KOMPIUTERINĖ LINGVISTIKA/ COMPUTATIONAL LINGUISTICS

Kompiuterinis vertimas į lietuvių kalbą: alternatyvos ir jų lingvistinis vertinimas

Inga Petkevičiūtė, Bronius Tamulynas

crossref http://dx.doi.org/10.5755/j01.sal.0.18.407

Anotacija. Šiuo metu Lietuvoje geriausiai žinomos dvi iš anglų kalbos į lietuvių kalbą verčiančios ir laisvai prieinamos kompiuterinio vertimo (KV) sistemos — *Google Vertėjas* ir VDU „Internetinė informacijos vertimo priemonė“ (VDU IIVP). Išsamus bei argumentuotas jų lingvistinis tyrimas yra naudingas tiek sistemų kūrėjams (sistemoms tobulinti), tiek eiliniams naudotojams (įvertinti pasirinkimo galimybes). Paviršutiniškų vertinimų, kritikos bei įvairaus pobūdžio atsiliepimų galima rasti internetiniuose dienoraščiuose, tačiau šie vertinimai paprastai yra subjektyvūs ir nepagrįsti. Straipsnyje pateikiami lingvistiniai kompiuterinio vertimo tyrimo rezultatai: išverstų tekstų analizė, kurią sudaro dažniausiai pasitaikančių vertimo klaidų pavyzdžiai, komentarai, abiejų sistemų lingvistinių klaidų aptarimas ir palyginimas vertimo kokybės atžvilgiu. Išvadose, atsižvelgiant į tekstų pobūdį, tipines klaidas bei KV sistemos architektūros tipą, duodamos rekomendacijos, kaip, vienu ar kitu atveju, būtų galima pagerinti jų veikimą.

Reikšminiai žodžiai: *Kompiuterinis vertimas, lingvistinės klaidos, morfologija, leksika, sisteminės klaidos, daugiareikšmiškumas.*

Įvadas

Dėl kalbų savitumo ir jų nuoseklios kaitos kompiuterinis vertimas apibūdinamas kaip pakankamai sudėtingas procesas, todėl mažai klystančios kompiuterinio vertimo sistemos kūrimas reikalauja kruopštaus ir imlaus darbo. Pagrindinė kliūtis — vertimo procesą lydinčios neišvengiamos sisteminės, sintaksinės bei semantinio pobūdžio klaidos. Šio tyrimo pagrindinis objektas yra kompiuterinio vertimo procese pasitaikančių tipinių lingvistinių ir kitokių klaidų kilmė ir jų sprendimo problemų analizė.

Kompiuterinio vertimo sritis Lietuvoje nėra plačiai tyrinėta, todėl darbų šia tema nėra daug. Bene geriausiai žinomi ir laisvai prieinami visuomenei yra 2007–2008 metų E. Rimkutės ir J. Kovalevskaitės straipsniai, kuriuose aptariamos kai kurios vertimo procese atsirandančios problemos. Šio darbo tema yra aktuali, nes Lietuvoje per kelis pastaruosius metus atsirado dvi į lietuvių kalbą verčiančios, laisvos prieigos sistemos, todėl jų išsamesnio ir argumentuoto tyrimo rezultatai būtų naudingi tiek sistemų kūrėjams (galėtų sistemas patobulinti), tiek eiliniams naudotojams (leistų susipažinti su vertimo problemomis, kurias vėliau tektų spręsti).

Tyrimo tikslas — išsiaiškinti kompiuterinio vertimo esmines praktines problemas, su kuriomis dažniausiai susiduria eiliniai kompiuterinio vertimo sistemų naudotojai.

Teorinės dalies pagrindą sudaro Didžiosios Britanijos lingvistų J. Hutchins‘o ir H. Somers‘o darbai. Remtasi ir kitų mokslininkų: E. Forsbom, A. Chitu, M. Blekhmano ir kt. darbais. Kompiuterinio vertimo sistemų tyrimui naudota informacija, gauta iš oficialių vertimo sistemų svetainių.

Tyrimo rezultatas — esminių kompiuterinio vertimo procese atsirandančių lingvistinių ir sisteminių klaidų sąvadas, klaidų kilmės aprašas, glausta jų analizė ir būdai kaip to būtų galima išvengti.

Kompiuterinio vertimo kokybės vertinimo aspektai

M. Riedel’is ir T. Schwarze’as vertimo problemas skiria į septynias grupes: *polisemija, homonimija* (polisemijos atveju žodis turi keletą panašių reikšmių, homonimijos atveju keletas individualias reikšmes turinčių žodžių turi tą pačią fonetinę išraišką), *sintaksinis daugiareikšmiškumas* (sakinio struktūra priklauso ne tik nuo žodžių, bet ir nuo semantikos), *referencijos daugiareikšmiškumas* (įvardžiai turi tam tikrą ryšį su žodžiais, tačiau ne visada aišku su kuriais žodžiais, todėl nuorodos gali išsiplėsti netgi per sakinio ribas), *neaiškios klaidos* (neaiškūs žodžiai, terminai ir posakiai), *sinonimai* (dažnai atsitinka, kad keletas žodžių turi labai panašią reikšmę, todėl pasirinkti tinkamą variantą yra sudėtinga), *metaforos ir simboliai* (priklauso nuo kultūrinių ir istorinių aspektų ir dažnai jie tiesiog negali būti verčiami), *nauji žodyno dariniai* (kalbos nuolat kinta, sukuriami nauji žodžiai, ypač naujų technologijų pavadinimai) (Riedel, Schwarze, 2001).

D. Arnold‘as, L. Balkan ir kt. skiria tris vertimo klaidų grupes: *klaidas, atsirandančias dėl struktūrinių ir leksinių kalbų skirtumų*, *daugiaprasmiškumo* ir *daugiažodžių vienetų:*

idiomų ir kolokacijų klaidas (Arnold ir kt., 1994). Minėtoji klasifikacija apima aspektus, nurodytus M. Riedel'io ir T. Schwarze'o vertimo klaidų skirstyme, tačiau šis skirstymas nėra labai išsamus ir apima ne visas vertimo problemų sritis. J. Hutchins'as ir H. Somersas klaidas skirsto į penkias grupes: *morfologijos*, *leksinio daugiareikšmiškumo* (kategorijos, perkėlimo daugiareikšmiškumo, polisemijos ir homografijos), *struktūrinio daugiaprasmiškumo* (tikrojo ir atsitiktinio struktūrinio daugiaprasmiškumo), *anaforų vertimo* ir *kiekybinių įvardžių daugiaprasmiškumo* klaidas (Hutchins, Somers, 1992). Jie taip pat mini laiko kategorijos, modalumo, tematikos, mandagumo ir formalumo lygmenų bei kitokias problemas. Šių mokslininkų pateikta klasifikacija yra pakankamai išsami ir apima visas vertimo problemų sritis: leksiką, morfologiją, struktūrinius aspektus ir kt.

KV kokybinis vertinimas reikalingas tam, kad būtų galima palyginti skirtingų KV instrumentų bei technologinių priemonių veikimą arba tiesiog pagerinti tam tikros sistemos vertimo galimybes (Hutchins, 1997). Atsižvelgiant į naudotojų poreikius, KV kokybės vertinimo atributai turėtų būti parenkami pagal tai, kas tą sistemą vertins (klientų agentūra, sistemos tvarkytojas, sistemos kūrėjas, patyręs ar eilinis naudotojas ir pan.) ir kaip vertinimo tikslai priklauso nuo sistemos užbaigtumo ar nuo to, kokiam naudotojui ji sukurta. Pavyzdžiui, diagnostinis vertinimo tipas skirtas surasti klaidas, progresyvus tipas — išsiaiškinti pokyčius skirtingose produkto versijose, adekvatumo tipas — išsiaiškinti, ar produktas atitinka iškeltą užduotį ir pan. (Forsbom, 2003).

Kompiuterinio vertimo technologijų tyrimo eiga. Buvo pasirinkta 15 skirtingų žanrų (penkių funkcinių stilių: mokslinio, publicistinio, buitinio, administracinio ir meninio) tekstų, kurių bendras žodžių skaičius yra 4723. Būtent tokia, manoma pakankamai reprezentatyvi, tekstų apimtis pasirinkta todėl, kad nagrinėjant vertimo rezultatus būtų žinomas teksto kontekstas, t.y, kad būtų galima tinkamai nustatyti vertimo klaidas. Kiekvienam funkciniam stiliui priskirta po tris skirtingo žanro tekstus, pavyzdžiui, meninio stiliaus tekstai buvo: romanas, esė ir eilėraštis; mokslinio stiliaus tekstai buvo: mokslinis straipsnis, monografija ir disertacija ir pan. (iš tikrųjų grožinio teksto KV kokybinis vertinimas nėra esminis dėl suprantamos šių tekstų pragmatinės paskirties ir dėl nėra esminis vienos ar kitos sistemos kokybinis rodiklis).

Tyrimo rezultatai pateikti lentelėse, kuriose kiekvienas sakinys įrašytas originalo kalba, o apačioje — išverstas KV sistemomis (*VDU IIVP* ir *Google Vertėju*). Taigi, tekstai nagrinėjami po sakinį, t.y. kiekviename sakinyje ieškoma klaidų/netikslumų/problemų, kurios atsirado vertimo procese. Pastarieji rezultatai patalpinti atskiroje lentelėje — apibendrintoje klaidų suvestinėje, kurioje kiekvienam KV instrumentui nurodytas klaidų skaičius pagal visus tyrimui naudotus tekstų žanrus (žr. 1 lentelę).

Kiekvienas išverstas sakinys šio straipsnio autorių buvo vertinamas pagal subjektyvią KV kokybės skalę, siekiant išsiaiškinti, kuri sistema kokio funkcinio stiliaus tekstus verčia geriau ir kokybiškiau. Visos pastebėtos klaidos ir trūkumai buvo grupuojami pagal gimininguskokybės vertinimo požymius, aprašomas jų pobūdis bei pateikiami vartosenos pavyzdžiai. Pagal šiuos eksperimentus buvo lyginamos minėtos KV sistemos, analizuojamos vertimo klaidų priežastys ir nustatoma vertimo kokybė pagal klaidų rūšį, skaičių bei vertimo imties tekstų semantinį adekvatumą.

Kompiuterinio vertimo klaidų tipai ir rūšys. Klaidingus KV atvejus galima suskirstyti į du tipus: lingvistines ir sistemines klaidas. **Lingvistines** klaidas sudaro du smulkesni tipai (potipiai): morfologijos ir leksikos klaidos. *Morfologijos* klaidos (žodžiai nesuderinami pagal skaičių, giminę, linksnį, veiksmažodžių asmens ir formos problemos, netinkamos kalbos dalies parinkimo ir prielinksnių bei prieveiksmių vertimo klaidos ir pan.; *leksikos* klaidos (neišversti posakiai, žodžiai, daugiareikšmiškumas, sutrumpintų žodžių, įvardžių, santrumpų, žodžių, sujungtų brūkšneliu vertimo klaidos, pažodinis posakių bei frazeologizmų ir pan. vertimas). **Sisteminės** klaidos (žodynų ar programos kodų klaidos) yra tokios, kurioms nėra lingvistinio, o kartais ir argumentuoto loginio paaiškinimo.

Reikia pripažinti, kad klaidų skaičius nėra absoliutus ar neginčijamas, nes tyrimo metu (tyrimas buvo atliktas 2008 metų pirmoje pusėje) buvo siekiama ne absoliutaus kiekybinio tikslumo, bet norima sužinoti apytikslį klaidų *santykį* abiejose sistemose (vertimo kokybės, atsižvelgiant į funkcinį stilių, palyginimui). Tiriamuose tekstuose rastos 23 rūšių klaidos: 7 rūšių morfologijos, 9 rūšių leksikos ir 7 rūšių sisteminės vertimo klaidos.

Lingvistinės klaidos KV tekstuose

Gramatikos klaidos. *Linksnių vartosena*. Tai viena dažniausiai pasitaikančių vertimo klaidų: *Google Vertėjas* padarė 380, o VDU IIVP — 163 linksnių vartosenos klaidas. Jų galima rasti net pačiuose trumpiausiuose sakiniuose. Pastebėta, kad neretai *Google Vertėjas* žodžius išverčia vardininko linksniu. Ypač sudėtingi atvejai būna tada, kai KV sistemai tenka susieti kelis žodžius, tarp kurių įsiterpia papildomų žodžių. Tuomet vienas žodis verčiamas taisyklingai, t. y., pagal sakinio prasmę, o kitas žodis — vardininko linksniu (arba kitu, atsitiktiniu linksniu).

Pagrindinė veiksmažodžio forma. Google Vertėjas padarė 61, o VDU IIVP — 12 pagrindinės veiksmažodžio formos vartosenos klaidų. Dažniausiai vertimo procese naudojama bendratis arba esamojo laiko III asmens veiksmažodžiai. Ypač sudėtinga situacija būna tada, kai KV sistema verčia veiksmažodžius tam tikromis išvestinėmis formomis. Kadangi minėtos KV programos neturi galimybės naudoti kokias nors kitas papildomas teksto suvokimo (semantikos ar pragmatikos) žinias, todėl labai sunku nustatyti, kuri veiksmažodžio forma yra tinkamesnė. Ypač jei šalia esantys žodžiai ar gramatinės konstrukcijos aiškiai nenurodo, kokia veiksmažodžio forma turi būti naudojama.

Skaičius. Vertimo procese pasitaiko atvejų, kai netaisyklingai nurodomas žodžio skaičius. Pastarosios klaidos vertimo procese pakankamai dažnos: *Google Vertėjas* padarė 43, o VDU

IIVP — 15 tokių klaidų. Neretai netaisyklingo skaičiaus vartojimą lemia šalia esantys įvardžiai, t. y., jei sistema juos išverčia ir netinkamai, toliau prie jų derina kitus gretimus einančius žodžius. Pasitaikė sakinių, kai įvardis išverčiamas taisyklingai, tačiau po jo einantis žodis nederinamas pagal skaičių. Šis atvejis galėtų būti priskiriamas ir programos kodo klaidų grupei. Retkarčiais ir pačių įvardžių sklaičius yra netinkamai suderintas vertimo tekstuose.

Asmuo. Netinkamai nurodytas asmuo yra viena retesnių klaidų, atsirandančių vertimo procese. *Google Vertėjas* padarė 22, o VDU IIVP — 2 tokias klaidas. Kaip ir netaisyklingai nurodytas skaičius, taip ir asmens nurodymas priklauso nuo greta jų esančių įvardžių. Jei įvardis netinkamai išverčiamas, po jo einantys žodžiai taip pat bus nesuderinti nei pagal skaičių, nei pagal giminę, nei pagal asmenį. Tačiau pasitaikė atvejų, kai sistema įvardžius išverčia taisyklingai, tačiau po jų einančius žodžius vis tiek verčia pagal esamojo laiko III asmenį.

Giminė. Google Vertėjas 48, o VDU IIVP 15 kartų neteisingai nurodė žodžių giminę. Jeigu sakinyje nėra aiškių giminės atributų (pavyzdžiui, įvardžių *she, he)*, tada sistema žodį verčia vyriškąja gimine.

Kalbos dalis. Google Vertėjas 54, o VDU IIVP 44 kartus žodį išvertė ne ta kalbos dalimi, kuri atitiko kontekstą. Tai labai dažna vertimo klaida, turinti didelės įtakos verčiamo sakinio prasmei. Tas pats anglų kalbos žodis gali turėti daiktavardžio, būdvardžio, prieveiksmio ar veiksmažodžio reikšmę, todėl kalbos dalies parinkimas yra artimai susijęs su semantine informacija. Jei prieš verčiamą žodį yra dalelytė *to*, tai pasikliaunama, kad tas žodis bus veiksmažodis. Jei minėtosios dalelytės nėra, tikėtina, kad verčiamas žodis bus daiktavardis arba būdvardis ir pan. Tačiau KV sistemose šių taisyklių paisoma ne visais vertimo atvejais.

Neigiami veiksmažodžiai. Verčiamuose tekstuose Google Vertėjas 4, o VDU IIVP 1 kartą sakinyje, kai prieš veiksmažodį buvo parašyti neigiami prieveiksmiai (dažniausiai *never)*, išverstame tekste (lietuvių) nesugeneravo neigiamo veiksmažodžio. Pavyzdžiui, sakinį "*He never stops*" reikėtų versti „*Jis niekada nesustoja*", o ne „*Jis niekada sustoja*". Iš tikrųjų, buvo pastebėta, jog KV dažnai paisoma šios taisyklės, tačiau problemų atsiranda tada, kai tarp neigiamo prieveiksmio ir po jo einančio veiksmažodžio įsiterpia papildomas žodis. Tuomet žodžiai tarpusavyje nesujungiami ir jie verčiami atskirai.

Leksikos klaidos. *Neišverstas posakis.* Verčiamuose tekstuose *Google Vertėjas* 22, o VDU IIVP 1 kartą neišvertė gana įprastų posakių, pavyzdžiui, *I guess* ir *I don't really*. Šie atvejai pasirodė keisti, todėl buvo nuspręsta tuos pačius posakius įterpti į naujus sakinius ir pakartoti vertimą. Iš pradžių buvo manyta, kad tai atsitiktinumas, tačiau vertimą pakartojus, kituose sakiniuose KV sistemos šių posakių vėl neišvertė (beje, tai gali būti siejama su sisteminiais sprendimais, — jeigu žodyne šie posakiai nebuvo įvesti, tai ji turėjo minėtus žodžius išversti kiekvieną atskirai, tačiau to nepadarė). Naudotojui tai gana didelis KV trūkumas.

Daugiareikšmiškumas. Google Vertėjas padarė 146, o VDU IIVP — 218 šios rūšies klaidų. Pagal klaidų skaičių, daugiareikšmiškumas būtų antroje sąrašo vietoje. Lietuvių kalbos žodžiai, kaip ir daugelio kitų kalbų žodžiai, gali turėti daug reikšmių, todėl KV sistemos nevisada gali parinkti tinkamą reikšmę ir dėl to gali iš esmės pasikeisti sakinio prasmė arba yra pavojus ją prarasti. Dauguma šiuolaikinių KV sistemų (pvz., *Google Vertėjas*) remiasi statistiniais duomenimis, kurie gaunami iš didelių tekstynų, t. y., pagal statistinį dažnį nustatoma, kaip dažnai ir kokiame kontekste tam tikras žodis vartojamas. Tačiau, jei žodis sutinkamas neįprastame kontekste, vertimo rezultatai nebus patikimi. VDU sistemos veikimas pagrįstas kitu metodu — taisyklėmis. Tačiau, nepaisant to, daugiareikšmiškumas yra viena iš nedaugelio klaidų rūšių, kurias VDU IIVP daro dažniau už *Google Vertėją*.

Neišverstas žodis. Verčiamuose tekstuose *Google Vertėjas* 59, o VDU IIVP 39 kartus paliko neišverstus žodžius, kurie tikrai nebuvo netipiški. Tuos pačius žodžius perkėlus į kitus sakinius ir pakartotinai verčiantt, paaiškėjo, kad kai kuriuos žodžius sistemos vis dėl to išverčia, o dalį palieka neišverstais. Manoma, taip atsitinka todėl, kad tam tikrų žodžių formos nėra įtrauktos į sistemos žodyną.

Sutrumpintų žodžių vertimas. Analizuojamuose tekstuose *Google Vertėjas* 8, o VDU IIVP 5 kartus neišvertė sutrumpintų žodžių. Kaip matome, tai nėra dažna vertimo klaida. Šio tipo žodžiai daugiausia naudojami buitinėje kalboje arba poezijoje. Jų daugiau pasitaikė meninio stiliaus tekste — eilėraštyje. Tiesa, gana keista, jog „Google" sistema neišvertė tokio populiaraus trumpinio, kaip *I'll* (*I will*), tuo tarpu trumpinį *I'd* (*I would*) verčia teisingai.

Įvardžiai. Tiriamuose tekstuose *Google Vertėjas* 30, o VDU IIVP 19 kartų įvardžius išvertė netinkamai. Dažniausiai taip verčiami asmeniniai įvardžiai. Pastebėta, jog angliškas įvardis *you* beveik visada verčiamas antruoju daugiskaitos asmeniu, t. y., *jūs*, tačiau pagal prasmę sakiniuose labiau tiktų vienaskaitos antrojo asmens įvardis *tu*. Neretai taip išverstas įvardis turi ir netinkamą linksnį.

Pažodinis posakių/ frazeologizmų/kolokacijų vertimas. Verčiamuose tekstuose *Google Vertėjas* padarė 7, o VDU IIVP 9 tokios rūšies klaidas. Tai vėlgi sietina su tuo, kad sistemos žodyne šie posakiai/frazeologizmai/pastovūs žodžių junginiai nebuvo įvesti kaip pastovūs žodžių junginiai, todėl sistemos juos ir verčia atskirai, t. y., kiekvieną žodį individualiai.

Žodžiai, sujungti brūkšneliu. Verčiamuose tekstuose *Google Vertėjas* 6, o VDU IIVP 5 kartus netinkamai išvertė žodžius, sujungtus brūkšneliu. Atliekant tyrimą, sistemoms buvo gana sunku susidoroti su tokiais žodžiais, todėl juos palikdavo neišverstus. Taip pat pasitaikė vienas įdomus atvejis, kai du žodžiai buvo sujungti brūkšneliu, o sistema išvertė tik pirmąjį žodį — antrąjį paliko neišverstą. Galbūt šį atvejį būtų galima pavadinti atsitiktinumu, nes tikriausiai antrasis žodžių junginio dėmuo tiesiog nebuvo įvestas į žodyną. Reikia pažymėti, kad KV sistemos blogai verčia arba neverčia ne visus brūkšneliu sujungtus žodžius, dauguma tokių žodžių būna išversti taisyklingai.

Santrumpos. KV sistemoms nevisada pavyksta taisyklingai išversti santrumpas, todėl kartais jos paliekamos neišverstos. Tokias klaidas *Google Vertėjas* padarė 2, o VDU IIVP — 3.

Tikriniai vardai/pavadinimai. Verčiamuose tekstuose *Google Vertėjas* 3, o VDU IIVP 7 kartus netaisyklingai išvertė tikrinius daiktavardžius. Reikia pastebėti, jog VDU IIVP pavardes išverčia, t. y., jei pavardė sutampa su žodžiu, esančiu sistemos žodyne, ji išverčiama pagal nustatytą reikšmę. Pavyzdžiui, sakinyje, paimtame iš šnekamosios kalbos teksto, *"Kelly Brook lookt awful" išverčiama pavardė: „Kelly Upokšnis lookt baisus"*.

Sisteminės klaidos

Praleidžiamas veiksmažodis. Šios klaidos nebuvo dažnos — tiriamuose tekstuose buvo rasti tik 27 atvejai, kai sakinyje praleidžiamas veiksmažodis. Tai išskirtina *Google Vertėjo* klaida. Pastebėta, kad vertimo procese neretai praleidžiamas tarinys, jeigu jis yra žodis *is* (ar jo forma *was, were*), tačiau šis atvejis nebūtinai kartojasi visuose sakiniuose. Gali būti taip, kad viename sakinyje tarinys bus praleistas, o kitame jau bus išverstas. Tai yra svarbu, nes dažnai tarinio nebuvimas sakinyje lemia viso sakinio prasmės suvokimą.

Didžiųjų/ mažųjų raidžių rašyba. Verčiamuose tekstuose *Google Vertėjas* 38, o VDU IIVP 4 kartus be reikalo žodžius išvertė didžiosiomis/mažosiomis raidėmis arba tik su pirma didžiąja raide. Galima įtarti, kad tai yra tipiška sistemos klaida. Kaip taisyklė šios klaidos netrukdo suprasti sakinio prasmės. Buvo ir tokių atvejų, kai originaliame tekste netikrinis žodis buvo parašytas pirma didžiąja raide (norint pabrėžti), o išverstame tekste didžiosios raidės jau nebeliko. Šio tipo klaidos labiau būdingos „Google" sistemai.

Praleistas žodis. Būta atvejų, kai originalo tekste tam tikras žodis yra, o vertime jo nebelieka, net jei jis buvo gana svarbus. Tokių klaidų verčiamuose tekstuose *Google Vertėjas* padarė net 51, o VDU IIVP tik 2, todėl galima tvirtinti, jog ši klaida labiau būdinga „Google" sistemai. Dažniausiai praleidžiamos nesavarankiškos kalbos dalys (jungtukai, prielinksniai, dalelytės). Šių žodelių praradimas nėra labai svarbus sakinio prasmei, tačiau toks vertimas negali būti laikomas pakankamai tiksliu.

Žodis išverstas kita kalba. Tai gana neįprastas vertimo sistemos trūkumas, tačiau vertimo procese taip atsitiko net 33 kartus, t. y., sistema net 33 žodžius išvertė kita kalba (manoma, kad lenkų). Šį fenomeną taip pat labai sunku paaiškinti, nes jei sistema nerastų verčiamo žodžio ekvivalento, tai jį tiesiog paliktų neišverstą. Galima spėti, jog tai yra gana grubi programos kodo klaida. Šios rūšies klaida išskirtinai būdinga *Google Vertėjui.*

Neatsižvelgiama į diakritinius ženklus. Verčiamuose tekstuose buvo du atvejai, kai originalo sakinyje apostrofu buvo sutrumpintas žodis. Kaip jau buvo minėta anksčiau, sistema verčia ne visus sutrumpintus žodžius, tačiau šį kartą žodį išvertė. Ši klaida gana reta ir verčiamuose tekstuose buvo būdinga tik VDU sistemai.

Žodis verčiamas nežodynine reikšme. Verčiamuose tekstuose pasitaikė 7 atvejai, kai sistema žodį išvertė tokia reikšme, kurios nėra žodyne. Šios rūšies klaidą būtų galima pavadinti programos kodo arba sistemos žodyno klaida. Tai išskirtinai *Google Vertėjo* klaidų rūšis.

Papildomas žodis. Nagrinėjant tekstus, buvo pastebėta, kad *Google Vertėjas* 8, o VDU IIVP 1 kartą išverstame sakinyje įterpė papildomą žodį, kurio originalo sakinyje nebuvo.

Aptarus rastas vertimo klaidas ir jų tipus, jas galima palyginti su kitų mokslininkų nurodytais galimų KV klaidų atvejais. Pavyzdžiui, lyginant mūsų nustatytas klaidas su M. Riedel'io ir T. Schwarze'o aprašytomis klaidomis, galima pastebėti, kad nemaža dalis klaidų tipų sutampa. Jų minimos polisemijos, homonimijos ir sinonimų klaidos šiame darbe apibendrintos ir pavadintos daugiareikšmiškumo klaidomis. Tą patį patvirtina ir klaidingi įvardžių vertimo atvejai. Tyrimo metu nebuvo aptikta klaidų, kurias minėtieji mokslininkai vadina „neaiškiomis kliūtimis", taip pat nerasta metaforų ir simbolių (jų tiesiog nebuvo tyrimui pasirinktuose tekstuose) klaidingo interpretavimo atvejų, nebuvo tiriama sakinio sintaksinė struktūra. Tyrimo metu buvo nustatyta nemažai daugiareikšmiškumo klaidų, netiksliai išverstų frazeologizmų, taip pat įvairių klaidų, atsirandančių dėl struktūrinių ar kalbų leksikos skirtumų. J. Hutchins'o ir H. Somers'o pasiūlytoje vertimo klaidų klasifikacijoje į tris atskirus punktus skirstomos daugiaprasmiškumo klaidos, dar skiriamos morfologinės ir įvardžių vertimo klaidos. Mes taip pat identifikavome visas jų minėtas klaidas, tačiau pagal pasirinktą kitokį grupavimą, kurį sudarė du stambūs klaidų tipai — lingvistinių (morfologijos ir leksikos) ir sisteminių, jos buvo kitaip interpretuojamos. Verta paminėti, kad nė vienas autorius savo klasifikacijoje neišskyrė sisteminių klaidų, kurios dažnai gali turėti nemažą įtaką teksto prasmei suvokti. Akivaizdu, kad vertimo klaidas klasifikuoti galima įvairiai, tačiau pagrindinėmis klaidomis visada liks daugiareikšmiškumo bei leksikos ir morfologijos klaidos.

Kompiuteriu verstų tekstų suprantamumo ir prasmingumo lyginimas

Kitas dviejų KV technologijų lyginimasis tyrimas buvo atliktas tekstų suprantamumo ir prasmės atžvilgiu. Lyginimas buvo atliekamas nagrinėjant sudarytas lenteles, kuriose pateikiama, koks sakinių procentas yra įvertintas tam tikru subjektyviu kokybės vertinimo balu (1–7). Sistemos lyginimos pagal funkcinius stilius, nes minėtosios lentelės sudarytos atskirai kiekvienam funkciniam stiliui. Dėl jau minėtų semantikos ir pragmatikos dalykų meninio stiliaus tekstų prasmingumo vertinimas yra santykinis. Tačiau tam tikru aspektu šio stiliaus tekstų tyrimas kartais padeda lengviau nustatyti klaidų atsiradimo priežastis ir paryškina KV sistemų gebėjimus.

Meninis stilius. Google Vertėjo verčiamo romano ištrauka buvo įvertinta įvairiai: nuo pačio mažiausio iki aukščiausio kokybės vertinimo balo. Apie 50 % visų įvertinimų sudarė 5 ir 6 balų įvertinimai, o tai reiškia, kad beveik pusė romano sakinių buvo išversti priimtinai ir patenkinamai. 13 % saki-

nių išversti pakankamai gerai, tačiau apie 35 % sakinių vertimo kokybė nebuvo pakankama. VDU IIVP romano ištraukos įvertinimai šiek tiek skiriasi: ši sistema net 52 % sakinių išvertė pakankamai gerai, apie 25 % sakinių išversti priimtinai ir patenkinamai. Nepatenkinamais balais (1–3) įvertinta tik 16 % sakinių. Iš šių skaičių galima daryti išvadą, jog romano ištrauką geriau išvertė VDU sistema, nes sakinių, įvertintų patenkinamais balais (5–7), procentas yra 76 %, tuo tarpu „Google“ sistema patenkinamai išvertė 57 % sakinių. *Google Vertėjo* romano ištraukos kokybės vertinimų vidurkis yra 4,4, o VDU sistemos — 5,4. Šie skaičiai tik dar kartą iliustruoja VDU sistemos pranašumą romano vertimuose.

Google Vertėjo verčiamo eilėraščio įvertinimai buvo prastesni: visi sakiniai įvertinti 1 ir 2 balais (žr. 25 lent.). 80 % sakinių įvertinti 1 ir 20 % sakinių — 2 balais. Tokia vertimo kokybė yra nepriimtina ir bloga. VDU sistemos rezultatai panašūs: 40 % sakinių įvertinti 1 ir 60 % sakinių — 2 balais. Šios sistemos vertimas truputį geresnis nei *Google Vertėjo*, nes jos verstų sakinių, įvertintų 2 balais yra 40 % daugiau. *Google Vertėjo* eilėraščio kokybės vertinimų vidurkis yra 1,2, o VDU sistemos — 1,6. Šie skaičiai parodo nežymų VDU sistemos pranašumą verčiant eilėraštį, tačiau, bet kuriuo atveju, vertimo kokybė nėra priimtina.

Google Vertėjo verčiamos esė ištraukos įvertinimai svyravo tarp 2 ir 4 balų, t. y., tarp blogo ir vidutiniško vertimo. 80 % sakinių įvertinti blogai ir nepakankamai ir tik 20 % sakinių išversti vidutiniškai. VDU IIVP įvertinimai įvairesni, nes svyruoja tarp 2 ir 6 balų: tarp blogo ir patenkinamo vertimo. 50 % sakinių išversti priimtinai ir patenkinamai, 30 % — vidutiniškai ir 20 % — blogai. Akivaizdu, kad esė ištrauką geriau išvertė VDU sistema. Šį faktą dar patvirtina ir tai, kad *Google Vertėjo* esė ištraukos kokybės vertinimų vidurkis yra 3, o VDU sistemos — 4,4.

Abi sistemos geriausiai išvertė romano ištrauką, o blogiausiai — eilėraštį. Visus meninio stiliaus tekstus kokybiškiau išvertė VDU KV sistema.

Mokslinis stilius. Google Vertėjo verčiamo mokslinio straipsnio kokybės balai svyravo tarp 1 ir 4 balų, t. y., tarp nepriimtino iki vidutinio vertimo. 70 % sakinių išversti nepriimtinai arba nepatenkinamai ir tik 30 % sakinių išversti vidutiniškai. VDU IIVP mokslinio straipsnio ištraukos įvertinimai šiek tiek skiriasi: jie svyruoja tarp 2 ir 7 balų, t.y., tarp blogo ir pakankamai gero vertimo. 40 % sakinių išversti priimtinai ir pakankamai gerai, tačiau 30 % sakinių vertimo kokybė nėra pakankama. Iš pateiktų duomenų galima daryti išvadą, jog mokslinio straipsnio ištrauką geriau išvertė VDU sistema, nes sakinių, įvertintų patenkinamais balais (5–7), procentas yra 40 %, tuo tarpu „Google“ sistema patenkinamai neišvertė nė vieno sakinio. *Google Vertėjo* mokslinio straipsnio ištraukos kokybės vertinimų vidurkis yra 2,7, o VDU sistemos — 4,3. Šie skaičiai dar kartą patvirtina VDU sistemos pranašumą prieš *Google Vertėją*.

Google Vertėjo verčiamos disertacijos ištraukos kokybės vertinimai svyruoja tarp 2 ir 6 balų, t. y., tarp blogo ir patenkinamo vertimo. 63 % sakinių išversti blogai/ nepakankamai/ vidutiniškai ir 36 % sakinių išversti priimtinai ir patenkinamai. VDU IIVP kokybės vertinimai svyruoja tarp 3 ir 7 balų, t. y., tarp nepakankamos ir gana geros kokybės. Net 72 % sakinių išversti priimtinai ir pakankamai gerai ir tik 9 % sakinių vertimas buvo nepakankamas. Šie rezultatai patvirtina, jog disertacijos ištrauką geriau išvertė VDU sistema. Tą pačią išvadą patvirtina ir tai, kad *Google Vertėjo* disertacijos ištraukos kokybės vertinimų vidurkis yra 4, o VDU sistemos — 5,4.

Google Vertėjo verčiamos monografijos ištraukos kokybės vertinimai svyruoja tarp 3 ir 6 balų, t. y., nepakankamo ir patenkinamo vertimo. 66 % sakinių išversti priimtinai ir patenkinamai, 17 % — nepakankamai ir 17 % — vidutiniškai. VDU IIVP kokybės vertinimai svyruoja tarp 3 ir 7 balų, t. y., tarp nepakankamo ir pakankamai gero vertimo. Net 83 % sakinių ši sistema išvertė priimtinai ir gana gerai, 16 % sakinių vertimo kokybė yra vidutiniška arba nepatenkinama. Nepaisant to, jog sistemų kokybės vertinimai svyravo panašiuose intervaluose, tačiau kokybiškiau monografijos ištrauką išvertė VDU IIVP. Šią išvadą patvirtina ir tai, kad *Google Vertėjo* monografijos ištraukos kokybės vertinimų vidurkis yra 4,7, o VDU sistemos — 5,5.

Mokslinio stiliaus tekstų imtyje tiriamos sistemos geriausiai išvertė monografijos, o blogiausiai — mokslinio straipsnio ištraukas. Visus mokslinio stiliaus tekstus kokybiškiau išvertė VDU sistema.

Buitinis stilius. Google Vertėjo verčiamo anekdoto kokybės balai apėmė visą skalę, t. y., nuo 1 iki 7 balų. Apie 50 % sakinių išversti priimtinai/patenkinamai/pakankamai gerai, 30 % — nepriimtinai/ blogai/ nepakankamai, 15 % — vidutiniškai. VDU IIVP išversto anekdoto įvertinimai šiek tiek skiriasi: jie svyruoja tarp 2 ir 7 balų, t. y., tarp blogo ir pakankamai gero vertimo. Net 72 % sakinių išversti priimtinai/patenkinamai/pakankamai gerai, tačiau 16 % sakinių vertimo kokybė nėra priimtina/pakankama arba bloga. Iš pateiktų duomenų galima daryti išvadą, jog anekdotą geriau išvertė VDU sistema, nes sakinių, įvertintų patenkinamais balais (5–7), procentas yra 72 %, o „Google“ sistemos — 50 %. Šį faktą patvirtina ir tai, kad *Google Vertėjo* anekdoto kokybės vertinimų vidurkis yra 4,3, o VDU sistemos — 5,4.

Google Vertėjo verčiamos internetinio dienoraščio (blog'o) ištraukos kokybės balai vėl apima visą skalę, tik šį kartą didžioji dalis sakinių įvertinta nepriimtinai/ blogai/ nepakankamai (70 % sakinių). 15 % sakinių įvertinta vidutiniškai ir tik 15 % sakinių įvertinta priimtinai/ patenkinamai/ pakankamai gerai. VDU sistema internetinio dienoraščio ištrauką išvertė ne taip kokybiškai, nes 50 % sakinių įvertinti nepriimtinai/ blogai/ nepakankamai ir tik 30 % sakinių įvertinti priimtinai ir pakankamai gerai. Nepaisant to, kad abiejų sistemų vertimas nebuvo aukštos kokybės, tačiau šiek tiek geriau internetinio dienoraščio ištrauką išvertė VDU sistema. Be to, *Google Vertėjo* internetinio dienoraščio (blog'o) ištraukos kokybės vertinimų vidurkis yra 2,8, o VDU sistemos — 3,5.

Google Vertėjo verčiamoje nesinchroninio internetinio pokalbio (forum'o) ištraukoje didžioji dalis sakinių įvertinti

pirmoje kokybės skalės pusėje esančiais balais: 66 % sakinių įvertinti nepriimtinai/ blogai/ nepatenkinamai, 19 % — vidutiniškai ir tik 15 % — priimtinai/ patenkinamai. Aukščiausiu balu nebuvo įvertintas nei vienas sakinys. VDU IIVP vertinimas buvo pakankamai panašus: 61 % sakinių įvertinti nepriimtinai/ blogai/ nepatenkinamai, 14 % — vidutiniškai ir tik 24 % — priimtinai/ patenkinamai/ gana gerai. Šie duomenys rodo, kad nors abiejų sistemų vertimai ir buvo panašūs, vis dėlto šiek tiek palankesnis buvo VDU sistemos vertimas. Verta išskirti, kad *Google Vertėjo* nesinchroninio internetinio pokalbio (forum'o) ištraukos kokybės vertinimų vidurkis yra 2,6, o VDU sistemos — 3,2.

Tiriamos sistemos geriausiai išvertė anekdotą, o blogiausiai — nesinchroninio internetinio pokalbio (forum'o) ištrauką. Visus buitinio stiliaus tekstus kokybiškiau išvertė VDU sistema.

Administracinis stilius. Google Vertėjo verčiamo protokolo ištraukoje didžioji dalis sakinių įvertinti antroje kokybės skalės pusėje esančiais balais: net 92 % sakinių įvertinti priimtinai/ patenkinamai/ pakankamai gerai, 8 % sakinių įvertinti vidutiniškai. VDU sistema šį tekstą išvertė geriau: visi sakiniai įvertinti priimtinai/ patenkinamai/ pakankamai gerai. Šie duomenys rodo, jog abi sistemos protokolo ištrauką išvertė labai gerai, tačiau VDU IIVP vertimas buvo nežymiai geresnis, tai patvirtina ir faktas, kad *Google Vertėjo* protokolo ištraukos kokybės vertinimų vidurkis yra 6,1, o VDU sistemos — 6,7.

Google Vertėjo ir VDU IIVP verčiamo įstato ištraukos įvertinimai labai panašūs: abiejų sistemų įvertinimai svyruoja tarp 3 ir 7 balų, t. y., tarp nepakankamo ir pakankamai gero vertimo. Abi sistemos po 7 % sakinių išvertė nepakankamai ir vidutiniškai, tačiau „Google" sistema net 79 % sakinių išvertė patenkinamai ir pakankamai gerai, tuo tarpu VDU IIVP taip išvertė 65 % sakinių, todėl galima tvirtinti, jog įstato ištrauką kokybiškiau išvertė *Google Vertėjas*. Tą rodo ir kokybės vertinimų vidurkiai: *Google Vertėjo* įstato ištraukos kokybės vertinimų vidurkis yra 5,9, o VDU sistemos — 5,7.

Google Vertėjo ir VDU IIVP verčiamos konvencijos ištraukos įvertinimai labai panašūs: abiejų sistemų įvertinimai svyruoja tarp 5 ir 7 balų, t. y., tarp priimtino ir pakankamai gero vertimo. Žvelgiant į vertinimo rezultatus, iš pirmo žvilgsnio sunku nustatyti, kuri sistema atliko kokybiškesnį vertimą. Iš tikrųjų, konvencijos ištraukos įvertinimai yra vienodi: *Google Vertėjo* konvencijos ištraukos kokybės vertinimų vidurkis yra 6,6, o VDU sistemos irgi 6,6.

Vieną administracinio stiliaus tekstą (protokolą) geriau išvertė VDU sistema, antrą tekstą (įstatą) — „Google" sistema, trečiąjį tekstą (konvenciją) abi sistemos išvertė vienodai.

Publicistinis stilius. Google Vertėjo verčiamos recenzijos ištraukos kokybės balai apima visą skalę, tačiau didžioji dalis sakinių įvertinta nepriimtinai/ blogai/ nepakankamai (71 % sakinių). Po 14 % sakinių išversta vidutiniškai ir priimtinai/ patenkinamai. VDU IIVP kokybiniai įvertinimai geresni: 57 % sakinių išversta priimtinai/ patenkinamai/ pakankamai gerai, po 21 % sakinių išversta vidutiniškai ir nepriimtinai/ nepakankamai. Akivaizdu, kad recenzijos ištrauką kokybiškiau išvertė VDU sistema. Šį faktą patvirtina ir tai, kad *Google Vertėjo* recenzijos ištraukos kokybės vertinimų vidurkis yra 2,8, o VDU sistemos — 5, taigi VDU sistemos vertimo kokybė beveik dvigubai aukštesnė, nei *Google Vertėjo*.

Google Vertėjo verčiamo straipsnio ištraukoje 51 % sakinių išversta priimtinai/ patenkinamai/ pakankamai gerai, 18 % — vidutiniškai ir 32 % — nepriimtinai/ nepakankamai. VDU vertinimai aukštesni: 73 % sakinių išversta priimtinai/ patenkinamai/ pakankamai gerai, 9 % — vidutiniškai ir 15 % — nepriimtinai/ blogai/ nepakankamai. Suprantama, kad kokybiškesnį vertimą atliko pastaroji sistema. Be kita ko, *Google Vertėjo* straipsnio ištraukos kokybės vertinimų vidurkis yra 4,6, o VDU sistemos — 5,6.

Google Vertėjo verčiamos interviu ištraukos kokybės įvertinimai pasiskirstę antroje skalės pusėje: net 85 % sakinių išversti priimtinai/ patenkinamai/ pakankamai gerai ir tik 15 % sakinių išversta blogai/ nepakankamai. VDU sistemos vertimas vėl pranašesnis: 90 % sakinių išversta priimtinai/ patenkinamai/ pakankamai gerai ir tik 10 % sakinių išversta nepriimtinai/ nepatenkinamai. Suprantama, kad VDU sistemos vidutinis kokybės įvertinimas yra aukštesnis už „Google" sistemos įvertinimą: *Google Vertėjo* interviu ištraukos kokybės vertinimų vidurkis yra 5,5, o VDU sistemos — 6,1.

Tiriamos sistemos kokybiškiausiai išvertė interviu, o prasčiausiai — recenzijos ištrauką. Visus publicistinio stiliaus tekstus kokybiškiau išvertė VDU sistema.

Išanalizavus ir palyginus įvairių funkcinių stilių tekstų kokybinius įvertinimus, galima teigti, jog iš 15 verstų tekstų, 13 tekstų kokybiškiau išvertė VDU sistema; vieną tekstą (įstatą) geriau išvertė „Google" sistema ir vieną tekstą (konvenciją) abi sistemos išvertė vienodai gerai. Pastebėta, jog VDU sistemos įvertinimai labiau koncentruojasi antroje kokybės vertinimo skalės pusėje, o „Google" sistemos įvertinimai išsidėsto visoje skalėje tolygiai. Vidutiniai VDU IIVP kokybiniai įvertinimai svyravo tarp 2 ir 7 balų, t. y., tarp blogo ir pakankamai gero vertimo, o *Google Vertėjo* kokybiniai įvertinimai svyravo tarp 1 ir 7 balų, t. y., tarp nepriimtino ir pakankamai gero vertimo. Abi sistemos geriausiai išvertė administracinio stiliaus tekstų (protokolo, įstato ir konvencijos) ištraukas, prasčiausiai — meninio stiliaus tekstų (romano, esė ir ypač eilėraščio) ištraukas. Įvertinus gautus duomenis, galima daryti išvadą, jog kokybiškesnį visų funkcinių stilių vertimą atlieka VDU *„Internetinė informacijos vertimo priemonė"*.

Išvados

Kompiuterinio vertimo sistemų analitiniam tyrimui buvo pasirinktos dvi iš anglų į lietuvių kalbą verčiančios sistemos: *Google Vertėjas* ir VDU *„Internetinė informacijos priemonė"*. Jos gerai žinomos aktyvioje informacinėje erdvėje, geba versti į lietuvių kalbą, turi patogią ir nemokamą prieigą. Tyrimui buvo naudotos skirtingų funkcinių stilių ir žanrų tekstų ištraukos. Remiantis tipinių lingvistinių bei sisteminių

vertimo klaidų analizės rezultatais, jos buvo palygintos tarpusavyje. Išverstų tekstų prasminis vertinimas atliktas naudojant subjektyvaus vertinimo skalę. Tyrimo eigoje nustatyta, kad:

- Duotos imties tekstuose *Google Vertėjas* padarė 1066, o VDU IIVP — 565 vertimo klaidas;
- Abiem sistemoms būdingos tos pačios išskirtinės klaidos — linksnių vartosenos, daugiareikšmiškumo, netinkamos kalbos dalies parinkimo, neišverstų žodžių, pagrindinės veiksmažodžių formos vertimo klaidos, kurios sudaro beveik 73 % visų aptiktų klaidų;
- Iš bendrojo 23 klaidų sąrašo, 19 klaidų rūšių skaičius buvo nepalankus „Google“ sistemai;
- Iš 15 testuotų įvairaus sudėtingumo tekstų, 13 tekstų kokybiškiau išvertė VDU sistema;
- Vidutiniai VDU kokybiniai įvertinimai svyruoja tarp 2 ir 7 balų, t. y., tarp blogo ir pakankamai prasmingo vertimo. *Google Vertėjui* šie įvertinimai svyruoja nuo 1 iki 7 balų, t. y., tarp nepriimtino ir pakankamai gero vertimo;

Abi sistemos kokybiškiau vertė administracinio stiliaus, prasčiau — meninio stiliaus tekstų ištraukas. Išnagrinėjus vertimo klaidas, sąlyginai galima įžvelgti kai kurias galimas jų tobulinimo kryptis:

- Sukurti arba papildyti frazeologizmų, posakių, pastoviųjų žodžių junginių bei santrumpų žodynus;
- Sudaryti arba papildyti šnekamosios kalbos ir žargono žodynus;
- Sistemų žodynus reguliariai pildyti naujais žodžiais bei jų formomis;
- Sukurti ir įdiegti didesnės apimties lygiagrečiuosius arba palyginamuosius tekstynus;
- Išspręsti tikrinių daiktavardžių vertimo problemą bei ištaisyti specifines sistemines klaidas;
- Vertimo varyklėse papildyti taisyklių sąrašus, leidžiančius lanksčiau nustatyti žodžių gramatines kategorijas bei prasmingesnius tikslo tekstų transformavimo atvejus.

Literatūra

1. Arnold, D., Balkan, L., Meijer, S., Humphreys, R., Sadler, L., 1994. Machine Translation: An Introductory Guide [interaktyvus]. Blackwells-NCC, London. Prieiga per internetą: http://www.essex.ac.uk/linguistics/clmt/MTbook/HTML/ book.html. [žiūr. 2009 03 14].
2. Forsbom, E., 2003. Machine Translation Evaluation. Uppsala University. Prieiga per internetą: http://stp.lingfil.uu.se/~evafo/fmo_eval.pdf [žiūr. 2009 02 10].
3. Hutchins, J., Somers, H., 1992. An Introduction to Machine Translation. London: Academic Press, pp. 147–149. Prieiga per internetą: http://www.hutchinsweb.me.uk/IntroMT–8.pdf. [žiūr. 2009 01 09].
4. Hutchins, J., 1997. Evaluation of Machine Translation and Translation Tools. Iš: Survey of the State of the Art in Human Language Technology, pp. 418–419. Prieiga per internetą: http://www.hutchinsweb.me.uk/HLT–1997.pdf [žiūr. 2009 02 26].
5. Riedel, M., Schwarze, T., 2001. Machine Translation: History, Theory, Problems and Usage [interaktyvus]. Prieiga per internetą: http://archiv.tu–chemnitz.de/pub/2001/0043/data/presentation–html/img0.htm [žiūr. 2009 03 14].
6. Rimkutė, E., Kovalevskaitė, J., 2007. Mašininis vertimas — greitoji pagalba globalėjančiam pasauliui. Iš: „Gimtoji kalba“ [interaktyvus]., nr. 9. Prieiga per internetą: http://www.apiekalba.lt/index.php?option=com_content&task=view& id=41 [žiūrėta 2008 12 01].

Inga Petkevičiūtė, Bronius Tamulynas

Computer-based Translation into Lithuanian: Alternatives and Their Linguistic Evaluation

Summary

In machine translation (MT) it is extremely complicated to create perfectly functioning system. The main problems in the systems are errors occurring during translation process. The topic of this research is relevant because in the last years two freely accessible MT systems, supporting the Lithuanian language, were introduced in Lithuania. Comprehensive and well-grounded analysis of these systems would be useful to the system developers and ordinary users. The object of this research are typical linguistic and systemic problems occurring during translation. Those problems are indicators determining translation quality. The aim of this paper is to explore the main practical translation problems that ordinary MT users commonly deal with. Analysis has shown that the *Google Translator* had made 1066 and VDU system 565 translation errors. Most translation errors are common to both systems: declensional, polysemy, non-translated words, not suitable parts of speech constituted about 70 % of all errors. Out of 15 tested texts, VDU system has translated 13 texts in good quality. Out of 23 types of errors 19 types errors were “produced” by *Google* system. Both systems demonstrated the best translation results in translating administrative text and the worst results in translating fictional texts. Following the conducted analysis such recommendations could be made: to create or supplement dictionaries of phraseological units, expressions, constant word combinations, abbreviations, jargon and spoken language; constantly update dictionaries with new words and their forms; create larger parallel and comparative corpora; solve proper noun and systematic problems, etc.

Straipsnis įteiktas 2010 01
Parengtas spaudai 2010 12

Apie autorius

Inga Petkevičiūtė, KTU Nuotolinio mokymosi informacinių technologijų magistrantė.
Adresas: Kauno technologijos universitetas, Informatikos fakultetas, Studentų g. 50–414, 51368, Kaunas.
El. paštas: inga.petkeviciute@gmail.com

Bronius Tamulynas, technikos mokslų daktaras, Kauno technologijos universiteto Kompiuterių tinklų katedros docentas.
Mokslinės veiklos sritys: kompiuterinio kalbų vertimo technologijos, kompiuterinė lingvistika, intelektualiųjų sistemų modeliavimas, lingvistinių duomenų struktūros.
Adresas: Kauno technologijos universitetas, Informatikos fakultetas, Studentų g. 50–414, 51368, Kaunas.
El. paštas: bronius.tamulynas@ktu.lt

1 PRIEDAS

1 lentelė. Tyrimo rezultatai — apibendrinta klaidų suvestinė.

Eil. Nr.	Klaida	Meninis		Mokslinis		Buitinis		Administracinis		Publicistinis	
		Google	VDU	Google	VDU	Google	VDU	Google	VDU	Google	VDU
Morfologijos klaidos											
1.	Linksniai	58	32	94	32	73	27	66	56	89	16
2.	Pagrindinė V forma	9	4	20	2	14	1	1	2	17	3
3.	Skaičius	2	2	14	2	11	7	7	1	9	3
4.	Asmuo	10	1	3	1	5	-	-	-	4	-
5.	Giminė	7	3	17	1	14	4	5	1	5	6
6.	Kalbos dalis	7	11	15	7	20	10	1	3	11	13
7.	Neigiami veiksmažodžiai	-	1	1	-	3	-	-	-	-	-
8.	Praleidžiamas veiksmažodis	6	-	5	-	8	-	2	-	6	-
Leksikos klaidos											
9.	Neišverstas posakis	2	-	-	-	8	-	3	1	9	-
10.	Neišverstas žodis	14	7	7	8	28	19	1	2	9	3
11.	Žodžiai, sujungti brūkšneliu	-	1	-	-	3	2	-	-	3	2
12.	Pažodinis posakių vertimas	-	4	4	-	-	-	-	-	3	5
13.	Sutrumpinti žodžiai	7	4	-	-	-	-	-	-	1	1
14.	Daugiareikšmiškumas	34	28	33	43	42	56	4	42	33	49
15.	Įvardžiai	15	10	4	-	5	2	-	1	6	6
16.	Santrumpos	-	-	1	1	1	1	-	-	-	1
17.	Tikriniai vardai	-	-	-	-	2	2	-	-	1	5
Sisteminės/ programinės klaidos											
18	Neatsižvelgiama į diakritinius ženklus	-	1	-	-	-	1	-	-	-	-
19.	Papildomas žodis	-	-	3	1	2	-	2	-	1	-
20.	Žodis išverstas nežodynine reikšme	-	-	1	-	6	-	-	-	-	-
21.	Žodis išverstas kita kalba	10	-	3	-	5	-	-	-	15	-
22.	Praleistas žodis	6	-	17	-	19	2	3	-	6	-
23.	Didžiųjų raidžių rašyba	4	-	6	-	11	1	11	3	6	-
Viso:		198	107	248	98	280	135	106	112	234	113
		305		**346**		**415**		**218**		**347**	