ISSN 1648-2824 KALBŲ STUDIJOS. 2002. NR. 3 * STUDIES ABOUT LANGUAGES. 2002. NO. 3

KOMPIUTERINĖ LINGVISTIKA/ COMPUTATIONAL LINGUISTICS

Palyginamojo tekstyno kūrimo principai, problemos ir panaudojimo galimybės

Jurgita Mikelionienė

Anotacija. Tekstynas (angl. *Corpus*) – programiškai apdorotas elektroninių tekstų rinkinys, skirtas atspindėti realią kalbos vartoseną. Dėl savo apimties, prieinamumo, lingvistinių ir enciklopedinių žinių ir kitų ypatybių pasaulinėje kalbotyroje pastaraisiais dešimtmečiais tekstynų sudarymas laikomas vos ne pagrindiniu, o svarbiausia, labai patikimu kalbos tyrimo metodu ir priemone. Tekstynų lingvistika tapo vienos iš perspektyviausių šiuo metu kalbotyros šakos – kompiuterinės lingvistikos – sudedamąja dalimi. Kauno technologijos universitete (KTU) pradėtas rinkti naujo lietuvių kalbotyroje tipo lyginamasis tekstynas. Tai dvikalbis lietuvių ir anglų kalbų tekstynas, atspindintis vieno funkcinio stiliaus – šiuolaikinės mokslo kalbos ypatybes. Jo tekstų tematika labai konkreti ir apibrėžta – tai technologijos (chemijos, transporto, elektros, informatikos ir kt.) mokslų kalba. Straipsnyje pateikiamas kuriamo tekstyno modelis, aptariamos pagrindinės problemos, akcentuojami medžiagos paieškos ir atrankos kriterijai, svarstomas tekstyno pragmatiškumo klausimas, bandoma nuspėti palyginamojo tekstyno pritaikymo gali-mybes (terminologijoje, leksikografijoje, vertimuose ir kt.).

Pagrindiniai dabartinės lietuvių kalbos tekstynai. Lietuvių kalbos ir matematikos instituto mokslininkų 1995 m. sukurtas pirmasis 1 mln 200 tūkst. žodžių bei Vytauto Didžiojo universiteto (VDU) 100 mln žodžių tekstynai, bendrieji jų sudarymo principai tekstynų kūrėjų yra gana išsamiai aprašyti (Grumadienė, 1995; Marcinkevičienė, 2001). Tai vienakalbiai tekstynai, atspindintys tam tikrų laikotarpių kalbą. Mažesnis tekstynas sudarytas pagal griežtą imčių sistemą (atspindėti keturi funkciniai stiliai – publicistinis, mokslinis, beletristinis, administracinis – po 300 imčių, kurių kiekvienos dydis 1000 žodžių). Panašios konstrukcijos ir dydžio buvo pirmieji tekstynai, pvz., 1964 m. JAV sukurtas Brauno tekstynas. VDU tekstyno apimtį galima palyginti su anglų kalbos 100 mln žodžių BNC *(British National Corpus*) ar dar didesniu vokiečių kalbos COSMAS tekstynu (daugiau nei 700 mln nekaitomų žodžių ir kaitomų žodžių formų pavartojimų). Didesniojo lietuvių kalbos tekstyno pagrindinis tekstų masyvas periodiniais leidiniais iliustruoja publicistikos kalbą, taip pat nemažai yra grožinės, mokslinės literatūros, mažiau administracinės kalbos pavyzdžių. Čia tekstai neskaidomi į imtis, neretai įtraukiamas visas tekstas (žurnalas, laikraštis, knyga ir pan.), todėl jį santykinai būtų galima laikyti ištisinių tekstų tekstynu (angl. *full text. c.*). Dalį pastarojo tekstyno medžiagos sudaro ir verstiniai tekstai. Tekstyno rengėjai dalyvavo keliuose projektuose, kuriuose buvo paralelinami Platono, Dž. Orvelo knygų vertimai, šiuo metu lygiagretinami Europos Sąjungos dokumentų tekstai.

Šio straipsnio pagrindinis tikslas – pristatyti naują projektą – KTU pradėtą rengti specialųjį dvikalbį (kol kas) tekstyną. Tai lietuvių – anglų kalbų palyginamasis tekstynas, sudarytas iš įvairių technologijos sričių mokslinio stiliaus kalbos tekstų. Tiek savo empirine medžiaga, tiek struktūra bei dvikalbe prigimtimi – tai naujo tipo tekstynas lietuvių kalbotyroje, o kalbant apie tai, kad tai yra mokslo kalbos tekstynas, jis dar labai retas tiek Europos, tiek ir apskritai viso pasaulio praktikoje (pvz., čekų–anglų kompiuterijos, italų–anglų medicinos sričių tekstynai). Todėl nenuostabu, kad tekstyno sudarytojams kyla nemažai klausimų, susijusių su jo sandara, medžiagos atrankos kriterijais ir apimtimi. Apskritai, reikia pastebėti, kad tekstynų lingvistikai (ypač tų šalių, kuriose ši kalbotyros šaka žengia pirmuosius žingsnius) skirti darbai dažniau yra metodologinio nei teorinio pobūdžio. Ne išimtis ir šitas straipsnis.

KTU tekstynas yra ruošiamas laikantis tekstynų kūrimų tradicijos. Jo rengimas susideda iš trijų etapų: vizijos kūrimo, medžiagos gavimo ir tekstų aprašymo bei žymėjimo (Kennedy, 1998:70). Šiame straipsnyje juos smulkiau ir paanalizuosime.

Tekstyno tipas. Tekstynų lingvistikoje, kai kalbama apie dvikalbius (angl. *bilingual corpora*) ar daugiakalbius tekstynus (angl. *multilingual corpora*), aptariamos dvi pagrindinės jų rūšys: lygiagretusis tekstynas (angl. *parallel c.*) ir palyginamasis (angl. *comparable c.*) tekstynas. Abiejų pagrindas – originalo ir vertimo tekstai. Skiriasi tų tekstų santykis ir pateikimo forma.

Lygiagrečiajame, dar vadinamame paraleliu ar vertimo tekstynu, pvz., anglų–norvegų kalbų tekstynas ENCP (*The English-Norvegian Parallel Corpus*), imamas būtinai tas pats literatūros šaltinis, pvz., Šv. Raštas. Paskui suvienodinama visų į tekstyną įeinančių kalbų tekstų struktūra:

sulygiagretinama teksto struktūra: pastraipų, sakinių kiekis. Anot W. Teuberto (1996:245), lygiagretusis tekstynas gali būti sudarytas iš:

1) originalaus teksto, parašyto A kalba, ir jo vertimo į B bei C kalbas;

2) vienodo dydžio originalių A ir B kalba parašytų tekstų ir jų atitinkamų vertimų;

3) tik vertimų A, B, C, kalbomis tų tekstų, kurių originalas parašytas Z kalba.

Nors dauguma lygiagrečių tekstynų sudaryti iš grožinės literatūros tekstų, tačiau pastebimiausi rezultatai matomi iš lygiagrečios techninės dokumentacijos, įvairiakalbių vartojimo instrukcijų tekstynų panaudojimo.

Palyginamojo tekstyno sandara ne tokia griežta. Svarbu tik, kad tekstai būtų artimos tematikos. Tekstynų lingvistikos teoretikai ir praktikai tokio tipo tekstynus apibūdina taip: tai rinkinys atskirų vienakalbių tekstynų, kurie vartojami tokių pačių ar panašių pavyzdžių analizei iš skirtingo turinio tekstų, bet visomis tekstyną sudarančiomis kalbomis (McEnery, Wilson, 1996:57). Palyginamasis tekstynas gali būti sudarytas tik iš įvarių kalbų vertimų, iš skirtingų kalbų originalių tekstų bei mišriuoju atveju – ir iš originalių, ir iš verstinių tekstų.

KTU tekstyno tipas balansuoja tarp bendrojo ir specialiojo. Į bendrąjį jis panašus tuo, kad gali būti skirtas daugialypei analizei, o į specialųjų – kad kai kurie jo pritaikymo tikslai yra visiškai aiškūs: tai gali būti ir kalbos mokymo priemonė. Vis dėlto aptariamasis tekstynas laikomas specialiuoju, daugiausia turint galvoje griežtai apibrėžtą jo tekstų specifiką.

Tekstyno sandara (angl. *corpus design*) yra labai svarbus tekstyno metodologijos aspektas, apimantis tris pagrindinius kriterijus: dydį, laiką, kada buvo parašytas tekstas, ir kalbos tipą – rašomosios ar šnekamosios kalbos elektroninis masyvas. Pastarasis požymis yra vienareikšmis. Imami tik parašyti ir tik publikuoti tekstai. Specialiojo tekstyno dizainui skiriamas ne toks didelis dėmesys kaip bendrojo. Tačiau vienas reikalavimas yra itin problemiškas. Tai reikalavimas į tekstyną įkelti visą leidinį (knygą, straipsnį ir kt.) (Pearson, 1998:59). Jis grindžiamas nuomone, kad jeigu į tekstyną bus įtraukiamos tik įvadinės kūrinio dalys, tai į jį nepateks nemažai sudėtinių terminų, o jei knygos dalis bus imama kur nors iš galo, nebus terminų definicijų, kurios labai reikalingos terminologams ir terminografams. Deja, tas reikalavimas kai kurioms mūsų leidykloms asocijuojasi su autorių teisių pažeidimu, nors iš tikro taip nėra (jei knyga jau išleista, ją galima ir nusiskanuoti, ir rankiniu būdu susivesti į kompiuterį).

Dydis. Kol kas numatoma surinkti 1 mln žodžių tekstyną. Galima svarstyti: daug tai, ar mažai? Jei šių dienų bendrojo pobūdžio tekstynui tai būtų juokingas skaičius (minimaliu laikomas 10-20 mln žodžių tekstynas), tai šio tekstyno, kaip specialiojo, mokslo kalbos specifika, tokią apimtį leidžia laikyti optimalia (Yang, 1986; Fang, 1991). Be to, anksčiau vyravusią nuomonę, kad tekstynas turi būti kiek įmanoma didesnis (Sinclair, 1991:18), dabar keičia nuostata, kad ne į dydį turi būti kreipiamas pagrindinis dėmesys, o į tekstyno reprezentatyvumą. Dydis yra tas kriterijus, kurį lemia pasirinkti tekstyno kūrimo tikslai (Biber, 1993:256; Kennedy, 1998:68).

Chronologinės ribos. Tekstyno kūrimo darbai turėtų trukti bent trejus metus, pradedant skaičiuoti nuo 2002 m. Tačiau tekstų parašymo pradžia nuspręsta laikyti 2000-uosius metus. Manoma, kad daugiausia terminologiniais tikslais kuriamas tekstynas negali būti senesnis nei 10 metų, vadinasi, jis nuolat turi būti pildomas ir nuolat peržiūrimas, atsisakoma šaltinių su pasenusia terminija.

Tekstų klasifikacija. Pagrindinis tekstyno sudedamasis vienetas yra tekstas. Požymiai, pagal kuriuos yra klasifikuojami į tekstyną dedami tekstai, yra nevienalyčiai. Nusistovėjęs jų skirstymas į kalbinius ir nekalbinius, dar kitaip į vidinius ir išorinius kriterijus (Atkins et al., 1992). Kalbiniams kriterijams yra priskiriami tema ir funkcinis stilius, o nekalbiniai apima leidinio tipą (žurnalas, knyga), žanrą, duomenis apie teksto autorių (kartais nurodoma ne tik vardas ir pavardė, bet ir įvairūs sociokultūriniai požymiai: amžius, lytis, tautybė, tarmė, mokslo laipsnis), leidyklą, vertėją ir pan.

Kauno technologijos universiteto, kuriame kaupiamas tekstynas, studijų ir atliekamų mokslinių tyrimų specifika, spartus technikos progresas, kalboje pasireiškiantis naujos terminijos gausa, apskritai, stygius tokio pobūdžio elektroninių tekstų rinkinio, lėmė tekstyno tematiką ir sandarą. Kartais juokaujama, kad kiek yra tekstynų, tiek ir nuomonių apie tai, kaip klasifikuoti tekstus pagal temas. Siūloma ne kurti kiekvieną kartą vis naujas klasifikacijas, o remtis jau esančiomis, pvz., UDK. KTU tekstyną sudarant remiamasi 1998 m. Lietuvos Respublikos Vyriausybės patvirtinta ir Europos Sąjungoje galiojančia mokslo krypčių ir šakų klasifikacija. Tekstyne technologijos mokslai suskirstyti į 10 grupių: elektros ir elektronikos inžinerija, statybos, transporto, chemijos, informatikos, medžiagų, mechanikos ir matavimų inžinerija, aplinkos inžinerija ir kraštotvarka bei energetika ir termoinžinerija. Kiekvienos grupės viduje klasifikuojama dar detaliau, pvz., transporto inžinerija apima kelių transporto technologiją, geležinkelių, vandens, oro transporto technologijas. Kadangi teksto temos nustatymas yra gana problemiškas dalykas, tai tekstyno duomenų bazėje kiekvienam šaltiniui stengiamasi pateikti ir prasminius žodžius (angl. *key-words*), pagal kuriuos taip pat galima vykdyti paiešką. Jau iš pateiktos tekstų klasifikacijos matyti, kad ruošiamame tekstyne atspindimos kalbos vartojimo sfera yra labai konkreti, turinti apibrėžtas ribas. Visi tekstai priklauso tik mokslo stiliui ir yra skiriami arba specialisto specialistui, arba specialisto specifinei, neretai akademinei auditorijai.

Aptariamame dvikalbiame palyginamajame tekstyne galima rasti šiuos mokslinio stiliaus žanrus: straipsnį, disertaciją ar jos santrauką, konferencijos pranešimą, tezes, monografiją, vadovėlį ar kitą metodinę priemonę aukštosioms mokykloms. Duomenų bazėje be žanrų atsispindės ir kai kurie kiti nekalbiniai kriterijai.

Medžiagos rinkimas (angl. *collection of corpora*). Tekstai į tekstyną renkami dviem būdais. Pirmasis, tradicinis, – iš leidyklų imamos elektroninės knygų versijos. Antrasis – šiuo metu pats efektyviausias (Holmes-Higgin, Ahmad,

1996) – medžiaga, daugiausia anglų kalba, renkama iš didžiausios tekstų saugyklos – Interneto. Tačiau Internetas kartais vaizdingai palyginamas su žinių jūra, iš kurios ne taip lengva tų žinių „atsigerti“. Lietuvių kalba parašytų tekstų apskritai nepakanka. Renkant specialųjį tekstyną nesilaikoma gana paplitusio medžiagos kaupimo principo „ką randu, tą dedu“, todėl trūkstant vienos ar kitos mokslo šakos tekstų kompiuterinių versijų, reikalinga medžiaga rankiniu būdu suvedama į kompiuterį arba nuskaitoma tekstų skaitytuvu (skanuojama). Šis būdas nėra praktiškas, reikalauja nemažų laiko sąnaudų, susijusių su klaidų taisymu. Renkant medžiagą labai svarbus kolektyvinis darbas. Čia negali būti vienos nuomonės. Todėl tekstyno rengėjai renka atskirų teminių grupių tekstus. Tekstų šaltinių įvairovė lemia ir jų formatų nevienodumą. Tai tampa problema tada, kai tekstai, buvę skirti skaityti, dabar tampa skirtais tirti. Todėl nuo medžiagos rinkimo neatsiejamas ir tekstų standartų vienodinimas.

Negali renkant tekstyną likti nuošalyje loginis medžiagos atrankos principas. Galimi du keliai: indukcinis arba dedukcinis (Holmes-Higgins, Ahmad, 1996). Pristatomo tekstyno sandara konstruojama remiantis dedukciniu metodu, t.y. einama nuo idėjos (stilius, žanrai) į tekstų, atitinkančių sumanymo kriterijus, paiešką. Renkantis indukcinį kompiliacijos kelią, reiktų turėti etaloninį tekstų branduolį, kuris, padedamas tam tikrų loginių indukcinių procedūrų, imituotų biologinį augimą. Ir vienu, ir kitu keliu einant galima sukurti pakankamai reprezentatyvų tekstyną.

Tekstyno reprezentatyvumas. Tai vienas labiausiai diskutuotinų tekstynų lingvistikos klausimų. Egzistuoja tam tikras suvokimas, koks tekstynas laikomas reprezentatyviu. Pirmiausia, jis turi būti kuo labiau subalansuotas tiek temų, tiek žanrų, tiek tekstų prieinamumo, tipiškumo ir kt. Atžvilgiu. Lietuvių kalbos tekstynų kūrėjai savųjų tekstynų reprezentatyvumą traktuoja nevienodai. Galima nujausti, kad mažojo tekstyno autoriai mano, kad subalansuotą tekstyną atspindi vienodo dydžio imčių vienodas kiekis kiekvienam stiliui. Didžiojo tekstyno kūrėjai atsargesni. Jie tiesiogiai nepasako, kad jų tekstynas yra tikrai reprezentatyvus, tačiau mano, kad jis pakankamai įvairus ir subalansuotas (Marcinkevičienė, 2001). KTU specialaus mokslo kalbos tekstyno pobūdis leidžia reprezentatyvumą traktuoti dar kitaip. Žinoma, siekiama, kad tekstynas atspindėtų visas dešimt technologijos mokslų šakas ir jų poklasius, tačiau tekstai dedami į tekstyną ne lygiomis dalimis, o stengiamasi atsižvelgti į leidybos tendencijas. Pvz., elektros ir elektronikos inžinerijos grupei priskirtinų tekstų 2000 m. „Technologijos“ leidyklos planuose buvo suplanuota išleisti apie 425 lankų, o aplinkos inžinerijai – tik 45, todėl ir tekstyne numatoma išlaikyti panašų santykį (10:1).

Svarbus reprezentatyvumo kriterijus yra teksto kokybė. Todėl specialaus tekstyno tekstų autoriams keliamas vienas pagrindinis reikalavimas: jie turi būti pripažinti, geri savo mokslinių tyrimų srities specialistai (pageidautina, kad turėtų mokslinį laipsnį), jų darbai turi turėti mokslinę vertę.

Tekstyno anotavimas (angl. *corpus annotation*). Kad tekstynas būtų parankus kompiuterinės lingvistikos tyrimams bei kitokiai statistiniu aprašu paremtai kalbos analizei, kad galėtų perteikti visą reikšminę informaciją, perduodamą morfologinėmis ir sintaksinėmis priemonėmis, jis turi būti anotuotas. Kaip rodo įvarių konferencijų, seminarų medžiaga, tekstynų lingvistikos baruose pagrindinis dėmesys šiuo metu skiriamas būtent tekstynų anotavimo problemai. Šiuo metu egzistuoja apie 20 didžiulių anotuotų pagrindinių Europos kalbų tekstynų, iš kurių bene didžiausias, apimantis kelis šimtus milijonų žodžių vartosenos atvejų, yra Pensilvanijos universitete sukurtas anglų kalbos tekstynas *Penn Treebank*. Lietuvoje visiškai anotuoto tekstyno dar nėra. Kaip suprantamas terminas tekstyno anotacija? Tai tekstyno papildymas aiškinamąja lingvistine informacija (Garside et al., 1997). Anotacija gali būti suvokiama ir kaip galutinis anotavimo proceso rezultatas: teksto vienetamas yra „prikabinami“ lingvistiniai žymekliai (angl. *tag*). Mūsų projektas numato kol kas sužymėti tik tekstyno branduolį, kuris taip pat turėtų būti ir suparalelintas: 1 mln žodžių tekstynas turėtų 100 tūkst. žodžių anotuotą tekstynėlį, sudarytą pagal analogiškas tekstų proporcijas visam tekstynui. Kadangi skiriasi tyrėjų tikslai, skirsis ir anotacijos lygmenys. Pirmiausiai tekstai bus sulemuoti, t. y. visoms žodžių formoms bus nurodyta jų pagrindinė forma (pvz., daiktavardžiams, būdvardžiams vienaskaitos vardininko linksnis, veiksmažodžiams – bendratis ir t.t.). Antrasis žymėjimo etapas susijęs su morfologinės informacijos pateikimu, t. y., be to, kas nurodoma lemuojant žodžius, dar pateikiama ir jo morfologinė charakteristika (kalbos dalis, giminė, skaičius, linksnis, nuosaka, laikas, laipsnis, asmuo ir kt.). Na, o sintaksinis aprašas be viso šito dar turėtų pasipildyti žymekliais apie teksto sintaksinę struktūrą. Sintaksinio anotavimo (parsingo) programų Lietuvoje dar nėra sukurta, na o lemuoklį – programą, atkuriančią žodžio pagrindinę formą – VDU tekstynui yra sukūręs Vyt. Zinkevičius. Sintaksinis mokslinio stiliaus tekstų žymėjimas yra paprastesnis nei beletristinių, nes čia praktiškai nepasitaiko dialogų, tiesioginės kalbos bei nutylėjimų (elipsės). Žinoma, tokia anotacija yra tik pusiau automatinė. Pirmuoju etapu tekstą analizuoja kompiuteris, tačiau kitą darbo pusę atlieka kalbininkai. Kartais jų darbas apsiriboja tik teisingai kompiuterių atlikto darbo patvirtinimu, o kartais turi lemiamą įtaką teisingo sprendimo buvimui (pvz., pasirenkant homoformą).

Reikia pasakyti, kad tekstynų lingvistikoje dar pasitaiko svarstymų apie tai, jog anotuotas tekstas – tai jau sugadintas tekstas. Šios nuomonės šalininkai prioritetus atiduoda „žaliam“, neapdorotam elektroninių tekstų masyvui (Mihailov, Tommola 2001:71), teigdami, kad anotuotą tekstą sunku skaityti, problemiškas bet koks pakeitimas jame. Manytumėme, kad KTU renkamo tekstyno anotacijos būtinumą pagrindžia net tokie argumentai:

1) jei bus tekstas nesužymėtas kalbos dalimis, ieškant tam tikrų leksinių vienetų, bus pateikiamos homoformos (pvz., *kiškis* – veiksmažodis ir daiktavardis);

2) šnekos sintezatoriui taip pat būtina informacija apie kalbos dalis, nes kai kada vienodo grafinio pavidalo žodžiai ne tik priklauso skirtingoms kalbos dalims, bet skiriasi ir jų tarimas (pvz., *artì – árti*);

3) ruošiant mašininio vertimo algoritmus, žymėtas tekstas yra daug parankesnis. Pagaliau, norint, kad tekstus būtų patogu skaityti, galima turėti ir neanotuotą tekstyno versiją.

Tekstyno pragmatiškumas. Savaime suprantama, jog tekstynas kuriamas ne dėl paties savęs: jam numatomas praktinis pritaikymas. Kadangi renkamas palyginamasis tekstynas bus dvikalbis, idealu būtų, jei jis būtų sudarytas taip, kad tiktų gretinamosioms kalbų studijoms. Tačiau negalima pamiršti, kad nacionalinio tekstyno formavimas turi atitikti savas filologines tradicijas (jos dar tik apie 10 metų tęsiasi), o svetimkalbių negalima imituoti, reikia tik jų paisyti (Rykov, 1999). Ir tai galioja ne tik tekstų anotavimui, bet ir pačiam tekstyno konstravimui. Tekstynas tampa pagrindine žaliava daugeliui kalbų tyrimų ir kalbų technologijoms (Teubert, 1996a). Panašią išvadą – bet kokių kalbos technologijų neįmanoma kurti be sąlyčio su tekstynų lingvistais – daro ir kiti kalbininkai, analizavę darbo su tekstynais privalumus (Marcinkevičienė, 2000:57). Tai būtina priemonė šiuolaikinei leksikografijai, terminologijai, kalbų mokymui ir vertimams. Apie kiekvienos srities, paremtos KTU tekstyno medžiaga, perspektyvą pakalbesime plačiau.

Apie galimybę versti pasiremiant tiek iš originalių, tiek iš verstinių tekstų sudarytais tekstynais bene pirmą kartą prabilta prieš mažiau nei 10 metų (Baker, 1993). Prognozės pasitvirtino: dabar ir palyginamieji, ir lygiagretieji tekstynai sėkmingai naudojami kaip šaltiniai gretinamosios kalbų studijoms bei kaip vertimų „treniruokliai“. Gausėjant informacijai, didėja poreikis ir jos vertimams. Įrodyta, kad didelis dvikalbis tekstynas visada geriau nei dvikalbis žodynas. Nenuostabu, kad ieškoma ne tik priemonių pagreitinti įprastą vertėjo darbą, bet apskritai greitesnių vertimo būdų. Turiu galvoje mašininį vertimą. Automatinis vertimas yra viena iš tų kompiuterinės lingvistikos sričių, kurios pastaruoju metu ypač aktualios. Tekstynų lingvistika siūlo naujus automatinio vertimo metodus. Kaip rodo kitų, pvz., čekų mašininio vertimo specialistų patirtis, nustačius vertimo vienetus ir vertimo ekvivalentus paraleliame technikos kalbos tekstyne, galima tikėtis 85 proc. vertimo efektyvumo (Čmejrek, Cuřín, 2001).

Kitas palyginamojo tekstyno uždavinys – tarnauti teminografinėms reikmėms. Pirmiausia – atskirų terminų, jų apibrėžimų, sinonimų ar vertimo ekvivalentų paieškai. Sukurtas programinis įrankis iš pradžių turėtų sudaryti originalo teksto terminų sąrašą, o paskui surasti tų terminų atitikmenis paraleliame tekstyne. Tekstynas turėtų būti ir pagrindine medžiaga atskirų žodynų rengimui. Sparčiai plėtojantis ar kuriantis naujoms įvairių sričių technologijoms, tenka pripažinti, jog dvikalbiai terminų žodynai yra arba pasenę, arba jų iš viso dar nėra. Šitas tekstynas galėtų pasitarnauti dviejų tipų žodynų rengimui. Pirmiausia, tai gali būti vienakalbis bendrojo pobūdžio technologijų terminų žodynas arba kurios nors specialiosios šakos terminų žodynas. Antras galimas tipas – dvikalbis specialusis žodynas. Tai irgi gali būti bendro pobūdžio ar vienos kurios nors srities žodynas. Ruošiant tekstynais paremtus žodynus, yra puiki galimybė išsiaiškinti ne tik terminų sinonimus, pagrindinius junginius, bet ir apibrėžimus. Šiuolaikinėje leksikografijoje vyrauja nuostata apie tai, kad terminų definicijos reikalingos tik vienakalbiuose žodynuose, o dvikalbiuose, kadangi jie skirti jau išmanančiam savo kalbos tam tikros srities terminus specialistui, užtenka vertimo atitikmenų ir gramatinių nuorodų. Drįstame pritarti tiems tyrėjams (Pearson, 1998), kurie mano, kad ir vertimo žodynuose definicijos tik padėtų surasti tinkamesnį vertimo atitikmenį. Kadangi technikos terminai būna vienažodžiai ir sudėtiniai, tai skiriasi ir jų paieškos tekstyne būdai. Vienažodžius terminus galima rinkti atsižvelgiant į jų pasikartojimo tekste dažnumą, o dvižodžius ar daugiažodžius pagal kolokacijų ypatybes. Identifikavimui kuriamas algoritmas.

Tekstynas daro perversmą ne tik kalbos tyrimo metodologijoje, bet ir pačiame kalbos suvokime. Galimybė praplėsti verčiamo kalbos vieneto ribas (ne žodis, o jau sakinys ar net daugiau), leidžia peržiūrėti nuostatą apie egzistuojančią skirtybę tarp sintaksės ir leksikos (Teubert, 1996a). Darbas su tekstynu dar tik pradėtas. Manome, kad rengiamas palyginamasis dvikalbis tekstynas, sudarytas iš technologijos mokslų kalbos tekstų, būdamas ganėtinai reprezentatyvus, iš dalies anotuotas, besiremiantis loginiais medžiagos atrankos principais, pasitarnaus tiek fundamentaliesiems, tiek taikomojo pobūdžio lietuvių kalbos tyrimams (pvz., įvairių sričių terminologiniams, terminografiniams darbams) bei gretinamosioms lietuvių–anglų kalbų studijoms.

Literatūra

1. Atkins, S., Clear, J., Ostler, N. (1992). Corpus Design Criteria// Literary and Linguistic Computing 7 (1). Oxford: Oxford University Press. P. 1-16.
2. Bacer, M. (1993). Corpus Linguistic and Translation studies: Implications and Applications//Text and Technology: in Honor of John Sinclair. Amsterdam/ Philadelphia: John Benjamins Publishing Company. P. 233-250.
3. Biber, D. (1993). Representativeness in Corpus Design//Literary and Linguistic Computing 8 (4). Oxford: Oxford University Press. P. 243-257.
4. Čmejrek, M. Cuřín, J. (2001). Automatic Ectraction of Terminological Translation Lexicon from Czech-English Parallel Texts// International Journal of Corpus Linguistics. Amsterdam/ Philadelphia: John Benjamins Publishing Company. P. 1-12.
5. Fang, C. Y. (1991). Building a Corpus of the English of computer science// English Language Corpora. Amsterdam: Rodopi. P. 73-78.
6. Garside R., Leech G., McEnery T. et al. (1997). Corpus annotation: linguistic information from computer text corpora. London and New York: Longman.
7. Grumadienė, L. (1995). Rengiamas lietuvių kalbos dažnumų žodynas// Kalbos kultūra (67), Vilnius. P. 91-96.
8. Holmes-Higgin, P., Ahmad, H. (1996). Assembling and Viewing a Corpus of Texts: Self-organisation, Logical Deduction and Spreadin Activation as Metaphors//Euralex'96 Proceedings. Stokholm, P. 250-269.
9. Yang, H. (1986). A New Technique for Identifying Scientific and Technical Terms and Describing Science Texts// Literary and Linguistic Computing, 1(2). Oxford: Oxford University Press. P. 93-103.
10. Kennedy, G. (1998). An Introdaction to Corpus Linguistics. London/New York: Longman.
11. Marcinkevičienė, R. (2000). Tekstynų ligvistika (teorija ir praktika)// Darbai ir dienos (24). Kaunas. P.7-63.
12. McEnery, T., Wilson A. (1996). Corpus Linguistics. Edinburgh: Edinburgh University Press.
13. Michailov, M., Tommola, H. (2001). Compiling Parallel Text Corpora// International Journal of Corpus Linguistics. Amsterdam/ Philadelphia: John Benjamins Publishing Company. P. 66-77.
14. Pearson,. J. (1998). Terms in context. Amsterdam/ Philadelphia: John Benjamins Publishing Company.

15. Рыков, В. В. (1999). Прагматически ориентированный корпус текстов// Тверской лингвистический меридиян. Теоретический сборник. Тверь. Вып. 3. С. 89-96.

16. Sinclair, J. M. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

17. Teubert, W. (1996). Comparable or Parallel Corpora?// International Journal of Lexicography (9)3. P. 238-264.

18. Teubert, W. (1996a). Editorial// International Journal of Corpus Linguistics 1(1). Amsterdam/ Philadelphia: John Benjamins Publishing Company. P. iii – x.

Jurgita Mikelionienė

Principles of Comparative Corpus Development, Its Problems and Use Possibilities

Summary

Corpus is a programmically processed collection of electronical texts; which aims at reflecting the real usage of language. Due to its volume availability, linguitic and encyclopaedic knowledge and other specific features the corpus has been considered next to the main and highly reliable method and means of language investigation in the world science of language. Corpus linguistics has become an inseparable component of one of the most perspective branches of language science – computer linguistics. Composing of comparative corpus that is of new type in the Lithuanian linguistics has been started at Kaunas University of Technology. This is a bilingual Lithuanian-English corpus that reflects the usage of one functional style – modern scientific language. Thematics of the texts is concrete and strictly limited – language of technology (chemistry, logistics, electricity, informatics, etc.) sciences. The paper presents the model of the composed corpus, discusses main problems, emphasizes the criteria of the material search and selection, analyzes the issue of the corpus pragmatism and applicability (terminology, lexicography, translation, etc).

Straipsnis įteiktas 2002 03
Parengtas spaudai 2002 11

Apie autorių

Jurgita Mikelionienė, dr., Vytauto Didžiojo universiteto Lietuvių kalbos katedra, Kompiuterinės lingvistikos centras, Kauno technologijos universiteto Lietuvių kalbos katedra.
Interesų sritys: tekstynų lingvistika, leksikologija, leksikografija, terminologija.
Adresas: Kauno technologijos universitetas, Humanitarnių mokslų fakultetas, Lietuvių kalbos katedra, Gedimino 43, LT-3000, Kaunas.
El.paštas: jurgam@eurotinklas.lt