KOMPIUTERINĖ LINGVISTIKA / COMPUTATIONAL LINGUISTICS

Lietuvių–latvių ir latvių–lietuvių kalbų lygiagretusis tekstynas LILA

Erika Rimkutė, Andrius Utka, Kristīne Levāne-Petrova

crossref http://dx.doi.org/10.5755/j01.sal.0.23.4582

Anotacija. Straipsnyje pristatomas naujas kalbos išteklius – lygiagretusis beveik iš 9 mln. žodžių sudarytas lietuvių–latvių, latvių–lietuvių kalbų tekstynas LILA, kurio tekstai sulygiagretinti pastraipų ir sakinių lygmeniu. Tekstynas yra su metaduomenimis, kuriuose pateikiama informacija apie autorius, leidimo metus ir pan. Tekstynas struktūriškai anotuotas: jame sužymėtos pastraipų ir sakinių ribos. Kol kas tai vienintelis dvikalbis šios kalbų poros tekstynas. Tekstynas parengtas 2011–2012 m. Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro (VDU KLC) darbuotojų kartu su Latvijos universiteto Matematikos ir informatikos instituto Dirbtinio intelekto laboratorijos (LU MII) mokslininkais. Pastraipoms ir sakiniams lygiagretinti naudotas VDU KLC sukurtas pusiau automatinis įrankis *Aligner 2.0.6.7*. Straipsnyje aprašyta, su kokiomis problemomis, rengdami tekstynus ir kitas kalbos priemones, susiduria rečiau vartojamų kalbų atstovai. Daugiausia problemų kelia ribotas tekstų pasirinkimas, dėl to sunku sudaryti norimos apimties ir pobūdžio tekstynus; ilgai užtrunka spausdintų tekstų skaitmeninimas. Pristatyta tekstyno sudarymo koncepcija, sandara, jo rengimo etapai; išsamiau aprašytas naudotas lygiagretinimo įrankis. Straipsnyje rašyta apie lygiagrečiojo tekstyno paieškos sistemą, šio ir kitų lygiagrečiųjų tekstynų panaudojimo galimybės, ypač kalboms mokyti ir mokytis, struktūrinių lietuvių ir latvių kalbų skirtumų analizei, vertimų kokybės lyginimui, keliakalbiams žodynams sudaryti, kalbų technologijų srityje (kuriant statistinio automatinio vertimo sistemas).

Reikšminiai žodžiai: *lygiagretusis tekstynas, lietuvių kalba, latvių kalba, baltų kalbos, mažai išteklių turinčios kalbos.*

Įvadas

Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centre (VDU KLC) ir Latvijos universiteto Matematikos ir informatikos instituto Dirbtinio intelekto laboratorijoje (LU MI) jau daugiau negu 15 metų rengiami ir kaupiami įvairūs kalbiniai resursai: vienakalbiai ir daugiakalbiai tekstynai, elektroniniai žodynai, duomenų bazės ir kiti resursai. Dažniausiai naudojami resursai yra *Dabartinės lietuvių kalbos tekstynas* ir *Dabartinės latvių kalbos tekstynas*. Mokslininkų, studentų, vertėjų, kalbų besimokančiųjų dėmesio sulaukia KLC lygiagretieji anglų–lietuvių, lietuvių–anglų, čekų–lietuvių, lietuvių–čekų kalbų tekstynai, taip pat domimasi ir latviškais specialiaisiais tekstynais – *Senie* – senųjų tekstų tekstynu ir *Saeima* – Latvijos Seimo stenogramų tekstynu.

Ilgą laiką buvo pasigendama artimų kalbų lygiagrečiojo tekstyno. Lygiagretusis tekstynas yra sudarytas iš vienos kalbos tekstų ir jų vertimų kitoje kalboje. 2011–2012 m. vykdant ES tarpvalstybinės Lietuvos–Latvijos bendradarbiavimo programos projektą *Humanitarinių mokslų švietimo infrastruktūra Rytų Latvijoje ir Lietuvoje (Kaunas),* parengtas likusių gyvų baltų, t. y. lietuvių–latvių ir latvių–lietuvių, kalbų lygiagretusis tekstynas LILA. Šio tekstyno tekstai sulygiagretinti pastraipų ir sakinių lygmeniu. Nuo 2012 m. tekstynas prieinamas internete VDU KLC ir LU MII svetainėse[1].

Šio straipsnio tikslas – remiantis tekstynų lingvistikos metodika aprašyti lygiagrečiojo tekstyno sudarymo principus, etapus, aptarti problemas, su kuriomis susiduria rečiau vartojamų kalbų atstovai, pristatyti tekstyno panaudojimo galimybes, internete prieinamą paieškos sistemą, paskatinti įvairių sričių specialistus (tiek teoretikus, tiek praktikus) naudotis šiuo nauju resursu, taikyti įvairaus pobūdžio tyrimuose.

Rečiau vartojamų kalbų kalbinių resursų sudarymo problemos

Dauguma tekstynų lingvistikos teorinių studijų ir praktinių tyrimų yra pritaikytos anglų kalbai ir dar kelioms dominuojančioms kalboms (pvz., vokiečių). Šiais tyrimais nustatyti tekstynų sudarymo reikalavimai ir principai taip pat labiau pritaikyti dažniau vartojamoms pasaulio kalboms. Pavyzdžiui, tekstyno reprezentatyvumo ir subalansavimo kriterijai, tinkantys didžiosioms pasaulio kalboms, labai sunkiai pasiekiami mažiau resursų turinčioms kalboms (Marcinkevičienė et al., 2012; plačiau apie daugiakalbių

[1] Žr. http://tekstynas.vdu.lt/page.xhtml?id=parallelLILA, http://www.korpuss.lv/LILA

tekstynų sudarymą, naudojimą ir taikymą žr. Rimkutė et al., 2006).

Kuriant dvikalbius lygiagrečiuosius tekstynus, subalansavimo problemų padvigubėja, nes tada reikia sukaupti ir suderinti ne vienos kalbos, o dviejų kalbų išteklius. Čia galimos keturios situacijos:

1) didelės kalbos[2] originalūs tekstai lygiagretinami su kitos didelės kalbos vertimais;

2) didelės kalbos originalūs tekstai lygiagretinami su mažos kalbos vertimais;

3) mažos kalbos originalūs tekstai lygiagretinami su didelės kalbos vertimais;

4) mažos kalbos originalūs tekstai lygiagretinami su mažos kalbos vertimais.

Pirmoms dviem kalbų porų situacijoms yra skirta dauguma tyrimų ir dėmesio. Tokia padėtis susiklostė natūraliai: su tokiomis kalbų poromis dirba didžiausios tyrėjų pajėgos ir būtent čia yra sukaupta daugiausia kalbinių išteklių. Trečia ir ketvirta situacijos yra gerokai sudėtingesnės.

Siekdama sukurti subalansuotą lietuvių–anglų kalbų lygiagretųjį tekstyną autorinių neologizmų vertimo tyrimui, J. Vaičenonienė (2012) susidūrė su lietuviškos literatūros vertimų į anglų kalbą trūkumu. Nors kai kurių lietuvių autorių daug kūrinių išversta į anglų kalbą (pvz., Ričardo Gavelio), tačiau, siekiant išvengti šių autorių kalbos dominavimo, autorei teko atsisakyti dalies medžiagos ir apsiriboti tik gana mažu 1,7 mln. žodžių lygiagrečiuoju tekstynu.

Šiame straipsnyje pristatomas lygiagretusis tekstynas LILA priklausytų ketvirtajai situacijai. Tiek lietuvių, tiek latvių kalbos turi nedidelį jas vartojančių žmonių skaičių, jų kalbiniai ištekliai ir mokslinis potencialas yra gana riboti. Todėl, norėdami pasiekti suplanuotus tekstyno sandaros ir dydžio tikslus, susidūrėme su ribotų išteklių problema ir turėjome priimti kompromisinius sprendimus. Tiesa, lietuvių ir latvių kalbų pora turi vieną privalumą – tai yra kaimyninių tautų kalbos, priklausančios vienai baltų kalbų šakai ir turinčios panašią istoriją.

Toliau vertinant ketvirtos situacijos kalbų poras, galima būtų teigti, kad geografiškai labiau atskirtų dviejų mažų kalbų poroms (pvz., lietuvių ir maltiečių) būtų dar sunkiau ar net neįmanoma rasti pakankamai lygiagrečių tekstų norint sukurti lygiagretųjį tekstyną.

Lygiagretusis tekstynas LILA

Toliau pateikti pagrindiniai duomenys apie tekstyną LILA:

- dviejų krypčių vertimai: iš latvių kalbos į lietuvių kalbą ir iš lietuvių kalbos į latvių kalbą;
- įvairių temų ir žanrų tekstai;
- tekstynas subalansuotas: tekstai įtraukti pagal konkrečias proporcijas;
- sulygiagretintas pastraipų ir sakinių lygmeniu;
- pateikiami metaduomenys apie autorius, leidimo metus ir pan.;
- struktūriškai anotuotas: sužymėtos pastraipų ir sakinių ribos.

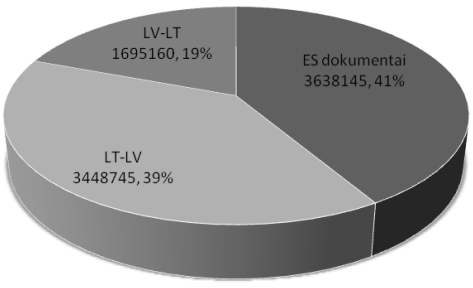
Bendras tekstyno dydis yra 8 782 050 žodžių: didžiausia yra lietuvių–latvių kalbų tekstų dalis (3 448 745 žodžiai), dvigubai mažiau yra latvių–lietuvių kalbų tekstų (1 695 160 žodžių). Tokia nesimetriška duomenų situacija susidarė dėl to, kad pastaraisiais metais daugiau išversta iš lietuvių kalbos į latvių, o ne iš latvių į lietuvių. Į latvių kalbą išversta daug produktyvių šiuolaikinių lietuvių rašytojų, pvz., Jurgos Ivanauskaitės, Sigito Parulskio, Kristinos Sabaliauskaitės, kūrinių. Vertimų iš latvių į lietuvių kalbą gerokai mažiau, pvz., Ingos Ābeles, Laimos Muktupāvelos, Nuoros Ikstenos.

Siekiant sukurti didesnės apimties tekstyną nebuvo galima išsiversti be ES tekstyno tekstų[3]. ES dokumentai sudaro nemažą lygiagrečiojo tekstyno dalį (3 638 145 žodžius).

Dokumentai, kaip tam tikras žanras, yra vertinga lygiagrečiojo tekstyno dalis. Vis dėlto šie tekstai neprilygsta originaliems lietuviškiems ar latviškiems tekstams, nes į lietuvių ir latvių kalbas išversti per tarpinę kalbą (greičiausiai per anglų kalbą), t. y. netiesiogiai. Lietuviški ir latviški tekstai gali būti paveikti anglų kalbos, todėl ši tekstyno dalis turbūt nėra itin naudinga mokslininkams, besigilinantiems į šių baltų kalbų ypatybes, lyginantiems originalą ir išverstą tekstą, besimokantiems kalbų ar analizuojantiems latvių ir lietuvių kalbų skirtumus.

LILA tekstyno paieškoje ES dokumentai sudaro atskirą tekstyno dalį tam, kad tyrėjai, kuriems svarbi vertimo kryptis ir autentiškumas, galėtų atskirti tiesioginius ir netiesioginius vertimus (plačiau žr. poskyrį *Paieškos galimybės*).

Tekstyne norėta atspindėti įvairius žanrus, tekstų tipus, todėl tekstyne yra grožinės literatūros, populiariosios literatūros, publicistikos ir administracinės literatūros tekstų. Šios proporcijos pateiktos 1 paveiksle.



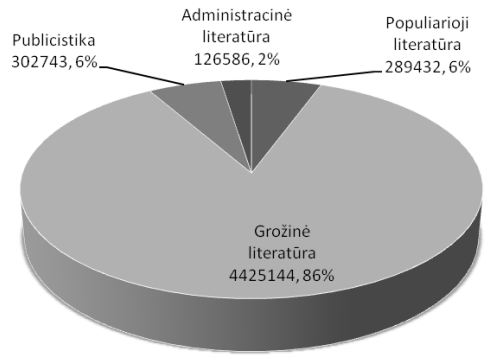
1 pav. Viso LILA tekstyno proporcijos

3 ir 4 paveiksluose atspindėtos proporcijos lietuvių–latvių ir latvių–lietuvių kalbų tekstyno dalyje. Juose matyti, kad tiek lietuvių–latvių kalbų, tiek latvių–lietuvių kalbų tekstynuose didžiausią dalį sudaro grožinė literatūra, todėl ateityje pildant tekstyną reikėtų didinti kitas dalis, kad taip nedominuotų grožinė literatūra. Kita vertus, grožinėje lite-

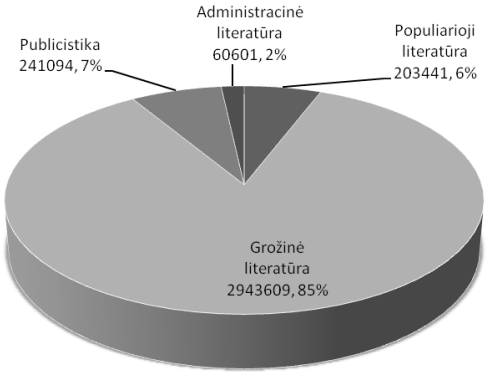
[2] Didelės kalbos – tai dažnai vartojamos, plačiai paplitusios kalbos, mažos kalbos – rečiau vartojamos, ne taip paplitusios kalbos.

[3] http://ipsc.jrc.ec.europa.eu/?id=198

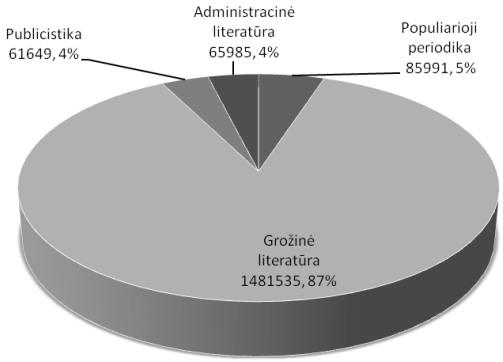
ratūroje galima rasti įvairesnės leksikos nei, tarkim, dažnai pastoviomis klišėmis pasižyminčioje administracinėje kalboje, todėl LILA tekstyno žodingumas (t. y. skirtingų žodžių kiekis) turėtų būti gana didelis.



2 pav. Viso LILA tekstyno proporcijos be ES tekstų



3 pav. LILA tekstyno lietuvių–latvių kalbų dalies proporcijos



4 pav. LILA tekstyno latvių–lietuvių kalbų dalies proporcijos

Tekstų atrankos kriterijai

Dar prieš pradedant rengti lygiagretųjį tekstyną LILA, buvo žinoma, kad tekstų rengiamam tekstynui nebus daug. Iš pradžių buvo nustatyti trys pagrindiniai tekstų atrankos kriterijai:

- **laikas:** norėta, kad tekstynas atspindėtų dabartinę kalbą, todėl įtraukti tik tie tekstai, kurie publikuoti ne anksčiau kaip 1991 m.;
- **publikavimo tipas:** tekstai turėjo būti elektroniniai (pvz., interneto svetainės) arba spausdinti, pvz., nepublikuotas romano rankraštis nebūtų įtrauktas į tekstyną;
- **originalūs tekstai:** tekstas turėjo būti originalus, t. y. išverstas iš lietuvių kalbos į latvių kalbą arba iš latvių kalbos į lietuvių. Vertimai į latvių ir lietuvių kalbas iš kitų kalbų į tekstyną neįtraukti (išimtis ES tekstai).

Surinkus duomenis apie visus prieinamus kalbinius išteklius, suprasta, kad tekstų nedaug, todėl netaikyti jokie kiti papildomi tekstų atrankos kriterijai.

Tekstyną sudaro tik **pilni** tekstai (pvz., į tekstyną įdėtas visas romanas, neatsižvelgiant į jo apimtį, o ne jo fragmentas).

Visi į tekstyną įtraukti tekstai yra **autentiški**: jie niekaip netaisomi, kad atspindėtų kalbą tokią, kokia yra. Jeigu į elektroninę formą perkeltas spausdintas tekstas ir tame spausdintame variante buvo klaidų, jos paliktos, nors buvo keletas išimčių, kai ištaisytos klaidos. Jeigu tekste pasitaikė kirilica parašytų intarpų, tai šios teksto dalys transliteruotos. Jei to nebūtų padaryta, tekstyno naudotojai matytų iškraipytą tekstą, nes dėl skirtingo kodavimo kirilica parašytas tekstas tekstyno sąsajoje nebūtų matomas. Viešai neprieinamuose dokumentuose, kurie įdėti į tekstyną, ištrinti asmeniniai duomenys (vardai, adresai ir pan.; plačiau žr. Levāne-Petrova, 2012).

Tekstyno kalbinių resursų atranka ir tvarkymas

Atsižvelgiant į nustatytus tekstyno tekstų atrankos kriterijus pagal kalbos žanrus ir temas, buvo apibendrinta informacija apie kalbinius resursus, įdėtus į tekstyną.

Spausdinti kalbiniai resursai

Remiantis bibliografine rodykle, publikuota Latvijoje 2008 m. (Rodyklis, 2008), pirmiausia buvo apibendrinti duomenys apie lietuvių ir latvių kalbomis parašytą arba į šias kalbas išverstą *grožinę literatūrą*, kurios knygos išleistos nuo 1991 m. Sudarius sąrašą apie visus tekstynui tinkamus grožinės literatūros šaltinius, paaiškėjo, kad vertimų iš latvių kalbos į lietuvių kalbą yra beveik dvigubai mažiau negu vertimų iš lietuvių kalbos į latvių kalbą, nors, lyginant su, pvz., latvių–estų–latvių kalbų vertimais, šaltinių yra daug daugiau. Į lietuvių–latvių kalbų tekstyno dalį įdėti 22 grožinės literatūros šaltiniai, t. y. knygos, išverstos iš lietuvių kalbos į latvių kalbą, o į latvių–lietuvių kalbų tekstyno dalį įkelta 15 knygų.

Svarbus žanras yra *administracinė literatūra*, nors viešai prieinamų lygiagrečių dalykinių tekstų yra labai nedaug. Išimtis – Europos Sąjungos dokumentai. Atsižvelgiant į tai, kad latvių ir lietuvių įmonės turi ryšių ir atstovų vienoje ar kitoje šalyje, bandyta gauti viešai neskelbiamų tekstų. Tam tikslui buvo pasirašyta sutartis su vertimo biuru „Skrivanek Latvia“ dėl lygiagrečių latvių ir lietuvių tekstų. Iš šio vertimo biuro gauta ne tik įvairių dokumentų, bet ir kitų žanrų tekstų, svarbių tekstyno žanrinei įvairovei.

Dar viena tekstyno dalis – *populiarioji literatūra*. Šią dalį sudaro atsiminimų knygos, kulinarinės knygos ir kt.

Į tekstyną įdėta kad ir nedaug publikuotų *publicistikos* tekstų, pvz., Lietuvos ir Latvijos forume paskelbtų tekstų,

žurnale *Šeimininkė* publikuotų straipsnių. Visiškai nerasta dvikalbių *mokslinių* tekstų.

Elektroniniai kalbiniai resursai

Išskyrus grožinę literatūrą, kitų stilių, įvairių žanrų ir temų tekstų daugiausia rasta internete. Dažniausiai internete randamos Latvijos ir Lietuvos įmonių svetainės su informacija apie įmonės istoriją, veiklą ir pan., žinios apie kaimynės šalies įvykius, sporto ir kultūros apžvalgos.

Nors buvo tikėtasi, kad kaimyninės šalies įvykiai internetiniuose portaluose, pvz., Delfi, bus plačiai atspindėti, bet reikia paminėti, kad kaimyninės šalies žinios verčiamos retai; paprastai jos yra lokalizuojamos, t. y. vienos kalbos tekstas neverčiamas tiesiogiai, o pranešamas įvykis atpasakojamas.

Kitų žanrų tekstų vertimai internete taip pat dažnai išversti ne pažodžiui. Tokio pobūdžio lokalizuoti tekstai į tekstyną neįtraukti, nes lietuvių ir latvių kalbų tekstai gana smarkiai skiriasi, pvz., praleistos pastraipos, sakiniai ir pan. Į tekstyną tokie tekstai nedėti ir dėl kitos priežasties: kartais neaišku, kuria kalba parašytas originalas, o kuria vertimas. Tai svarbu, nes šiame straipsnyje pristatomas tekstynas yra dvikryptis ir lygiagretinant nurodoma, kuri kalba originalo, o kuri – vertimo.

Tekstų tvarkymas

Kad tekstus būtų galima įkelti į tekstyną, jie turi būti specialiai paruošti: jei neturima elektroninių tekstų variantų, tekstai skenuojami, kopijuojami iš interneto, tada lygiagretinami, registruojami tekstyno duomenų bazėje.

Nemaža tekstyno šaltinių dalis buvo spausdinti, jie buvo prieinami popieriniu pavidalu, todėl tokius tekstus reikėjo suskaitmeninti. Nors skaitmeninimas yra brangus ir ilgas procesas, nebuvo kitokių galimybių gauti seniau publikuotų darbų, ypač grožinės literatūros.

Paprastai reikalingas kalbinis šaltinis skenuojamas ir atpažįstamas 98 % tikslumu, bet norint kelti tekstus į tekstyną reikalingas dar didesnis tikslumas, todėl skenavimo ir atpažinimo metu atsiradusios klaidos buvo ištaisytos naudojantis klaidų tikrintuvu, be to, žmonių peržiūrėtos ir ištaisytos.

Savaime suprantama, kad elektroniniai tekstai galutinai sutvarkomi ir parengiami kelti į tekstyną daug greičiau, bet, kaip minėta, tokių tekstų rasta gana nedaug (išskyrus Europos Sąjungos dokumentus).

Lygiagretinimo įrankis

LILA tekstynas sulygiagretintas pastraipų ir sakinių lygmeniu naudojant VDU KLC sukurtą pusiau automatinį įrankį *Aligner 2.0.6.7*. Šis įrankis jau buvo naudotas kuriant ir kitus dvikalbius lygiagrečiuosius tekstynus: anglų–lietuvių (Utka et al., 2008), lietuvių–anglų (Vaičenonienė, 2012) ir vokiečių–lietuvių (Kovalevskaitė, 2012). Apie LILA tekstyno lygiagretinimą rašyta Utka et al. (2012).

Aligner 2.0.6.7 veikimo algoritmas remiasi tik struktūriniais tekstų požymiais (sakinių ir pastraipų ribų žymomis). Kaip ir kai kurie kiti lygiagretinimo įrankiai, *Aligner 2.0.6.7* naudoja Geilo ir Čerčo algoritmą (Gale et al., 1993), kuris lygiagretina sakinius remdamasis lygiagrečių segmentų dydžiu išreikštu simbolių skaičiumi.

Lygiagretinimo procesas susideda iš kelių etapų:

1) tekstai automatiškai sulygiagretinami pastraipų lygmeniu;
2) žmogus patikrina lygiagretinimo rezultatą ir jei yra klaidų, jas ištaiso;
3) tekstai automatiškai sulygiagretinami sakinių lygmeniu;
4) žmogus patikrina lygiagretinimo rezultatą ir jei yra klaidų, jas ištaiso;
5) automatiškai nustatomos ir ištaisomos žmogaus padarytos teksto struktūros klaidos.

Toks lygiagretinimo procesas gana lėtas ir brangus (vieno teksto lygiagretinimas užima tiek pat laiko ar net daugiau negu to teksto skaitmeninimas), lyginant su visiškai automatiniais lygiagretinimo metodais. Kita vertus, jis turi ir vieną privalumą – lygiagretinimo rezultatas yra labai geros kokybės, nes kiekvienas lygiagretusis segmentas yra peržiūrėtas žmogaus.

Paieškos galimybės

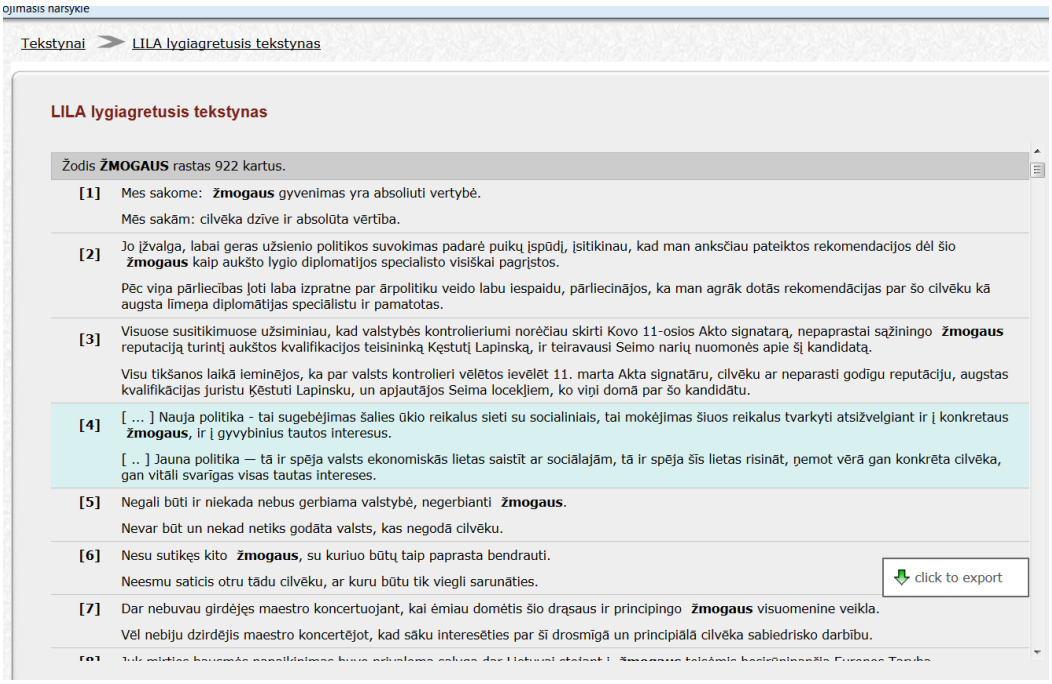
Internete prieinamoje lygiagrečiojo tekstyno LILA paieškos sistemoje galima atlikti paiešką tiek lietuvių, tiek latvių kalba. Galima atskirai analizuoti tik tuos tekstus, kurių aiški originalo ir vertimo kalba, o nesirinkti tų tekstų, kurie išversti per kalbą tarpininkę (Europos Sąjungos dokumentų).

Tekstyne galima ieškoti konkretaus žodžio arba žodžio dalies. Ieškant žodžio dalies reikia įrašyti žodžio dalį ir žvaigždutę, pvz., ieškant *apsirik**, gaunami tokie rezultatai: *apsiriko, apsirikti, apsirikę, apsirikdavo* ir kt. Toliau galima ieškoti visų rastų žodžių arba pasirinkti vieną ar kelis iš jų. Prie rastų žodžių nurodomas ir jų dažnumas lygiagrečiajame tekstyne.

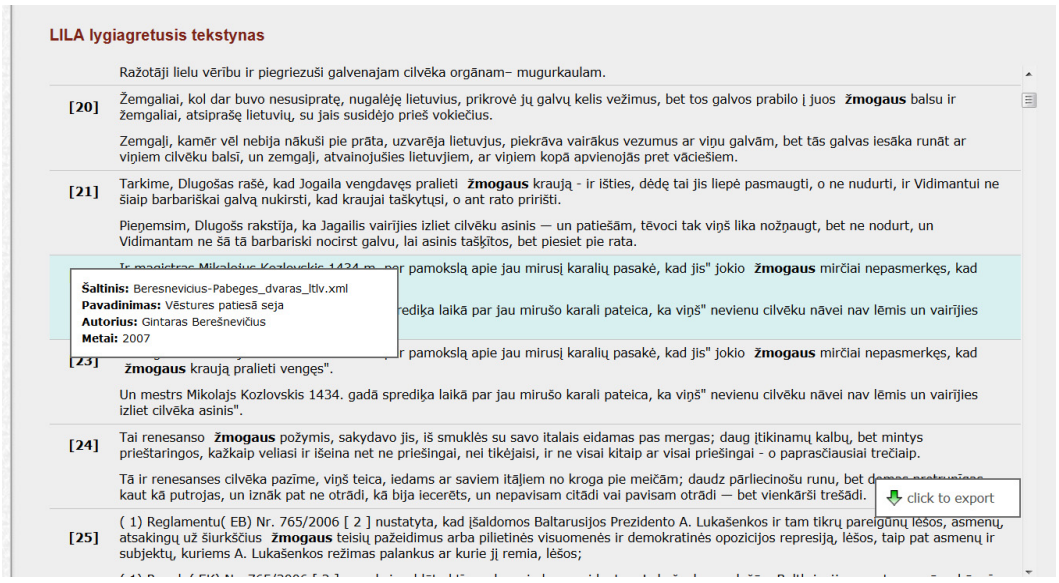
Išplėstinėje paieškoje galima pasirinkti konkordanso pateikimo formą (kad būtų pateiktas lygiagretus ar vertikalus konkordansas); radus daugiau nei vieną žodį, galima rinktis, ar iš karto pateikti tų visų žodžių konkordansą, ar pirma pateikti dažninį žodžių sąrašą. Žodžiai gali būti surūšiuoti pagal dažnumą arba pagal abėcėlę.

Rastuose rezultatuose ieškomas žodis yra pajuodintas (žr. 5 pav.). Spustelėjus kitą dominantį žodį (ar kelis žodžius), tas žodis visuose rastuose sakiniuose bus pažymėtas ta pačia spalva. Taip lygiagrečiojo tekstyno naudotojams lengviau analizuoti rezultatus, ypač pastoviuosius junginius, nes kol kas nėra galimybės paieškos langelyje nurodyti du ar daugiau žodžių. Gautus duomenis galima išsisaugoti atskiroje rinkmenoje ir susirūšiuoti pagal savo poreikius.

Prie kiekvienos lygiagrečios eilutės nurodomi šaltiniai. Palaikius pelę ant eilučių numerių, parodoma, iš kokio teksto paimtas sakinys. Visų rezultatų pabaigoje pateikiama išsami informacija apie šaltinius: teksto identifikacinis pavadinimas, autorius, teksto pavadinimas, originalo metai (žr. 6 pav.).



5 pav. Paieškos rezultatai ieškant konkretaus žodžio ar žodžio formos



6 pav. Metaduomenų pateikimo pavyzdys

Lygiagrečiųjų tekstynų panaudojimo galimybės

Lygiagretieji tekstynai gali būti naudojami daugelyje sričių:

- verčiant;
- mokant ar mokantis kalbų;
- analizuojant vertimus;
- kuriant statistinio automatinio vertimo sistemas;
- lyginant kalbas;
- kuriant žodynus;
- kitose srityse.

Pirmiausia lygiagretieji tekstynai labai pravartūs vertėjams ar besimokantiems kalbų, norintiems rasti tinkamą vertimo atitikmenį, kurio galbūt nėra dvikalbiame žodyne, siekiant palyginti to paties vieneto vertimą. Pvz., *Lietuvių–latvių kalbų žodyne*, išleistame 1995 m. (jo internetinė versija prieinama *www.letonika.lv*), matyti, kad lietuviškas žodis *lakrica*[4] yra pateiktas žodyne su latvišku atitikmeniu *lakrica*, *lakricsakne*. Tekstyne LILA latviškas žodis *lakrica* pavartotas 12 kartų, o jo lietuviški atitikmenys yra *saldymedis* ar *saldišaknis*, o ne *lakrica,* kaip teigiama žodyne, pvz.:

[4] LKŽ (http://www.lkz.lt/) nurodyta, kad *lakrica* reiškia „saldymedžio šaknys“ ir „saldymedžio šaknų ekstraktas“. DLKŽ (http://dz.lki.lt/) šis žodis nepateiktas.

LV: *...ogļu melni saldās un sāļās* ***lakricas*** *stienīši...*

LT: *...anglių juodumo saldieji ir sūrieji* ***saldymedžio*** *pagaliukai...*

LV: *...****lakricas*** *ekstrakts, kas satur vairāk...*

LT: *...****saldišaknės*** *ekstraktas, kuriame sacharozės kiekis...*

Taigi tekstyne rasti tikslesni atitikmenys negu žodyne, išleistame beveik prieš 20 metų. Kaip matyti, per šį laiką kalboje įsitvirtino nauji aptariamos realijos pavadinimai, kurias galima rasti tekstyne.

Žodynuose ne visada pavyksta rasti žodžių junginius, kuriuose pavartotas ieškomas žodis. Be to, dažnai žodis, kuris funkcionuoja pastoviame žodžių junginyje, verčiamas visai kitaip negu vienas savarankiškas žodis. Pvz., lietuviško žodžio *kodas* latviškas atitikmuo yra *kods*; žodyne pateikti keli žodžių junginiai su šituo žodžiu. Dažnai vertėjai susiduria su pastoviaisiais žodžių junginiais, į kuriuos įeina *kodas*, pvz., *mokesčių mokėtojo kodas*. Tam, kad tokio junginio neišverstų pažodžiui ir įsitikintų, ar tikrai parinko tinkamą atitikmenį, vertėjai galėtų pasinaudoti tekstynu, pvz.:

LT: ***Mokesčių mokėtojo kodas:*** *BDL MMD 65M10 Z352S.*

LV: ***Nodokļu maksātāja reģistrācijas numurs:*** *BDL MMD 65M10 Z352S*

LT: *...pareiškėjo pavadinimas, adresas ir* ***PVM mokėtojo kodas****...*

LV: *...pieteikuma iesniedzēja vārds, uzvārds/nosaukums, adrese un* ***PVN maksātāja numurs****...*

Iš pavyzdžių matome, kad žodžių junginiai su žodžiu *kodas* į latvių kalbą verčiami kitaip, o šie termino atitikmenys žodyne nepateikti.

Be to, lygiagrečiame tekstyne galima pamatyti vertimo ypatumus, įvertinti vertimo kokybę, t. y. naudoti vertimo studijose. Tai atskleidžia, pvz., neišverstų, sujungtų, suskaidytų sakinių kiekis, pagal originalo sakinį išlaikyta ar pagal vertimo kalbos ypatybes pertvarkyta žodžių tvarka, pažodinis ar kalbos, į kurią verčiama, ypatybes atspindintis vertimas; galima matyti, ar tinkamai perteiktos reikšmės, stilistiniai niuansai, žodžių žaismas, vaizdingi posakiai, terminai ir pan.

Turint didelį lygiagrečių tekstų kiekį ir matant metaduomenis, galima daryti išvadas apie vertėjo darbo kokybę, (ne)išradingumą, kalbos, iš kurios verčia, ir kalbos, į kurią verčia, mokėjimą.

Toliau pateiktuose pavyzdžiuose matyti, kaip vertėjas vaizdingus posakius iš latvių kalbos į lietuvių išvertė atsižvelgdamas į kalbų specifiką, tai kalbai būdingus palyginimus, pastoviuosius junginius, o ne pažodžiui:

LV: *Pašai* ***jānorij tas krupis*** (pažodžiui: *reikia nuryti tą rupūžę).*

LT: *Pačiai reikia* ***išgerti tą nuodų taurę****.*

LV: *Emīlijai treji brāļi, vīri kā* ***brieži*** (pažodžiui: *kaip elniai).*

LT: *Emilija turi tris brolius – vyrai* ***ąžuolai****.*

Analizuojant lygiagrečius sakinius, galima mokytis kitų kalbų, nes mokantis vartoti kitos kalbos žodžius reikalingas konkretus arba ilgesnis kontekstas, kurį galima rasti lygiagrečiuosiuose tekstynuose.

Kaip pavyzdys toliau straipsnyje pateikti keli lietuvių ir latvių kalboje plačiai vartojami žodžių junginiai.

Lietuvių kalboje yra veiksmažodžių, kurie vartojami su objektu, išreikštu dalies kilmininku, pvz., žodžių junginyje *nebaigėme gerti arbatos* daiktavardis yra kilmininko linksnio, o analogiškame latvių kalbos žodžių junginyje *nebijām beigušas dzert tēju* objektas yra pasakytas galininku. Tokią vartoseną lietuvių ar latvių kalbos besimokantis žmogus gali dažnai pamatyti tekstyne, kur randami tokie žodžių junginiai, pvz.:

LT: *Dar nebaigėme* ***gerti*** *arbatos.*

LV: *Vēl nebijām beigušas* ***dzert*** *tēju.*

Latvių kalboje dažnai vartojamos prielinksninės konstrukcijos, o lietuvių kalboje dažnai analogiškos konstrukcijos pasakomos kitaip – su įnagininku, pvz.:

LT: *Minkštu sniegu važiuoti* ***dviračiu*** *tai mirtinas triukas.*

LV: *Mīkstā sniegā braukt* ***ar velosipēdu*** *ir nāves triks.*

LT: *Dabar trokšta važinėtis siaurais* ***keliukais****...*

LV: *Tagad vēlas braukt* ***pa*** *mazajiem* ***celiņiem****...*

Žinoma, galima rasti atvejų, kai lietuvių kalboje vartojama prielinksninė konstrukcija, o latvių kalboje neprielinksninė, pvz.:

LT: - *Jis man* ***šovė į galvą****, jau prasižiojus.*

LV: - *Tas man* ***iešāvās galvā****, kad jau biju pavēris muti.*

LT: - *Man iškart ji pasirodė kaip kekšė* ***su*** *ta trumpa* ***suknele****.*

LV: - *Man viņa uzreiz izskatījās pēc maukas tajā īsajā* ***kleitiņā****.*

Besimokantys kitos kalbos gali pasitikrinti tekstyne, kokia konstrukcija taisyklinga ir įprasta. Tekstynas pravartus besimokantiems kitos kalbos tada, kai reikia palyginti, išmokti tai, kas nesutampa su gimtąja kalba, pvz., anksčiau minėti linksnių ir prielinksnių junglumo atvejai.

Lygiagretieji tekstynai labai naudingi lyginamiesiems kalbų tyrimams, tiriant konkrečius kalbinius aspektus. Toliau pateiktuose pavyzdžiuose pateikti lietuvių kalbos neveikiamosios rūšies dalyviai ir jų atitikmenys latvių kalba. Latvių kalboje yra 5 nuosakos – viena iš jų debityvas. Lietuvių kalboje nėra tokios nuosakos ir poreikis išsakomas kitaip. Tekstyne galima analizuoti, kaip lietuvių kalboje išreiškiamas poreikis tais atvejais, kai latvių kalboje vartojamas debityvas, pvz.:

LT: *Žinojau, kaip tai* ***daroma****.*

LV: *Es zināju, kā tas* ***ir jādara****.*

LT: *...yra tiesiogiai* ***pripažįstamas*** *ir* ***vykdomas*** *kitoje valstybėje narėje.*

LV: *...šis lēmums* ***ir*** *tieši* ***jāatzīst*** *un* ***jāizpilda*** *citā dalībvalstī.*

Kaip minėta, lygiagrečiajame tekstyne galima analizuoti ir terminus. Kaip matyti toliau pateiktuose pavyzdžiuose,

lietuvių kalboje vartojamas deminutyvinis terminas *lauro lapelis*, o latvių kalboje šio termino atitikmuo yra ne deminutyvinis daiktavardis:

> LT: *Į troškinį krėskite pomidorų pastą, dėkite kornišonus, cukrų, druską, pipirus,* ***laurų lapelius****, gvazdikėlius ir kaitinkite dar 15 minučių.*
>
> LV: *Sautējumam pievienojiet tomātu pastu, pielieciet kornišonus, cukuru, sāli, piparus,* ***lauru lapas****, krustnagliņas un karsējiet vēl 15 minūtes.*

Kita lygiagrečiųjų tekstynų panaudojimo sritis – kalbų technologijos. Šiuolaikinės statistinio automatinio vertimo sistemos kuriamos lygiagrečiųjų tekstynų pagrindu. Kuo didesnė lygiagrečiųjų tekstų apimtis, tuo geresnę vertimo kokybę galima pasiekti tokiomis sistemomis. Nors LILA tekstynas dar nėra tokio dydžio, kad užtikrintų gerą lietuvių–latvių ar latvių–lietuvių kalbų automatinio vertimo kokybę, bet šis tekstynas yra gera būsimų darbų pradžia.

Lygiagretieji tekstynai ypač reikalingi kuriant dvikalbius žodynus ar vykdant lyginamuosius kalbų tyrimus. Tokiuose tekstynuose leksikografai gali rasti ne tik žodžių atitikmenis, bet ir dažnines žodžių vartojimo charakteristikas, autentiškus vartojimo pavyzdžius. Šiam tikslui gali būti sukurti specializuoti įrankiai, pavyzdžiui, Ciuricho universiteto *Bilingwis* (Volk et al., 2011).

Apibendrinamosios pastabos

Straipsnyje pristatytas naujas 2011–2012 m. parengtas lietuvių–latvių ir latvių–lietuvių kalbų tekstynas LILA, sudarytas iš beveik 9 mln. žodžių. Šis tekstynas pusiau automatiškai, naudojant VDU KLC sukurtą įrankį *Aligner 2.0.6.7,* sulygiagretintas pastraipų ir sakinių lygmeniu. Prie kiekvienos konkordanso eilutės nurodomi metaduomenys apie autorių, teksto pavadinimą, teksto parengimo metus.

Rengiant šį tekstyną susidurta su mažai resursų turinčioms kalboms būdingomis problemomis: trūksta vertimų iš lietuvių į latvių kalbą ir atvirkščiai, sunku gauti reikalingų tekstų, ilgai užtrunka spausdintų tekstų skaitmeninimas ir parengimas dėti į tekstyną. Dėl šių priežasčių pristatomo tekstyno dydis ir pobūdis priklauso nuo turimų kalbinių resursų.

LILA tekstynas, kaip ir kiti lygiagretieji tekstynai, pravartus besimokantiems kalbų, nes tekstyne išryškėja leksiniai ir gramatiniai kalbų skirtumai; lengviau išmokti ar suprasti žodžius ilgesniame realiai vartojamos kalbos kontekste.

Šis tekstynas turėtų tapti svarbiu pagalbininku ir vertėjams, nes jie galės matyti daug vienoje vietoje ieškomo žodžio pavartojimo atvejų, vertimo atitikmenų.

Vertimus analizuojantys teoretikai ir praktikai LILA tekstyne galės įvertinti vertimo kokybę, palyginti pastoviųjų junginių, vaizdingų posakių, terminų vertimo ypatybes. Šiuo aspektu pristatomas tekstynas svarbus ir leksikografams.

LILA tekstynas bus svarbus resursas kalbų technologijų srityje, ypač kuriant automatinio vertimo sistemas.

Literatūra

1. Gale, W. A., Church, K. W., 1993. A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics,* 19 (1), pp. 75–102.
2. Kovalevskaitė, J., 2012. *Lietuvių kalbos samplaikos.* Daktaro disertacija. Kaunas: VDU leidykla.
3. Levāne-Petrova, K., 2012. Latviešu-lietuviešu-latviešu paralēlo tekstu korpusa izveide. *Vārds un tā pētīšanas aspekti: rakstu krājums*, Nr. 16 (2). Liepāja: LiePA, pp. 180–188.
4. Marcinkevičienė, R., Kovalevskaitė, J., Utka, A., Vaičenonienė, J., 2012. An Overview of Multilingual Processing for Lithuanian. In: C. Vertan, W. v. Hahn (eds.). *Multilingual Processing in Eastern and Southern EU Languages: Low-Resourced Technologies and Translation*. Cambridge Scholars Publishing, pp. 119–143.
5. Rimkutė, E., Kovalevskaitė, J., Daudaravičius, V., 2006. Daugiakalbių tekstynų naudojimas ir taikymas. *Darbai ir dienos*, 45, pp. 41–62.
6. Rodyklis – Latvieši, igauņi un lietuvieši: literārie un kultūras kontakti. Bibliogrāfiskie rādītāji. Latvijas Universitāte: Literatūras, folkloras un mākslas institūts, 2008.
7. Utka, A., Kovalevskaitė, J., Rimkutė, E., Daudaravičius, V., 2008. Bilingual Parallel Corpora for English, Czech and Lithuanian. *Proceedings of the Third Baltic Conference on Human Language Technologies 2007*. Kaunas, pp. 319–326.
8. Utka, A., Levāne-Petrova, K., Bielinskienė, A., Kovalevskaitė, J., Rimkutė, E., Vēvere, D., 2012. Lithuanian-Latvian-Lithuanian Parallel Corpus. In: A. Tavast, K. Muischnek, M. koit (eds.). *Human Language Technologies. The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012.* Amsterdam, Berlin, Tokyo, Washington, DC: IOS Press, pp. 260–264.
9. Vaičenonienė, J., 2012. *Lietuvių literatūra anglų kalba: tekstynu paremtas autorinių neologizmų vertimo tyrimas*. Daktaro disertacija. Kaunas: VDU leidykla.
10. Volk, M., Göhring, A., Lehner, S., Rios, A., Sennrich, R., Uibo, H., 2011. Word-aligned Parallel Text: A New Resource for Contrastive Language Studies. *Supporting Digital Humanities, Conference 2011*, Copenhagen, Denmark.

Erika Rimkutė, Andrius Utka, Kristīne Levāne-Petrova

Lithuanian-Latvian, Latvian-Lithuanian Parallel Corpus (LILA)

Summary

The paper presents a new linguistic resource, LILA, which is the Lithuanian-Latvian-Lithuanian parallel corpus aligned on paragraph and sentence level. The total size of the LILA corpus is 9 m words. So far it is a unique resource for this language pair. The corpus contains metadata with bibliographical information (title, author, year of publishing, etc.). The corpus contains the structural annotation, which includes boundaries of aligned segments, paragraphs, and sentences. The alignment of paragraphs and sentences has been done by the semi-automatic alignment tool *Aligner 2.0.6.7.* The corpus was compiled during 2011-2012 by scientists of the Vytautas Magnus University's Centre of Computational Linguistics (VMU CCL) and the Latvian University's Mathematical and Informatics Institute's Laboratory of Artificial Intelligence (LU MII). The paper describes problems and challenges that need to be solved, when a parallel corpus for two small languages is created. The limited choice of appropriate parallel material poses the most difficult obstacle, as then it is difficult to compile a corpus of desired size. The paper presents: the conception and structure of the LILA corpus, phases of its compilation, the alignment tool, the query system, and examples of usage. The corpus is especially useful for teaching and learning languages, for comparing languages, for compilation of dictionaries, and for developing language technology tools (e. g. statistical machine translation systems).

Straipsnis įteiktas 2013 06
Parengtas spaudai 2013 11

Apie autorius

Erika Rimkutė, humanitarinių mokslų daktarė, Vytauto Didžiojo universiteto Lietuvių kalbos katedros docentė, Kompiuterinės lingvistikos centro vyresnioji mokslo darbuotoja.
Mokslinės veiklos sritys: tekstynų lingvistika, kompiuterinė lingvistika, automatinė morfologinė analizė ir sintezė, morfologinis daugiareikšmiškumas, automatinė sintaksinė analizė.
Adresas: Vytauto Didžiojo universitetas, Humanitarinių mokslų fakultetas, Kompiuterinės lingvistikos centras, K. Donelaičio g. 52–206, 44244 Kaunas.
El. paštas: e.rimkute@hmf.vdu.lt

***Kristīne Levāne-Petrova**,* Latvijos universiteto Matematikos ir informatikos instituto Dirbtinio intelekto laboratorijos mokslo darbuotoja.
Mokslinės veiklos sritys: tekstynų lingvistika, kompiuterinė lingvistika, automatinė morfologinė analizė, gretinamieji ir lyginamieji kalbų tyrimai, vertimas.
Adresas: Latvijos universitetas, Matematikos ir informatikos institutas, Dirbtinio intelekto laboratorija, Raiņa bulvāris 29, Rīga, LV-1058.
El. paštas: kristine.levane-petrova@lumii.lv

***Andrius Utka**,* humanitarinių mokslų daktaras, Vytauto Didžiojo universiteto Lietuvių kalbos katedros docentas, Kompiuterinės lingvistikos centro vadovas.
Mokslinės veiklos sritys: tekstynų lingvistika, kompiuterinė lingvistika, automatinis ir automatizuotas vertimas, statistinė kalbos analizė.
Adresas: Vytauto Didžiojo universitetas, Humanitarinių mokslų fakultetas, Kompiuterinės lingvistikos centras, K. Donelaičio g. 52–206, 44244 Kaunas.
El. paštas: a.utka@hmf.vdu.lt