ISSN 1648-2824 KALBŲ STUDIJOS. 2002. NR. 2 * STUDIES ABOUT LANGUAGES. 2002. NO. 2

Anglų kalbos tekstų kompiuterinio vertimo į lietuvių kalbą technologijos

Gediminas Misevičius, Bronius Tamulynas, Marius Žemaitis

Anotacija. Straipsnyje aptariami bendrieji kompiuterizuotų vertimo technologijų principai ir jų realizavimo galimybės. Analizuojami anglų kalbos tekstų alternatyvūs kompiuterizuoto vertimo metodai į lietuvių kalbą ir siūlomas konceptualusis hierarchinis adaptyviosios kompiuterinio vertimo sistemos modelis. Apžvelgiami tiesioginis, transformacinis ir tarpinės kalbos *Interlingua* vertimo principai ir jų realizavimo būdai siūlomo modelio architektūroje. Apibrėžiami funkciniai reikalavimai žinių šaltiniams, vertimo proceso valdymo blokui ir sąsajai su vartotoju. Aptariami vertimo terpės, teksto skirstymo, vertimo proceso, sakinio grupių išskyrimo ir gramatinės analizės programiniai moduliai. Vertimo principas grindžiamas tekstų transformavimo technologija, įjungiant pažodinio vertimo procedūras. Pateikiama trumpa pirmosios kompiuterinio vertimo sistemos versijos charakteristika. Ją sudaro vartotojo sąsaja, specializuotas virtualus kompiuterinis žodynas, anglų kalbos teksto skirstymo, vertimo varyklės (angl. *engine*) ir žinių šaltinių valdymo moduliai.

Įvadas

Šiuolaikinių kompiuterizuoto vertimo (KV) sistemų veikimas grindžiamas trimis vertimo principais: *tiesioginiu (pažodiniu), transformaciniu (perstatomuoju) arba per tarpinę, vadinamąją „interlingua“* kalbą (Trujillo, 1999; Arnold, 1994; Mitamura, 1991). Šie principai iš esmės atspindi tik atitinkamą vertimo proceso varyklės charakteristiką ir mažai priklauso nuo bendrosios KV sistemos architektūrinės koncepcijos. Straipsnyje pateikiamas KV modelis, orientuotas į hierarchinę modulinę sistemos architektūrą, kurios atitikmuo artimas intelektualiųjų ekspertinių sprendimų priėmimo lentos koncepcijai. Apibendrintą sprendimų lentos modelio schemą sudaro trys pagrindiniai elementai – *sprendimų lenta* (bendras veiksmų laukas, kuris skaidomas į dvi dalis išeities ir tikslinės kalbos tekstams saugoti)*, žinių šaltiniai* (aktyvūs tekstų analizės ir vertimo varyklės moduliai, kurie gali tiesiogiai „veikti“ bendroje atminties srityje) *ir valdymo posistemė,* kuri stebi, aktyvina ir valdo vertimo procesą. Tokiu būdu šios architektūros KV sistema pasižymi adaptyvumo savybe, nes žinių šaltiniai, būdami nepriklausomi, gali įsijungti į vertimo procesą ir veikti pagal valdymo komandą, kai jiems veiksmų lauke sukuriama palanki situacija, atitinkanti jų žinių lygį.

Vertimo varyklė siūlomo modelio kontekste gali būti realizuota vienu ar keliais žinių šaltiniais, papildant juos autorine tekstų koregavimo galimybe vartotojo lange. Išvardytus principinius varyklių privalumus ir trūkumus sunku vertinti vienareikšmiškai. Gali pasirodyti, kad pažodinis vertimas yra netobulas. Tačiau yra žinoma nemažai pažodinio KV sistemų, kurios sėkmingai platinamos. Iš tikrųjų, grynai pažodinio vertimo sistemų praktiškai ir nėra, nes kiekviena KV sistema realizuoja atitinkamas metodų kombinacijas. Pavyzdžiui, verčiant *perstatymo* metodu dažnai tenka pasinaudoti ir *tiesioginiu* vertimu, nes sakinio analizė ne visada gali būti patenkinama. Todėl, atsižvelgiant į tai, kokio rezultato mes siekiame, galime pasirinkti vieną iš minėtų vertimo principų.

KV sistemose verčiant tekstus iš anglų į lietuvių kalbą ypač opią problemą sudaro sinoniminis ir kontekstinis daugiaprasmiškumas. Be to, kiekviena kalba turi tik jai būdingą vienokią ar kitokią žodžių tvarką sakinyje, kai kurie žodžiai neturi atitikmens tikslinėje kalboje ir pan. Dėl šių ir kitų priežasčių vertimas interpretuojamas kaip tam tikras kūrybinis aktas. Nepaisant minėtų išskirtinių atvejų ir naudojant sudėtingesnius originalaus ir tikslinio teksto semantinės analizės programinius modulius, vertimo procesas su tam tikromis išlygomis gali būti kompiuterizuotas. Su panašiomis problemomis lanksčiau susidoroja KV sistemos, kurios naudoja kalbos žinių šaltinius, kaip nepriklausomus programinius vienetus ar komponentus (Tamulynas, 2001). Kaip antai: teksto ir sakinio analizės, žodžių analizės moduliai, gramatinės ir semantinės analizės komponentai, žinių gavyba iš tekstyno ir pan. Teksto analizės arba skirstymo šaltinis išskiria sakinius, o juose žodžius. Žodžių analizės šaltinis pateikia informaciją apie žodžių semantines, sintaksines savybes ir ieško frazės ar žodžių atitikmens žodynuose. Gramatinės analizės modulis normuoja žodžių tvarką sakiniuose ir pan. Jo tikslas sutvarkyti linksnius, laikus ir teisingai surikiuoti žodžius sakinyje. Sintaksinių sakinio grupių sudarymas ir jų panaudojimas gramatinės analizės komponente yra vienas iš būdų, kuris leidžia pasiekti geresnę kompiuterinio vertimo kokybę (Žemaitis, 2002).

Straipsnyje pristatoma naujos architektūros KV sistema ir jos pirmoji programinė realizacija, kuri verčia nesudėtingus specializuotus anglų kalbos tekstus į lietuvių kalbą.

Kompiuterizuoto vertimo proceso funkcinė struktūra

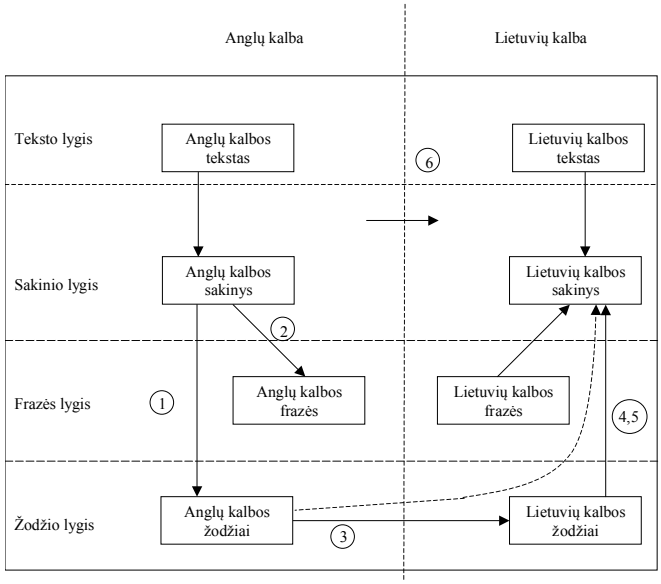
Vertimo proceso eigą galima aiškiau suvokti panagrinėjus elementaraus anglų kalbos sakinio (išeities teksto) *“A book is on the table.”* vertimą į lietuvių kalbą (žr. 1 lent.). Pirmojoje vertimo fazėje originalo (išeities) anglų kalbos sakinys skaidomas į atskirus žodžius. Pažodžiui išvertus visus žodžius, nustatoma jų prasmė, forma, kalbos dalis ir pan. Kituose žingsniuose, identifikavus žodžių funkciją sakinyje, formuojamas semantiškai adekvatus originalui sakinys bei gramatiškai derinami žodžiai.

1 lentelė. Anglų kalbos sakinio vertimo fazės

Sakinys originalo klb.	***A***	***Book***	***Is***	***on***	***the***	***table***
Reikšmė		*Knyga*	*yra (būti)*	*ant*		*stalas*
Kalbos dalis	žym. artikelis	Daiktavardis	veiksmažodis	prielinksnis	žym. artikelis	daiktavardis
Forma			esam. l., vnsk., 3 asm.			viet. linksnis
Funkcija sakinyje		Veiksnys	Tarinys			papildinys
Sakinys tikslinėje klb.		*Knyga*	*Yra*	*ant*		*stalo*

Šis paprastas pavyzdys, nors ir neatskleidžia vertimo proceso sudėtingumo, tačiau pakankamai vaizdžiai iliustruoja pagrindinius vertimo principus bei etapus.

Pirmajame paveikslėlyje vaizduojami apibendrinti pagrindiniai vertimo žingsniai ir atitinkama funkcinė schema.



1 pav. Vertimo proceso funkcinė schema

Funkcinėje schemoje skaičiais parodyta vertimo proceso pagrindinės fazės arba etapai:

1 – pradinis sakinio paruošimas (skirstymas, analizė, žodžių atributų išskyrimas);
2 – frazių (frazeologizmų) paieška, analizė, vertimas;
3 – pažodinė analizė (kalbos dalies nustatymas, originalo kalbos žodžio formos nustatymas, pažodinis vertimas);
4 – sakinio vertimas (žodžių atitikmens ir funkcionalumo sakiniuose nustatymas, jų suderinimas);
5 – galutinis tikslinės kalbos sakinio sutvarkymas (žodžių tvarka, skyrybos ženklai);
6 – semantinių ir pragmatinių žinių gavyba iš tekstynų ir tekstų suderinamumo „filtrų“ panaudojimas.

Vertimo proceso analizė leidžia skirti kai kurias būdingas jo ypatybes. Tai sudėtingas uždavinys: gausu žodžių ir frazių vertimo variantų, prasmė priklauso nuo konteksto, reikalingos papildomos probleminės žinios ir pan. Nesunku pastebėti, kad vertimas iš vienos kalbos į kitą turi aiškią hierarchiją, t. y. vertimo objektai (sakiniai, frazės, žodžiai) yra išsidėstę skirtinguose hierarchiniuose duomenų pateikimo lygmenyse. Siekiant paspartinti vertimo procesą, efektyviau išnaudoti kompiuterio resursus, dalis kompiuterizuoto vertimo žingsnių gali būti atliekami lygiagrečiai (pvz., žodžių ir frazių apdorojimas), kiekvienas vertimo etapas arba fazė gali būti nagrinėjami kaip atskiri uždaviniai, o juos vykdantieji programiniai moduliai gali būti aprašomi ir realizuojami nepriklausomai vienas nuo kito. Įvertinus išvardytus ypatumus, kompiuterinio vertimo konceptualiajam modeliui siūloma taikyti hierarchinę architektūrą, kuri žinių inžinerijoje atitinka, vadinamąjį sprendimų lentos modelį (Sidorov, Tamulynas, 2001).

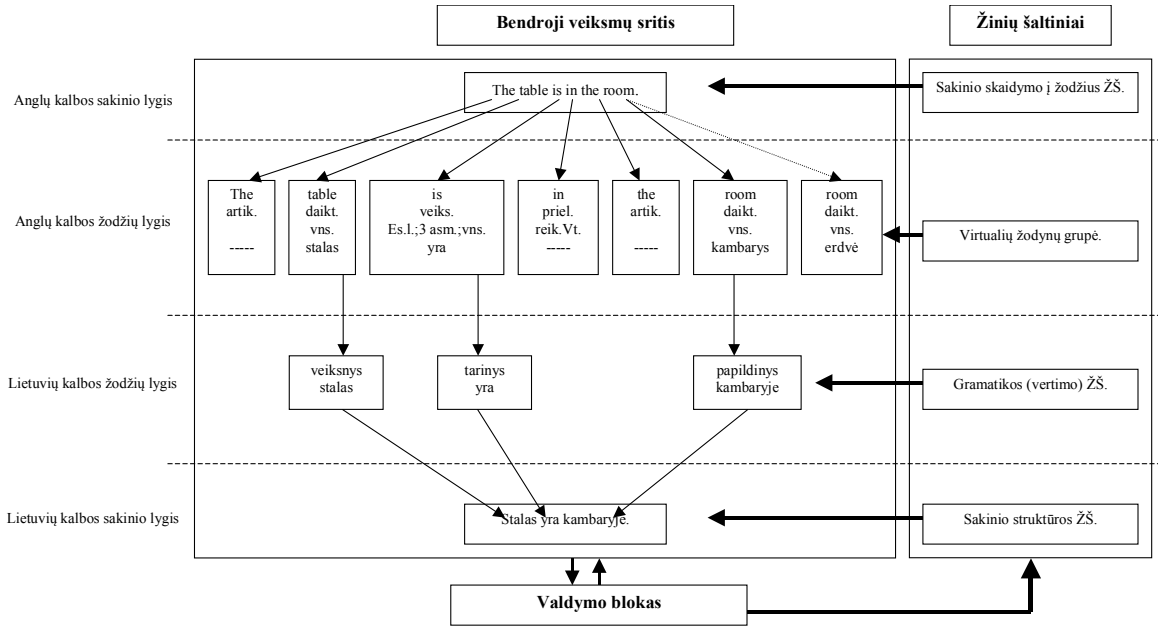
Pirmieji straipsniai intelektikoje apie hierarchinę sprendimų modelio architektūrą pasirodė 1962 metais, kada A.Newell aprašė grupinį problemos sprendimo būdą, kuris 1976 m. buvo pritaikytas kompiuteriniam šnekamosios kalbos atpažinimui (Sidorov, 2001). Ši sistema, turinti kai kuriuos išskirtinius žmogaus intelekto imitavimo bruožus, informatikoje žinoma **Hearsay II** vardu. Apibendrintą sprendimų lentos modelio schemą sudaro pagrindiniai elementai: *sprendimų lenta, žinių šaltiniai ir valdymo posistemė*.

Sprendimų lenta. Ją galima interpretuoti kaip didelės apimties duomenų bazę ar atminties sritį, kurios turinys yra pasiekiamas kiekvienam žinių šaltiniui ir tik jie gali daryti pakeitimus šioje srityje. Bazę sudaro hierarchiniai sluoksniai, kuriuose atsispindi skirtingi uždavinio abstrakcijos lygiai. Šioje bazėje gali būti saugomi įvairūs objektai, su savais, tik jam būdingais specifiniais atributais. Tarpusavyje objektai siejami ryšio atributais. Pavyzdžiui, **Hearsay II** balso atpažinimo sistemoje žemiausiame lygyje yra kvantuoti signalai, o aukščiausiame – atpažintas ir sintezuotas sakinys.

Žinių šaltiniai (ŽŠ). Tai išskaidytos probleminės srities žinios, kurias galima interpretuoti kaip nepriklausomus

programinius ekspertinių sprendimų išvedimo modulius. Kiekvienam žinių šaltiniui būdingos dvi sprendimo priėmimo fazės: išankstinių sąlygų įvertinimas ir elgsenos (veiksmų) fazė. Žinių šaltiniai gali būti realizuojami įvairiai: naudojant loginių taisyklių ar procedūrų rinkinį, semantinius ar neuroninius tinklus ir pan. Tokiu būdu žinių šaltinį galima interpretuoti kaip nukreiptą į problemas ekspertinę sistemą, turinčią atskirą žinių bazę bei sprendimo išvedimo varyklę.

Valdymo blokas. Pagrindinė valdymo funkcija – išrinkti naudingiausią sprendimo procesui žinių šaltinį ir jį aktyvinti. Vienu metu gali būti tenkinamos kelių žinių šaltinių aktyvumo sąlygos, t. y. pretendentų gali būti keli ir tokiu būdu turėsime konfliktinę veiksmo pasirinkimo situaciją. Kiekvienas žinių šaltinis teikia savo pasiūlymus, kuriuose nurodo, kokius pakeitimus sprendimo lentoje ruošiasi atlikti. Pagal šiuos pasiūlymus valdymo posistemė išsirenka, kurį žinių šaltinį aktyvinti. Paprasčiausiu atveju valdymo modulis aktyvina tą žinių šaltinį, kuris pirmasis pranešė, kad gali priimti sprendimą esant duotai situacijai sprendimo lentoje. Sistemos darbo efektyvumą galima padidinti vienu metu nagrinėjant kelias hipotezes ar priimant sprendimus keliuose hierarchijos lygiuose.



2 pav. Kompiuterizuotos vertimo sistemos konceptualusis modelis

Taikant aprašytą sprendimų lentos modelį ir atsižvelgiant į kompiuterinio vertimo proceso aiškiai išreikštą hierarchinę struktūrą, galima skirti atitinkamas žinių šaltinių grupes:

- virtualiųjų žodynų grupė: bendros paskirties virtualus žodynas, frazeologinis virtualus žodynas, teminiai (specialieji) virtualūs žodynai, specialieji daugiakalbiai tekstynai;
- sintaksinė sakinio analizė;
- gramatinės analizės ŽŠ grupė;
- pagalbiniai ŽŠ, kurie „išmano“ ir yra atsakingi už vertimo konteksto įvertinimą, originalo ir jo vertimo semantinį adekvatumą bei kt.;
- probleminių žinių gavyba iš specialiųjų tekstynų ir jų panaudojimas tekstų semantinei erdvei formuoti.

Kompiuterinės vertimo sistemos vartotojai (vertėjai) bei korektoriai (ekspertai) sprendimų lentos modelyje gali būti interpretuojami kaip specialieji ŽŠ, jei jie yra aktyvūs dialogo dalyviai ir jų veiksmai turi įtakos kompiuterinio vertimo proceso kokybei. Elementarus konceptualiojo kompiuterinio vertimo sistemos hierarchinis modelio variantas su specifikuotais ŽŠ vaizduojamas 2 paveikslėlyje. Valdymo bloko strategija šiuo atveju yra artimesnė veiksmų planavimui ir grįžtamojo ryšio palaikymui. Pagrindinės valdymo bloko veiksmų vykdymo funkcijos: žinių šaltinių veiksmų koordinavimas, tekstų adekvatumo kokybės kriterijų įvertinimas bei vertimo variantų skaičiaus reguliavimas.

Alternatyvūs kompiuterinio vertimo principai

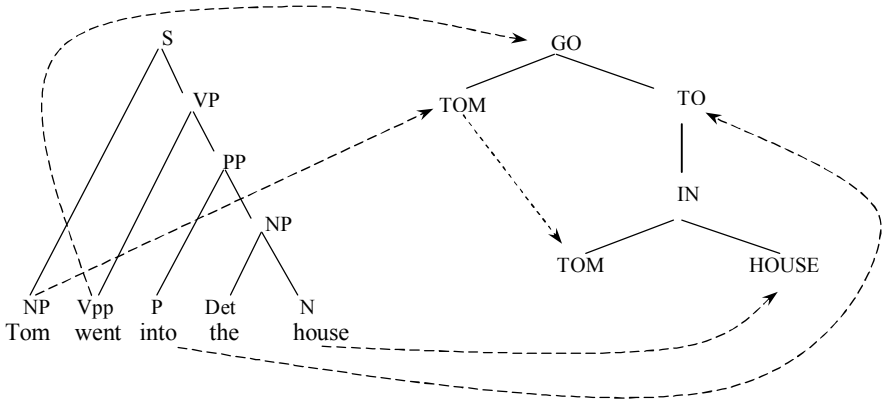
Šiuo metu kompiuterinio vertimo (KV) sistemos naudoja tris vertimo principus (Misevičius, 2001): *tiesioginį (pažodinį), transformacinį (perstatomąjį) ir tarpinės kalbos „interlingua“. Tiesioginis* vertimas – tai pažodinis vertimas, taikomas tarp dviejų konkrečių kalbų, neanalizuojant sakinio, kurį reikia išversti. Transformavimo ar *perstatymo* metodas veikia atvirkščiai: atlikus sakinio analizę, generuojamas vertimo tekstas. Tai reiškia, kad sistema pirma analizuoja originalo kalbos sakinį, sukuria atitinkamą semantinį jo įvaizdį ir po to jį transformuoja į tikslinės kalbos sakinį. *Interlingua* metodas yra šiek tiek panašus į perstatymo procedūrą. Iš tikrųjų, tai speciali tarpinė kalba, kuria galima adekvačiai interpretuoti verčiamų kalbų sakinio struktūras, semantinius ryšius ir ją naudoti formuojant tikslinės kalbos tekstą.

Išvardytųjų metodų privalumus ir trūkumus sunku vertinti vienareikšmiškai. Pažodinis vertimas nėra tobulas, tačiau galima įvardyti keletą pažodinio KV sistemų, kurios

sėkmingai platinamos. Iš tikrųjų, kiekviena KV sistema realizuoja atitinkamas metodų kombinacijas. Tarkime, verčiant *perstatymo* metodu dažnai tenka pasinaudoti ir *tiesioginiu* vertimu, nes sakinio analizė ne visada gali būti patenkinama.

Universalumu ir lankstumu išsiskiriantis *Interlingua principas* vertimo kokybe nepasižymi, tačiau jis atveria formalias galimybes patogiai išreikšti teksto prasmę. Kaip pavyzdį galima paminėti leksinę konceptualiąją struktūrą (LKS), kuri buvo panaudota KV UNITRAN realizacijoje (anglų, vokiečių, ispanų kalboms). Speciali programa analizuoja sakinį, duotą kuria nors iš minėtų kalbų ir pavaizduoja jį LKS. Iš LKS formos galima vienareikšmiškai generuoti sakinį į bet kokią kitą kalbą. 3 pav. matome schemą, kuri iliustruoja trumpo anglų kalbos sakinio transformavimo procesą į *interlingua,* o po to iš jos – į lietuvių kalbą.

Angliškai: Tom went into the house.
LKS: GO(TOM, TO(IN(HOUSE)))
Lietuviškai: Tomas įėjo į namą.



3 pav. Sakinio vaizdavimas LKS

LKS yra abstrakti semantinės informacijos vaizdavimo forma. Atlikus struktūrinę pradinio teksto analizę ir naudojant specialias taisykles, nustatoma priklausomybė tarp teksto ir jį atitinkančio LKS tipo. Suprantama, kad LKS turi būti nepriklausoma nuo kalbos ir išpildyti principinius reikalavimus *interlinguai* pavaizduoti. Paveikslėlyje santrumpos žymi: **S** – sakinys (sentence); **N** – daiktavardis (noun) – veiksnys; **V –** veiksmažodis (verb) – tarinys; **P** – prielinksnis (preposition); **Det** – pažymimasis žodis (determinator); **VP** – veiksmažodinė frazė (verb phrase) – tarinys ir frazės kartu su juo, pvz., *gerai* **miegu***, sparčiai* **einu**; **PP** – prielinksninė frazė (prepositional phrase), pvz., **ant** *stalo***, po** *3 minučių;* **NP** – daiktavardinė frazė (noun phrase), veiksmai, objektai ir įvardžiai, pvz., *kiekvieną* **dieną***,* **jie***,* **namas**.

LKS tipai nustato probleminės srities esybių rūšis, kurios sudaro tam tikrą sistemą. 2 lentelėje pateikiamas galutinis visų tipų sąrašas, kuris naudojamas aprašant LKS forma. Baziniai elementai yra pritaikomi pridedant lauko indikatorių, nurodantį sritį, kurioje elementas turi būti interpretuojamas. Tai yra svarbu norint nustatyti, kokio tipo yra nagrinėjamas elementas ir kartu labai naudinga informacija generuojant iš LKS į tikslo kalbą. 3 lentelėje pateikiamas išsamus bazinių elementų sąrašas. Klausiamųjų sakinių struktūra lietuvių ir anglų kalbose yra kitokia nei paprastuose sakiniuose, todėl juos atitinkanti LKS skiriasi. LKS yra rekursyvi struktūra, sudaryta iš neskaidomų dalių bei komponentų, kurie vėl gali būti LKS, t. y. vėl skaidomi kaip ir visas sakinys. Bendru atveju ši struktūra atrodytų taip:

$$[Event\ GO_{Location}(Thing,\ Path\ TO_{Location}(Position\ IN_{Location}(Thing,\ Thing)))].$$

Kadangi kiekviena kalba interpretuoja tą pačią LKS skirtingais būdais, įvedus atitinkamus parametrus, tą pačią LKS galima realizuoti skirtingomis semantinėmis struktūromis.

2 lentelė. LKS aprašyme naudojami tipai

Tipas	*Atitikmuo*	*Paaiškinimas*	*Pavyzdys*	*Kalbos ar sakinio dalis*
Event	Įvykis	Atitinka veiksmažodžius, atsakančius į klausimą „Kas atsitiko?“	Eiti, valgyti, sportuoti	Veiksmažodis/tarinys
State	Būsena	Veiksmažodžiai, neatsakantys į klausimą „Kas atsitiko?“	Būti, norėti	Veiksmažodis/tarinys
Position	Pozicija	Nusako situaciją su esybe abstrakčioje arba fizinėje vietoje	Ant, už, po	Prielinksnis/ vietos apl.
Path	Kelias	Nurodo konceptualią arba fizinę kryptį	Į, iš, link	Prielinksnis/ vietos apl.
Thing	Daiktas	Daiktavardžiai, kurie žymi objektus	Stalas, obuolys	Daiktavardis/ veiksnys
Property	Savybė	Kokybė, priskiriama *Daiktui*	Naujas, gražus	Būdvardis/ Pažyminys
Location	Vieta	Fizinė vieta	Namas, gatvė.	Daiktavardis/ aplinkybė
Time	Laikas	Nusako laiko trukmę, tarpsnį	Po 3 val., vakar	Laiko aplinkybė
Manner	Būdas	Kokybė, priskiriama *Įvykiui* arba *Būsenai*	Meiliai, gražiai	Prieveiksmis/ būdo apl.
Intensifier	Intensy-vumas	Nusako didesnį ar mažesnį *Savybės*, *Būdo* arba kito *Intensyvumo* laipsnį	Labai, truputį, vos	Būdvardis/ Pažyminys
Purpose	Tikslas	*Būdo* ar *Įvykio* priežastis	Sotumas, turtingumas	Tikslo aplinkybė

Tokiu būdu parametrai papildo LKS duomenų struktūras tai kalbai būdinga informacija, kuri yra pasiekiama analizės ir generavimo moduliams. Dauguma parametrų iš esmės pakeičia LKS interpretavimą, t. y. nustato, kuri LKS dalis, keičiant standartines sujungimo taisykles, turi būti interpretuojama ir kaip interpretuojama. Šie parametrai ir sudaro pagrindinį konkrečiai kalbai būdingą žinių šaltinį. Pagal nutylėjimą standartinės LKS taisyklės yra tokios kaip ir anglų kalboje, tačiau LKS galima pritaikyti konkrečiai kalbai, kuri turi savas sintaksės bei gramatikos taisyklės. LKS pagrindiniai parametrai:

- **Žvaigždutė (*)** – nurodo, ar LKS dalis turi būti realizuojama sintaksiškai, ar ne. Jei * yra, reikia, o jei nėra – ne.
  Anglų k. LKS: enter=[*State* GO(*Thing* *X,*Path* TO(*Position* IN(*Thing* X, *Thing* *Z)))]; I entered the house.
  Lietuvių k. LKS: įeiti=[*State* GO(*Thing* *X,*Path* TO(*Position* *IN(*Thing* X, *Thing* *Z)))]; Aš įėjau į namą.
- **:INT ir :EXT** – pakeičia standartinius loginių subjektų ir argumentų interpretavimo metodus.
  :INT priverčia taip pažymėtą objektą būti veiksmažodinės frazės (VP) viduje, o :EXT – išorėje.
- **:CAT –** pakeičia sintaksinę objekto kategoriją (vietoj reikšmės pagal nutylėjimą). Pavyzdžiui, *Event* ar *State* tipai pagal nutylėjimą LKS atvaizduoja į veiksmažodžius, o *Thing* į daiktavardžius. Tačiau konkrečiose kalbose kalbos/sakinio dalį dažnai reikia keisti. Todėl numatyta galimybė įgyti kitą formą pagal konkrečios kalbos taisykles.
  Anglų k. LKS: BE=[*Event* $BE_{Possesional}$(*Thing* *X, $TO_{Property}$(*Thing* X, *Property* *Y))]; – I was told.
  Lietuvių k. LKS: BE=[*Event* $BE_{Possesional}$(*Thing* *X:EXT, $TO_{Property}$(*Thing* X, *Property* *Y:CAT(VERB):INT))];
  Man buvo pasakyta.
- **:DOMOTE ir :PROMOTE –** sakinio temos keitimas, norint pabrėžti arba sumažinti tam tikro LKS leksikinio vieneto tipo įtaką sakinyje. Pvz., Vaikai rinko uogas *miške./* Miške vaikai rinko *uogas.*
- :**CONFLATED –** reikšminiai komponentai yra įtraukiami į LKS, tačiau jų nereikia interpretuoti sintaksiškai.

Šie parametrai leidžia išvengti vertimo nukrypimų. Jie saugo nuo tematinių (gramatinių objektų ir subjektų vietos keitimas skirtingose kalbose), struktūrinių (įvairių prielinksnių bei prieveiksmių įterpimas arba išmetimas) ir kategorinių (kalbos dalies pasikeitimas kitoje kalboje) bei kitokių sintaksinių nukrypimų.

3 lentelė. Baziniai LKS elementai

Pavadinimas	*Apibūdinimas*	*Pavyzdžiai*
Locational	Nurodantis vietą	Petras guli ligoninėje. Jie nuvažiavo į parduotuvę.
Possesional	Teigiantieji	Aš turiu butą. Jis gavo 100 litų.
Identificational	Identifikuojantys	Ji tapo motina. Antanas išliko romus.
Temporal	Laiko	Susitikimas įvyko 13 valandą. Paskaita truko nuo 18 val. iki vėlyvo vakaro.
Circumstantial	Aplinkybės	Dainius ir toliau groja gitara. Remigijus neberašo eilėraščių.
Existential	Egzistavimo	Jonas sukūrė eilėraštį.
Perceptional	Suvokimo/supratimo	Jie pamatė vaivorykštę. Jis pastebėjo dūmus.
Intentional	Ketinimo	Aš gėriau, nes buvau ištroškęs. Įrašiau muzikinį diskelį.
Instrumental	Instrumentinis	Tomas važiavo su žmona. Martynas apšvietė kambarį su žibintu.

Generavimui iš LKS į lietuvių kalbą naudojami apjungimo metodai, kurie susieja LKS komponentus su atitinkamais sintaksiniais dublikatais. Kadangi LKS yra rekursyvi struktūra, todėl gali būti naudojamas nedaug pakeistas tiesinių struktūrų generavimo algoritmas SHDG (*Semantic Head-Driven Generation*). Būtinas papildymas susijęs su skirtumų sąrašu, t. y. parametrizavimu, naudojamu atvaizduoti išplėstą LKS. SHDG įgyvendina tris pagrindines generatoriaus užduotis – leksikalizaciją (vertimo žodyno naudojimas), struktūrizuotų išsireiškimų generavimą ir lingvistinę realizaciją. Ypač patogus gali būti generatorius, kuris efektyviai naudoja gramatiką iš kito galo – pradedant reikšmės pavaizdavimu ir baigiant gramatinių žodžių konstravimu, atitinkančių verčiamo vieneto kontekstinę prasmę. Kadangi *interlingua* atveju reikšmė atvaizduojama per LKS formą, todėl generatoriui belieka paimti verčiamo teksto LKS formos atitikmenį kalboje, į kurią verčiame, ir, naudojant vertimui orientuotą žodyną, sudaryti kitos kalbos sakinį.

Verčiant iš anglų į lietuvių kalbą ir atvirkščiai, reikia suformuoti po vieną LKS kiekvienai kalbai, bei parašyti po du algoritmus kiekvienai kalbai: sakinio vertimui iš šaltinio kalbos į LKS ir generavimui iš LKS į tikslo kalbą. Naudojant pažodinį ar transformacinį vertimą, reikalingi tik du algoritmai. Tačiau augant papildomų kalbų skaičiui, *interlingua* vertimo principas reikalauja mažesnio algoritmų skaičiaus (Misevičius, 2002).

Lietuvių kalboje nėra griežtos sakinio tvarkos (Dobrovolskis, 2000; Ambrazas, 1997; Piesarskas, 1999), nes tą pačią mintį galima pasakyti keliais būdais. LKS būdinga griežta sakinio struktūra vertimui į lietuvių kalbą netrukdo. Nepaisant minėtų *interlingua* privalumų, jos programinis realizavimas yra sudėtingas ir pakankamai painus dėl kelių priežasčių:

- sudėtinga LKS specifikacija, reikalaujanti lingvistinės analizės;
- generavimo problemos iš LKS į tikslo kalbą;
- netaisyklingų sakinių vertimas (geriau tinka pažodinis vertimas);
- painus frazeologizmų vertimas į LKS ir iš jos į tikslo kalbą.

Interlingua kalbos panaudojimas būtų paprastesnis tuo atveju, jei lietuvių kalbą būtų galima įjungti į jau esamas vertimo sistemas, naudojančias LKS. Tokiu būdu ji taptų lygiavertė su visomis kitomis vertimo sistemos kalbomis, nes būtų galima versti ne tik iš visų tos sistemos palaikomų kalbų, bet ir atvirkščiai. Deja, KV sistemos, paremtos *interlingua* vertimo principu, nepasižymi pakankama vertimo kokybe. Kitas *interlingua* panaudojimo atvejis gali būti intelektualaus teksto skirstymo algoritmo kūrimas, kuris jau pradinėje anglų kalbos teksto analizės fazėje leistų gauti gilesnes kalbos žinias ir leistų sumažinti semantinį daugiaprasmiškumą. Apibendrinus pasakytas mintis, peršasi išvada, kad norint greičiau turėti bent patenkinamos kokybės vertimo sistemą, geriausias būdas naudoti kombinuotą tiesioginį-transformacijos metodą, t. y. sudaryti kiekvieno sakinio gramatinį medį ir jį konvertuoti į atitinkamą lietuvių kalbos medį.

Sintaksinės sakinio grupės

Kompiuterizuoto vertimo sistemų gramatinės analizės komponentai, manipuliuodami gramatinėmis taisyklėmis (Trujillo, 1999; Arnold, 1994; Blekhman, 1998), turi sukurti tikslinėje kalboje sintaksiškai taisyklingą ir pakankamai prasmingą sakinį. Gausiausias taisyklių rinkinys taikomas sakinio dalių išskyrimui ir analizei (veiksniui, tariniui, papildiniui, aplinkybėms ir kt.). Sakinio dalis gali sudaryti vienas ar daugiau žodžių, kurie įvardijami kaip žodžių junginiai. *Sintaksinės sakinio grupės* yra tokie žodžių junginiai, kurie sukuria savarankišką sintaksiškai taisyklingą prasmę sakinyje. Jos nėra vienareikšmis atitikmuo paprastoms sakinio dalims, nors sudaromos pagal tas pačias taisykles. Kartu jos atspindi ir išreiškia tam tikrus prasminius ryšius sakinyje ir palengvina teksto analizę bei daro paprastesnį ir aiškesnį programinių modulių kūrimą. Verčiant tekstą iš anglų į lietuvių kalbą, sintaksinės sakinio grupės sudaromos pagal anglų, t. y. šaltinio kalbos sakinį. Prieš sudarant sintaksines grupes, sintaksinės analizės komponentas iš kitų kompiuterinio vertimo sistemos žinių šaltinių privalo gauti suskaidytą į sakinius šaltinio kalbos tekstą, kurie, savo ruožtu, turi būti išskaidyti į "žodžio" tipo elementus su atitinkamais atributais.

Šią informaciją turi pateikti žinių šaltinis, kuris atlieka paiešką žodyne. Ketvirtojoje lentelėje pavaizduota vertimo proceso teksto transformavimo informacinė struktūra. Gramatinės analizės modulis, suskaidytą sakinį į "žodžio" tipo vienetus, pagal gautą iš žodyno informaciją ir nustatęs žodžio kalbos dalis, gali kurti sintaksines grupes, jas analizuoti, taikydamas atitinkamas gramatines taisykles. Kadangi įvairiose kalbose kalbos dalių skaičius skiriasi, sintaksinių grupių tipų skaičius taip pat gali skirtis. Į sintaksines grupes žodžiai įtraukiami ir jungiami pagal prasmę ir vaidmenį sakinyje, todėl ne visoms kalbos dalims būtina sudaryti tokias grupes. Pavyzdžiui, lietuvių kalboje netgi prieveiksmis turi ryšių su kitais sakinio žodžiais:

$$\underset{\text{prieveiksmis}}{daug} \xrightarrow{ko?} \underset{\text{daiktav.}}{laiškų}$$

Todėl prieveiksmiui lietuvių kalboje tenka sudaryti atskirą sintaksinę grupę. Pagal tai, kuri kalbos dalis yra pagrindinis sintaksinės grupės dėmuo, skiriamos *daiktavardžio, veiksmažodžio, prielinksnio, prieveiksmio* sintaksinės grupės (Dobrovolskis, 2000; Ambrazas, 1997).

Trumpai charakterizuosime pagrindines sintaksines sakinio grupes (anglų–lietuvių kalbų vertimo atveju).

- Daiktavardžio grupė (Noun phrase) – daiktavardis ir įeinančios į grupę kalbos dalys: būdvardis, artikelis ir įvardis (determinantai), skaitvardis. *Pastaba:* jei sakinyje iš eilės eina dvi daiktavardžio grupės atskirtos jungtuku ar prieveiksmiu, jos gali būti sujungtos į vieną, įtraukiant ir jungiamąjį žodį.
- Veiksmažodžio grupė (Verb phrase) – veiksmažodis ir įeinančios į grupę kalbos dalys: papildantis veiksmažodis anglų kalboje (pvz., "could", "may"). *Pastaba:* jei sakinyje iš eilės eina dvi veiksmažodžio grupės, atskirtos jungtuku, jos gali būti sujungtos į vieną, įtraukiant ir jungiamąjį žodį.
- Prielinksnio grupė (Preposition phrase) – prielinksnis. *Pastaba:* prielinksnio grupė seka prieš daiktavardžio grupę.
- Prieveiksmio grupė (Adverb phrase) – prieveiksmis. Jų grupės gali būti sujungtos į vieną, jei jos atskirtos jungtuku.
- X grupė skirta jungti kitų kalbos dalių žodžius ar frazes, kurie nepatenka į kitas grupes. Tai gali būti pavadinimai, ženklai ar nekaitomos frazės sakinyje, taip pat nepažymėti kita kalbos dalimi žodžiai ankstesniame lygmenyje.

4 lentelė. Žodžio atributai

Atributas	*Paaiškinimas*
Žodis šaltinio kalboje (anglų)	Žodis suskaidytame šaltinio sakinyje
Vertimas tikslinėje kalboje (lietuvių)	Šaltinio kalbos žodžio vertimas gautas iš žodyno
Žodžio kamienas tikslinėje kalboje	Žodžio kamienas tikslinėje kalboje, be galūnės
Kalbos dalis	Sintaksinių grupių sudarymui reikia žinoti: daiktavardį, įvardį ir artikelį (determinantus), veiksmažodį, skaitvardį, prieveiksmį, būdvardį, prielinksnį, jungtuką, skyriklį. Kitos kalbos dalys gali būti priskirtos neįvardytoms kalbos dalims, jei jos ypatingos reikšmės sakinio analizei neturi.
Linksnis	Daiktavardis ir būdvardis lietuvių kalboje kaitomas linksniais.
Skaičius	Skaičius – būdvardžiui ar daiktavardžiui (keičiant žodžio formą)
Giminė	Giminė – būdvardžiui ar daiktavardžiui (keičiant žodžio formą)
Laikas	Lietuvių ir anglų kalboje veiksmažodis laikui žymėti turi skirtingas sudarymo taisykles
Asmuo	Veiksmažodis asmenuojamas.
Nuoroda į žodyną	Nuoroda į žodyną, spartesnei žodžio paieškai.

Visi sakinio žodžiai turi būti priskirti kokiai nors grupei. Sintaksinių grupių sudarymo pavyzdys, verčiant iš anglų į lietuvių kalbą, pateiktas penktoje lentelėje. Daiktavardžio grupės gali būti dviejų tipų *objektas* arba *subjektas*. Grupė *subjektas* atlieka veiksnio vaidmenį sakinyje, grupė *objektas* – papildinio, pažyminio ar aplinkybių vaidmenį.

Sudarytoms sintaksinėms sakinio grupėms yra paprasčiau taikyti gramatines, morfologines ir sakinio skyrybos taisykles. Pvz., pagal sakinio darybos taisyklę sintaksiškai teisingą sakinį būtinai sudaro daiktavardžio ir veiksmažodžio grupės, kitos grupės yra papildomos. Šiai taisyklei yra nesvarbu kiek sakinyje yra žodžių ir kokia jų tvarka, tačiau ji sako, jog sakinyje būtinai turi būti bent vienas daiktavardis, įeinantis į daiktavardžio grupę, ir bent vienas veiksmažodis, t. y.

5 lentelė. Sintaksinių grupių sudarymo ir taisyklių taikymo joms pavyzdys

Šaltinio sakinys:	Little boys went to a big school.							
Informacija iš žodyno	Little	Boys	Went	to	a	big	school	.
	būdvardis	daiktavardis	veiksmažodis	prielinksnis	det	būdvardis	daiktavardis	Skyr
	Mažas	Berniukai	Ėjo	į		didelis	mokykla	
Grupių sudarymas	daiktavardžio grupė		veiksma–žodžio grupė	prielinksnio grupė	daiktavardžio grupė			x gr.
Taisyklių grupėms taikymas	Little	Boys	Went	to	a	big	school	.
	būdvardis	daiktavardis	veiksmažodis	prielinksnis	det	būdvardis	daiktavardis	skyr
	daiktavardžio gr. subjektas		veiksm. gr.	priel. gr.	daiktavardžio. gr. objektas			
	Maži	Berniukai	Ėjo	į		didelę	mokyklą	.
Tikslo sakinys	Maži berniukai ėjo į didelę mokyklą.							

S–> D.Gr.+V.Gr. arba S–>D.Gr+V.Gr+[Priel. Gr.].

Šioms grupėms patogu taikyti sintaksinio derinimo, valdymo, šliejimo ryšių taisykles:

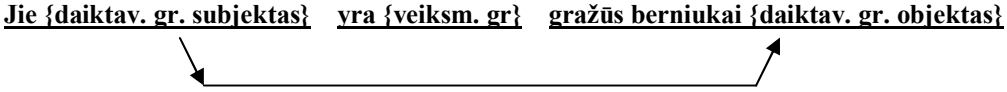
- derinimas – grupės priklausomojo žodžio giminės, skaičiaus ir linksnio derinimas pagal pagrindinį žodį. Pvz., daiktavardžio grupėje linksniai, giminė derinami pagal pagrindinį žodį – daiktavardį: {maži berniukai} d.gr., {didelę mokyklą};
- valdymas – pagrindinis žodis reikalauja tam tikro priklausomųjų žodžių linksnio. Pavyzdžiui:

$$\underset{veiksm.gr.}{rašo} \xrightarrow{ką} \underset{daiktav.gr.}{laišką}\ ,\quad \underset{priel.gr.}{į} \xrightarrow{ką} \underset{daiktav.grupė}{didelę\_mokyklą}\ ,$$

- šliejimas, kai su pagrindiniu žodžiu pagal prasmę susiejamas nekaitomas žodis. Pavyzdžiui:

$$\underset{prieveiksm.gr.}{daug} \xrightarrow{ko?} \underset{daiktav.gr.}{mandagių\_laiškų}$$

Sintaksinėms grupėms taikyti taisykles, kurios susiję su daiktavardžių grupėmis, subjektais arba objektais yra paprasčiau, kai subjektas gali valdyti objekto linksnį, giminę ar skaičių:



Penktoje lentelėje aprašytos sintaksinės sakinio grupės buvo realizuotos prototipinėje vertimo iš anglų kalbos į lietuvių kalbą KV programoje, kuri skirta nesudėtingiems techniniams tekstams (pvz., dokumentų) vertimui. Įdiegtas gramatinės analizės komponentas su sintaksinių sakinio grupių taikymu, žymiai pagerino lietuviškų tekstų vertimo kokybę. Šiuo metu gramatinės analizės komponentas jau naudoja per 20 taisyklių įvairioms sintaksinėms grupėms. Programoje įdiegtas universalus virtualaus žodyno modelis (Pacevičius, 2001), kuris gali būti papildomas atitinkamais atributais. Nepriklausomi išskirtinių žodžių ir morfologinės analizės komponentai reglamentuoja lietuviškų žodžių linksniuočių ir asmenuočių taikymo tvarką. Planuojama įdiegti frazeologizmų analizės komponentą.

Išvados

1. Straipsnyje apžvelgiami kompiuterinio vertimo pagrindiniai principai, analizuojamos alternatyvios anglų kalbos tekstų vertimo į lietuvių kalbą technologijos ir pristatomas konceptualusis hierarchinis kompiuterinio vertimo modelis, pagrįstas sprendimų lentos paradigma.
2. Atlikta lankstaus *Interlingua* vertimo principo analizė teikia vilčių, kad, naudojant panašią metodiką, bus galima sukurti tekstų skirstymo ir kalbos vertimo algoritmus su efektyviau veikiančiais kalbos žinių šaltiniais.

3. Sintaksinių sakinio grupių panaudojimas gramatinės analizės kompiuterinio vertimo sistemos moduliuose leidžia realizuoti sudėtingesnius ir kokybiškesnius vertimo sistemos lygmenis.
4. Sukurta vertimo terpė specializuotiems nesudėtingos struktūros tekstams versti iš anglų į lietuvių kalbą.

Literatūros sąrašas

1. Ambrazas, V. ir kt. (1997) Dabartinės lietuvių kalbos gramatika, Vilnius.
2. Arnold, D.L., Balkan L. and others (1994) Machine Translation: An Introductory Guide.
3. Blekhman, M.S. (1998) Machine Translation: Professional Experience.
4. Dobrovolskis, B., Kniūkšta, P., Kučinskaitė, A. ir kt. (2000) Lietuvių kalbos žinynas, Kaunas: Šviesa.
5. Misevičius, G. (2002) *Interlingua* principai kompiuterinio vertimo sistemose į lietuvių kalbą// KTU konf. „Informacinės technologijos 2002“, p. 311-316.
6. Misevičius, G., Tamulynas, B. (2001). Kalbos vertimo kompiuterizavimo alternatyvios technologijos// Mokslinė–techninė konferencija "Garso korta 2001 05 04": [Kompaktinis diskas], Kaunas: Technologija, ISBN 9955-09-063-4.
7. Mitamura, T., Nyberg, E.H., Carbonell, J.G. (1991) An Efficient Interlingua Translation System for Multi-lingual Document Production // Proc. of Machine Translation Summit III, Washington, DC.
8. Pacevičius, P., Tamulynas, B. (2001) Konceptualus virtualaus žodyno modelis kompiuterizuoto vertimo sistemose// KTU konf. „Informacinės Technologijos 2001“, Kaunas, p. 122-124.
9. Piesarskas, B. (1999) Didysis anglų-lietuvių kalbų žodynas, Vilnius.
10. Sidorov, S., Tamulynas, B. (2001) Konceptualus hierarchinis kompiuterinio tekstų vertimo modelis// KTU konf. „Informacinės Technologijos 2001“, Kaunas, p. 91-95.
11. Tamulynas, B., Žemaitis M. (2002) Sintaksinių sakinio grupių sudarymas ir jų panaudojimas kompiuterinio vertimo programose. Kaunas, KTU konf. „Informacinės Technologijos 2002“, p. 317-320.
12. Trujillo, A. (1999) Translation Engines: Techniques for Machine Translation: Springer.

Gediminas Misevičius, Bronius Tamulynas, Marius Žemaitis

Technologies of Computer-based Translation from English into Lithuanian

Summary

The article gives short overview of language processing technology and focuses on developing the computer-based translation (CBT) from English to Lithuanian. Details of *direct, transfer and interlingua* translation principles and its application for English-Lithuanian translation problem are discussed. Common machine translation models are presented and conceptual hierarchical model for CBT system from English to Lithuanian is proposed. According to the CBT model paradigm a system for specialized English translation into Lithuanian is created. It includes user interface, virtual dictionary, text prasing, translation engine and several modules of knowledge sources. Direct translation strategy with some transfer elements of syntactic sentence groups is used. It allows implementing better translation quality for more complicated sources. The features of the first version of CBT is presented .

Straipsnis įteiktas 2002. 04
Parengtas publikuoti 2002.05

Apie autorius

Gediminas Misevičius, dokt., Kauno technologijos universitetas.
Interesų sritis: intelektualios informacinės sistemos, kompiuterinė lingvistika, kompiuterių tinklai.
Adresas: Kauno technologojos universitetas, Studentų g. 50, LT-3028 Kaunas, Lietuva.
El. paštas: gediminas.misevicius@hansa.lt

Bronius Tamulynas, doc. dr., Kauno technologijos universitetas.
Interesų sritis: intelektualios informacinės sistemos, kompiuterinė lingvistika, kompiuterizuoto mokymosi sistemos.
Adresas: Kauno technologojos universitetas, Studentų g. 50, LT-3028 Kaunas, Lietuva.
El. paštas: bronius@pit.ktu.lt

Marius Žemaitis, magistr., Kauno technologijos universitetas.
Interesų sritis: intelektualios informacinės sistemos, kompiuterinė lingvistika.
Adresas: Kauno technologojos universitetas, Studentų g. 50, LT-3028 Kaunas, Lietuva.
El. paštas: mzemai@fortas.ktu.lt