

JOLANTA KOVALEVSKAITĖ, ERIKA RIMKUTĖ,

JURGITA VAIČENONIENĖ

Vytauto Didžiojo universitetas

ARBITRALIŲJŲ LIETUVIŲ KALBOS KOLOKACIJŲ NUSTATYMAS

ESMINIAI ŽODŽIAI: arbitraliosios kolokacijos, lietuvių kalba, ribotas leksinis junglumas, vektorių metodas, automatizuotas nustatymas, lingvistinis neautomatinis nustatymas.

ANOTACIJA

Straipsnyje pristatomo tyrimo objektas – lietuvių kalbos arbitraliosios kolokacijos. Jų junglumo pagrindas yra ne semantiniai, o leksiniai ribojimai, tokios kolokacijos tiesiogiai neatspindi nekalbinės tikrovės. Ryšys tarp jų žodžių – arbitralus: nors pagal panašią reikšmę teoriškai galima rinktis daugiau negu vieną junglumo partnerį, bet linkstama vartoti kurį nors vieną. Straipsnyje aprašomi arbitraliųjų kolokacijų nustatymo metodai, išbandyti vykdant projektą „Neautomatinis lingvistinės analizės ir automatizuotas kompiuterinės lingvistikos metodas (ARKA)“. Medžiaga tyrimui imta iš Lietuvių kalbos pastoviųjų žodžių junginių duomenų bazės.

ABSTRACT

The article aims to describe the methodological approaches of arbitrary collocation (AC) identification developed within the framework of the project “Arbitrary Collocations of Lithuanian: Identification, Description and Usage (ARKA)”. The source of the research was the “Database of Lithuanian Multiword Expressions”. Arbitrary collocations were determined by combining manual and semi-automatic methods of computational linguistics.

ĮVADAS

Didelę pastoviųjų žodžių junginių dalį, kaip žinoma, sudaro kolokacijos, tačiau, jas automatiškai nustatant iš tekstynų statistiniais ar hibridiniais metodais (plačiau žr. Bielinskienė ir kt. 2019: 15–22), nustatoma ir nemažai tokių junginių, kurie kolokacijoms

priskiriami dėl to, kad juos sudarančių žodžių vartojimo drauge dažnis automatiškai įvertinamas kaip statistiškai reikšmingas, pvz.: *šešeri metai, saulėta diena, graži moteris* ir pan. Taip tarp kolokacijų atsiranda netikrųjų, kurios sudaromos kaip paprasti laisvieji žodžių junginiai remiantis semantiniais, o ne leksiniais apribojimais (Marcinkevičienė 2010: 90–91).

Kai kurie tyrėjai kolokacijomis laiko tik tikrąsias kolokacijas (kitaip dar vadinamas nemotyvuotomis, arbitraliosiomis) – įprastinius žodžių junginius, kurie tiesiogiai neatspindi nekalbinės tikrovės, ryšys tarp jų žodžių – arbitralus (išsamiau apie arbitralumą žr. Kovalevskaitė ir kt. 2021).

Šio tyrimo objektas yra būtent arbitraliosios kolokacijos (toliau – AK), t. y. tokios, kurioms būdingas ribotas leksinis junglumas (angl. *lexical restrictedness*): nors teoriškai pagal panašią reikšmę galima rinktis daugiau negu vieną junglumo partnerį, linkstama vartoti kurį nors vieną, pvz., *suprasti užuominą*, o ne **suvokti užuominą, ranka pasiekiamas*, o ne **ranka paimamas*. Rūta Petrauskaitė rašo: „Tiesioginio ryšio su pavadinamąja tikrove nebuvimas lemia tai, kad ne tik skirtingose kalbose to paties žodžio junglumas smarkiai skiriasi, bet net ir toje pačioje kalboje vieno kolokacijos nario negalima pakeisti kitu, jam sinonimišku, nerizikuojant, kad pasakyta frazė skambės neįprastai, pvz., *laikas eina, bėga, slenka*, bet ne **žingsniuoja, skuodžia, šliaužia*“ (Marcinkevičienė 2010: 91). Tikrųjų (arbitraliųjų) kolokacijų pagrindinio žodžio junglumo partneriai, pasak R. Petrauskaitės, yra nenuspėjami, tokių kolokacijų junglumo pagrindas – leksiniai apribojimai (Marcinkevičienė 2010: 90).

AK iki šiol nėra sulaukusios didelio lietuvių kalbos tyrėjų dėmesio, nors jų nustatymas padėtų išskirti šias kolokacijas kaip lietuvių kalbos kolokacijų branduolį, ypač naudingą vertimo, lietuvių kalbos kaip svetimiosios mokymo, leksikografijos tikslams. Šiais tikslais AK analizuojamos vykdant projektą „Arbitraliosios lietuvių kalbos kolokacijos: atpažinimas, aprašymas, vartojimas (ARKA)“¹, kuriuo siekiama aprašyti lietuvių kalbos AK ir parengti AK leksikografinius bei mokomuosius išteklius.

Straipsnio tikslas – aprašyti, kokie metodai taikyti vykdant projektą ARKA lietuvių kalbos AK atpažinti, ir palyginti skirtingais metodais gautus rezultatus. AK analizuotos neautomatiniu lingvistinės analizės ir automatizuotu kompiuterinės lingvistikos metodais.

¹ Projektą finansuoja Lietuvos mokslo taryba (Nr. S-LIP-20-18), išsamiau žr. <https://arka.pastovu.vdu.lt>.

Šaltiniai: 1) Lietuvių kalbos pastoviųjų žodžių junginių duomenų bazė (toliau – PŽJDB)², kurioje pateikiama apie 12 tūkst. kolokacijų; 2) *Delfi.lt* tekstynas³, sudarytas iš 70 mln. žodžių, apimantis 2014–2016 m. publikuotus tekstus iš įvairių *Delfi.lt* naujienų portalo rubrikų (plačiau apie tekstyną Bielinskienė ir kt. 2019: 14–15).

Tiriamoji medžiaga – kolokacijos, sudarytos iš 97 daiktavardžių (*vieta, aplinka, būdas, vyras, moteris, dalykas* ir kt.⁴). Jų struktūra – būdvardis / būdvardiškai vartojamas dalyvis + daiktavardis, veiksmazodis + daiktavardis, daiktavardis + daiktavardis. Atlikus tyrimą, apie 2400 kolokacijų pripažintos kaip AK. Didžioji dalis jų yra dvižodės, bet dalis – trižodės; pastarosios paprastai sudarytos iš veiksmazodžių, būdvardžių ir daiktavardžių (gali skirtis dėmenų išsidėstymo tvarka).

Pirmoji tiriamosios medžiagos dalis, t. y. kolokacijos, kurių struktūra būdvardis + daiktavardis, yra aptarta ankstesniame straipsnyje (Kovalevskaitė ir kt. 2021). Jame pristatytas bandymas AK atpažinti taikant vieną iš kompiuterinės lingvistikos metodų – automatizuotą (vektorių) metodą, šis metodas taip pat išsamiai pristatytas. Šiame straipsnyje apibendrinamas visas AK atpažinimo tyrimas. Pirmame skyriuje pristatomi abu taikyti metodai – neautomatinis (žr. 1.1 sk.), kartu aprašant ir taikytus kriterijus bei testus, ir automatizuotas, jo taikymui atskleisti pasitelkiant kolokacijų grupę, kurios struktūra – veiksmazodis + daiktavardis (žr. 1.2 sk.). Antrame skyriuje, remiantis visų tyrimo etapų duomenimis, aptariami ir palyginami metodų taikymo rezultatai, jų galimybės ir trūkumai, svarbūs vertinant nustatytų AK sąrašą ir planuojant tolesnius AK tyrimus.

1. ARBITRALIŲJŲ KOLOKACIJŲ NUSTATYMO METODAI

Rašant apie AK nurodoma, kad ribą tarp arbitraliųjų ir vadinamųjų trivialiųjų kolokacijų būtų galima brėžti ties abstrakčiųjų daiktavardžių riba: „<...> kolokacijos, kurių pagrindą sudaro abstraktieji daiktavardžiai, gali būti laikomos tikrosiomis, nemotyvuotomis,

² <http://resursai.pastovu.vdu.lt/paieska/paprastoji>

³ <https://klc.vdu.lt/pastovuSearch.html>

⁴ Tai tiek konkretieji, tiek abstraktieji, dažniausiai pastoviuosius žodžių junginius, nustatytus iš *Delfi.lt* tekstyno, sudarantys daiktavardžiai, išsamiau apie juos žr. Bielinskienė ir kt. 2019.

o pastovūs junginiai su konkrečiais daiktavardžiais paprastai remiasi nekalbine tikrove ir yra trivialūs“ (Marcinkevičienė 2010: 91). Taigi, vienas iš darbuose apie kolokacijas įvardytų kriterijų yra daiktavardžių semantika. Atliekant šį tyrimą paaiškėjo, kad didžioji dalis AK išties yra sudarytos su abstraktų pavadinimais, tačiau pastebėta, kad arbitralumas būdingas ir kolokacijoms su konkrečiais daiktavardžiais, plg.: *apriboti greitį, atlikti mankštą, imti kyšį, palikti aikštę, pasiimti kreditą, pasikalbėti telefonu, pastatyti spektaklį* ir kt.

Šio tyrimo iššūkis buvo ne tik pritaikyti AK atpažinimo kriterijus, bet ir pasirinkti tokią metodiką, kuri tiktų dirbti su dideliu duomenų kiekiu, kad, remiantis ja, būtų galima sudaryti lietuvių kalbos AK sąrašus ištyrus didelį kolokacijų skaičių (tam naudota PŽJDB medžiaga). Vykdamas projektą, išbandyta ir pritaikyta dvejopa AK nustatymo metodika: neautomatinė ir automatizuota. Pirmiausia taikytas neautomatinis būdas, pasirinkus AK pripažinimo kriterijus (žr. 1.1 sk.). Kadangi PŽJDB yra daug duomenų, vienas iš projekto uždavinių buvo kartu su neautomatiniu metodu išbandyti ir tokį metodą, kuris leistų AK pripažinimo procesą bent iš dalies automatizuoti (žr. 1.2 sk.).

1.1. Neautomatinis AK nustatymas

Neautomatiškai nustatant AK, taikyti du pagrindiniai kriterijai:

- 1) riboto leksinio junglumo kriterijus, kuriam įvertinti taikytas sinonimų pakeičiamumo ir (arba) daiktavardžių leksinių semantinių grupių metodas;
- 2) perkeltinės reikšmės kriterijus.

Jei kolokacija atitiko bent vieną kriterijų, ji laikyta arbitraliąja. Toliau aptariama, kaip taikytas kiekvienas kriterijus siekiant įvertinti, ar kolokaciją laikyti arbitraliąja.

1.1.1. AK nustatymas pagal ribotą leksinį junglumą

Kaip minėta įvade, arbitraliosiomis laikomos kolokacijos, turinčios ribotą leksinį junglumą, t. y. būdvardžiui ar veiksmažodžiui, vartojamam konkrečia reikšme junginyje su daiktavardžiu, būdingas tam tikras, daugiau ar mažiau arbitralus, pakeičiamumo ribojimas (angl. *arbitrary restriction on substitutability*; Nesselhauf 2003, 2005). Taigi, remiantis riboto leksinio junglumo samprata, bet kuris kolokacijos dėmuo tam tikra reikšme negali būti

pavartotas su visais sintaksiškai ir semantiškai įmanomais kolokatais. Nors teoriškai ir galima rasti kolokatų pakaitų, bet jie arba visai netiktų, arba būtų neįprasti, pvz.: *palanki* vs. **naudinga aplinka*; *daugiasluoksnis* vs. **sluoksniuotas romanas*.

Riboto leksinio junglumo kriterijų lingvistai tikrino dviem testais:

1) pagal veiksmažodžio ar būdvardžio sinonimus (t. y. ar galima pakeisti kitais artimos reikšmės žodžiais, laikytiniais sinonimais);

2) pagal daiktavardžių leksines semantines grupes.

Kadangi AK laikomos tokios kolokacijos, kurių vieno iš dėmenų paprastai negalima pakeisti kitu žodžiu – artimos reikšmės sinonimu, AK atpažinti taikytas sinonimų pakeičiamumo testas (adaptuotas iš Nesselhauf 2003): jeigu vienas kolokacijos dėmuo negali būti pakeičiamas kitu artimos reikšmės sinonimu, tokia kolokacija laikyta AK. Pavyzdžiui, vertinant kolokacijų *nepaklusti įstatymui*, *diegti sistemą*, *įdėtas darbas* arbitralumą, tikrinta pagal galimus veiksmažodžių sinonimus **nusižengti įstatymui*, **įvesti sistemą*, **įkištas / įkeltas darbas* ir spęsta, ar pakeičiamumas galimas; jeigu ne – kolokacijos pripažintos kaip AK. Testuota trims lingvistams taikant introspekcijos metodą, bet remtasi iš *Delfi.lt* tekstyno nustatytais kolokacijomis. Vis dėlto reikia pripažinti, kad rasti sinonimą neretai buvo gana sunku, todėl dažnai taikytas ir antrasis – daiktavardžių leksinių semantinių grupių – testas.

Ribotas leksinis junglumas taip pat vertintas pagal tai, kokia yra leksinių semantinių grupių įvairovė: remiantis PŽJDB pateiktomis kolokacijomis, spęsta, su kelių leksinių semantinių grupių daiktavardžiais jungiasi kolokacijos būdvardis ar veiksmažodis, ir taip įvertinta leksinio junglumo apimtis. Kad būtų kuo tiksliau nustatyta, su kokiais žodžiais gali būti vartojamas daiktavardis, analizuota visa jo paradigma, t. y. visos kolokacijos, sudarytos su tuo daiktavardžiu. Bet, nors visą tiriamąją medžiagą sudaro apie 12 tūkst. kolokacijų, prieinamų PŽJDB, kartais būdavo matyti, kad informacijos nepakanka. Jeigu tiriamojoje medžiagoje pasitaikydavo tik viena kolokacija su tam tikru būdvardžiu / veiksmažodžiu, lingvistai paradigmą užpildydavo remdamiesi savo kaip gimtakalbių kalbos jausmu; jeigu bazėje būdavo kelios kolokacijos, riboto leksinio junglumo testai buvo taikomi turimai paradigmai.

Remiantis daiktavardžių leksinėmis semantinėmis grupėmis, junglumo apimtis buvo vertinama atsižvelgiant į tokias sąlygas:

a) kolokacijos nelaikytos AK, jeigu su tam tikru būdvardžiu / veiksmažodžiu jungiasi daug įvairių leksinių semantinių grupių daiktavardžių ir tos grupės gausios. Toks būdvardžio / veiksmažodžio junglumas laikytas plačiu, be aiškesnių leksinių apribojimų, plg. kolokacijų su *geras* gausą: *geras rezultatas, darbas, pasaulis, žmogus, laikas, gyvenimas, būdas, dalykas, šalis, komanda, diena, sprendimas, metai, oras, vieta, pusė, sezonas, galimybė, kaina, rungtynės, automobilis, maistas, ženklas, sveikata, kelias, vyras, vaikas, produktas, klausimas, pradžia, priemonė* ir kt.; dar plg. *pirkti: pirkti automobilį, maistą, produktą, turtą, žemę*;

b) jeigu su būdvardžiu / veiksmažodžiu jungiasi daug įvairių leksinių semantinių grupių daiktavardžių, bet viena grupė išsiskiria arba vartojami tik keli tam tikros semantinės grupės daiktavardžiai, kolokacija galėjo būti laikoma AK. Pavyzdžiui, nors *statyti* junglumas platus, bet šį veiksmažodį galima pavartoti ne su visų transporto priemonių pavadinimais, plg. *statyti automobilį* (kaip ir *statyti dviratį, motociklą*), bet *prišvartuoti laivą, nutupdyti lėktuvą*. Vadinasi, kolokacija *statyti automobilį* dėl riboto junglumo priskirtina AK. Veiksmažodis *turėti* laikomas plataus leksinio junglumo žodžiu, nes gali sudaryti daug ir įvairių kolokacijų, tačiau visoje jo paradigmoje išsiskiria kolokacija *turėti reikalą* – ji laikoma arbitraliąja. Su *subjurti* vartojami dviejų leksinių semantinių grupių – gamtos reiškinių grupės ir jausmų – žodžiai, bet kolokacijos sudaromos tik su daiktavardžiais *subjuro oras* ir *subjuro nuotaika*, todėl abu atvejai yra AK;

c) AK nelaikomos vienai didelei semantinei grupei priskiriamos kolokacijos. Pavyzdžiui, PŽJDB rasta tik viena kolokacija *nuskaičiuoti mokesčius*, bet lingvistai, remdamiesi gimtakalbio kalbos jausmu, gali užpildyti paradigmą kitais žodžiais (nuskaičiuoti galima pinigus, paskolą ir kitus finansinius dalykus), todėl *nuskaičiuoti mokesčius* nepripažinta kaip AK. Tai PŽJDB iš vienos kolokacijos sudaryta, bet teoriškai gausi leksinė semantinė grupė.

Ribotas leksinis junglumas tikrintas vertinant abu kolokacijos žodžius: būdvardžio / veiksmažodžio negalima pakeisti kitais sinonimais (1 testas) arba (ir) daiktavardžių paradigma ribota (2 testas). Pavyzdžiui, kolokacija *patiekiamas maistas* rodytų, kad gali būti daug maisto grupės daiktavardžių (galima patiekti maistą, gėrimus, vakarienę, pietus...). Taigi, ji neatitinka 2 testo a sąlygos, nes semantinė grupė viena, tačiau atitinka b sąlygą, nes *patiekti* negalima pakeisti į *paduoti*. Taigi, pagal 2 testą kolokacija pripažįstama arbitraliąja. Kitas pavyzdys:

veiksmazodis *surengti* vartojamas su renginius, įvykius įvardijančiais daiktavardžiais (*surengti rinkimus, koncertą...*), vadinasi, tokių kolokacijų nereikėtų laikyti arbitraliosiomis pagal semantinės grupės gausos kriterijų, bet prie AK jas galima priskirti pagal 1 testą, nes *surengti* šiuo atveju negalima pakeisti sinonimu *suruošti*.

Sinonimijos testą ne visada galima pritaikyti. Pavyzdžiui, jo negalima pritaikyti, kai yra keli AK variantai, plg.: *kainos kyla* ir *kainos auga*, *didėja*; *klausimas kamuoja* ir *klausimas kankina*; *duoti galimybę* ir *suteikti galimybę*; *įgyti teisę* ir *įgauti teisę*. Taip pat šio testo negalima pritaikyti tokiais atvejais, kaip *sumokėti mokesčius*, *apgyvendintas miestas* – žodžiams *sumokėti*, *apgyvendinti* sinonimą sudėtinga pasiūlyti net ir hipotetiškai. Jeigu nerandama sinonimo, negalima patikrinti pakeičiamumo, nes turi būti bent jau teorinio pasirinkimo tarp dviejų (ar daugiau) sinonimų galimybė.

Riboto leksinio junglumo patikrą tiek sinonimijos (1 testas), tiek leksinių semantinių grupių testu (2 testas) atlikę lingvistai rėmėsi savo kaip gimtakalbių kompetencija. Tačiau tokio darbo rezultatams neišvengiamą įtaką daro subjektyvumo veiksnys (apie šią sinonimų nustatymo problemą žr. Divjak, Gries 2006: 24–26). Todėl sinonimų pakeičiamumui vertinti išbandytas ir automatizuotas metodas (žr. 1.2 sk.).

1.1.2. AK nustatymas pagal perkeltinę reikšmę

Kai kuriuose ankstesnio laikotarpio tyrimuose (plg. Howarth 1998) tokios kolokacijos, kurių vienas iš dėmenų – perkeltinės reikšmės, vadinamos ribotosiomis kolokacijomis (angl. *restricted collocations*). Pasak Annos Reder (2006: 82), kolokacijos su perkeltinės reikšmės žodžiu yra riboto junglumo junginiai, kuriuose nenuspėjamą ryšį tarp kolokacijos narių lemia konvencija, o metaforiškumas yra konvencijos dalis. Todėl šios kolokacijos svarbios kalbos mokymui (Lewis 2000: 29). Tarkim, su anglų kalbos būdvardžiu *heavy* vartojami įvairūs daiktavardžiai, kurie reiškia dalykus su požymiu „sunkus“, bet žodžių junginyje *heavy smoker* (liet. *užkietėjęs rūkalius*) minimas ne objektas, o asmuo, todėl mokant kalbos šią kolokaciją verta išskirti kaip įsidėmėtiną žodžių junginį.

Būdvardžiai, vartojami perkeltinėmis reikšmėmis, paprastai pasižymi ir platesne junglumo apimtimi, o kolokacijos su platesnio junglumo būdvardžiais yra sudėtingesnės besimokantiejiems (Nesselhauf 2005: 238). Dėl aktualumo kalbos mokymui ir vertimui

(Volungevičienė 2010; Stulpinaitė ir kt. 2016) nustatant lietuviškąsias AK panaudotas perkeltinės reikšmės kriterijus.

Kolokacijas analizuojant neautomatiškai, AK požymiu laikytas vieno iš kolokacijos dėmenų (būdvardžio, veiksmažodžio ar daiktavardžio) vartojimas netiesiogine reikšme. Tai nustatyta analizuojant visą tam tikro žodžio paradigmą, generuotą iš *Delfi.lt* tekstyno duomenų (žr. 1 lentelės 1 stulpelį). Iš tokios paradigmos matyti, su kokiais kitais būdvardžiais ar veiksmažodžiais kolokacijas sudaro konkretus daiktavardis ir kokia reikšme tose kolokacijose vartojamas tiek jis, tiek ir jo kolokatai – būdvardžiai ir veiksmažodžiai.

Kartais nustatyti, ar žodis vartojamas tiesiogine ar perkeltine reikšme, leidžia tik platesnis kontekstas – tada pasitelkta PŽJDB, *Delfi.lt* tekstynas, plg. *žalias + automobilis*:

„Unicredit Leasing“ Lietuvoje atstovas pastebi dar porą **žaliųjų automobilių** rinkos tendencijų.

Mariaus Slaboševičiaus pilotuotas **žaliasis automobilis** atsiliko pusę sekundės <...>.

„Tesla Model S“ yra laimėjęs kelis geriausio automobilio apdovanojimus, kurių vienas yra 2013 m. Pasaulio metų **žaliasis automobilis**.

Įdomūs žaliųjų automobilių koncepciniai modeliai.

Iš pavyzdžių matyti, kad būdvardis *žalias* apibūdina automobilį kaip ekologišką, bet nenusako jo spalvos⁵. Tokia kolokacija laikyta arbitraliaja. Žinoma, *Delfi.lt* tekстыne, iš kurio duomenų sudaryta PŽJDB, galima rasti nemažai atvejų, kai *žalias* apibūdina automobilio spalvą.

Daiktavardžių junginiai su būdvardžiais ir veiksmažodžiais laikyti AK, jeigu bent vienas iš dėmenų perkeltinės reikšmės. Tokiu atveju aprašytieji testai (žr. 1.1.1 sk.) netaikyti. Būdvardžių ir veiksmažodžių vartojimas perkeltine reikšme itin būdingas abstrakčiųjų daiktavardžių, taip pat ir keletui konkrečių dalykų reiškiančių daiktavardžių kolokacijoms, pavyzdžiui, su daiktavardžiais *amžius*, *aplinka*, *apsauga*, *būdas*, *dalykas*, *dalys*, *dėmesys*, *duomenys*, *galimybė*, *gyvenimas*, *informacija*, *įstatymas*, *įvykis*, *kalba*, *karas*, *kartas*, *klausimas*, *laikas*, *metas*, *pabaiga*, *pagalba*, *pradžia*, *reikalas*, *rezultatas*, *saugumas*, *sprendimas*, *sveikata*, *teisė*, *turtas* sudarytoms kolokacijoms.

⁵ Atkreipiame dėmesį, kad apie išskirtinę vartoseną signalizuoja ir įvardžiutinės formos.

Kolokacijos su abstrakčiais daiktavardžiais, kaip nustatyta R. Petrauskaitės (žr. Marcinkevičienė 2010), gali sudaryti konceptualiąsias metaforas (pvz.: *švelni priemonė, aštri kritika, sausas rezultatas, nutekinta informacija, pakurstyti kalbas, duomenų nutekimas, įstatymo spraga*)⁶. Kaip teigia R. Petrauskaitė, „svarbiausias metaforiškumo kriterijus – reikšmės perkėlimas, išryškėjantis iš konkretaus ir abstraktaus daiktavardžio priešpriešos. Kai konkretaus daikto savybės imamos pagrindu abstraktui įtraukti į kontekstą, sudaryti žodžių junginį su juo, tuomet gimsta metafora <...> vienas, dažniausiai abstraktus, dalykas suvokiamas kaip kitas, konkretus, pasitelkus pastarojo terminiją <...>“ (Marcinkevičienė 2010: 178).

Perkeltine reikšme pavartoti būdvardžiai (rečiau veiksmažodžiai) junginiuose su konkrečiais daiktavardžiais suteikia jiems semantiškai nebūdingų savybių (pvz.: *greitas maistas, ryški moteris, matomas žmogus*). Kaip ir junginiuose su abstrakčiais daiktavardžiais, dažnai žodžių junginio reikšmė suprantama ne iš pavienių žodžių, o iš viso žodžių junginio (pvz.: *geležinis žmogus, auksinis amžius, laukinė moteris*, taip pat žr. anksčiau pateiktus pavyzdžius su daiktavardžiu *automobilis*). Kitų kalbų kolokacijų tyrimuose (pvz., Strakatova, Hinrichs 2019) leksikalizuotoms kolokacijoms priskiriamos perkeltinės reikšmės kolokatą turinčios kolokacijos, kuriose perkeltinė reikšmė išryškėja tik su ribotu konkrečių leksinių semantinių grupių žodžių skaičiumi. Pavyzdžiui, būdvardžio *heavy* reikšmė junginiuose su daiktavardžiais *smoker, rain, traffic* (liet. *užkietėjęs rūkalius, stiprus lietus, intensyvus eismas*) nėra prototipinė, nes žymi ne svorį, o intensyvumą.

Nustatyta, kad didelį metaforizacijos potencialą turi vidutinio leksinio junglumo⁷ žodžiai, pavyzdžiui, veiksmažodžiai *paralyžiuoti* (*paralyžiuoti eismą*), *sekti* (*senka laikas*), *sugriauti* (*sugriauti gyvenimą*). Tarp būdvardžių daug tokių, kurie, vartojami tiesioginėmis

⁶ Apie erdvines, ontologines ir struktūrines metaforas kolokacijose taip pat žr.: Drūlienė 2008a, b; Šilinytė 2008. Šio straipsnio ir viso projekto tikslas nėra suklasifikuoti nustatytas metaforas, todėl į tai labiau nesigilinama, nors suprantama, kad AK gali būti kognityvinės lingvistikos specialistams įdomus tyrimo objektas.

⁷ Remiantis Susane Handl 2008 m. pateikta formule, *Delfi.lt* tekstyno pagrindu nustatyti analizuojamų kolokacijų dėmenų kolokatyvumo įverčiai. Šie tyrimai bus tęsiami. Kol kas galima pasakyti, kad AK turėtų būti vidutinio stiprumo junginiai, t. y. kad nei vieno, nei antro (ar trečio) dėmens kolokatyvumo įvertis neturėtų būti labai mažas ar labai didelis. Jei, tarkim, pirmo dėmens įvertis didelis, o antro mažas, tai pirmasis dėmuo gali būti vartojamas su mažai žodžių, o antrasis – su labai daug žodžių, pvz.: *psichoaktyvi medžiaga, autistiškas vaikas, gėlas vanduo, pavėsinga vieta, trišalė sutartis, mažametis vaikas, gazuotas / distiliuotas vanduo, pakeleivingas automobilis*.

reikšmėmis, žymi fizinius požymius (pvz.: *slidus dalykas, švarus darbas, kruvinas darbas*), bet gali būti ir kitų reikšmių būdvardžių (*brangus laikas, kamuojantis klausimas, ryškus įvykis*).

Perkeltinės reikšmės kriterijaus taikymas iliustruojamas būdvardžio *(ne)švarus*, veiksmažodžio *nutekėti* ir daiktavardžio *euras* paradigmų analize (žr. 1 lentelę):

1 LENTELĖ. Perkeltinės reikšmės kriterijaus taikymas

Paradigma	AK nustatymas
1. švarus sportas, darbas; nešvarus darbas, dalykas, reikalas 2. švarus (-i) vanduo, aplinka, oras, miestas, vieta, automobilis; nešvarus vanduo	1. Laikoma AK dėl perkeltinės būdvardžio reikšmės junginiuose su abstrakčios reikšmės daiktavardžiais. 2. Nelaikoma AK, nes šiose kolokacijose būdvardis vartojamas tiesiogine reikšme.
1. nutekėti informacija 2. nutekinta (-i) informacija, medžiaga, duomenys	1–2 laikoma AK, nes junginiuose su abstraktais realizuota perkeltinė veiksmažodžio reikšmė.
1. euro silpnėjimas, stiprėjimas 2. euro kursas, moneta, centas 3. euro keitimas	1. Laikoma AK, nes eurui priskiriamos gyvam objektui būdingos savybės. 2. Tai terminai; nelaikoma AK. 3. Nelaikoma AK, nes nėra perkeltinės reikšmės.

Kaip matyti iš 1 lentelėje pateiktų pavyzdžių, ne visada kolokacijų paradigmoje būdvardis ar veiksmažodis, pavartotas tiesiogine reikšme, o kartais ir ne visa kolokacija su tam tikru būdvardžiu / veiksmažodžiu yra perkeltinės reikšmės, bet perkeltinė reikšmė aiški tokiais atvejais: *skausmingas sprendimas / dalykas / įvykis; uždaras žmogus / būdas / gyvenimas; atsiveria kelias / rinka / vaizdas; išsikovoti teisę / vietą* ir t. t. Tokios kolokacijos laikomos arbitraliosiomis.

Apibendrinant galima teigti, kad perkeltinės reikšmės kriterijus ypač aktualus analizuojant iš būdvardžių ir daiktavardžių sudarytas kolokacijas – dažnai juo remtasi kaip

pagrindiniu, aiškiausiu arbitralumo rodikliu. O analizuojant veiksmažodžio ir daiktavardžio kolokacijas dažniau taikytas riboto leksinio junglumo kriterijus (ir semantinių grupių testas). Kadangi iš dviejų daiktavardžių sudarytų AK rasta labai nedaug, šios struktūros kolokacijų ypatybės sunku apibendrinti (žr. 1.1.1 sk.).

1.2. Automatizuotas AK nustatymas

Vienas iš AK kriterijų – ribotas leksinis junglumas – buvo tikrinamas ir automatizuotai, t. y. vektorių metodu buvo nustatyti pakeičiamumo testui reikalingi sinonimai. Išsamiau apie pakeičiamumo testo, vektorių metodo taikymą rašoma anksčiau paskelbtame straipsnyje (Kovalevskaitė ir kt. 2021). Jame aprašytas vektorių metodo taikymas būdvardžio + daiktavardžio kolokacijų arbitralumui įvertinti, todėl šiame straipsnyje išsamiai aprašomas tik veiksmažodžio + daiktavardžio kolokacijų arbitralumo vertinimas.

Pirmiausia trumpai priminsime pagrindinius vektorių metodo taikymo etapus (1.2.1 sk.), tada aptarsime veiksmažodžių sinonimų atranką vektorių metodu (žr. 1.2.2 sk.) ir apibendrinsime veiksmažodžių ir daiktavardžių kolokacijų pripažinimo arbitraliosiomis rezultatus (žr. 1.2.3 sk.).

1.2.1. AK nustatymas pagal ribotą leksinį junglumą taikant vektorių metodą

Taikant vektorių metodą, siekta AK nustatymą automatizuoti, sumažinti subjektyvumo vertinant duomenis veiksnį. Vektorių metodas ribotam leksiniam junglumui nustatyti taikytas tokia seka. Pirmiausia iš viso *Delfi.lt* tekstyno automatiškai išrinkti susiję žodžiai⁸. Tam pritaikytas *GloVe* – žodžių vektorių (angl. *embeddings*) modelis (Pennington ir kt. 2014). Vektorių metodas (angl. *Word Embedding Approach*) priklauso metodų, skirtų žodžiams automatiškai analizuoti pagal jų distribucijos, arba konteksto, panašumus (angl. *distributional similarity measures*), grupei. Šie metodai taikomi įvairiems ryšiams (hiperonimijai, sinonimijai, antonimijai, partonimijai) automatiškai atpažinti ir yra grindžiami

⁸ Vykdam ankstesnį projektą („Lietuvių kalbos pastoviųjų žodžių junginių automatinis atpažinimas (PASTOVU)“, žr. <https://pastovu.vdu.lt>) nustatyti vektoriai pateikti čia: <https://clarin.vdu.lt/xmlui/handle/20.500.11821/26>. Jie šiame tyrime naudojami sinonimams, kurių reikia AK nustatyti, atrinkti.

hipoteze, kad artimos reikšmės ir (arba) semantiškai susiję žodžiai yra vartojami panašiam kontekste (žr. Grigonytė 2010).

Pirmiausia automatiškai sudarytos PŽJDB kolokacijose pavartotų būdvardžių ir veiksmažodžių vektorių eilės (toliau VE)⁹. Taikant automatinę susijusių žodžių atranką, atsisakyta plataus leksinio junglumo ar nesavarankiškai vartojamų būdvardžių ir veiksmažodžių, pvz.: *geras, naujas, puikus; būti, reikėti, pradėti*. VE sugrupuoti būdvardžiai ar veiksmažodžiai atrinkti pagal konteksto panašumą. Sudaryta apie 800 VE su būdvardžiais, daugiau nei 3800 – su veiksmažodžiais. Nors VE buvo pateikiami susiję žodžiai, tarp duomenų buvo ir kitais semantiniais ryšiais (pvz., antoniminiais) susijusių ar net nesusijusių žodžių, todėl reikėjo lingvistinio redagavimo siekiant nustatyti, kurie iš susijusių žodžių laikytini sinonimais, galimais naudoti vertinant, kiek ribotas vieno ar kito sinonimo junglumas su tam tikru daiktavardžiu. Sinonimais laikyti tos pačios kalbos dalies žodžiai, kurių denotatinė (dalykinė) reikšmė labai panaši, bet leksinis junglumas iš dalies arba visiškai skiriasi¹⁰.

Sutvarkytos veiksmažodžių VE naudotos veiksmažodžio + daiktavardžio kolokacijų arbitralumui įvertinti. Norėta nustatyti, kuriose kolokacijose veiksmažodžiškasis dėmuo negali būti pakeičiamas artimos reikšmės veiksmažodžiu. Prie kiekvienos veiksmažodžių sinonimų poros ar sinonimų grupės buvo pateiktos kolokacijos su tam tikru daiktavardžiu. Jeigu su tuo daiktavardžiu buvo pateikta kolokacija tik su vienu sinonimų grupės ar poros sinonimu, daryta išvada, kad analizuojamas daiktavardis jungiasi tik su vienu tos sinonimų grupės veiksmažodžiu, ir kolokacija laikyta arbitraliąja. Jeigu analizuojamas daiktavardis sudarė kolokacijas su visais sinonimų poros ar grupės veiksmažodžiais, kolokacijos nepriskirtos prie arbitraliųjų (žr. 2 lentelę).

⁹ Vektorių metodu nustatytus susijusius žodžius tiksliausia būtų vadinti *vektorių metodu nustatytų galimų sinonimų eilėmis*. Dėl trumpumo toliau vartojame *vektorių eilės*, tiksliau, santrumpa VE.

¹⁰ Šiame straipsnyje taikoma sinonimo samprata gali būti siejama su lietuvių leksikografijoje labiau paplitusiu terminu *daliniai sinonimai* (žr. Jakaitienė 2010).

2 LENTELĖ. AK pripažinimo etapai naudojant VE

Pripažinta AK	Nepripažinta AK
1 etapas: imama VE su dviem sinonimų poromis <i>gauti–įgyti</i> ir <i>duoti–teikti</i>; 2 etapas: pagal šią VE tikrinama, su keliais iš šių sinonimų daiktavardis <i>apdovanojimas</i> sudaro kolokacijas; nustatomos dvi kolokacijos: <i>gauti apdovanojimą</i> (vienas poros <i>gauti–įgyti</i> sinonimas) ir <i>teikti apdovanojimą</i> (vienas poros <i>duoti–teikti</i> sinonimas); 3 etapas: priimamas sprendimas, kad <i>gauti apdovanojimą</i> ir <i>teikti apdovanojimą</i> laikytini AK.	1 etapas: imama VE su dviem sinonimais <i>įkurti</i> ir <i>įsteigti</i>; 2 etapas: pagal šią VE tikrinama, su keliais iš šių sinonimų daiktavardis <i>asociacija</i> sudaro kolokacijas; nustatomos dvi kolokacijos: <i>įkurti asociaciją</i> ir <i>įsteigti asociaciją</i>; 3 etapas: priimamas sprendimas, kad <i>įkurti asociaciją</i> ir <i>įsteigti asociaciją</i> nėra AK.

Pavyzdžiui, su daiktavardžiu *apdovanojimas* buvo pateiktos kolokacijos *gauti apdovanojimą*, *teikti apdovanojimą*, bet nebuvo **įgyti apdovanojimą*, **duoti apdovanojimą*. Vadinasi, su sinonimų poromis *gauti–įgyti* ir *duoti–teikti* daiktavardis *apdovanojimas* sudaro tik po vieną parinktį, taigi kolokacijos *gauti apdovanojimą* ir *teikti apdovanojimą* priskirtos prie arbitraliųjų. Iš kolokacijų su veiksmažodžiais arbitraliosiomis dažnai laikytos tos, kuriose vartojami išblukusios reikšmės (desemantizuoti) veiksmažodžiai (plačiau Kovalevskaitė ir kt. 2020), pvz.: *atlikti tyrimą*, *duoti rezultatą*, *imti mokestį*. Daiktavardžio *asociacija* kolokacijų arbitralumas vertintas pagal VE, kurioje yra du sinonimai – *įkurti* ir *įsteigti*; kadangi šis daiktavardis kolokacijas sudarė su abiem sinonimais, jos atmestos kaip nearbitralios.

Kaip ir vertinant kolokacijas su būdvardžiais, atliekant eksperimentinę tyrimą veiksmažodžių sinonimai testuoti neautomatiškai, nes reikia atsižvelgti į sinonimo ir į testuojamo daiktavardžio reikšmę. Jeigu VE esančių sinonimų reikšmė nesutampa su vertinamosios kolokacijos veiksmažodžio reikšme, jos veiksmažodžiui pakeičiamumo testas

negali būti taikomas. Pavyzdžiui, kolokacijos *statyti automobilį* arbitralumui įvertinti VE su *statyti* netiko dėl reikšmės „statyti pastatą“ (166 VE¹¹).

Taigi, automatizuoto metodo taikymo etapai yra tokie: a) automatinis (vektorių metodu sudaromos VE), b) neautomatinis (lingvistas peržiūri ir, jeigu reikia, patikslina); c) iš dalies automatinis (kolokacijos automatiškai susiejamos su konkrečiomis VE ir po peržiūros patvirtinamos arba nepatvirtinamos kaip AK).

1.2.2. Veiksmažodžių sinonimų atranka vektorių eilėse

Kadangi generuotose VE buvo daug perteklinių, neaktualių duomenų, vadinamojo triukšmo (angl. *noise*), kiekvieną VE peržiūrėjo lingvistas ir nereikalingus, sinoniminiais ryšiais nesusijusius žodžius ištrindavo, o kai kurias VE iš viso pašalindavo. Toliau aptarsime skirtingus VE lingvistinio redagavimo atvejus.

Veiksmažodžių VE skaičius buvo daug didesnis nei būdvardžių (atitinkamai 3800 ir 800 VE), nes ir pačių veiksmažodžių sąrašas. Automatiškai VE sudaroma kiekvienam veiksmažodžiui, todėl kai kurios VE yra panašios, tačiau pagal tai, koks veiksmažodis yra atraminis, gali skirtis panašioje VE esančių sinonimų skaičius. Atraminis veiksmažodis – tai žodis, į kurį panašūs kiti pateikti žodžiai, kitaip sakant, toks, kuriam ieškota galimų sinonimų (panašumas gali būti nustatytas pagal hiperoniminį, sinoniminį, antoniminį ar partoniminį ryšį; šis žodis sinonimų grupėse pabraukiamas). Pavyzdžiui, 151 VE-ėje atraminis veiksmažodis yra *organizuoti*, o 40 VE daug ilgesnė, jos atraminis žodis yra *vykdyti*:

151 VE: ***organizuoti***, *rengti*, *vykdyti*

40 VE: ***vykdyti***, *įgyvendinti*, *rengti*, *organizuoti*, *atlikti*, *įvykdyti*, *nutraukti*

Kaip jau minėta, automatiškai sudarytos VE dar turėjo būti peržiūrėtos lingvisto. Atlikus šį darbą, pastebėta, kad kaip su būdvardžiais, taip ir su veiksmažodžiais automatiškai surinkta daug ilgų VE, kuriose yra kelios veiksmažodžių sinonimų poros, ir mažai tokių, kur būtų tik viena sinonimų pora arba trys sinonimai. Kartais po peržiūros paaiškėdavo, kad atraminis veiksmažodis, pagal kurį sudaroma VE, neturi sinonimų, nors toje VE gali būti kita

¹¹ Čia ir toliau nurodomi vektorių metodu nustatytų galimų sinonimų, kuriais remtasi šiame straipsnyje, VE numeriai.

tinkama sinonimų pora, pavyzdžiui, atraminio veiksmažodžio *kisti* VE-ėje yra sinonimų pora *paaštrėti, pablogėti*.

Kaip ir būdvardžių, veiksmažodžių VE „triukšmą“ lemdavo antonimai, plg. VE *pasivoyti–aplenkti, sutrumpinti–prailginti, griežtinti–švelninti* sudarydavo vienos leksinės semantinės grupės nariai, kurie nėra sinonimai, pvz.: **pasnigti**: *palynoti, nugriaudėti; nepasisiekti*: *pargriūti, išsausėti, patinti, pakrypti, sušlubuoti, sudužti; vakarienauti*: *pietauti; pavirti*: *troškinti*. Gryninant VE, dažnai paaiškėdavo, kad kuris nors ar keli veiksmažodžiai netinka, nes įvardija pernelyg konkretų panašaus veiksmo aspektą, pavyzdžiui, žodžio *remtis* VE-ėje pateikti tokie žodžiai: **vadovautis**, **pasikliauti**, **apsiriboti**, **kliautis**. Ištrintas *apsiriboti* ir sinonimais laikyti tik *vadovautis, pasikliauti, kliautis*.

Dažni ir tokie atvejai, kai vienas VE žodis yra platesnės reikšmės, pavyzdžiui, *paveikti*, o kiti VE priskirti žodžiai konkrečiau įvardija neigiamo poveikio aspektus, pavyzdžiui, *pakenkti, sutrikdyti*. Todėl tik jie laikyti sinonimiškais. Iš VE *pagrįsti, įvardinti, paaiškinti, pateisinti* sinonimais pripažinti *pagrįsti* ir *paaiškinti*, o *pateisinti* ir *įvardyti* sinonimais nelaikyti, nes papildomais semantiniais požymiais pernelyg nutolsta nuo pirmų dviejų žodžių reikšmės. Dėl tokių atvejų kartais tekdavo atsisakyti visos VE, pavyzdžiui, **atmesti**, *ignoruoti, uždrausti*.

Kartais būdavo visai netinkamos ir atraminių veiksmažodžių VE, kurioms, atrodytų, tikrai turėtų būti galima automatiškai nustatyti sinonimus, pvz.: *suprastėti* (galimas sinonimas *pablogėti*, bet jo VE nebuvo); *sąlygoti* (galimas sinonimas *lemti*, bet jis nenurodytas, buvo *pakentėti*). Veiksmažodžių, turinčių neiginį, VE nebuvo tikslios, todėl atmestos, pavyzdžiui, atmesti atraminiai veiksmažodžiai *negydyti, nesuskaičiuoti, ne pridėti*.

Paminėtina, kad VE dažnai tekdavo tikslinti ir dėl veiksmažodžių darybinių ypatybių. Pavyzdžiui, sinonimais nelaikyti eigos ir įvykio veiksmo veiksmažodžiai (*tirti* ir *ištirti*), taip pat tokie skirtingą reikšmę dėl priesagos įgiję veiksmažodžiai, kaip *palyti* ir *palynoti, naršyti* ir *panaršyti*, sangrąžiniai ir nesangrąžiniai veiksmažodžiai (*klausyti* ir *klausytis*). Jeigu, tarkim, atraminis VE veiksmažodis būdavo sangrąžinis, tai pagal jį buvo atsisakoma nesangrąžinio. Pavyzdžiui, VE **išsirinkti**, **atsirinkti**, **pasirinkti**, **nuspręsti**, **rinktis** atsisakyta veiksmažodžių *išrinkti, parinkti*. Darybiniai variantai (pvz., *vardinti* ir *vardyti, lukterėti* ir *luktelėti*) laikyti sinonimais.

Gryninant VE išryškėjo ir dar vienas automatinės atrankos ypatumas: kadangi vektorių metodu žodžiai surenkami pagal jų konteksto panašumą, VE gali sudaryti tik iš pirmo žvilgsnio netinkami sinonimai. Kitaip sakant, nežinant konkrečių daiktavardžių, dėl kurių du ar keli veiksmažodžiai sudėti į vieną VE, kai kurių VE kokybę galima įvertinti klaidingai. Pavyzdžiui, 60-ojoje VE buvo tokie žodžiai: *rodyti*, *demonstruoti*, *jausti*. Iš karto ištrynus žodį *jausti* būtų suklysta, nes jis čia atsidūrė dėl šiems veiksmažodžiams bendro konteksto, kuriame kalbama apie jausmus, plg. *demonstruoti*, *rodyti jausmus* ir, pavyzdžiui, *jausti pyktį*, *jausti nusivylimą* ir pan. Veiksmažodžiai *sumažėti*, *pigti* arba *padidėti*, *ūgtelėti* gali būti sinonimiški, jeigu kalbama apie kainas. Visi trys VE *apsispręsti*, *priimti*, *pasirinkti* veiksmažodžiai yra sinonimai, o veiksmažodis *imti*, kuris dažniausiai ištrinamas, čia atsirado todėl, kad žymi *priimti sprendimą*, kitaip sakant, šią VE reikia skaityti taip: *apsispręsti*, *priimti sprendimą*, *pasirinkti*¹².

Iš tokių pavyzdžių matyti, kad automatiškai galima nustatyti daugiau ir (arba) kitokių sinonimų, nes vektorių metodu vertinamas konteksto, o ne atskirų žodžių reikšmės panašumas. Vargu ar imdami atskirus žodžius sinonimiškai susietume veiksmažodžius *atsikratyti*, *numesti*, *sulieknėti*, tačiau juos randame vienoje VE, nes jie pavartoti bendrame kontekste, kuriame buvo ir daiktavardis *svoris*. Galima pateikti ir daugiau pavyzdžių, plg.: *trukti*, *užtrukti*, *gyvuoti*, *egzistuoti*, *užsitęsti*; *domėtis* ir *užsiimti*; *gerti* ir *vartoti*; *nusiteikti* ir *pasiruošti*; *skambinti*, *kreiptis*, *pranešti*; *atkurti*, *atstatyti*, *atgaivinti*; *peržiūrėti*, *koreguoti*, *atnaujinti*, *keisti*; *išprovokuoti*, *lemti*, *peraugti*; *šokiruoti*, *sukrėsti*, *priblokšti*; *atlaikyti*, *ištverti*, *įveikti*.

Sporto kontekste visi keturi veiksmažodžiai – *užsitikrinti*, *iškovoti*, *pasiekti*, *išsaugoti* – yra sinonimiški, nes kartais iškovojimas nusakomas taip: *užsitikrino sau vietą finale* arba *išsaugojo titulą*. Tokie rezultatai rodo, kad vertinant sinonimiją svarbu apimti daugiau negu vieną žodį, nes sinonimų grupės nariai gali būti ir pavieniai žodžiai, ir pastovieji žodžių

¹² VE-ėse būna tokių žodžių, kurie susiję sintagmiškai, pvz., *įžeisti* ir *nenorėti*. Kadangi *nenorėti* dažnai vartojamas bendrame su *įžeisti* kontekste (kai atsiduria junginyje *nenorėjo įžeisti, bet...*), jie sudėti į vieną VE. Kitas atvejis: **liautis**: *nustoti*, *verkti*. Veiksmažodis *verkti* čia atsirado dėl to, kad atrinktas iš jam dažno konteksto *liovėsi verkęs / verkusi*. Kitaip sakant, labai ilgos (ir netinkamos) VE būna surinktos tiems atraminiams veiksmažodžiams, kurie vartojami kaip pagalbiniai ir dėl savo gramatinių ypatybių yra plataus junglumo, pvz.: *daryti*, *imti*, *pradėti*, *norėti*, *reikti*, *privalėti*, *reikėti*, *stengtis*, *nepajėgti*.

junginiai. Taikant vektorių metodą, kai lyginamas konteksto panašumas, kaip tik ir išryškėja pavienių žodžių sąsajos su kontekstu. Lingvistui aprėpti tiek duomenų būtų sudėtinga (daugiau apie plataus distribucijos rodiklių spektro panaudojimą automatinei sinonimų analizei žr. Divjak, Gries 2006).

Pastebėta, kad dažnai specifiskesnės reikšmės ir siauresnio junglumo veiksmažodžio VE yra trumpesnė ir švaresnė, tikėtina, dėl siauresnio junglumo. Palyginus sinonimų *nerimauti, jaudintis, būgštauti* ir *nuogąstauti* VE, matyti, kad *nerimauti* VE ilgesnė ir ne tokia gryna, joje yra platesnės reikšmės žodžių, pavyzdžiui, *bijoti* (***nerimauti***: *jaudintis, baimintis, bijoti*), o *būgštauti* VE surinkti tik tie keturi žodžiai: ***būgštauti***: *nerimauti, jaudintis, nuogąstauti*. Dar pastebėta, kad VE-ėse, kurios pagrįstos konkretnės reikšmės veiksmažodžiu (*duoti, bėgti, išimti, atsistoti, medžioti, kvėpuoti, sugesti, kalėti*), gali būti mažai arba visai nebūti sinonimų, nes pagal konkrečios reikšmės veiksmažodį surinkti žodžiai susiję temišškai, o ne sinonimiškai, plg.: ***išimti***: *išmesti, išbraukti, išnešti*; ***medžioti*** ir *žvejoti*; ***kvėpuoti*** ir *vaikščioti*; ***virti*** ir *kepti*; ***kalėti*** ir *nuteisti*; ***sugesti***: *skilti, perkaisti, sprogti, užsikimšti, užsidegti*. Tokios VE sudarytos ir iš dviejų žodžių, bet gali būti ilgesnės. Kartais būna labai ilgos, pavyzdžiui, su atraminiu veiksmažodžiu *sumenkti* buvo net 26 veiksmažodžiai.

Apibendrinant galima paminėti, kad patikslinus gauta 2634 iš 3800 VE (70 proc.). Taigi, automatiškai sugeneruotos VE nėra pakankamai tikslios, kad jas būtų galima neredaguojant taikyti tolesniam AK nustatymo etapui. Tokios paklaidos gali būti nulemtos *Delfi.lt* tekstyno ypatumų – nors jo žodžių apimtis didelė (apie 70 mln.), vieno stiliaus tekstuose leksinė įvairovė gali būti mažesnė. Tikslesnėms VE generuoti reikėtų didesnės funkcinių stilių įvairovės tekstyno.

1.2.3. Rezultatai: VE kokybė ir arbitralumo pripažinimas

Palyginus būdvardžių ir veiksmažodžių VE prieš peržiūrą ir po jos, paaiškėjo, kad abiem atvejais gana daug VE atsisakyta kaip neinformatyvių:

- 1) iš 1927 būdvardžių VE po peržiūros liko 800 VE (41 proc.);
- 2) iš 3800 veiksmažodžių VE po peržiūros liko 1166 VE (30 proc.).

Likusias VE taip pat reikėjo redaguoti. Tiesa, neatliktas vertinimas, kiek žodžių dažniausiai ištrinama vienoje VE. Dažnai tekdavo ištrinti pagalbinius, ne visai savarankiškus

veiksmažodžius: *daryti, imti, gyventi, vertėti, reikti, privalėti, pradėti, gauti, norėti, stengtis, pavykti*. Daug išstrintinų žodžių būdavo VE pabaigoje, kur atsidurdavo žodžiai su mažiausia taikyto įverčio riba (0,5; didžiausias įvertis yra 1), bet nemažai buvo ir tokių atvejų, kai buvo ištrinamas žodis su aukštu įverčiu, tam tikroje VE pateikiamas 2–5 pozicijoje (VE daugiausia gali sudaryti 20 žodžių).

Pabrėžtina, kad vertinant pagal VE ir apsiribota tik tais žodžiais, kurie yra toje VE. Jeigu gali būti kitų sinonimų, bet jų toje VE, pagal kurią vertinama, nebuvo (jie neatrinkti vektorių metodu, nepavartoti *Delfi.lt* tekстыne ar kt.), kai kurios kolokacijos galėjo būti priskirtos prie nearbitraliųjų (arba atvirkščiai). Pavyzdžiui, vertinant kolokacijas *krintanti žvaigždė* ir *kylanti žvaigždė*, naudota VE, kurioje yra veiksmažodžiai *augti, kristi, kilti, mažėti, vystytis, didėti*. Šioje VE nėra *griūti*, kuris galėtų būti sinonimiškas *kristi*, todėl *krintanti žvaigždė* nepripažinta AK.

Tyrimo metu pastebėta ir automatinės duomenų atrankos privalumų: 1) kaip minėta 1.2.2 poskyryje, vektorių metodu galima nustatyti daugiau ir kitokių sinonimų nei pasitelkus vien tik introspekcijos metodą; 2) vektorių metodas tinka dideliame duomenų kiekiui apdoroti, o rankiniu būdu vargu ar būtų įmanoma įvertinti tokios gausybės žodžių konteksto panašumą.

Veiksmažodžių sinonimų VE-ėms pritaikius pakeičiamumo testą, 291 kolokacija pripažinta kaip arbitralioji. Akivaizdu, kad būtent VE esantys sinonimai lemia, kurios kolokacijos bus pripažintos kaip AK. Viena vertus, tai šio metodo ribotumas, kita vertus, tai galimybė automatizuoti arbitralumo nustatymo procesą.

Kai arbitralumo pripažinimas priklauso tik nuo konkrečios VE, dažnai gali atsitikti taip, kad kai kurios kolokacijos neatrenkamos pagal vieną, bet atrenkamos pagal kitą VE. Pavyzdžiui, vertinant kolokacijų *kurti įvaizdį* ir *formuoti įvaizdį* arbitralumą pagal 368 VE ***kurti, formuoti***, nė viena iš dviejų kolokacijų nebūtų priskirta AK. Bet veiksmažodžiai *kurti, formuoti* yra ir kitose VE, prie kurių priskirtas tas pats daiktavardis *įvaizdis*, todėl kolokacija *kurti įvaizdį* kaip AK pripažįstama pagal kitą VE (45): ***kurti, sukurti, vystyti, plėtoti, įgyvendinti, rengti, įkurti***.

Galimas ir kitas atvejis: tos pačios kolokacijos gali būti pripažintos kaip arbitraliosios daugiau nei pagal vieną VE, pavyzdžiui, *atlikti mankštą, daryti mankštą* galima įvertinti kaip AK ir pagal 11-ąją (***atlikti, vykdyti, daryti***), ir pagal 15-ąją VE (***padaryti, atlikti, sukurti***;

nuveikti). Problema kyla tada, kai arbitralioji kolokacija nelaikoma arbitraliąja, jeigu to negalima nustatyti pagal nė vieną VE. Taip gana dažnai pasitaikydavo vertinant kolokacijų variantus. Pavyzdžiui, su daiktavardžiu *mokesčiai* vartojamas ir *kelti*, ir *didinti*, abu šie veiksmažodžiai yra 52-oje VE, pagal kurią vertinamas arbitralumas, todėl šios kolokacijos nebus priskirtos prie AK. Tačiau jos yra AK variantai (plg. *auga*, *kyla*, *didėja kainos*). Toks pats atvejis – *atlikti* / *vykdyti* / *daryti tyrimą* (pagal 11-ąją VE šie junginiai nepripažinti kaip AK). Toje pačioje VE buvo *gauti* ir *įgyti*, todėl *gauti diplomą* ir *įgyti diplomą* nepripažintos kaip AK. Tokie atvejai rodo, kad vertinant kolokacijų arbitralumą pagal riboto leksinio junglumo kriterijų svarbu ir taikomos testavimo priemonės, ir tai, kad AK variantams nustatyti labiau tinka neautomatinis metodas.

Kadangi sinonimų pakeičiamumo testavimas taip labai susijęs su konkrečioje VE esančiais žodžiais, gali atsitikti ir taip, kad panašių, tik priešingų reikšmių kolokacijos neatrenkamos. Pavyzdžiui, kolokacija *kyla nepasitenkinimas* atrinkta, o *mažėja nepasitenkinimas* – ne. Neautomatiškai jos abi būtų įvertintos kaip AK. Vadinasi, kai arbitralumui įvertinti naudojami lingvistinės introspekcijos būdu įvertinti duomenys, išsaugoma dalis 2 metodu prarandamų AK. Tiesa, taikant lingvistinį neautomatinį metodą, esama subjektyvumo. Automatizuotas metodas objektyvesnis, nors tenka pripažinti, kad duomenys fragmentiškesni (testuoti naudojama tik tiek ir tokių VE, kiek yra surinkta). Be to, ne visos kolokacijos, kurios yra AK, kaip tokios pripažįstamos.

2. AK NUSTATYMO REZULTATAI

Žiūrint į nustatytų AK skaičius, akivaizdžiai matyti, kad neautomatiniu ir automatizuotu metodais nustatytų sutampančių AK yra labai mažai (išsamiau žr. 2.1 ir 2.2 sk.), vadinasi, ir vienu, ir kitu metodu atskirai gaunama svarbių duomenų.

Neautomatinės analizės metodu surinktų duomenų daug daugiau negu nustatytų automatizuotai, bet taikant neautomatinę analizę taikytas ne vienas, o du atrankos kriterijai (riboto leksinio junglumo ir perkeltinės reikšmės), o taikant vektorių metodą – tik riboto

leksinio junglumo. Taigi, taikant daugiau kriterijų, surenkama daugiau kolokacijų su įvairesniais būdvardžiais ir veiksmažodžiais.

Remiantis konkrečių AK pavyzdžiais, šiame skyriuje bus paaiškinama, kodėl gauti būtent tokie rezultatai, kaip juos nulėmė taikyti kriterijai ir naudoti duomenys. Pirmiausia aptarsime, ką gauti duomenys rodo. Visos AK su vienu konkrečiu būdvardžiu arba veiksmažodžiu gali būti nustatytos skirtingai: a) tik vienu iš dviejų metodų; b) dalis kolokacijų – vienu, dalis – kitu; c) dalis kolokacijų – vienu, dalis – kitu, dalis – abiem metodais. Tolesniuose poskyriuose skirtingos struktūros AK aptarsime komentuodami abiem metodais gautus rezultatus.

2.1. Arbitraliosios kolokacijos, sudarytos iš būdvardžių ir daiktavardžių

Iš viso nustatytos 1048 AK, kurių struktūra būdvardis + daiktavardis. Į šias kolokacijas įeina 320 skirtingų būdvardžių. Tiesa, su vienais būdvardžiais (*didelis, mažas, atviras, aukštas, modernus, išskirtinis, pagrindinis* ir pan.) sudarytos kelios ar keliolika AK, o su kitais (pvz., *netrumpas, nuodugnus, permainingas, pikantiškas*) – pavienės AK (vidutiniškai su tuo pačiu būdvardžiu sudaryta 3,3 kolokacijos; išsamiau žr. 3 lentelę).

3 LENTELĖ. Skirtingais metodais nustatytos AK (būdvardis + daiktavardis)

Neautomatiniu metodu nustatytos AK		Automatizuotu metodu nustatytos AK		Abiem metodais nustatytos AK	
Sk. (proc.)	Skirtingų būdvardžių skaičius (vidutinis AK skaičius, kurios sudarytos su tuo pačiu būdvardžiu)	Sk. (proc.)	Skirtingų būdvardžių skaičius (vidutinis AK skaičius, kurios sudarytos su tuo pačiu būdvardžiu)	Sk. (proc.)	Skirtingų būdvardžių skaičius (vidutinis AK skaičius, kurios sudarytos su tuo pačiu būdvardžiu)
577 (55,06)	230 (2,5)	306 (29,2)	86 (3,56)	165 (15,74)	42 (3,93)

Kaip matyti iš 3 lentelės, daugiausia AK nustatyta neautomatiniu metodu – daugiau nei pusė visų AK (pvz.: *žiaurus vaizdas*, *vergiškas darbas*, *tiesmukas klausimas*, *skaudus dalykas*); apie trečdalis AK nustatyta automatizuotai (pvz.: *vertingas turtas*, *unikalus projektas*, *sudėtinga diena*, *optimalus būdas*), o mažiausią dalį (15,74 proc.) sudaro tos AK, kurios nustatytos abiem metodais (pvz.: *broliškas karas*, *intensyvi prekyba*, *milžiniškas projektas*, *žema kaina*). Panašios proporcijos matyti ir iš AK su veiksmažodžiais (žr. 2.2 sk.).

Analizuojant AK būdvardžius, matyti, kad dalis sudaro tik kuriuo nors vienu metodu nustatytas AK (žr. anksčiau pateiktus pavyzdžius), o kiti sudaro abiem metodais nustatytas AK (pvz.: *didelis*, *mažas*, *stiprus*).

Tyrimo pradžioje buvo iškelta hipotezė, kad AK, nustatomose neautomatiniu metodu, būdvardžių įvairovė didesnė, o AK, nustatytose taikant automatizuotą vektorių metodą, – mažesnė. Ši hipotezė pasitvirtino: pirmuoju metodu nustatytose AK pavartoti 235 skirtingi būdvardžiai, vadinasi, tas pats būdvardis sudaro vidutiniškai 2,5 AK. Vektorių metodu nustatytose AK yra 86 skirtingi būdvardžiai, taigi vidutiniškai 3,56 AK sudaryta su tuo pačiu būdvardžiu. Mažiausia būdvardžių įvairovė (42) tarp AK, kurios nustatytos abiem metodais, bet ir tokių kolokacijų mažiausia. Šioje AK grupėje su tuo pačiu būdvardžiu sudaryta vidutiniškai 3,93 kolokacijos. Manytina, kad tokia nedidelė būdvardžių įvairovė nulemta palyginti mažo duomenų skaičiaus.

Didelė tikimybė, kad kolokacija bus arbitralioji, jei ją sudaro tokie būdvardžiai, kaip *didis*, *didelis*, *didžiulis*, *milžiniškas*, *mažas*, *įtemptas*, *pavojingas*, *rimtas*, *stiprus*, *sunkus* su abstrakčiais daiktavardžiais, nes šie būdvardžiai bus pavartoti netiesiogine reikšme.

Aptariant, kaip taikomi metodai nulėmė būtent tokius rezultatus, kaip pavyzdį norėtume aptarti AK su būdvardžiu *protingas*, nes iš jo matyti visa įvairovė, t. y. dalis kolokacijų nustatyta vienu, dalis – kitu, dalis – abiem metodais. Iš aštuonių būdvardžio *protingas* kolokacijų (*protingas sprendimas*, *protingas miestas*, *protingas vaikas*, *protingas vyras*, *protingas laikas*, *protingas būdas*) kaip AK pripažintos šešios: šio būdvardžio kolokacijos su daiktavardžiais *sprendimas* ir *miestas*, atrinktos neautomatiškai, kolokacijos su *vaikas* ir *vyras*, atrinktos automatizuotai, o kolokacijos su *laikas* ir *būdas* atrinktos abiem metodais.

Lingvistai nustatė, kad semantinė grupė „asmenys“ yra įprasta šiam būdvardžiui. Su šios grupės žodžiais gali būti pasirinktinai derinami panašios reikšmės būdvardžiai (gali būti

ne tik *protingas*, bet ir *išmintingas žmogus*), todėl *protingas vaikas* ir *protingas vyras* tarp AK pateko tik kaip automatinės atrankos rezultatas (kolokacija *išmintingas žmogus Delfi.lt* tekстыne neužfiksuota), vėliau, peržiūrint duomenis, jos atmestos. Lingvistai pamatė, kad kiti daiktavardžiai (*sprendimas, laikas, būdas*) daugiausia priskirtini leksinei semantinei „abstraktų“ grupei, bet vis dėlto kolokacijos sudaromos tik su tam tikrais abstrakčiaisiais daiktavardžiais, vadinasi, pasirinkimo galimybės šioje grupėje ribotos, o tai rodo arbitralumą. Todėl visos šios kolokacijos pripažintos arbitraliosiomis. Tik dvi iš jų (*protingas laikas* ir *protingas būdas*) kaip AK nustatytos automatizuotai, nes, taikant vektorių metodą, netaikomas leksinių semantinių grupių testas.

Neautomatiniu metodu nustatytos AK. Pavyzdžiui, su būdvardžiu *aukštas* arbitraliosiomis pripažintos devynios kolokacijos: *aukštas rezultatas, aukštas pareigūnas, aukšta kaina* ir kt. Čia būdvardis vartojamas netiesiogine reikšme (ir beveik tik su abstrakčiaisiais daiktavardžiais), taigi perkeltinės reikšmės kriterijus lėmė, kad lingvistai šias kolokacijas priskyrė prie AK. Automatizuotu metodu labai dažnai šių kolokacijų neaptikta, nes sinonimų pakeičiamumo testui neturėta VE, kuriose būtų atitinkamas būdvardis (*aukštas, uždaras, skaudus, silpnas, platus, patogus, lengvas, laisvas, ilgas ar greitas*), todėl nebuvo galima įvertinti pakeičiamumo. Kaip jau minėta ankstesnėje dalyje, automatizuoto metodo rezultatai labai priklauso nuo to, ar turima sinonimų pakeičiamumui patikrinti ir kokia sinonimų eilės kokybė. Šis atvejis iliustruoja, kad, taikant tik automatizuotą metodą, dalis AK gali likti neatpažintos, o lingvistinė kompetencija gali kompensuoti tekstyno duomenų trūkumą.

Dar viena priežastis, kodėl kai kurios AK automatiškai gali būti neatpažintos, – su tam tikros reikšmės sinonimų eile gali būti susieti ne visi daiktavardžiai, nes dalies daiktavardžių kontekstas gali skirtis. Pavyzdžiui, tarp būdvardžio *ryškus* kolokacijų daug tokių, kurios nustatytos neautomatiškai (pvz.: *ryški moteris, ryškus žmogus*) ir tik viena AK nustatyta abiem metodais – *ryškus įvykis*. Tai reiškia, kad VE su šiuo būdvardžiu buvo, bet su ja automatiškai susieti ne visi trys daiktavardžiai, o tik *įvykis*.

Automatizuotu metodu nustatytos AK. AK su būdvardžiu *efektyvus* nustatytos tik vektorių metodu, pvz.: *efektyvi priemonė, efektyvus rezultatas, efektyvi sistema, efektyvus darbas*. VE buvo daugiau sinonimų (pvz., *veiksmingas*), tačiau kolokacijos *veiksminga priemonė, veiksmingas rezultatas* ir kt. tekstųne neužfiksuotos, todėl, pritaikius negalimo

pakeičiamumo testą, kolokacijos su *efektyvus* pripažintos AK. Čia svarbu pridurti, kad lingvistai, remdamiesi gimtakalbių kompetencija, šiuo atveju įvertino kitaip: pakeičiamumas yra galimas (įmanoma vartoti ir *efektyvi priemonė*, ir *veiksminga priemonė*), todėl nuspręsta šių kolokacijų nelaikyti AK. Automatizuotai AK pripažinta daug tokių kolokacijų, kurių būdvardis yra vertinamasis, plg.: *išskirtinis* (nes sinonimiškai susietas su *ypatingas*), *keistas* (susietas su *neįprastas*), *modernus* (susietas su *šiuolaikiškas*), *pagrindinis* (susietas su *esminis*), *sudėtingas* (susietas su *sunkus*). Lingvistai tokių atvejų dažniau buvo linkę nepriskirti prie AK.

Kolokacijos su *stabilus*, pavyzdžiui, *stabilus gyvenimas*, *stabili kaina*, *stabili savaitė*, *stabili valstybė*, kaip arbitraliosios pripažintos tik vektorių metodu, nes pakeičiamumas tikrintas turint sinonimą *pastovus*, kuris kaip kolokatas nepavartotas su atitinkamais daiktavardžiais. Vis dėlto lingvistai šių kolokacijų kaip arbitraliųjų nepripažino, nes, žiūrint į *stabilus* junglumą, matyti, kad jis tikrai platus, ir apribojimų nenustatyta, o *pastovus* nelaikytas *stabilus* sinonimu, plg.: *stabilus darbas*, *stabili valstybė*, *stabili kaina*, *stabilus rezultatas*, *stabilus gyvenimas*, *stabili savaitė*.

Abiem metodais nustatytos AK. Tarp turimų duomenų nėra nė vieno atvejo, kad visos to paties būdvardžio kaip AK pripažintos kolokacijos būtų nustatytos abiem metodais. Visais atvejais abiem metodais nustatyta tik jų dalis, o tarp tų būdvardžių dažniausiai yra tokių, kurie gali būti vartojami netiesiogine reikšme, pvz.: *garsus*, *greitas*, *liūdnas*, *mažas*, *nelengvas*, *patrauklus*, *pavojingas*, *rimtas*, *stiprus*, *sunkus*, *tvirtas*. Kaip jau minėta, daugumos tokių būdvardžių AK nustatytos tik neautomatizuotu metodu, pavyzdžiui, kolokacijos su būdvardžiu *silpnas* (taip pat žr. anksčiau minėtus pavyzdžius). Taip yra todėl, kad, AK nustatant neautomatiškai, tokie atvejai atrinkti pagal perkeltinės reikšmės kriterijų, o taikant 2 metodą jie tiko tik jeigu buvo sinonimų eilės, naudojamos pakeičiamumo testui, nariai.

Patikrinus, kiek atskirais metodais surenkama kolokacijų vienam iš produktyviausių daiktavardžių *sprendimas*, matyti panašios kaip ir iš 3 lentelės tendencijos: neautomatiniu metodu gauta daugiausia AK, tačiau yra nemaža dalis AK, kurios nustatytos automatizuotai.

4 LENTELĖ. Nustatytos AK su daiktavardžiu *sprendimas* ir būdvardžiais

Neautomatiniu metodu nustatytos AK	Automatizuotu metodu nustatytos AK	Abiem metodais nustatytos AK
1) atsakingas	1) aiškus	1) greitas
2) beprecedentis	2) kategoriškas	2) logiškas
3) drastiškas	3) keistas	3) nelengvas
4) drąsus	4) konkretus	4) rimtas
5) formalus	5) kūrybiškas	5) rizikingas
6) griežtas	6) naudingas	6) sunkus
7) išmanus	7) netikėtas	7) žalingas
8) išmintingas	8) optimalus	8) tvirtas
9) konstruktyvus	9) paprastas	
10) kūrybingas	10) prastas	
11) kvailas	11) realus	
12) laikinas	12) sudėtingas	
13) lemiamas	13) unikalus	
14) lemtingas		
15) lengvas		
16) nelogiškas		
17) neracionalus		
18) pagristas		
19) patogus		
20) pažangus		
21) praktiškas		
22) priešingas		
23) protingas		
24) racionalus		
25) sąmoningas		
26) saugus		
27) savarankiškas		
28) skandalingas		
29) skausmingas		
30) skubotas		
31) skubus		
32) sumanus		
33) taikus		
34) teisingas		
35) tvarus		
36) vieningas		

Pateikiant daiktavardžio *sprendimas* pavyzdį, svarbu paminėti dar vieną su automatizuotu metodu susijusį aspektą: dalis arbitraliosiomis priskirtų kolokacijų gali būti diskutuotinos, t. y. gali atrodyti lyg esančios ant ribos, arčiau netikrųjų. Pavyzdžiui, gali kilti klausimas, ar tokios kolokacijos kaip *aiškus sprendimas* ar *keistas sprendimas* tikrai laikytinos arbitraliosiomis. Atrodytų, kad būdvardžiai *aiškus*, *keistas*, *kūrybiškas*, *naudingas*, *netikėtas* ir kiti gali būti derinami ne tik su *sprendimas*, bet ir su daug kitų daiktavardžių, su kuriais jie gali būti jungiami pagal reikšmę, plg. visas šių būdvardžių kolokacijas, kurios nustatytos *Delfi.lt* tekstyne: *aiškus ženklas*, *aiškus vaizdas*, *aiški informacija*, *aiškus rezultatas*, *aiški sistema*, *aiškus dalykas*; *keistas dalykas*, *keistas būdas*, *keistas klausimas*, *keista vieta*, *keistas sprendimas*, *keistas įvykis*, *keistas vaizdas*. Lingvistai tikrino ribotą leksinį junglumą ne tik sinonimijos testu, bet ir pagal leksines semantines grupes, todėl kolokacijų *aiškus sprendimas* ir *keistas sprendimas* prie arbitraliųjų nepriskyrė. Vertinant leksinį ribotumą automatizuotai, šios kolokacijos prie AK priskirtos, nes tikrinant sinonimų pakeičiamumą nustatyta, kad daiktavardis *sprendimas* sudarė kolokacijas tik su vienu iš sinonimų poros *aiškus*, *suprantamas* ir sinonimų poros *keistas*, *neįprastas* narių. Šie pavyzdžiai rodo, kad tikėtis tikslesnių AK rezultatų galima ribotam leksiniam junglumui patikrinti taikant ne tik sinonimijos, bet ir leksinių semantinių grupių testą.

2.2. Arbitraliosios kolokacijos, sudarytos iš veiksmažodžių ir daiktavardžių

Nustatyta 1171 AK, sudaryta iš veiksmažodžio ir daiktavardžio. Tai gana panašus skaičius į AK, sudarytą su būdvardžiais, skaičių – 1048. Taip pat gana panašiai, palyginti su kolokacijomis, sudarytomis su būdvardžiais (žr. 2 lentelę), pasiskirsto skirtingais metodais nustatytos AK (žr. 3 lentelę): maždaug pusė AK nustatyta neautomatiniu būdu (pvz.: *apniukusi diena*, *aptarti galimybes*, *iškviesti apsaugą*), apie trečdalis – automatizuotai (pvz.: *suskubti į pagalbą*, *žvelgti į gyvenimą*, *sugadintas turtas*, *kaupiti turtą*), o beveik penktadalis nustatytos abiem metodais (pvz.: *atlikti tarnybą*, *duoti komandą*, *matyti reikalą*).

Analizuota ir šios struktūros AK leksinė įvairovė, t. y. kiek skirtingų veiksmažodžių sudaro tam tikru metodu nustatytas AK. Su tuo pačiu veiksmažodžiu sudaryta vidutiniškai 2,05 šiame poskyryje aptariamos AK (žr. 5 lentelę).

5 LENTELĖ. Skirtingais metodais nustatytos AK (veiksmažodis + daiktavardis)

Neautomatiniu metodu nustatytos AK		Automatizuotu metodu nustatytos AK		Abiem metodais nustatytos AK	
Sk. (proc.)	Skirtingų veiksmažodžių skaičius (vidutinis skaičius AK, kurios sudarytos su tuo pačiu veiksmažodžiu)	Sk. (proc.)	Skirtingų veiksmažodžių skaičius (vidutinis skaičius AK skaičius, kurios sudarytos su tuo pačiu veiksmažodžiu)	Sk. (proc.)	Skirtingų veiksmažodžių skaičius (vidutinis AK skaičius, kurios sudarytos su tuo pačiu veiksmažodžiu)
561 (47,91)	376 (1,5)	382 (32,62)	189 (2,02)	228 (19,47)	102 (2,24)

AK veiksmažodžių sintaksinės funkcijos įvairesnės nei būdvardžių: gali būti vartojamos asmenuojamosios formos, prisijungiančios objekto, subjekto ar aplinkybių funkciją atliekančius daiktavardžius (pvz.: *apipilti klausimais, kelti kainą, parodyti laikas*¹³, *pastatyti į vietą*), arba neasmenuojamosios, dažniausiai dalyviai, atliekantys atributo funkciją (pvz.: *išpūsta kaina, einamieji metai, neapgalvotas sprendimas, klestinti šalis, negirdėtas dalykas, vykdomas darbas*). Trižodėse AK gali varijuoti žodžių tvarka, pvz.: *atskubėti į pagalbą / į pagalbą atskubėti; eiti į rinkimus / į rinkimus eiti*.

Lyginant su kolokacijomis, kurias sudaro būdvardžiai, matyti didesnė skirtingų veiksmažodžių įvairovė: neautomatiniu metodu nustatytose AK tas pats veiksmažodis sudaro vidutiniškai 1,5 kolokacijos, automatizuotu vektorių metodu nustatytose AK – vidutiniškai 2,02 kolokacijos, o abiem metodais nustatytose AK veiksmažodžių įvairovė mažiausia – tas pats veiksmažodis vidutiniškai sudari 2,24 AK. Tikėtina, kad tokį rezultatą lėmė tai, kad abiem metodais nustatytų kolokacijų yra mažiausiai. Kita vertus, tendencijos panašios: didžiausia leksinė įvairovė pasižymi neautomatiniu metodu nustatytos AK.

¹³ Kartais subjektinių kolokacijų lemos atrodo neįprastos, bet pasirinkta pateikti bendratį, nes gali būti vartojamos labai įvairios asmenuojamosios formos, pvz.: *parodys laikas, parodė laikas, parodytų laikas*.

Analizuojant duomenis, matyti didelė veiksmazodžių įvairovė (skirtingais veiksmazodžiais laikyti žodžiai su priešdėliu ir be priešdėlio, su skirtingais priešdėliais (pvz.: *saugoti, apsaugoti, išsaugoti*), sangražinės ir nesangražinės formos, jei nekeičia reikšmės, laikytos tuo pačiu veiksmazodžiu, bet yra junglių veiksmazodžių, su kuriais sudaryta po kelias ar net keliolika AK, pvz.:

suteikti: *suteikta apsauga, suteikta informacija, suteikti galimybę, suteikti galimybę rinktis, suteikti informacijos apie;*

skirti: *skirti mažai dėmesio, skirti laiko, skirti dėmesį, skirti daug dėmesio, skirtas namų areštas, didelį dėmesį skirti, daug laiko skirti, daug dėmesio skirti, dėmesį reikėti skirti;*

ateiti / ateinantys: *ateiti į rinkimus, ateiti euras, ateiti laikas, ateiti metas, ateinantys metai, ateinantys rinkimai, ateinančių metų biudžetas, ateinančių metų pradžia, ateinantis mėnuo, ateinančios savaitės pradžia, ateinantis sezonas, ateinantys metai, ateinanti diena, ateinanti savaitė.*

Matyti, kad nemaža su vienu veiksmazodžiu sudarytų AK dalis yra vienos kolokacijos variacijos, plg. anksčiau pateiktus AK *skirti dėmesį* variantus. Vis dėlto bet kokios leksinės, gramatinės kolokacijų variacijos (keičiasi žodžių tvarka, pridedama papildomų žodžių) skaičiuojamos kaip atskiros kolokacijos. Atskiromis kolokacijomis laikoma, jei tas pats veiksmazodis vartojamas asmenuojamąja (tokiu atveju kaip lema dažniausiai pateikiama bendratis) arba neasmenuojamąja forma (pvz.: *kovoti dėl vietos, kovoti iki pabaigos, kovojanti grupė*). Taigi, kolokacijose su veiksmazodžiais didesnę leksinę įvairovę (palyginti su būdvardžiais) lemia veiksmazodžio kaip kalbos dalies ypatumas – dažnas tos pačios šaknies veiksmazodžio vartojimas su skirtingais priešdėliais. Viena vertus, tokius veiksmazodžius galima vertinti kaip skirtingus žodžius, kita vertus, tokius duomenis ne visai įmanoma palyginti su būdvardžių, kurie visais atvejais yra skirtingi žodžiai, o ne darybiniai to paties žodžio variantai, duomenimis.

Neautomatiniu metodu nustatytos AK. Kiek kitaip nei būdvardžių + daiktavardžių kolokacijos, kolokacijos su veiksmazodžiais dažnai pripažintos kaip arbitraliosios taikant ne tik perkeltinės reikšmės kriterijų (pvz.: *gadinti gyvenimą, atverti galimybę*), bet ir riboto leksinio junglumo kriterijų ir leksinių semantinių grupių testą. Pavyzdžiui, kolokacijos su *vaizdas, būdas, informacija* vertintos turint tokias paradigmas:

iškraipyti informaciją

iškraipyti rinką

iškreipta informacija

iškreiptas būdas

iškreiptas vaizdas

Veiksmažodžių *iškraipyti*, *iškreipti* (vartojamų ir atributiškai) junglumas nėra platus: prie kolokacijų su *iškraipyti* paradigmos dar būtų galima pridėti *iškraipyti žodžius*, bet daiktavardis *žodžiai* patektų į vieną grupę su *informacija*, todėl priimtas sprendimas abi kolokacijas priskirti prie AK. Su atributiškai vartojamu dalyviu *iškreiptas* sudarytos trys kolokacijos ir visos laikytos AK, nes, nors galima išplėsti paradigmą (pvz., pridėti *iškreiptas mąstymas*), ji vis tiek būtų ribota. Daiktavardžiai *informacija*, *vaizdas*, *būdas*, kad ir priskirti vienai abstraktų leksinei semantinei grupei, yra vos trys, o nedidelis žodžių skaičius yra viena sąlygų taikant leksinių semantinių grupių testą. Pritaikius sinonimų pakeičiamumo testą, paaiškėja, kad vietoj *iškraipyti*, *iškreiptas* netinka pavartoti, tarkim, *išdaryti* / *išdarytas*.

Automatizuotu metodu nustatytos AK. Svarbu pabrėžti vieną automatizuoto metodo ribotumą: dalis AK gali būti neatpažinta, nes tokios kolokacijos nenustatytos *Delfi.lt* tekстыne (nors pats žodžių junginys galėjo būti pavartotas). Tarkim, trižodė kolokacija *aiškintis įvykio aplinkybes* pripažinta AK pagal į VE patekusius sinonimus *analizuoti*, *aiškintis*, *tirti*. Tekстыne rasta tik kolokacija su veiksmažodžiu *aiškintis*, todėl pripažinta, kad ji yra AK. Kita vertus, kitame tekстыne galėtų būti ir kolokacija *tirti įvykio aplinkybes*, taigi, testuojant pagal tą pačią sinonimų grupę, arbitraliosiomis būtų pripažintos dvi, o ne viena kolokacija.

Tarp automatizuotu metodu nustatytų kolokacijų yra tokių atvejų, kurie lingvistų nepripažinti kaip AK kaip tik dėl to, kad, remiantis kalbos jausmu, nemanyta, jog esama polinkio rinkti vieną sinonimą iš kelių galimų. Pavyzdžiu galėtų būti kolokacijos *suprasti pasaulį*, *pažinti pasaulį*, *žiūrėti į pasaulį*. Nė vienos iš šių kolokacijų lingvistai nepripažino kaip arbitraliosios. Automatizuotu metodu nustatyta, kad visos trys yra AK, nes tarp kolokacijų iš *Delfi.lt* tekstyno nebuvo kolokacijų su kitais atitinkamų sinonimų grupių nariais, t. y. buvo *suprasti pasaulį*, bet nebuvo *suvokti pasaulį*; buvo *pažinti pasaulį*, bet nebuvo *išmokti pasaulį*; buvo *žiūrėti į pasaulį*, bet nebuvo *žvelgti į pasaulį*. Lingvistiškai vertinant nustatyta, kad šių trijų veiksmažodžių junglumas neturi riboto junglumo požymių.

Abiem metodais nustatytos AK. Reikia pasakyti, kad šios grupės AK nustatytos nebūtinai pagal tą patį kriterijų. Vertinant arbitralumą neautomatiniu metodu, AK galėjo būti nustatyta pagal kurį nors vieną testą, nebūtinai pagal sinonimų pakeičiamumą. Kaip rašyta apie būdvardžių ir daiktavardžių kolokacijas, vienodos AK neautomatiniu būdu atrinktos pagal perkeltinės reikšmės kriterijų, o automatizuotu metodu – pagal riboto leksinio junglumo kriterijų (sinonimų pakeičiamumo testą), plg. *sutriko eismas* ir *sutriko sveikata* lingvistų atrinktos dėl to, kad *sutrikti* jungiasi su ribota daiktavardžių grupe, o 2 metodu šios kolokacijos pripažintos AK dėl to, kad iš sinonimų *sugedo* ir *sutriko* vertinamosiose kolokacijose užfiksuotas tik vienas sinonimas. *Apgalvotas sprendimas* pripažinta AK, nes nebūta kolokacijos su kitu sinonimu – *apmąstyta sprendimas*. Taigi, svarbu turėti omenyje, kad šios grupės kolokacijos nėra pagal tą patį kriterijų atpažintos kolokacijos.

6 LENTELĖ. Nustatytos AK su daiktavardžiu *sprendimas* ir veiksmazodžiais

Neautomatiniu metodu nustatytos AK	Automatizuotu metodu nustatytos AK	Abiem metodais nustatytos AK
1) apskūsti (teismo) sprendimą 2) atidėti sprendimą 3) daryti sprendimą 4) grįsti sprendimą 5) ieškoti sprendimo 6) įsiteisėjęs sprendimas 7) neapgalvotas sprendimas 8) nepriimti sprendimo 9) nevykdyti teismo sprendimo 10) sprendimas pasiteisino 11) priimti atitinkamą sprendimą priimti reikiamą sprendimą 12) priiminėti sprendimą 13) pritarti sprendimui 14) skelbti sprendimą 15) skūsti sprendimą (teismui) 16) vykdyti teismo sprendimą	1) priimti sprendimą 2) reikalauti sprendimo 3) siūlomas sprendimas 4) laukti sprendimo 5) pasirašyti sprendimą	1) įvykdyti teismo sprendimą 2) padarytas sprendimas 3) pasiūlytas sprendimas 4) paskelbti sprendimą 5) apgalvotas sprendimas

Įvertinus, kiek skirtingais metodais nustatoma vieno iš produktyviausių daiktavardžių *sprendimas* kolokacijų veiksmažodis + daiktavardis (žr. 6 lentelę), matyti, kad daugiausia jų nustatyta neautomatiniu metodu, o abiem metodais ir automatizuotu metodu nustatyta tik po penkias AK.

APIBENDRINAMOSIOS PASTABOS

Išanalizavus dvižodes ir trižodes AK, sudarytas iš būdvardžių ir daiktavardžių, taip pat iš veiksmažodžių ir daiktavardžių, galima teigti, kad neautomatiniu metodu jų nustatyta apie pusė, automatizuotu metodu – apie trečdalis, o abiem metodais – maždaug penktadalis. Sutampa tik nedidelė šiais metodais nustatomų AK dalis, o tai leidžia teigti, kad abiem metodais galima nustatyti daugiau AK. Vadinasi, siekiant atpažinti AK, verta naudotis ir neautomatinio, ir automatizuoto metodo teikiamomis galimybėmis.

Pabrėžtina, kad atliekant neautomatinę analizę taikytas ne vienas, o du atrankos kriterijai, todėl neautomatiškai nustatytų duomenų daug daugiau negu nustatytų automatizuotai. Neautomatinė analizė apėmė ne tik perkeltinės reikšmės kriterijų, bet ir riboto leksinio junglumo kriterijų, tikrintą dviem testais (leksinių semantinių grupių ir sinonimų pakeičiamumo), o taikant automatizuotą metodą taikytas tik riboto leksinio junglumo kriterijus pagal sinonimų pakeičiamumo testą.

Dėl skirtingo kriterijų skaičiaus skirtingais metodais gautų rezultatų negalima palyginti, nes pirmuoju metodu AK bet kuriuo atveju nustatyta daugiau, o sutampančių kolokacijų palyginti mažai. Būtų įdomu palyginti tik pagal bendrą riboto junglumo kriterijų (sinonimų pakeičiamumo testą) automatizuotai ir neautomatiškai nustatytas AK. Toks palyginimas galėtų parodyti, kad dalis nustatytų AK yra tos pačios, bet būtų galima įvertinti, kuriuo metodu AK atrinkta daugiau, kuo skiriasi nesutampančios AK.

Neautomatiniu metodu surinkta ne tik daugiau, bet ir leksiškai įvairesnių duomenų. Šiuo metodu nustatytas AK sudaro įvairesni būdvardžiai ir veiksmažodžiai, o automatizuotu metodu gaunamos mažesnės leksinės įvairovės AK.

Išbandant automatizuotą metodą, labiausiai siekta sumažinti subjektyvumo veiksnį vertinant arbitralumą ir remtis tik tekстыne užfiksuotais duomenimis. Turint sinonimų poras ar grupes, nustatytas vektorių metodu iš konkretaus tekstyno, vertinimo procedūra supaprastinama, nes lengviau nuspręsti, ar kolokacijos dėmuo gali būti pakeičiamas, ar ne: jeigu tiriamojoje medžiagoje yra kolokacijų su visais sinonimų poros ar grupės žodžiais, pakeičiamumas galimas.

Žinoma, išryškėjo ir automatizuoto metodo trūkumų. Kadangi vertinamos kolokacijos arbitralumas nustatomas remiantis VE sinonimais, dalis AK gali būti neatpažinta, nes arbitralumo pripažinimas labai priklauso nuo turimos VE. Testuoti naudojama tik tiek ir tik tokių sinonimų, kokie surinkti VE, dėl to kaip AK pripažįstamos ne visos kolokacijos.

Nors AK analizės procesą galima automatizuoti, pirminiame šio proceso etape reikalingas lingvistų indėlis – vektorių metodu sudarytose VE atskirti sinonimus nuo kitais semantiniais ryšiais susijusių žodžių, pavyzdžiui, antonimų. Patikrinus, kiek po lingvistinės patikros lieka automatiškai gautų duomenų, matyti, kad iš VE pašalintų būdvardžių ir veiksmažodžių yra daug, o likę atitinkamai 41 procentas ir 30 procentų duomenų taip pat redaguotini, tvarkytini. Kadangi vektorių metodu nustatomi ne tik sinoniminiais ryšiais susiję žodžiai, paruošti sinonimų duomenis pakeičiamumo testui užtrunka ilgiau. Todėl reikėtų išbandyti kitus metodus, skirtus sinonimijos ryšiams nustatyti.

Vertinant, kiek tinkamas sinonimų atrankai vektorių metodas, minėtini tokie automatiškai surinktų duomenų vertę įrodantys argumentai. Viena, vektorių metodu galima nustatyti daugiau ir kitokių sinonimų nei vien tik pasitelkiant introspekciją. Antra, šiuo metodu nustatyti sinonimai (žinoma, prieš tai atrinkus nesinonimus) atskleidžia sinonimijos požiūriu svarbų reiškinį – sinonimiški ne tik pavieniai žodžiai, bet ir žodžių junginiai. Trečia, vektorių metodas tinka dideliame duomenų kiekiui apdoroti, t. y. nustatyti sinonimus iš didelio tekstyno, o rankiniu tokios gausybės žodžių konteksto panašumą būdu sudėtinga įvertinti.

LITERATŪRA

- Bielinskienė Agnė, Boizou Loic, Bumbulienė Ieva, Kovalevskaitė Jolanta, Krilavičius Tomas, Mandravickaitė Justina, Rimkutė Erika, Vilkaitė-Lozdienė Laura (sud.) 2019: *Lietuvių kalbos kolokacijų žodynas*, Kaunas: Vytauto Didžiojo universitetas. Prieiga internete: <https://doi.org/10.7220/9786094673733>.
- Divjak Dagmar, Gries Stefan T. 2006: Ways of Trying in Russian. Clustering and Comparing Behavioral Profiles. – *Corpus Linguistics and Linguistic Theory* 2(1), 23–60.
- Drūlienė Viltė 2008a: *Erdivinės metaforos Daiktavardinių frazių žodyne*: magistro darbas, Kaunas: Vytauto Didžiojo universitetas.
- Drūlienė Viltė 2008b: Lietuviškosios metaforikos vertikalumo matas. – *Darbai ir dienos* 50, 73–90.
- Grigonytė Gintarė 2010: *Building and Evaluating Domain Ontologies: NLP Contributions*: daktaro disertacija, Berlin: Logos Verlag Berlin.
- Handl Susanne 2008: Essential Collocations for Learners of English: The Role of Collocational Direction and Weight. – *Phraseology in Foreign Language Learning and Teaching*, eds. F. Meunier, S. Granger, Amsterdam, Philadelphia: John Benjamins Publishing Company, 43–66.
- Howarth Peter 1998: Phraseology and Second Language Proficiency. – *Applied Linguistics* 19 (1), 24–44.
- Jakaitienė Evalda 2010: *Lietuvių kalbos leksikologija*, Vilnius: Vilniaus universiteto leidykla.
- Kovalevskaitė Jolanta, Rimkutė Erika, Vilkaitė Laura 2020: Light Verb Constructions in Lithuanian: Identification and Classification. – *Kalbų studijos* 36, 5–16.
- Kovalevskaitė Jolanta, Rimkutė Erika, Vaičenonienė Jurgita 2021: Automatizuotas arbitraliųjų kolokacijų atpažinimas: būdvardžių ir daiktavardžių kolokacijos. – *Kalbų studijos* 39, 71–84.
- Lewis Michael 2000: *Teaching Collocation: Further Developments in the Lexical Approach*, London: Language Teaching Publications.
- Marcinkevičienė Rūta 2010: *Lietuvių kalbos kolokacijos*, Kaunas: Vytauto Didžiojo universitetas.
-

- Nesselhauf Nadja 2003: The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. – *Applied Linguistics* 24 (2), 223–242.
- Nesselhauf Nadja 2005: *Collocations in a Learner Corpus*, Amsterdam: John Benjamins.
- Pennington Jeffry, Socher Richard, Manning Christopher D. 2014: GloVe: Global Vectors for Word Representation. – *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNL)*. Association for Computational Linguistic. Doha, Qatar, October 25–29, 2014, 1532–1543.
- Reder Anna 2006: *Kollokationen in der Wortschatzarbeit*. Praesens Verlag.
- Strakatova Yana, Hinrichs Erhard 2019: Semantic Modelling of Adjective–Noun Collocations Using FrameNet. – *Proceedings of the Joint Workshop on Multiword Expressions and WordNet*, 104–113.
- Stulpinaitė Miglė, Horbačauskienė Jolita, Kasperavičienė Ramunė 2016: Issues in Translation of Linguistic Collocations. – *Kalbų studijos* 29, 31–41.
- Šilinytė Justina 2008: *Ontologinės ir struktūrinės metaforos Daiktavardinių frazių žodyne*: magistro darbas, Kaunas: Vytauto Didžiojo universitetas.
- Volungevičienė Skaistė 2010: Metaforinių kolokacijų vertimo problemos. – *Kalbų studijos* 16, 23–27.

Gauta 2021 11 14

Priimta 2021 12 15

ARBITRALIŲJŲ LIETUVIŲ KALBOS KOLOKACIJŲ NUSTATYMAS*Santrauka*

Šiame straipsnyje sprendžiama arbitraliųjų lietuvių kalbos kolokacijų atpažinimo problema. Tam skirtas projektas „Arbitraliosios lietuvių kalbos kolokacijos: atpažinimas, aprašymas, vartojimas (ARKA)“. Arbitraliosiomis kolokacijomis laikomos kolokacijos, kurių žodžių junglumo pagrindas yra ne semantiniai, o leksiniai ribojimai. Tokios kolokacijos tiesiogiai neatspindi nekalbinės tikrovės, o ryšys tarp jų žodžių – arbitralus: nors pagal panašią reikšmę teoriškai galima rinktis daugiau negu vieną junglumo partnerį, bet linkstama vartoti kurį nors vieną, pvz., *suprasti užuominą*, bet ne **suvokti užuominą*.

Tyrimo šaltinis yra Lietuvių kalbos pastoviųjų žodžių junginių duomenų bazė, kurioje pateikiama apie 12 tūkst. kolokacijų; ištyrus kolokacijas su 97 daiktavardžiais, arbitraliosiomis priskirta apie 2400 kolokacijų, kurių struktūra yra būdvardis / būdvardiškai vartojamas dalyvis + daiktavardis, veiksmažodis + daiktavardis, daiktavardis + daiktavardis.

Straipsnyje aprašyti du pagal minėtą projektą išbandyti arbitraliųjų kolokacijų nustatymo metodai: neautomatinis lingvistinės analizės ir automatizuotas kompiuterinės lingvistikos metodas. Nustatant AK neautomatiškai, taikyti du pagrindiniai kriterijai: perkeltinės reikšmės kriterijus ir riboto leksinio junglumo kriterijus, kuriam įvertinti taikyti veiksmažodžio ar būdvardžio sinonimų pakeičiamumo ir (arba) daiktavardžių leksinių semantinių grupių testai. Vienas iš AK nustatymo kriterijų – riboto leksinio junglumo – buvo tikrinamas ir automatizuotai, vektorių metodu nustačius pakeičiamumo testui reikalingus sinonimus. Automatizuoto metodo taikymą sudarė šie etapai: 1) automatinis, kai vektorių metodu buvo surenkamos vektorių eilės su panašų kontekstą turinčiais žodžiais; 2) neautomatinis – vektorių eilių tvarkymas paliekant tik sinoniminiais ryšiais susijusius žodžius; 3) iš dalies automatinis etapas, kai kolokacijos buvo automatiškai susiejamos su konkrečiomis sinonimų eilėmis ir po peržiūros lingvistų patvirtinamos arba nepatvirtinamos kaip AK.

Neautomatiniu metodu nustatyta apie pusė AK, automatizuotu metodu – apie trečdalis. Neautomatinei analizei taikytas ne vienas, o du atrankos kriterijai, todėl neautomatiškai nustatytų duomenų daug daugiau negu nustatytų automatizuotai. Nors

automatizuotu metodu galima sumažinti subjektyvumą vertinant duomenis ir supaprastinti arbitralumo vertinimo procedūrą, tačiau pastebėta ir tai, kad arbitralumo pripažinimą lemia vektorių eilėje esantys sinonimai, taigi dalis AK gali likti neatpažintos. Vektorių metodu surinktus sinonimus reikia atskirti nuo kitais semantiniais ryšiais susijusių žodžių, tam reikalinga lingvisto patikra. Tačiau, nepaisant šio ribotumo, vektorių metodas yra tinkamas dideliame duomenų kiekiui apdoroti, o rankiniu įvertinti gausybės žodžių konteksto panašumą būtų sunku.

Abiem metodais atpažintos kolokacijos sudaro tik penktadalį visų nustatytų AK, todėl yra pagrindo manyti, kad siekiant atpažinti AK svarbu taikyti tiek automatizuoto, tiek ir neautomatinio metodo teikiamas galimybes.

IDENTIFICATION OF LITHUANIAN ARBITRARY COLLOCATIONS*Summary*

In this article, we describe the methodological approaches of arbitrary collocation (AC) identification developed within the framework of an ongoing project “Arbitrary Collocations of Lithuanian: Identification, Description and Usage (ARKA)”. The object of this research is arbitrary collocations of the Lithuanian language the collocability of which is determined by lexical rather than semantic constraints. Such collocations do not directly reflect the non-linguistic reality and the relation between the collocates is arbitrary: although there may be several close synonyms, a particular one is preferred in a certain word combination (e.g., *suprasti užuominą* vs. **suvokti užuominą*).

The source of the research was the Database of Lithuanian Multiword Expressions encompassing over 12.000 collocations. The structural composition of the identified 2400 arbitrary collocations was as follows: adjective (participle)+noun, verb+noun, and noun+noun.

Arbitrary collocations were determined by combining the manual linguistic analysis and semi-automatic methods of computational linguistics. The manual method included two major AC identification criteria: lexical restrictedness and (or) meaning transfer. Lexical restrictedness was measured using two tests: the synonym substitution of pre-modifier and (or) semantic field comparison of the head noun. Moreover, lexical restrictedness was also assessed using a semi-automatic approach by analyzing synonym pairs in the automatically generated vector strings. The semi-automatic approach consisted of three stages: (1) the automatic generation of vector strings with potential synonyms; (2) the manual vector string editing to reduce the noise, and (3) the semi-automatic process during which the collocations were compared to particular synonym pairs in vector strings and approved or not approved by linguists as arbitrary.

Approximately half of ACs were detected by using the manual method based on two AC identification criteria, whereas about one-third of ACs were identified using the semi-automated method based on one criterion. On the one hand, the semi-automatic method can reduce the subjectivity in data evaluation and simplify the AC identification procedure. On

the other hand, as arbitrariness is determined on the basis of the available synonyms in the vector string, some ACs may not be identified because of the insufficient scope of the generated synonyms. Also, the automatically generated synonyms need to be manually analyzed to eliminate semantically unrelated words. Despite these limitations, the method is suitable for processing large amounts of the data. It would be difficult to manually assess the contextual similarity of such a large number of words.

About one fifth of ACs were identified using a combination of both methods. These results suggest that different methodological approaches allowed the researchers to detect more arbitrary collocations. It is maintained that taking into consideration the mentioned advantages and disadvantages, the best results in the Lithuanian arbitrary collocation identification are acquired when combining both manual and semi-automatic methods.

KEYWORDS: arbitrary collocations, Lithuanian, lexical restrictedness, vector space model, semi-automatic approach, linguistic manual approach.

JOLANTA KOVALEVSKAITĖ, ERIKA RIMKUTĖ,

JURGITA VAIČENONIENĖ

Vytauto Didžiojo universitetas

Kristijono Donelaičio g. 58, 44248 Kaunas

[*jolanta.kovalevskaite@vdu.lt*](mailto:jolanta.kovalevskaite@vdu.lt)

[*erika.rimkute@vdu.lt*](mailto:erika.rimkute@vdu.lt)

[*jurgita.vaicenoniene@vdu.lt*](mailto:jurgita.vaicenoniene@vdu.lt)