedžerės / managerės, multipleksinės, neformatės, puzlės / puzzlės / pūzlės (žr. 1 lentelę), skaipės / skypės, sponsorės, šoumenės;

(i)u kamieno daiktavardžių vns. vard. ir dgs. gal., pvz.:

aspiratorius, blokatorius, epiliatorius / epiletorius;

(i)u kamieno daiktavardžių vns. gal. ir dgs. kilm., pvz.:

aspiratorių, sponsorių.

Šie MD atvejai dar gali būti vadinami *gramatiniais homonimais*, kurie lietuvių kalbotyroje apibrėžiami kaip „fonetiškai sutampančios to paties žodžio formos, turinčios skirtingas gramatines reikšmes“ (Jakaitienė 1980: 42). Šioms formoms bendra tai, kad atliekant paiešką *Google* sistemoje jos atsiduria toje pačioje paieškos eilutėje, tarpusavyje yra sujungtos loginiu jungtuku OR (arba) ir jų dažnumas nėra pakartotinai sumuojamas. Tokiais atvejais paieškos eilutę dėl ekonomiškumo vis tiek patogu trumpinti, tačiau rezultatų tai nelemia.

Kita MD rūšis yra ortografiškai sutampančios to paties žodžio, bet skirtingų rašybos variantų kaitybos formos:

vyriškosios ir moteriškosios giminės vardažodžių dgs. kilm., pvz.:

baikerių, blogerių / bloger'ių / blog'erių, gliancinių, lūzerių / looserių / looser'ių / loser'ių / loserių, looserišku / lūzerišku, menedžerių / managerių, multipleksinių, puzlių / pūzlių / puzzlių (žr. 1 lentelę), spamintojų, sponsorių;

(i)a kamieno vardažodžių vns. vard. ir (i)o kamieno vardažodžių dgs. gal., pvz.:

looseriškas / lūzeriškas, spamintojas;

(i)a kamieno daiktavardžių dgs. vard. ir (i)o kamieno daiktavardžių vns. naud., pvz.:

spamintojai;

(i)a kamieno daiktavardžių vns. viet. ir *ė* kamieno daiktavardžių vns. įnag., pvz.:

gūgle / google'e, internaute, puzzle / pūzle (žr. 1 lentelę), skaipe, šoumene;

(i)o ir (i)a kamienų vardažodžių vns. gal., pvz.:

looserišką / lūzerišką, spamintoją.

Pastarieji MD atvejai yra sudėtingesni ir reikalauja daugiau pastabumo. Ortografiškai sutampančios skirtingų to paties žodžio rašybos variantų kaitybos formos atsiduria skirtingose paieškos eilutėse, tai reiškia, kad tokių formų dažnumas, skaičiuojant bendrą atskiro žodžio dažnumą, sumuojamas pakartotinai. Todėl tiriant žodžių dažnumą šio tipo sutampančias formas reikėtų palikti tik kurioje nors vienoje paieškos eilutėje.

Verta paminėti, kad atliekant santykinio NS dažnumo tyrimą teko susidurti su dviem dar neaprašytais MD rūšimis, kai ortografiškai sutampa (i)a kamieno daiktavardžių dgs. vard. su (i)o kamieno daiktavardžių vns. naud. (plg. *tie varnai* ir *tai várnai*) ir (i)a kamieno daiktavardžių vns. viet. su *ė* kamieno daiktavardžių vns. įnag. (plg. *tame draugè*⁵ ir *su draugè*). Nors šios dvi MD rūšys aptiktos tiriant NS vartoseną, iš pateiktų pavyzdžių matyti, kad yra būdingos ir bendrinei lietuvių kalbai.

Sistemiški, tačiau už lietuvių kalbos ribų patenkantys MD atvejai – su netiriamomis nekaitomomis autentiškos rašybos NS formomis ortografiškai sutampančios morfologiškai adaptuojamų NS kaitybos formos:

(i)a kamieno daiktavardžių vns. viet., pvz.:

drive (drivas), google (googlas), puzzle (puzzlas, žr. 1 lentelę), skype (skypas);

(i)o kamieno daiktavardžių vns. vard. ir vns. įnag., pvz.:

media;

ė kamieno daiktavardžių vns. įnag., pvz.:

google (googlė), karaoke (karaokė), puzzle (puzzlė, žr. 1 lentelę), skype (skypė).

Kadangi dėl jau anksčiau minėtų priežasčių autentiškos rašybos NS variantai sunkiai tiriami, pastarojo tipo morfologiškai daugiareikšmės formos iš paieškos eilučių šalintinos visais atvejais.

⁵ Nors kalbos kultūros požiūriu ši forma nevartotina, tačiau teoriškai įmanoma.

Morfologiškai daugiareikšmės skirtingos leksinės reikšmės žodžių formos

Kartais vieno žodžio visos ar dalis kaitybos formų savo rašyba gali sutapti su kito žodžio kaitybos formomis. Akivaizdu, kad, priešingai nei anksčiau aprašyta MD atvejais, skirsis leksinė reikšmė. Be to, šie MD atvejai yra atsitiktiniai ir nepasižymi sistemingumu. Toliau išvardijami tyrimo metu rasti tokio MD tipo pavyzdžiai, nurodant NS rašybos varianto formą ir bendrinės kalbos žodį, su kuriuo ta forma sutampa:

- *blogas*, *blogas* (būdv.), *blogai* (prieiv.) – paieškos eilutė sutrumpinta iki „blogui“ OR „bloge“ OR „blogams“ (nurodytos tik nuo bendrinės kalbos žodžio besiskiriančios formos – vns. naud., vns. viet., dgs. naud.);
- *smailas*, *smailas* (būdv.), *smailai* (prieiv.) – paieškos eilutė trumpinama kaip ir *bloga* atveju: „smailui“ OR „smaile“ OR „smailams“;
- *bodis*, *bodis* (būdv.), *bodžiai* (prieiv.) – paieškos eilutė trumpinama iki „bodis“ OR „bodžio“ OR „bodžiui“ OR „bodį“ OR „bodyje“ OR „bodžiams“ (nuo bendrinės kalbos žodžio besiskiriančios formos – vns. vard., vns. kilm., vns. naud., vns. gal., vns. viet., dgs. naud.);
- *brendas*, *brendo* (bręsti, bręsta, brendo), *brendu* (bristi, brenda, brido);
- pasitaikė ir tokių atvejų, kai tiriamas NS rašybos variantas visiškai sutapo su bendrinės kalbos žodžiu – plg. *bòtas*, *bòtas* (aukštas moteriškas kaliošas), *bòtas* (toks laivas) bei *sáitas* ir *saĩtas*.

Kaip elgtis tokiais homonimijos atvejais, padeda nuspręsti VDU tekstyno ar DDRKŽ teikiami duomenys apie žodžių paplitimą. Jei sutampantis bendrinės kalbos žodis yra retas, paieškos eilutės galima ir netrumpinti, nes paieškos rezultatams tai neturi didelės įtakos. Be to, kartais pakanka ir vieno rašybos varianto rezultatų laikyti tiriamą žodį dažnu.

Dėl kitų priežasčių

Pasitaiko atvejų, kai tiriamas žodis gali būti pavartotas kokiame nors plačiai vartojamame įmonės, įstaigos ar kitokiame pavadinime arba pasirinktas kaip slapyvardis. Interneto pavadinimų paprastai apstu, tai labiau sietina su reklamos strategijomis ar diskusijų forumų struktūros principais, nei su tiems pavadinimams sudaryti pavartotų žodžių vartosenos įpročiais. Todėl atliekant paiešką nuo tokių atvejų stengiamasi atsiriboti.

Paieškos laukas gali būti siaurinamas dviem būdais. Pirmą – pagal kontekstą, kai paieškos eilutėje nurodoma neįtraukti tokio žodžio, kuris pagal vartosenos pavyzdžius tame pačiame kontekste dažnai vartojamas su tiriamuoju kalbos reiškiniu, tačiau su juo nesudaro žodžių junginio. Antra – pagal kolokaciją⁶, kai paieškos eilutėje nurodoma neįtraukti tokio žodžio, kuris su tiriamuoju reiškiniu sudaro vartosenoje dažną žodžių junginį. Atliktame tyrime pasitaikę ypatingesni paieškos lauko siaurinimo atvejai pateikiami 2 lentelėje, nurodant žodžius, kuriais pasinaudota, bei NS rašybos variantus ir formas, kuriems tas siaurinimas buvo taikomas (kur reikia, nurodytos ir kolokacijos).

2 LENTELĖ. Paieškos lauko siaurinimas pagal kontekstą ir kolokaciją

Pagal kontekstą	-nikas -autorius -vartotojas -narys -blogas -uab -biuro -sklypas	<i>Harissa / harisa, googlis, sasha funke, esesas, twisteris, neformatas, mixas, extremas</i> <i>topas, saitas, respectas, funkas, autletas, bugo</i> <i>grafitas</i> <i>skypas</i>
Pagal kolokaciją	-la -centras -eesti -area -baltijos +kava -vakarų	„La speme“ (<i>spemas</i>) „Topo centras“ (<i>topas</i>) „Eesti post“ (<i>post'as</i>) „Cluster area“ (<i>cluster'is</i>) „Baltijos autletas“ (<i>autletas</i>) <i>Ekspreso kava, Vakarų ekspresas</i>

Daugiausia sunkumų kyla dėl tų žodžių, kurie eina slapyvardžiu (pirmoji ir antroji lentelės eilutės). Atsižvelgiant į paieškos rezultatus, tenka rinktis vieną ar kelis nurodytus apribojimo žodžius. Lentelėje išvardyti tik veiksmingiausi apribojimo žodžiai, tačiau galutinį pasirinkimą lemia kiekvieno tiriamo atvejo specifika.

⁶ *Kolokacija* – labai įprastas žodžio vartojimo kontekstas, prasmingas ir gramatiškas, dažnai kartu vartojamų mažiausiai dviejų žodžių junginys (Jakaitienė 2005: 25).

Šiame kontekste labai įdomus ir kiek neįprastas *skaipto* atvejis. Tiriant šio NS rašybos variantą *skypas* teko siaurinti paieškos lauką, nes internete buvo rasta nemažai skelbimų apie parduodamus žemės sklypus, parašytų su klaida (praleista raidė *l*).

IŠVADOS

1. Sudarant paieškos eilutes visų pirma svarbi reprezentatyvių žodžių formų atranka, kuri atliekama remiantis lietuvių kalbos gramatinių kategorijų dažnumo tyrimų rezultatais. Tiriant vardažodžius į paieškos eilutę telpa beveik visa kaitybos paradigma, veiksmazodžių – tik dalis. Parankiausia į paieškos eilutę įrašyti veiksmazodžio bendratį ir tiesioginės nuosakos asmenuojamąsias veiksmazodžio laikų (esamojo, būtojo kartinio, būsimosio ir būtojo dažninio) formas. Taip galima aprėpti kiek daugiau nei pusę veiksmazodžių gramatinių kategorijų distribucijos.

2. Dėl skirtingos gramatinių kategorijų distribucijos aprėpties, kurią lemia ribotas *Google* paieškos eilutės ilgis, netenkama svaraus veiksmazodžių ir vardažodžių lyginimo pagrindo, todėl parankiau yra atskirti vardažodžių ir veiksmazodžių grupes.

3. Neretai paieškos lauką tenka siaurinti dėl rezultatų anglų kalba pertekliaus ir daugiareikšmiškumo problemos lietuvių kalboje. Tai daroma naudojantis logine funkcija „išskyrus“, kuri *Google* paieškos sistemos sintaksėje žymima ženklu „-“, arba į paieškos eilutę neįtraukiant tam tikrų daugiareikšmių formų.

4. Atliekant paiešką atkreiptinas dėmesys į skirtingą poveikį paieškos rezultatams turinčius daugiareikšmiškumo atvejus, kuriems atitinkamai taikytinas skirtingas paieškos lauko apribojimas. Ortografiškai sutampančias to paties rašybos varianto kaitybos formas iš paieškos eilutės galima šalinti, tačiau rezultatams tai įtakos neturi. Sutampančias to paties žodžio, bet skirtingų rašybos variantų kaitybos formas patartina palikti tik kurio nors vieno rašybos varianto paieškos eilutėje. Morfologiškai daugiareikšmes skirtingos leksinės reikšmės žodžių formas reikėtų palikti arba šalinti, atsižvelgiant į VDU teksto ir DDRLKŽ duomenis.

5. Aptartas paieškos metodas pravartus tiek renkant įvairius kalbos reiškinių pavyzdžius, tiek tiriant jų paplitimą vartosenoje. Be to, santykinio NS dažnumo nustatymo metodas gali būti taikomas ir kalbos norminimo veikloje, pavyzdžiui, sudarant didžiosiomis kalbos klaidomis laikomų NS sąrašus.

ŠALTINIAI

Paieškos sistema *www.google.lt*

The 500 Most Commonly Used Words in the English Language: *http://www.world-english.org/english500.htm*

TŽŽ 2001 – *Tarptautinių žodžių žodynas*, Vilnius: Alma littera.

VDU Dabartinės lietuvių kalbos tekstynas: *http://donelaitis.vdu.lt*

LITERATŪRA

DDRŁKŽ 1998: L. Grumadienė, V. Žilinskienė. *Dažninis dabartinės rašomosios lietuvių kalbos žodynas*, Vilnius: Lietuvių kalbos institutas.

Gaivenis K., Keinys S. 1990: *Kalbotyros terminų žodynas*, Kaunas: Šviesa.

Jakaitienė E. 1980: *Lietuvių kalbos leksikologija*, Vilnius: Mokslo.

Jakaitienė E. 2005: *Leksikografija*, Vilnius: Mokslo ir enciklopedijų leidybos institutas.

LKE 1999: *Lietuvių kalbos enciklopedija*, Vilnius: Mokslo ir enciklopedijų leidybos institutas.

Rimkutė E. 2002: Homoformos dabartinės lietuvių kalbos tekстыne. – *Lituanistica* 2(50), 86–101.

Rimkutė E. 2003: Morfoliginio daugiareikšmiškumo tipologija. – *Lituanistica* 4(56), 60–78.

Rimkutė E. 2006: *Morfoliginio daugiareikšmiškumo ribojimas kompiuteriniame tekстыne*. Daktaro disertacija, Kaunas.

Rimkutė E., Grybinaitė A. 2004: Dažniausios lietuvių kalbos morfoliginio daugiareikšmiškumo rūšys ir jų automatinis vienareikšminimas. – *Kalbų studijos* 5, 74–78.

Ryklienė A. 2001: *Elektroninis diskursas: kalbos ypatybės ir stilius*. Daktaro disertacijos santrauka, Vilnius.

Vaicekauskienė L. 2007: *Naujieji lietuvių kalbos svetimžodžiai. Kalbos politika ir vartosena*, Vilnius: Lietuvių kalbos institutas.

Žilinskienė V. 2002: Žodžių formų vartojimas lietuvių kalbos dalykinio ir publicistinio stilių duomenimis. – *Lituanistica* 1(49), 106–116.

Gauta 2008 07 08

POSSIBILITIES OF USING THE SEARCH SYSTEM *GOOGLE* IN LINGUISTIC RESEARCH

S u m m a r y

The paper introduces the methodology which can be used to implement *Google* as a tool for data accumulation and storage and investigating the spread of a linguistic phenomenon in the internet. The method is described in the framework of an investigation aimed at identifying relative frequency count of new loanwords between 2007 and 2008. The first section of the article describes the main principles of filling in the slot of a query: general as well as specific for a particular descriptive research. Rather than entering only the main word form, a proposal has been launched to enter the whole grammatical paradigm of the word or its major part considering that in the Google query slot the maximum number of key words is 32.

The second section identifies the ways to narrow down the search domain and reasons why it should be done. This type of research is bound to face a problem of word and word-form ambiguity. The term of ambiguity has been selected to avoid confusion of more specific terms like homonyms, homographs, homophones etc. Moreover, the explication of the above terms often emphasises the phonetic aspect of the word, which for internet search is irrelevant. The above terms are not sufficiently precise for some cases of ambiguity, when, for example, the spelling of one word form coincides with another form of the same word.

The method of search is instrumental in collecting various examples of language use as well as in investigating their frequency. Also the method of relative frequency of loanwords could be applied in language standardisation, e.g. in compiling lists of new loanwords which are considered grave mistakes.

SKAISTĖ ALEKSANDRAVIČIŪTĖ

Lietuvių kalbos institutas

P. Vileišio g. 5, LT-10308 Vilnius

skaiste.aleksandraviciute@lki.lt