ISSN 2335-2019 (Print), ISSN 2335-2027 (Online)
Darnioji daugiakalbystė | Sustainable Multilingualism | *22*/2023
https://doi.org/10.2478/sm-2023-0010

Giedrė Valūnaitė Oleškevičienė
Mykolo Romerio universitetas, Lietuva
Vitalija Karaciejūtė
Mykolo Romerio universitetas, Lietuva
Dalia Gulbinskienė
Vilniaus Gedimino technikos universitetas, Lietuva

LIETUVIŲ KALBOS DISKURSO ŽYMIKLIAI IR JŲ RYŠIAI DAUGIAKALBIAME TEKSTYNE

Santrauka. Diskurso ryšiais anotuotų tekstynų kūrimas ir tyrimai yra gana nauja sritis, todėl Lietuvos mokslininkai siekia papildyti esamus tekstynų resursus ir ieško būdų, kaip būtų galima tyrinėti diskurso ryšius siejant ir lyginant juos su kitomis kalbomis, nes tam tikrais atvejais skirtingose kalbose diskurso ryšiai realizuojami skirtingomis kalbinėmis priemonėmis. Šio straipsnio tikslas – remiantis užsienio mokslininkų patirtimi pristatyti lietuvių kalbos diskurso žymiklių ir ryšių anotavimą daugiakalbiame tekstyne ir aptarti diskurso žymiklių vertimo tyrimų gaires, verčiant iš anglų kalbos į lietuvių kalbą. Todėl pirmiausia aptariamos diskurso ryšių ir juos išreiškiančių diskurso žymiklių raiškos galimybės lietuvių ir anglų kalbose, atskleidžiami galimi vertėjų pasirinkimai atsižvelgiant į diskurso ryšius vertime ir skirtingų kalbinių priemonių vartojimą pusiau paruoštoje monologinėje kalboje. Straipsnyje pristatomas lygiagretusis daugiakalbis tekstynas TED-MBD (angl. *Multilingual discourse-annotated corpus*), kuris yra anotuotas diskurso lygmeniu, laikantis PDTB (angl. *Penn Discourse Treebank*) diskurso anotavimo tikslų ir principų. Straipsnyje išsamiai aptariama PDTB diskurso žymiklių anotavimo sistema, skaitytojas supažindinamas su diskurso ryšių reikšmių hierarchija, anotavimo principais ir PDTB schemos taikymo įžvalgomis. Taip pat aprašomi lietuviškosios tekstyno dalies anotavimo principai, pritaikyti laikantis PDTB diskurso anotavimo taisyklių; aptariami pirmieji rezultatai, susiję su diskurso ryšių raiška ir diskurso žymiklių vartojimu. Straipsnyje taip pat pristatomos pirmosios tyrimų gairės, kaip galima palyginti diskurso ryšiais anotuotus lietuviškus ir angliškus tekstus, siekiant suprasti vertimo tendencijas diskurso lygmeniu.

Pagrindinės sąvokos: diskurso žymikliai; diskurso ryšiai; daugiakalbis tekstynas; anotavimas; PDTB diskurso ryšių hierarchija.

Įvadas

Diskurso ryšiais anotuotų tekstynų kūrimas, tyrimas ir taikymas yra nauja sritis, reikalaujanti naujų mokslininkų kompetencijų kuriant ir anotuojant tekstynus bei tiriant diskurso ryšiais anotuotų tekstynų taikymo galimybes. Efektyvus diskurso valdymas bet kurioje kalboje charakterizuojamas aiškiais ryšiais tarp diskurso vienetų ir rišlia, nuoseklia kalbos struktūra. Tačiau tam tikrais atvejais skirtingose kalbose diskurso ryšiai ir struktūra užtikrinama

skirtingomis kalbinėmis priemonėmis. Taip pat reikia pastebėti, kad didžioji dalis diskurso tyrimų iš pradžių buvo pagrįsta gramatiniais jungtukų sąrašais, apibūdinančiais tam tikras teksto jungimo ir rišlumo užtikrinimo funkcijas (Hunston, 2002). Vėliau tyrimai plėtėsi ir rėmėsi pragmatinėmis kategorijomis, diskurso žymikliai analizuoti pragmatiniais, sociolingvistiniais aspektais ir interpretuoti kaip pragmatiniai vienetai, turintys pragmatikalizacijai būdingų bruožų (Beeching, 2012; Degand, Evers-Vermeul, 2015; Brinton, 2017; Šinkūnienė et al., 2020). Taigi, norėdamos užpildyti šį žinių lauką, įvairių šalių mokslininkų komandos kuria savo kalbų diskurso žymiklių leksikonus, pavyzdžiui, Ch. Roze ir kt. (2010) sukūrė prancūzų kalbos diskurso žymiklių leksikoną. Lietuvos mokslininkai taip pat pradeda gilintis į lietuvių ir kitų kalbų lyginamuosius diskurso ryšių ir diskurso žymiklių (angl. *discourse markers*) tyrimus, remdamiesi tekstynų duomenimis (Šolienė, 2018; 2020a; 2020b; Šinkūnienė et al., 2020). Taigi šiame straipsnyje pristatomas diskurso ryšiais anotuotas daugiakalbis tekstynas TED-MDB (angl. *Multilingual discourse-annotated corpus*), sukurtas bendradarbiaujant su tarptautine mokslininkų bendruomene; siekiama sudominti Lietuvos mokslininkus ir suinteresuotus asmenis diskurso ryšių tyrimais ir tikimasi, kad atsiras tyrėjų, norinčių kurti lietuvių kalbos diskurso žymiklių leksikoną, kaip tai jau atlikta kitose kalbose. Toks tyrimas būtų žymus indėlis kuriant modernius išteklius lietuvių kalba, kurie būtų naudingi vertėjams bei vertimo studijų programų studentams. TED-MDB tekstyne pusiau paruoštą monologinę kalbą įvairia tematika iš anglų kalbos į lietuvių kalbą verčia savanoriai vertėjai, prižiūrimi institucijų, atsakingų už vertimo kokybę.

Diskurso ryšiai tekstyne

Diskurso žymikliai, arba jungiamieji diskurso elementai, sudaro funkcinę leksinių elementų kategoriją, kuri naudojama tarp teksto ar diskurso vienetų žymėti ryšiams, užtikrinantiems teksto rišlumą, pavyzdžiui, tokius kaip paaiškinimas, kontrastas ar kt. (Hallidėjus & Hassan, 1976; Mannas & Thomson, 1988; Knottas & Dalas, 1994; Sandersas, 2000). Dauguma kalbų turi tokių elementų rinkinius, bet labai skiriasi jungiamųjų elementų skaičius, jų vartojimas ir išreiškiami diskurso ryšiai. Be to, žinoma diskurso žymiklių

savybė, kad jie dažnai yra daugiafunkciai ir gali perteikti kelis diskurso ryšius. Kai kuriais atvejais tą patį jungiamąjį ryšį perteikia įvairūs diskurso žymikliai. M. Baker (2011), aptardama įvairių kalbų skirtumus, teigia, kad vienose kalbose diskurso ryšiai išreiškiami sudėtingomis, kitose – paprastesnėmis struktūromis, kur diskurso ryšiai tarp struktūrų yra aiškiai išreikšti. Kitaip tariant, vienose kalbose labiau priimtina informaciją pateikti mažesnėmis dalimis, naudojant išreikštus diskurso žymiklius diskurso ryšiams signalizuoti, o kitose – didelėmis diskurso dalimis, naudojant mažiau išreikštų diskurso žymiklių. Taigi kyla klausimas, kaip vertėjai susitvarko su diskurso ryšiais, kai šaltinio tekste yra daugybė išreikštų diskurso žymiklių, arba, atvirkščiai, kaip jie pateikia diskurso ryšius, kai šaltinio tekste yra ribotas diskurso žymiklių skaičius.

Diskurso žymikliai susiję su teksto logika ir interpretacija, taigi diskurso žymiklių derinimo procesas, atsižvelgiant į tikslinės kalbos specifiką ir tikslinės kalbos teksto tipą, yra sudėtingas procesas. Vertėjai gali pasirinkti tam tikras strategijas. Norėdami sklandaus ir aiškaus vertimo, jie gali bandyti įterpti papildomų diskurso žymiklių, net jeigu originaliame tekste jie nevartojami, arba gali pasirinkti originalaus teksto diskurso žymiklių vertimą pažodžiui, nors vertimas tiksline kalba gali atrodyti svetimas tos kalbos raiškai. Praktiškai vertėjai yra linkę pasirinkti arba dažniau naudoti vieną iš minėtų strategijų, arba ieškoti balanso ir naudoti šiek tiek visų paminėtų būdų (Baker, 2011).

Klasikinis diskurso žymiklių anotavimo metodas susideda iš nepriklausomo kelių anotatorių anotavimo, priskiriant reikšmę iš diskurso ryšių sąrašo tam tikram diskurso žymikliui. Paprastai tokias anotacijas vykdo daugiau nei vienas anotatorius, o vertinimo etape įvertinamas anotacijos patikimumas, išmatavus kelių anotavimų sutapimą.

Diskurso ryšiai gali būti anotuojami remiantis keliais žinomais diskurso modeliais, pavyzdžiui, retorinės struktūros teorija (angl. *Rhetorical Structure Theory* (RST)) (Mannas & Thompson, 1988) ir segmentinio diskurso pateikimo teorija (angl. *Segmented Discourse Representation Theory* (SDRT); Asheris & Lascarides, 2003). Tačiau šie du modeliai siekia pateikti išsamų teorinį diskurso santykių vaizdą, o anglų mokslininkų (Prasadas et al., 2000) sukurtas „Penn Discourse Treebank" (PDTB) leidžia labiau atsižvelgti į jungiamųjų elementų prasmę. Ši sistema leksiniu požiūriu remiasi į diskurso ryšius arba leksine

diskurso žymiklių reikšme grįstu požiūriu (net numanomi diskurso ryšiai yra išreiškiami galimu diskurso žymikliu) ir nedaromos prielaidos apie globalią diskurso struktūrą, todėl jis dar vadinamas teoriškai neutraliu požiūriu.

Vienas svarbiausių šaltinių, turinčių diskurso žymiklių anotaciją, yra „Penn Discourse Treebank" (Prasadas et al., 2008). PDTB pateikia „Wall Street Journal Corpus" (toliau – WSJ) tekstyno anotavimą diskurso lygmeniu. Diskurso anotavimas susideda iš rankiniu būdu anotuotų diskurso ryšių reikšmių – apie 100 rūšių tekste esančių diskurso žymiklių ir numanomų diskurso ryšių, siejančių diskurso argumentus. Visame WSJ tekstyne, apimančiame 1 000 000 anotuotų žymių, yra 18 459 anotuotų diskurso žymiklių, esančių tekste, ir 16 053 anotuotų numanomų diskurso ryšių. Reikšmės, kurias gali signalizuoti diskurso žymikliai, sudėtos į hierarchinę reikšmių struktūrą, kurią sudaro trys detalumo lygiai su keturiomis aukščiausio lygio reikšmėmis (laiko, priežasties, lyginamąja ir išplečiamąja), po kurių yra 16 potipių antrame lygmenyje ir 23 išsamios antrinės reikšmės trečiame lygmenyje.

Diskurso žymikliai PDTB anotavimo schemoje apima keletą diskurso žymiklių kategorijų. Pirmiausia aptariami išreikšti diskurso žymikliai, priklausantys aiškiai apibrėžtoms sintaksinėms klasėms, ir numanomi diskurso žymikliai, kurie gali būti įterpiami tarp pastraipų ar sakinių arba sudėtinių sakinių viduje tarp vidinių sakinių porų ir kurie nėra tiesiogiai susiję su apibrėžtomis sintaksinėmis klasėmis ir apibrėžtais diskurso žymiklių rinkiniais. Numanomų diskurso žymiklių atveju skaitytojas ar anotatorius turi mėginti įminti diskurso ryšį tarp gretimų sakinių ar diskurso dalių. Anotaciją sudaro jungiamojo diskurso žymiklio įterpimas, kuris geriausiai perteikia numanomą diskurso ryšį. Taip įterpti jungiamieji diskurso žymikliai vadinami numanomais diskurso žymikliais. B. Webber ir kt. (2008) taip pat aptaria daugybinius diskurso ryšius, kai skaitytojas ar anotatorius gali įžvelgti kelis diskurso ryšius ir siūlyti įterpti kelis numanomus diskurso žymiklius. Gretimos sakinių ar didesnės diskurso elementų poros, tarp kurių skaitytojas ar anotatorius neįžvelgia numanomo diskurso žymiklio, toliau skirstomos taip: a) AltLex (vadinamoji alternatyvi leksikalizacija), kai diskurso ryšys numanomas, bet bandymai įterpti kokį nors diskurso žymiklį yra pertekliniai dėl to, kad numanomas diskurso ryšys jau yra išreikštas kita leksine išraiška ar forma, kuri

gali būti vadinama alternatyvia leksikalizacija; b) EntRel (vadinamieji vientisumo ryšiai), kai negalima daryti jokių išvadų apie konkretaus diskurso ryšio egzistavimą, bet antrasis sakinys ar didesnis diskurso elementas yra skirtas tik tam tikram tolesniam pirmojo elemento aprašymui pateikti; c) NoRel (vadinamasis ryšio nebuvimas), kai nėra jokio diskurso ryšio tarp gretimų sakinių ir net negalima identifikuoti vientisumo ryšio, tada daroma išvada apie ryšio nebuvimą.

Kadangi vadinamiesiems argumentams (sakiniams ar diskurso dalims) klasifikuoti nėra visuotinai priimtų abstrakčių semantinių kategorijų, todėl du diskurso žymiklio jungiami elementai, arba argumentai, paprasčiausiai žymimi Arg2 ir Arg1. Arg2 – tai argumentas, sintaksiškai susijęs su diskurso žymikliu, Arg1 – tiesiog kitas argumentas. Arg1 ir Arg2 apibrėžiami kaip pažymėta teksto medžiaga, aktuali ir minimaliai reikalinga diskurso ryšiui paaiškinti. Kita papildoma teksto medžiaga nežymima.

Išreikštų diskurso žymiklių ir jų argumentų anotavimą sudaro atitinkamų teksto, su kuriuo dirbama, dalių parinkimas bei priskyrimas Arg1 ir Arg2 ir diskurso ryšio reikšmės priskyrimas atitinkamam diskurso žymikliui. Numanomų diskurso žymiklių anotavimas pradedamas pirmiausia pasirinkus Arg2 teksto dalį numanomam diskurso žymikliui, tada pasirenkama teksto atkarpa Arg1 ir galiausiai identifikuojama diskurso ryšio reikšmė, išreiškianti Arg1 ir Arg2 ryšį, teikiant žodį ar frazę šiam ryšiui išreikšti. AltLex atveju, užuot pateikus žodį ar frazę diskurso ryšiui išreikšti, pasirenkama ir pažymima teksto atkarpa, esanti Arg2, kuri išreiškia diskurso ryšį. EntRel ir NoRel atvejais anotavimas apima pirmiausia Arg2 teksto dalies pasirinkimą ir tada gretimų sakinių ar teksto dalių parinkimą ir žymėjimą kaip Arg1.

Taigi diskurso ryšiai pažymimi išreikštais diskurso žymikliais, numanomais diskurso žymikliais ir Altlex, vadinamąja alternatyvia leksikalizacija. EntRel ir NoRel atvejais nėra identifikuojami jokie diskurso ryšiai. Diskurso žymiklių reikšmės ar etiketės parenkamos iš hierarchinės klasifikacijos trijų lygių hierarchinio diskurso ryšių reikšmių grupavimo, kai diskurso žymikliai pagal išreiškiamą diskurso ryšį skirstomi į klases, tipus ir potipius bei anotavimo metu parenkamos reikšmės iš visų trijų hierarchijos lygių diskurso žymikliui apibūdinti.

Kalbant apie išreikštus daugybinius diskurso žymiklius, esančius šalia

vienoje vietoje, reikėtų akcentuoti, kad jie visi anotuojami atskirai. Kai toje pačioje vietoje yra keli išreikšti diskurso žymikliai (pvz., du ar keli diskurso žymikliai, išreikšti keliais prieveiksmiais arba jungtukų ir prieveiksmių samplaikomis ir kt.: *taip, pavyzdžiui*; *bet tada*; *ir dar daugiau*; *anksčiau, pavyzdžiui* ir kt.), tada kiekvienas diskurso žymiklis žymimas atskirai, atsižvelgiant į du jo argumentus. Tačiau reikia pažymėti, kad PDTB schemoje neatsižvelgiama į galimybę, jog diskurso žymikliai gali būti priklausomi vienas nuo kito ir turėti skirtingus argumentus. PDTB anotavimo schemoje nėra numatytas atskyrimas tarp priklausomų ir nepriklausomų diskurso žymiklių ir kaip sakinyje išdėstomi jų argumentai. Kalbant apie numanomus diskurso žymiklius, net jeigu anotatorius ir nori įterpti daugybinį diskurso žymiklį, toks atvejis anotuojamas viena diskurso ryšio reikšme iš diskurso reikšmių hierarchijos. Apibendrinant, moksliniai tyrimai rodo, kad PDTB anotavimo schema suteikia įžvalgų apie diskurso ryšius tekste ir diskurso žymiklius, identifikuojančius šiuos diskurso ryšius.

PDTB anotatoriams leidžiama laisvai pasirinkti reikšmes iš visų lygių, įskaitant galimybę anotuoti dviem reikšmės ženklais (iš bet kurio hierarchijos lygio), kad būtų galima atsižvelgti į dviprasmiškus atvejus. Taigi iš principo galimi 129 reikšmių deriniai. Panaši metodika buvo įgyvendinta norint anotuoti daugelio kitų kalbų, pavyzdžiui, hindi, čekų, arabų ir italų, diskurso ryšius (Webber & Joshas, 2012). Be to, S. Zufferey ir kt. (2017) atliko daugiakalbio anotavimo eksperimentą su penkiomis indoeuropiečių kalbomis, priklausančiomis germanų ir romanų kalbų šeimoms: anglų, prancūzų, vokiečių, olandų, italų. Atliekant visus šiuos tyrimus buvo pastebėta, kad nesutapimų tarp skirtingų anotatorių atvejai yra panašūs ir jų skaičius nėra didelis. Šie rezultatai rodo, kad PDTB metodika ir rezultatai gali būti pakartoti ir pritaikyti kitoms kalboms.

Lietuvių kalbos diskurso ryšių anotavimas TED-MBD tekstyne

TED daugiakalbis diskurso tekstynas (toliau – TED-MDB) yra lygiagretusis tekstynas, anotuotas diskurso lygmeniu, laikantis PDTB diskurso anotavimo tikslų ir principų (Zeyrek et al., 2018). TED-MDB sukurtas B. Webber ir kt. (2016) PDTB diskurso ryšių hierarchijos pagrindu ir jau apima 7 kalbas: turkų,

anglų, lenkų, vokiečių, rusų, portugalų, lietuvių[1]. Taigi šis tekstynas leidžia palyginti diskurso ryšiais anotuotus tekstus su angliškais diskurso ryšiais anotuotais tekstais siekiant suprasti vertimo tendencijas (TED-MDB tekstyno originalūs angliški tekstai yra išversti į kitas tekstyno kalbas), taip pat leidžia atlikti įvairių tekstyno kalbų analizę. Pagal TED-MDB projekto principus lietuviški tekstai buvo anotuoti pagrindiniais diskurso ryšių tipais (išreikštas (diskurso žymiklis, esantis tekste), numanomas (neišreikštas tekste), alternatyvi leksikalizacija, vientisumo ryšys, nėra ryšio) ir jų aukščiausio lygio reikšmėmis (laiko, priežasties, lyginamąja ir išplečiamąja), taip pat antrojo ir trečiojo lygio reikšmėmis, remiantis PDTB anotavimo schema. Bendras žodžių skaičius tekstyne – 53 305; iš jų anglų kalbos tekstuose – 8 094, vokiečių kalbos tekstuose – 8 472, lietuvių kalbos tekstuose – 5 857, lenkų kalbos tekstuose – 7 953, portugalų kalbos tekstuose – 9 298, rusų kalbos tekstuose – 7 696, turkų kalbos tekstuose – 5 935 žodžiai.

Vadovaujantis PDTB anotavimo schema, lietuvių kalboje išreikšti (aiškūs) diskurso žymikliai apima leksinius vienetus iš keturių gramatinių klasių: prijungiamieji jungtukai, pavyzdžiui, *kai*, *kol*, *nes*, *kadangi*; sujungiamieji jungtukai – *ir*, *bei*, *o*, *tačiau*; jungiamosios žodžių samplaikos (kai prie jungtukų šliejasi įvardžiai ar dalelytės) – *tam kad*, *taip kad*, *bet gi*; ir prieveiksmiai – *faktiškai*, *galiausiai*. Pagrindinė anotavimo užduotis yra išsiaiškinti, ar anotuojami žodžiai ir frazės veikia kaip diskurso žymikliai, nes jie gali atlikti ir kitas funkcijas tekste. Kaip ir PDTB, nustatomi ir anotuojami penki diskurso ryšių tipai: išreikšti diskurso ryšiai, numanomi diskurso ryšiai, alternatyvios leksikalizacijos, vientisumo ryšiai ir jokių ryšių. Žymint diskurso argumentus, tiek išreikštų diskurso žymiklių, tiek ir alternatyvių leksikalizacijų atveju, vadovaujamasi taisykle, kad Arg2 etiketė priskiriama argumentui, kuris yra sakinyje, sintaksiškai susijusiame su diskurso žymikliu; kitas argumentas žymimas Arg1. Ir PDTB schemoje, ir TED-MDB tekstyne prieveiksmiai vadinami „diskurso struktūros žymikliais“ (Hirschberg & Litman, 1987), nėra anotuojami, nes nurodo diskurso organizacinę struktūrą, o ne diskurso ryšius, siejančius du argumentus semantiškai, pavyzdžiui, lietuviškas *dabar* ir jo angliškas atitikmuo *now* (žr. 1 ir 2 pvz.):

[1] Jis yra atvirai prieinamas mokslininkams adresu: https://github.com/MurathanKurfali/Ted-MDB-Annotations

> *1. Dabar, kaip matote, įtampa, apie kurią girdėjome San Fransiske, kalbant apie žmonių susirūpinimą dėl būsto kainų ir gyventojų išstūmimo ir technologijų kompanijų, kurios atneša daug turto ir įsikuria, yra tikra*.
>
> 2. *Now you can see, though, that the tensions that we've heard about in San Francisco in terms of people being concerned about gentrification and all the new tech companies that are bringing new wealth and settlement into the city are real.*

Lietuvių kalboje, remdamasis PDTB anotavimo gairėmis ir anotuodamas numanomus diskurso žymiklius, anotatorius turi įterpti diskurso žymiklį, kuris geriausiai išreiškia numanomą dviejų gretimų sakinių santykį (žr. 3 pvz.) (visuose pavyzdžiuose Arg1 parodytas kursyvu, Arg2 paryškintas):

> 3. *Jie tokie sudėtingi ir gali atrodyti mums tolimi, kad galime būti linkę daryti štai ką: slėpti galvą smėlyje ir negalvoti apie tai*. [Numanomas = Bet] **Jei tik galite, priešinkitės tam**. (Numanomas (Implicit)) (Palyginimas: kontrastas (Comparison: Contrast)).

Lietuvių kalboje, remiantis PDTB gairėmis, alternatyvi leksikalizacija (AltLex) apima numanomus diskurso žymiklius tarp gretimų sakinių, kur atsiranda perpildymas, jei bandoma įterpti išreikštą diskurso žymiklį. Šio perpildymo priežastis yra ta, kad diskurso ryšys jau yra išreikštas tam tikra alternatyvios leksikalizacijos forma, į kurią lietuvių kalboje pateko dalelytės, pavyzdžiui, *na*, *va*, ir kitos kalbinės formos, pavyzdžiui, *vadinasi*, *vienas iš pavyzdžių* ir kt. (žr. 4 pvz.):

> 4. *Sėkmė mus motyvuoja, bet beveik pasiekta pergalė skatina mus leistis į nuolatinius ieškojimus*. [Vieną iš ryškiausių to pavyzdžių pastebime], **kai žvelgiame į skirtumą tarp olimpinio sidabro laimėtojų ir bronzos laimėtojų rungtynėms pasibaigus**. (AltLex) (Išplėtimas: pavyzdys (Expansion: Instatiation)).

Šiuo atveju galėtume bandyti įterpti diskurso žymiklį *pavyzdžiui*, bet toks bandymas būtų perteklinis.

Vientisumo ryšiai (EntRel) yra anotuojami tarp gretimų sakinių, kai subjektas ar objektas viename argumente toliau plačiau aprašomas kitame argumente (žr. 5 pvz.):

> 5. *Jie turėtų įvertinti ir tuos efektyvumo rodiklius, kuriuos vadiname ASV: aplinkosauga, socialiniai klausimai ir valdymas*. **Aplinkosauga apima energijos vartojimą, prieigą prie vandens, atliekų tvarkymą ir taršą ir ekonomišką išteklių naudojimą**. (EntRel).

Šiuo atveju matome platesnį aplinkosaugos aprašymą antrame diskurso vienete (sakinyje), tačiau negalime pritaikyti jokio diskurso ryšio iš PDTB ryšių hierarchijos, todėl anotuojame vientisumo ryšiu.

Nėra jokio ryšio (NoRel), jei anotatorius (skaitytojas) nemato jokio diskurso ryšio tarp gretimų sakinių (žr. 6 pvz.):

> 6. *Tai 4 milijardai viduriniosios klasės žmonių, kuriems reikia maisto, energijos ir vandens*. **Dabar jūs turbūt klausiate savęs: gal tai tik pavieniai atvejai**. (NoRel).

Šiuo atveju negalime identifikuoti jokio diskurso ryšio remdamiesi PDTB diskurso ryšių hierarchija.

TED-MDB prideda naują aukščiausio lygio kategoriją prie PDTB diskurso ryšių hierarchijos, vadinamąją hipoforą. Ši kategorija skirta užfiksuoti retorines klausimų–atsakymų poras, kai užduodamas retorinis klausimas ir pats kalbėtojas į jį atsako. TED-MDB anotuoja hipoforą kaip AltLex atvejį, išreikštą klausiamuoju žodžiu. Jei įmanoma ir reikalinga, gali būti pridėtas dar kitas papildomas klausimo–atsakymo poros diskurso ryšys. Pagal TED-MDB anotavimo instrukcijas lietuvių kalboje anotuojame klausimą kaip Arg2, atsakymą – kaip Arg1. Klausimas žymimas Arg2, nes AltLex išreiškiantis žodis yra klausimo dalis. Klausiamasis žodis (arba specialus žodis, arba žodis *ar*, vartojamas Taip / Ne klausimuose, kuris taip pat gali būti vartojamas kaip

išreikštas diskurso žymiklis lietuvių kalboje (žr. 7 pvz.), žymimas kaip AltLex, nes išreiškia diskurso ryšį tarp klausimo ir atsakymo. Pateiktame 7 pavyzdyje *ar* funkcionuoja kaip diskurso žymiklis, parodantis išreikštą diskurso ryšį:

> 7. *Niekas nepasikeis*, [ar] **mes bandysime pakeisti**, [ar] **tu nieko nebandysi** (Išreikštas) (Explicit) (Išplėtimas: atskyrimas) (Expansion: Disjunction).

Tolesniuose pavyzdžiuose iliustruojama, kaip hipofora anotuojama lietuvių kalboje (žr. 8 ir 9 pvz.):

> 8. [Ar] **įmonės, atsižvelgiančios į tvarumą, išties finansiškai sėkmingos**? *Galintis nustebinti atsakymas yra „taip*". (Išreikštas) (Explicit) (Altlex: Ar; Išplėtimas: detalizavimas: Arg1-kaip-detalė; Hipofora (Expansion: Level-of-detail: Arg1-as-detail; Hypophora)).

> 9. [Kodėl] **kas nors apskritai rinktųsi tokį gyvenimą** – *Atsakymas į šį klausimą gali skirtis, kaip skiriasi ir žmonės, sutinkami kelyje, bet keliautojai dažnai atsako vienu žodžiu: laisvė*. (Išreikštas) (Explicit) (Altlex: Kodėl; Priežastinis: priežastis: pagrindas; Hipofora (Contingency: Cause: Reason; Hypophora).

Tekstyno naudojimo tyrimui pavyzdžiai

Daugiakalbį tekstyną galima naudoti vertimo tyrimams. Pradžioje galima apžvelgti ir palyginti visą anotuotų tekstų anglų ir lietuvių kalbomis rinkinį ir pateikti anotuotų diskurso ryšių tipų dažnius bei PDTB aukščiausio lygio diskurso ryšių reikšmes lentelėse (žr. 1 ir 2 lentelę).

1 lentelė

Anotuotų ryšių tipų dažnis anglų ir lietuvių kalbose

Ryšio tipas	Anglų kalboje	Lietuvių kalboje
Alternatyvi leksikalizacija (AltLex)	33	7
Nėra ryšio (NoRel)	38	24
Išreikštas ryšys (Explicit)	225	297
Numanomas ryšys (Implicit)	132	177
Vientisumo ryšys (EntRel)	43	44

2 lentelė

Anotuotų aukščiausio lygio diskurso ryšių dažnis anglų ir lietuvių kalbose

Aukščiausio lygio reikšmės	Anglų kalboje	Lietuvių kalboje
Laiko	24	25
Lyginamoji	57	66
Hipofora	9	13
Išplečiamoji	213	262
Priežastinė	94	127

2 lentelėje matomas nedidelis anotuotų AltLex dažnis lietuvių kalboje galėtų reikšti tam tikrą tendenciją, atspindinčią vertėjų pasirinkimus verčiant diskurso žymiklius. Atrodo, kad vertėjai nebuvo linkę vartoti alternatyvios leksikalizacijos ir demonstravo tendenciją diskurso žymiklius perteikti žodynų pateiktais variantais. Tai siejasi su M. Baker (2011) pastebėjimais, kad vertėjai gali pasirinkti derinti diskurso žymiklius su tikslinės kalbos (kalbos, į kurią verčiama) pobūdžiu.

Kitas įdomus pastebėjimas – lietuvių kalboje yra daugiau išreikštų diskurso žymiklių nei angliškame variante. Tai gali būti paaiškinta vertėjų pastangomis perteikti angliškus numanomus diskurso ryšius aiškiai išreikštais diskurso žymikliais lietuvių kalboje (žr. 10 pvz.):

> 10. *Neblogai, tiesa* [Bet] **mes norim daugiau**. (Išreikštas) (Explicit) (Palyginimas: nuolaida: Arg2_kaip_paneigimas) (Comparison: Concession: Arg2_as_denier).
> <...> *that's okay, right*. [Numanomas (Implicit) = But] **We want more**. (Numanomas) (Implicit) (Palyginimas: nuolaida: Arg2_kaip_paneigimas) (Comparison: Concession: Arg2_as_denier).

Tačiau taip pat yra atvejų, kai angliškasis diskurso žymiklis perteikiamas netiesioginiu diskurso žymikliu lietuvių kalboje ir kartais dėl to lietuviškame vertime prarandama pradiniame angliškame tekste anotuoto diskurso ryšio prasmė (žr. 11 pvz.):

> 11. <...> *žiūrėti tik į rasę nepadeda bandant prisidėti prie įvairumo vystymo*. [Implicit = Taigi] **Bandome įvairumą naudoti sprendžiant kai kurias sudėtingesnes problemas, turime**

pradėti kitaip galvoti apie įvairumą. (Numanomas) (Implicit) (Priežastinis: priežastis: rezultatas (Contingency: Cause: Result)).
<...> *only looking at race doesn't really contribute to our development of diversity*. So if **we're trying to use diversity as a way to tackle some of our more intractable problems, we need to start to think about diversity in a new way**. (So (Išreikštas) (Explicit) (Priežastinis: priežastis: rezultatas (Contingency: Cause: Result)) if (Išreikštas) (Explicit) (Priežastinis: sąlyga: Arg2_kaip_sąlyga (Contingency: Condition: Arg2_as_condition)).

Taigi 11 pavyzdys rodo, kad vertėjas pasirinko neperteikti angliškųjų išreikštų diskurso žymiklių *So* ir *if* , ir nors „rezultato" diskurso ryšys išlieka numanomas, galima pastebėti „sąlygos" diskurso ryšio reikšmės praradimą.

Šiame straipsnyje pateikta tik keletas pavyzdžių, atskleidžiančių vertimo tyrimų galimybes naudojant diskurso ryšiais anotuotą daugiakalbį tekstyną. Taip pat, kaip minėta pradžioje, remiantis šiuo tekstynu galima tirti ir lietuviškų diskurso žymiklių raiškos ypatumus.

Išvados

Diskurso ryšiais anotuotas TED-MDB daugiakalbis tekstynas leidžia tirti lietuviškus diskurso žymiklius bei palyginti juos su kitų tekstyno kalbų diskurso žymikliais. Straipsnyje pateikti lietuvių ir anglų kalbos lyginamieji pavyzdžiai rodo, kad lietuviški diskurso žymikliai kartais išreiškiami vietoj angliškųjų numanomų – tai būtų galima paaiškinti vertėjų pastangomis išversti numanomą diskurso ryšį į ryšį su išreikštu diskurso žymikliu. Be to, pastebima, kad išreikštų diskurso žymiklių perteikimas netiesiogiai gali pakenkti diskurso ryšių prasmės perteikimui tekste. Tačiau tam reikėtų platesnių tyrimų ir įžvalgų apie konkrečius vertėjų pasirinkimus. Viena vertus, tokius pasirinkimus galima būtų paaiškinti sinchronizacijos reikalavimais verčiant transkribuotą tekstą, antra vertus, reikėtų nepamiršti, kad tam tikros stilistinės nuostatos gali būti individualus vertėjų pasirinkimas, pavyzdžiui, vieni vertėjai mėgsta dažniau vartoti išreikštus diskurso žymiklius, kiti – rečiau.

Ateityje gilinantis į lietuviškų tekstų ir TED-MDB daugiakalbio tekstyno

lyginamuosius tyrimus galima tikėtis atskleisti ir patikslinti daugiau vertimo tendencijų. Anotuodami daugiau lietuviškų tekstų pagal TED-MDB schemą ir tyrinėdami diskurso ryšiais anotuotus tekstus galime sukurti lietuvių kalbos diskurso žymiklių leksikoną.

Literatūra

Asher, N., & Lascarides, A. (2003). *Logics of conversation*. Cambridge University Press.

Baker, M. (2011). *In other words: A coursebook on translation (2nd edition)*. Routledge.

Beeching, K. (2016). *Pragmatic Markers in British English: Meaning in Social Interaction*. Cambridge University Press.

Brinton, L. (2017). *The Evolution of Pragmatic Markers in English: Pathways of Change*. Walter de Gruyter.

Degand, L., & Evers-Vermeul, J. (2015). Grammaticalization or Pragmaticalization of Discourse Markers? More than a terminological issue. *Journal of Historical Pragmatics*, *16(1)*, 59–85. https://doi.org/10.1075/jhp.16.1.03deg

Halliday, M., Kirkwood, A., & Hasan, R. (1976). *Cohesion in English*. Longman.

Hirschberg, J., & Litman, D. (1987). Now let's talk about now: Identifying cue phrases intonationally. *Proceedings of the 25th annual meeting on Association for Computational Linguistics* (pp. 163–171). Association for Computational Linguistics.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge University Press.

Knott, A., & Dale, R. (1994). Using linguistic phenomena to motivate a set of coherence relations. *Discourse Processes, 18*, 35–62.

Mann, W. C., Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization, *Text* 8, 243–281.

Roze, Ch., Danlos, L., & Muller, Ph. (2010). LEXCONN: a French Lexicon of Discourse Connectives. *Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010), Moissac, France*, https://journals.openedition.org/discours/8645

Sanders, T. J. M., & Noordman, L. G. M. (2000). The role of coherence relations

and their linguistic markers in text processing. *Discourse Processes, 29*, 37–60.

Šinkūnienė, J., Jasionytė-Mikučionienė, E., Ruskan, A., & Šolienė, A. (2020). Diskurso žymikliai lietuvių kalboje: reikšmės ir funkcijų kaitos aspektai. *Lietuvių kalba*, *14*. http://www.lietuviukalba.lt/index.php/lietuviu-kalba/article/view/310.

Šolienė, A. (2018). Diskurso žymikliai: ar viskas išverčiama? *Gimtoji kalba, 11*, 7–10.

Šolienė, A. (2020a). Lithuanian Discourse Markers *Na* and *Nu*: A Glimpse at Lithuanian-English Parallel Corpus Data. In S. Granger, & M.-A. Lefer (Eds.), *Translating and Comparing Languages: Corpus-based Insights. Selected Proceedings of the Fifth Using Corpora in Contrastive and Translation Studies Conference* (pp. 237–255). Presses universitaires de Louvain.

Šolienė, A. (2020b). They are *kind of / sort of* similar: A parallel corpus-based analysis of English *kind of* and *sort of* and their Lithuanian correspondences. *Lietuvių kalba*, *14*. http://www.lietuviukalba.lt/index.php/lietuviu-kalba/article/view/313

Webber, B., & Joshi, A. (2012). Discourse structure and computation: Past, present and future. *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries* (pp. 42–54). Association for Computational Linguistics.

Webber, B., Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., & Joshi, A. (2008). The Penn Discourse TreeBank 2.0, *Proceedings of the 6th International Conference on Language Resources and Evaluation* (pp. 2961–2968). European Language Resources Association, Marrakech.

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2016). A Discourse-Annotated Corpus of Conjoined VPs. *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)* (pp. 22–31). Association for Computational Linguistics.

Zeyrek, D., Mendes, A., & Kurfalı, M. (2018). Multilingual extension of PDTB-style annotation: The case of TED Multilingual Discourse Bank. *Proceedings of the Eleventh International Conference on Language*

Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA).

Zufferey, S., & Degand, L. (2017). Annotating the meaning of discourse connectives in multilingual corpora. *Corpus Linguistics and Linguistic Theory, 13(2)*, 399–422.

Giedrė Valūnaitė Oleškevičienė
Mykolas Romeris University, Lithuania
gvalunaite@mruni.eu
Vitalija Karaciejūtė
Mykolas Romeris University, Lithuania
vitalija.karaciejute@mruni.eu
Dalia Gulbinskienė
Vilnius Gediminas technical university (VILNIUS TECH), Lithuania
dalia.gulbinskiene@vilniustech.lt

LITHUANIAN DISCOURSE MARKERS AND THEIR RELATIONS IN A MULTILINGUAL CORPUS

Summary. The development and research of discourse-annotated corpora is a relatively new field, therefore Lithuanian researchers seek to supplement the existing corpora resources and look for ways to study discourse relations by linking and comparing them with their counterparts in other languages because in different languages discourse relations are realized by different linguistic means. The aim of the article is to present the developing available corpora resources drawing on the experience of foreign scholars and to discuss guidelines for translation research at the discourse level. Therefore, the article first deals with the possibilities of expressing discourse relations by using discourse markers as their linguistic realization in different languages, discussing possible choices of translators, taking into account discourse realtions in translation and the use of different linguistic means. The article presents the parallel multilingual corpus TED-MBD (*Multilingual discourse-annotated corpus*), which is annotated at the discourse level, in accordance with the objectives and principles of PDTB (*Penn Discourse Treebank*) discourse annotation. The article discusses in detail the annotation system of PDTB discourse markers, the reader is introduced to the hierarchy of senses of discourse relations, the principles of annotation and insights into the application of the PDTB scheme. It also describes the annotation principles of the Lithuanian part of the corpus in accordance with the PDTB discourse annotation rules; the first results related to the expression of discourse relations and the use of discourse markers are discussed. The article also presents the first research guidelines on how to compare Lithuanian and English discourse-annotated texts in order to understand translation tendencies at the discourse level.

Keywords: discourse markers; discourse relations; multilingual corpus; annotation; PDTB sense hierarchy.