DANIELIUS ALGIRDAS RALYS
Vilniaus universitetas

MAŠININIS VERTIMAS LIETUVIŲ KALBAI[1]

ESMINIAI ŽODŽIAI: lietuvių kalba, istorija, kompiuterinė lingvistika, mašininis vertimas, neuroniniai tinklai, dirbtinis intelektas.

ĮVADAS

Šiandien vis daugiau tekstų verčia kompiuteriai. Toks automatinis kompiuterinis vertimas dažniausiai vadinamas mašininiu vertimu (MV). Paprastai mašininis vertimas atliekamas siekiant kuo greičiau susivokti milžiniškame įvairiakalbės informacijos sraute, taip pat norint kuo plačiau paskleisti savo informaciją. Keitimasis elektronine informacija tarp tautų ir kalbų tampa svarbia kasdienio gyvenimo dalimi, atveria naujų galimybių mūsų kalbai, kelia jai naujų iššūkių ir problemų.

Terminą „mašininis vertimas“ (angl. *machine translation*) praėjusio amžiaus viduryje pirmasis pavartojo Warrenas Weaveris, pasiūlęs pritaikyti ką tik atsiradusius kompiuterius tekstų vertimui (Weaver 1949: 1). W. Weaveris siūlė pažvelgti į kita kalba parašytą tekstą kaip į šifrogramą, kurią galima dekoduoti. Britų ir amerikiečių specialistų kriptografiniai pasiekimai neseniai pasibaigusiame Antrajame pasauliniame kare žadino optimizmą. Tuomet nuoširdžiai buvo tikima, jog elektroninės mašinos per keletą metų padės įveikti Babelio bokšto prakeiksmą.

Deja, vertimas tobulėjo lėtai, o mašininio vertimo kokybė nė iš tolo neprilygo žmogaus vertimo kokybei. Po septyniasdešimties metų, po neabejotinų sėkmių ir nusivylimų, po daugybės nuveiktų darbų mašininio vertimo bendruomenėje vėl vyrauja optimistinės nuotaikos. Šįkart viltys dedamos į dirbtinius neuroninius tinklus ir greitą pažangą kuriant dirbtinį intelektą.

Mašininis vertimas neaplenkė ir lietuvių kalbos. Šiandien lengvai prieinamos internetinės mašininio vertimo priemonės verčia iš populiariausių kalbų į lietuvių kalbą ir atvirkščiai. Daugumos priemonių vertimo kokybė dar gana prasta, jų išverstus tekstus būtina redaguoti. Geresne kokybe išsiskiria tik vertimo priemonė *Google Translate*[2], kurioje pradėtos naudoti

[1] Straipsnis parengtas pagal pranešimą „Mašininis vertimas lietuvių kalbai“, skaitytą Vilniuje, 24-ojoje tarptautinėje Jono Jablonskio konferencijoje „Skaitmeniniai kalbos ištekliai, jų plėtros kryptys ir panaudos galimybės“ 2017 m. rugsėjo 29 d.

[2] Prieiga internete https://translate.google.com.

neuroninio vertimo technologijos. Dauguma šia priemone išverstų sakinių išlaiko originalo kalboje turėtą prasmę.

Mašininis vertimas plačiai naudojamas informacijai kitomis kalbomis skleisti . Dažnai mašininio vertimo rezultatai neatsakingai įkeliami į elektroninę erdvę visai neredaguoti. Tokių veiksmų pasekmės jau matomos: internete mirgėte mirga prasti lietuviški tekstai su iškraipyta informacija, kuriuos automatiškai sugeneravo vienokia ar kitokia mašininio vertimo programa[3]. Palaipsniui tokie tekstai gali prasiskverbti ir į tekstynus. Mašininis vertimas jau tapo kasdiene, kartais stebinančia, o kartais ir akis badančia realybe, kuri nusipelno ne tik įdėmesnio kalbininkų žvilgsnio, bet ir reikšmingesnio jų įsitraukimo į mašininio vertimo plėtotę.

Šiame straipsnyje trumpai pateikiama mašininio vertimo (MV) istorija, pristatomos kai kurios idėjos, turėjusios įtakos MV raidai. Pateikiami įvairių vertimo technologijų pasiekti rezultatai lietuvių kalbai. Analizuojamos MV pritaikymo lietuvių kalbai galimybės taisykliniais ir statistiniais bei neuroniniais metodais. Dėl vietos stokos neaprašomi mėginimai sujungti keletą šių metodų vienoje vertimo sistemoje.

1. MAŠININIO VERTIMO IŠTAKOS

Geras vertimas iš kitos kalbos visuomet yra tam tikras intelektinis iššūkis žmogui, o vertimas niekada nebuvo lengvas užsiėmimas. Nenuostabu, jog buvo siekiama vertimą kaip nors automatizuoti. 1932–1935 m. Prancūzijoje inžinierius Georges Artsrouni išrado ir sukonstravo universalų mechaninį prietaisą, kurį vadino mechaninėmis smegenimis (pranc. *cerveau mécanique*), kuris šalia kitų funkcijų turėjo galimybę versti renkamą tekstą į keletą kalbų (Hutchins 2004a: 12). Šiame elektromechaniniame prietaise buvo sukama ilga (iki 40 metrų ilgio) atminties juosta, pagaminta iš lankstaus kartono. Joje buvo surašytas daugiakalbis žodynas, galintis apimti net kelias dešimtis tūkstančių žodžių. Klaviatūroje surinktas prancūziškas žodis buvo mechaniškai koduojamas, o pagal tą kodą į reikiamą vietą pasisukdavo atminties juosta. Po keliolikos sekundžių prietaisas parodydavo įvesto žodžio vertimą (Daumas 1965: 294–295). Prietaisas buvo demonstruojamas Paryžiaus pasaulinėje parodoje 1937 m., sulaukė didelio susidomėjimo ir buvo apdovanotas Didžiuoju prizu. Faktiškai tai buvo pirmasis pasaulyje

[3] Džonatanas Sviftas (Jonathan Swift) savo *Guliverio kelionėse* sarkastiškai aprašo mašiną, dėliojančią beprasmes žodžių kombinacijas, kurios buvo nuolat stropiai surašomos į knygas. Nūdien toks pat niekalas gali būti pagamintas daug greičiau...

veikiantis mechaninis žodynas. G. Artsrouni vėliau dar tobulino savo prietaisą, tačiau mechaninės problemos ir prasidėjęs Antrasis pasaulinis karas nutraukė darbus.

1933 m. Sovietų Sąjungoje Piotras Smirnovas-Trojanskis gavo autorinį liudijimą elektromechaninei vertimo mašinai[4]. Nors ši mašina taip ir nebuvo pagaminta, tačiau išradėjas suformulavo daug vertingų mašininio vertimo idėjų, kurios buvo realizuotos tik po daugelio metų, atsiradus kompiuteriams. P. Smirnovas-Trojanskis kalbėjo rusiškai, todėl gerai suvokė fleksinių kalbų ypatybes. Jis pasiūlė vertimo procesą suskirstyti į tris etapus: pirmajame etape žmogus tekstą turi redukuoti į lemas ir jas anotuoti morfologinėmis-sintaksinėmis žymomis (analizės etapas); tada antrajame etape mašina automatiškai suranda vienos ar kelių vertimo kalbų lemas ir priskiria joms tas pačias žymas (transformavimo etapas); trečiajame etape redaktorius pagal vertimo lemas ir jų žymas turi paruošti sklandų vertimo tekstą (generavimo etapas). P. Smirnovas-Trojanskis taip pat siūlė įvairias sinonimų, homonimų ir idiomų vertimo metodikas. P. Smirnovo-Trojanskio idėjos nebuvo įvertintos jo gimtinėje, jis taip ir nesulaukė kompiuterių eros[5], tačiau išradėjo idėjos neliko užmirštos, vėliau tapo žinomos ir Vakaruose.

Praktinio mašininio vertimo era prasidėjo atsiradus kompiuteriams. 1947 m. W. Weaveris laišku kreipėsi į genialų kalbininką ir kompiuterių kūrėją Norbertą Wienerį klausdamas, ar naujausi kriptografijos pasiekimai ir ką tik atsiradę kompiuteriai[6] gali būti pritaikyti kalbų vertimui. W. Weaveris teigė, jog bet kurį rusišką tekstą galima traktuoti kaip kirilicos raidėmis užkoduotą anglišką pranešimą. Jam atrodė, kad kompiuteriai gali tą pranešimą tiesiog dešifruoti. Deja, N. Wienerio atsakymas buvo skeptiškas.

1949 m. W. Weaveris platesnei auditorijai išsiuntė memorandumą, kuriame pasiūlė panaudoti kompiuterius tekstams versti (Weaver 1949: 7)[7]. Tuo metu įvairiose amerikiečių ir britų laboratorijose vyravo paprasto pažodinio mašininio vertimo idėjos. Mokslininkas buvo tikras, kad kompiuteriai gali versti daug tobuliau. Memorandume W. Weaveris pasiūlė keturias vertimo problemų sprendimo ir tyrimo kryptis. Visų pirma, jis pasiūlė statistiškai analizuoti verčiamojo žodžio artimiausią aplinką ir taip išspręsti polisemijos problemas. Ši įžvalga šiandien sudaro statistinio mašininio vertimo pagrindą. Antrasis pasiūlymas rėmėsi prielaida, jog kalboje

[4] Autoriniame liudijime ši priemonė vadinama „mašina žodžiams parinkti ir spausdinti verčiant iš vienos kalbos į kitą arba keletą kalbų vienu metu“.

[5] P. Smirnovas-Trojanskis mirė 1950 m.

[6] Pirmasis britų elektroninis kompiuteris *Colossus* pradėjo veikti 1944 m. pradžioje, pirmasis amerikiečių elektroninis skaitmeninis kompiuteris ENIAC buvo pagamintas 1945 m. pabaigoje. Pirmasis sovietinis kompiuteris MESM buvo pagamintas 1950 m.

[7] Šiame W. Weaverio memorandume pavartotas terminas *mašininis vertimas* (angl. *machine translation*) veikiai prigijo.

yra loginių elementų, todėl naudojant neuroninių tinklų modelius (McCulloch, Pitts 1943: 115–133), turėtų būti galima dedukuoti vertimą iš originalo kalboje esančių prielaidų. Laikas parodė šios įžvalgos genialumą – šiandien neuroninis mašininio vertimo metodas verčia kokybiškiausiai. Trečią problemų sprendimo kryptį turėtų sudaryti kriptografijos metodų taikymas kalbų vertimui. Po keturiasdešimties metų tai buvo realizuota statistinio mašininio vertimo algoritmuose. Ketvirtajam pasiūlymui iliustruoti W. Weaveris nupiešė alegorinį paveikslą, kuriame žmonės apgyvendinti aukštuose uždaruose bokštuose, simbolizuojančiuose skirtingas kalbas. Žmonėms šiaip taip pavyksta bendrauti susišūkaujant per storas bokštų sienas. Laimei, bokštai turi bendrą pamatą ir požemį, tereikia ten nusileisti ir tada turėtų būti lengviau bendrauti. W. Weaveris teigė, jog pasaulio kalbose glūdi tam tikra bendra giluminė struktūra, kurią dar reikės atskleisti. Pastarasis mokslininko pasiūlymas paskatino imtis universalios kalbos – mašininio vertimo interlingvos[8] – paieškos ir konstravimo.

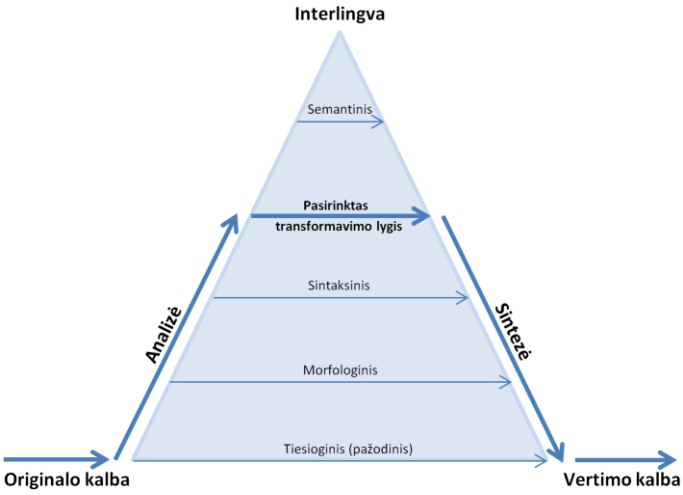
Po W. Weaverio memorandumo mašininis vertimas netruko įsibėgėti. 1954 m. Niujorke buvo viešai pademonstruota mašininio vertimo sistema, sukurta bendradarbiaujant IBM korporacijai ir Džordžtauno universitetui. Buvo verčiama iš rusų kalbos į anglų kalbą. Sistemoje buvo suprogramuotos tiktai šešios gramatinės taisyklės, o vertimui naudota apie 250 žodžių (Hutchins 2004b: 102). Sistema buvo orientuota į organinės chemijos tematiką, vertimo pavyzdžiai iš anksto kruopščiai apgalvoti. Po eksperimento viltasi, kad po kelerių metų mašinos vers puikiai, o darbams tęsti buvo paskirtas dosnus finansavimas. MV sparčiai imamas plėtoti Jungtinėse Amerikos Valstijose (JAV) ir Sovietų Sąjungoje, siekiant įgyti strateginį pranašumą šaltajame kare. Populiariausios verčiamos kalbos buvo rusų ir anglų.

2. TAISYKLINIS MAŠININIS VERTIMAS

Praėjusio amžiaus šeštojo dešimtmečio viduryje atsirado mašininio (kompiuterinio) vertimo sistemos, kurias imta vadinti taisyklinėmis (angl. *rule-based*). Jos buvo kuriamos laikantis požiūrio, kad kalbą galima aprašyti naudojant tam tikrų taisyklių (taip pat ir gramatinių) rinkinį. Taikant šį metodą, originalo kalbos sakinys yra analizuojamas iki tam tikro pasirinkto lygio, nustatoma sakinio struktūra, kuri transformuojama pagal susikurtas taisykles į vertimo kalbos ekvivalentinę struktūrą, pagal kurią sugeneruojamas vertimo kalbos sakinys. Įvairių kalbų sakinio struktūros gali labai skirtis, turėti vienų ar kitų ypatybių, kurios turi būti atspindėtos MV

[8] Nereikėtų maišyti su natūralios kalbos interlingvomis – esperanto, ido, volapiuko ir pan.

sistemoje. Lietuvių kalbos sakinio sintaksinių struktūrų pavaizdavimo, tinkamo mašininiam vertimui, problemas tyrinėjo Daiva Šveikauskienė (2005: 411–417). Yra įvairių taisyklinio MV sistemų atmainų, kurios skiriasi pagal analizavimo gylį (1 pav.[9]). Kuo gilesnė sakinio analizė, tuo paprastesnis transformavimas. Idealiu atveju, analizuojant tekstą iki interlingvos gylio, nebereikėtų jokios transformacijos – interlingvinis teksto atvaizdas turėtų būti vienodas visose kalbose.



1 PAV. Taisyklinio mašininio vertimo Vauquois trikampis

Taisyklinio MV problemas mėginta spręsti taikant tiek empirinius, tiek ir lingvistinius metodus. Empirikų stovykloje populiariausias buvo tiesioginio vertimo metodas. Taip verčiant tarp dviejų kalbų naudojamas tiktai žodynas ir paprastos programines taisyklės, praktiškai neatliekama kalbų analizė ar sintaksinis reorganizavimas (toks metodas atitinka žemiausią lygį 1 pav.). JAV šeštajame praėjusio amžiaus dešimtmetyje Erwino Reiflerio vadovaujama tyrėjų grupė kūrė specialius žodynus mašininiam vertimui, kuriuose šalia leksinių ekvivalentų buvo įdėtos ir lokalios žodžių tvarkos keitimo taisyklės. Taip pat buvo pateikiami daugelio frazių vertimai ir apibendrinamosios sąvokos (angl. *cover terms*) polisemijos problemoms spręsti. 1964 m. pradėjo veikti *Mark II* rusų–anglų tiesioginio vertimo sistema, sukurta JAV karinių oro pajėgų užsakymu. Joje panaudotas E. Reiflerio žodynas, apimantis per 170 000 žodžių (Reifler 1960: 312).

Lingvistiniai metodai plačiau buvo taikomi Džordžtauno universitete, kuriame susibūrė gausios mašininio vertimo tyrėjų pajėgos, pasiskirsčiusios į keletą skirtingus metodus taikančių

[9] Tokios iliustracijos dažnai vadinamos Vauquois trikampiu, šį iliustravimo metodą pasiūlė Prancūzijos mašininio vertimo pradininkas Bernarďas Vauquois (1976: 335).

grupių. Michaelo Zarechnako vadovaujamoje grupėje buvo sukurta Džordžtauno automatinio vertimo (angl. GAT – *Georgetown Automatic Translation*) sistema. Joje buvo atliekama trijų lygių analizė: morfologinė (apimanti ir idiomų identifikavimą), sintaksinė bei sintagminė. Modifikuota GAT sistema sėkmingai buvo pradėta naudoti Euratome (Isproje, Italijoje) 1963 m. ir JAV Atominės energijos komisijoje 1964 m. Daug ilgiau užtruko sukurti interlingvos pagrindu veikiančią MV sistemą. Komercinę vertę turinti MV sistema KANT pasirodė tik paskutiniame praėjusio amžiaus dešimtmetyje (Mitamura ir kt. 1991: 105–118). Sistema buvo naudojama kai kuriems techniniams tekstams versti. Bendrinei kalbai išreikšti interlingva taip ir nebuvo sukurta.

Visą dešimtmetį po IBM–Džordžtauno MV prezentacijos buvo labai stengiamasi esmingai pagerinti MV vertimo kokybę[10], tačiau pažanga buvo labai menka. 1966 m. JAV įkurtas ALPAC (angl. *Automatic Language Processing Advisory Committee*) komitetas nusprendžia, jog MV artimiausiu metu neturi perspektyvų, nes yra prastos kokybės, jį dar turi redaguoti vertėjas, tam sugaišdamas daugiau laiko, negu versdamas be jokių mašinų. ALPAC komitetas nepaisė to fakto, kad mašininis vertimas padeda išversti milžiniškus informacijos kiekius, o daugelį vartotojų tenkina prastas, vos prasmę perteikiantis vertimas. Po neigiamų komiteto išvadų MV projektų finansavimas JAV buvo praktiškai nutrauktas keliems dešimtmečiams, finansavimas sumenko ir kitose šalyse.

Nepaisant susidariusio skeptiško požiūrio į MV, šioje srityje vyko nuolatinė pažanga. 1968 m. JAV pradėta kurti taisyklinio vertimo sistema SYSTRAN (Toma 1977: 569–581) buvo nuolat tobulinama ir vėliau tapo komercinė. Ši sistema buvo pradėta naudoti JAV Gynybos departamente, NASA, Euratome, Europos Komisijoje ir daug kur kitur. Pagal 1975 m. komercinį susitarimą Europos Komisija gavo teisę tobulinti sistemą savo reikmėms. Sistema pakankamai gerai vertė kai kurių kalbos sričių tekstus į daugelį kalbų. Nuo 2010 m. SYSTRAN ėmė naudoti hibridines technologijas. Gana gerai veikiančios MV sistemos atsirado ir kitose šalyse: ARIANE (Grenoblyje, Prancūzijoje), į ją panaši MU sistema (Kiote, Japonijoje), taip pat SUSY (Sarbriukene, Vokietijoje). Europinis EUROTRA projektas (1982–1992 m.), kainavęs daugiau kaip 50 000 000 ECU[11], baigėsi nesėkme – šimtai specialistų taip ir nesukūrė veikiančios MV sistemos. Taip prasidėjo rimta taisyklinio MV krizė. Vėliau daug metų buvo trypčiojama vietoje.

[10] JAV per šį laikotarpį MV sistemų kūrimui ir tobulinimui išleista apie 20 mln. dolerių.

[11] 1994 m. Europos Bendrijų Komisijos (angl. CEC – *Commission of the European Communities*) pranešime pateikti EUROTRA finansavimo duomenys – Komisija 1982–1992 m. projektui skyrė 37,5 mln. ECU (CEC 1994: 3.10), įvairių šalių vyriausybės šešiolikai EUROTRA centrų papildomai skyrė daugiau kaip 20 mln. ECU (CEC 1994: A4.2).

Taisyklinis mašininis vertimas Lietuvoje atsirado tik šiame amžiuje. 2003–2004 m. mašininio vertimo sistema *Česílko* (Hajič ir kt. 2000: 10–12) buvo pritaikyta lietuvių kalbai (Homola, Rimkutė 2004: 77–81). Tuo laikotarpiu Petras Homola ir Erika Rimkutė tyrinėjo šios sistemos mašininio vertimo galimybes tarp čekų ir lietuvių kalbų. 2005–2007 m. Vytauto Didžiojo universitetas vykdė Europos Sąjungos Struktūrinių fondų finansuojamą projektą „Internetinė informacijos vertimo priemonė“. Projektuojamos MV sistemos struktūra aprašyta Vido Daudaravičiaus (2006: 7–18). Buvo sukurta vieša internetinė vertimo priemonė iš anglų kalbos į lietuvių kalbą[12]. Vertimo variklį pateikė Rusijos kompanija PROMT. Dalis sistemos kalbinių komponentų buvo paruošta Lietuvoje. Sistema puikiai vertė tik kai kuriuos sakinių tipus, geri vertimai sudarė mažumą. 2007 m. buvo atliktas šios MV sistemos kalbinis įvertinimas (Rimkutė, Kovalevskaitė 2008: 257–264). Audronė Daubarienė ir Greta Ziezytė (2013: 55–61) įvertino VDU MV sistemos vertimo atitiktį tekstualumo standartams (De Beaugrande, Dressler 1981: 3). Savo darbe autorės nustatė, jog ištirtieji skirtingų žanrų mašininio vertimo tekstai dėl daugybės semantinių klaidų neatitiko koherencijos ir kai kurių kitų tekstualumo standartų, todėl tokie vertimai negali būti priimtini ir informatyvūs skaitytojui.

Slenkant dešimtmečiams ir, iš esmės, tūpčiojant vietoje, buvo pradėta suvokti, jog kalba yra itin sudėtingas reiškinys, kurį labai sunku aprašyti taisyklėmis. Daugiareikšmiškumo, anaforų vertimo problemoms spręsti buvo siūloma naudoti ekstralingvistinę informaciją apie pasaulio sandarą ir jo logiką (Bar-Hillel 1958: 197–207). Tačiau tokios susistemintos informacijos tiesiog nebuvo, o formalus teksto prasmės aprašymas rodėsi sunkesnis uždavinys nei pats vertimas. Buvo akivaizdu, jog mašininio vertimo problemos reikalavo revoliucinių sprendimų.

2. STATISTINIS MAŠININIS VERTIMAS

1988 m. IBM mokslininkų grupė (Brown ir kt. 1988: 71–76) pasiūlė mašininiam vertimui naudoti lygiagrečiuose tekstynuose glūdinčią informaciją, ieškant tikimiausių vertimo variantų. Pasiūlytas vertimo modelis remiasi prielaida, jog triukšmingame kanale įvyksta originalo kalbos sakinio **f** pasikeitimas į sakinį **e** vertimo kalboje. Triukšmas kanale statistiškai iškraipo originalo kalbos sakinį **f**, todėl kanalo išėjime galime gauti įvairius sakinius ***e***, kuo įvairiausio ilgio, turinčius vienokią ar kitokią prasmę vertimo kalboje ar jos visai neturinčius. Kadangi tas kanalas stengiasi išversti teisingai, tai pro triukšmą vis dažniau prasiskverbia teisingi vertimo variantai.

[12] Prieiga internete http://vertimas.vdu.lt/twsas/.

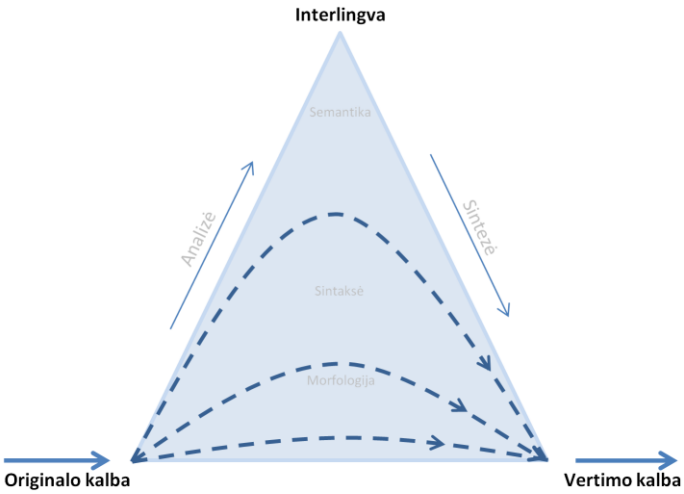
Šiame modelyje laikoma, kad kai į kanalą pasiunčiamas sakinys **f**, tinkamiausias vertimo variantas **ê** iš visų **e** yra tas, kurio pasirodymo kanalo išėjime sąlyginė tikimybė p(**e**|**f**) yra didžiausia, t. y.

$$\hat{\mathbf{e}} = \operatorname{argmax}_{\mathbf{e}} p(\mathbf{e} \mid \mathbf{f}).$$

Čia susiduriama su problema, jog norint rasti geriausią vertimo variantą **ê** reikia peržiūrėti ne tik visus galimus sakinius vertimo kalboje, bet ir visokiausius žodžių kratinius joje. Suskaičiuoti tokį kiekį tikimybių yra labai sunku. Pritaikius Bajeso teoremą, gaunama lengviau suskaičiuojama ekvivalentinė išraiška (Koehn ir kt. 2003: 127):

$$\operatorname{argmax}_{\mathbf{e}} p(\mathbf{e} \mid \mathbf{f}) = \operatorname{argmax}_{\mathbf{e}} p(\mathbf{f} \mid \mathbf{e})\, p(\mathbf{e})$$

Ši lygtis yra pamatinė statistinio mašininio vertimo lygtis (Brown ir kt. 1993). Joje sąlyginė tikimybė $p(\mathbf{f}|\mathbf{e})$ parodo tikimybę, kad kanalu buvo pasiųstas sakinys **f**, jei išėjime pasirodė sakinys **e**. Šis lygties narys vadinamas *vertimo modeliu*, kuris apskaičiuojamas iš dvikalbių tekstynų. Antrasis lygties narys $p(\mathbf{e})$ parodo sakinio **e** tikimybę vertimo kalboje. Šis narys aprašo *kalbos modelį*, kuris apskaičiuojamas iš vienkalbių vertimo kalbos tekstynų[13].



2 PAV. Statistinio mašininio vertimo Vauquois trikampis

IBM mokslininkų pasiūlytas vertimo būdas pagal vertimo kokybę iškart pradėjo konkuruoti su taisykliniu vertimu. Statistinių MV sistemų kūrėjus žavėjo galimybė versti į daugybę kalbų, neturint supratimo nei apie vertimą, nei apie gramatiką, užtekdavo kuriamas sistemas apmokyti iš turimų tekstynų. Klasikinio statistinio vertimo metu neatliekama jokia

[13] Kalbos modelis sumažina skaičiavimų apimtį, nes padeda iškart atmesti tekstynuose nerandamus sakinius, padeda parinkti sklandžiausią vertimą, taip pat netiesiogiai parodo, kodėl grožinę literatūrą geriau verčia vertėja(s), kurio(s) vertimo kalba yra gimtoji.

originalo kalbos sakinių analizė ar vertimo kalbos sakinių sintezė (2 pav.), verčiama tik pagal sistemos apmokymo metu nustatytas tikimybes. Peterio Browno ir jo kolegų pasiūlyti penki žodžių išdėstymo vertimuose modeliai nebuvo pakankamai lankstūs, jie leido originalo kalbos žodį išversti vienu ar keliais žodžiais arba jo išvis neversti. 2003 m. buvo pasiūlytas (P. Koehn ir kt. 2003: 127–133) lankstesnis mašininio vertimo modelis, galintis versti ir originalo kalbos frazes.

Šiuo laikotarpiu buvo sparčiai tobulinami mašininio vertimo kokybės automatinio vertinimo metodai, kadangi buvo labai svarbu greitai sužinoti, kuri derinamos mašininio vertimo sistemos versija verčia geriausiai. Taip pat buvo siekiama, kad toks vertinimas leistų objektyviai palyginti skirtingų sistemų vertimo kokybę. Ilgainiui viena populiariausių tapo BLEU (angl. *bilingual evaluation understudy* akronimas) metrika (Papineni ir kt. 2002: 311–318), kurioje automatiškai palyginama, kiek toli mašininio vertimo tekstas yra nutolęs nuo žmogiškojo vertimo. Šioje metrikoje neatsižvelgiama į teksto suprantamumą ar gramatinį taisyklingumą, joje skaičiuojami tiktai tiksliai sutampantys žodžiai ar frazės. Skaičiuojant metriką sutampančios frazės turi santykinai didesnį statistinį svorį nei paskiri sutampantys žodžiai. Priklausomai nuo uždavinio, MV kokybei vertinti dažnai naudojamos ir kitokios, vienokią ar kitokią gerą ypatybę turinčios, metrikos*: F-measure*, METEOR (Banerjee, Lavie 2005: 62–72), NIST (Doddington 2002: 138–145), TER (Snover ir kt. 2006: 223–231) ir dar daug kitų.

Nepaisant statistinio MV populiarumo, vis labiau aiškėjo šio metodo ribotumas verčiant į morfologiškai turtingas (fleksines) kalbas. Į fleksinių kalbų lygiagrečius tekstynus dažniausiai nėra patekusios visos galimos frazių ar net atskirų retesnių žodžių formos. Jeigu kurios nors gãlimos fòrmos nebuvo įtráuktos į MV sistemos apmokymą, tai tókios fòrmos ir nebus veřčiamos. Šį klasikinio MV trūkumą sukelia duomenų nepakankamumas, dėl kurio apmokius sistemą gaunama išretinta vertimo tikimybių matrica. Siekdami įveikti šį trūkumą, Philippas Koehnas ir Hieu Hoangas pasiūlė faktorizuotą statistinį mašininį vertimą (Koehn, Hoang 2007a: 868–876). Buvo pasiūlyta į vertimo sistemą įlieti papildomos lingvistinės informacijos anotuojant lingvistinės analizės žymomis lygiagrečius tekstynus, kurie vėliau naudojami sistemai apmokyti. 2007 m. P. Koehnas su kolegomis plačiajai visuomenei pateikė išbaigtą statistinio MV paketą *Moses* atvirajame kode (Koehn ir kt. 2007b: 177–180). Faktorizuoto statistinio MV pritaikymas šiek tiek pagerindavo fleksinių kalbų vertimo kokybę, paprastai 1–2 BLEU procentais (Bojar 2007: 235–236; Skadiņš ir kt. 2010: 128–129).

Nuo 2008 m. rugsėjo 25 d. *Google Translate* statistinio MV sistema ėmė versti ir lietuviškai. 2012–2014 m. Vilniaus universitetas vykdė Europos Sąjungos Struktūrinių fondų

finansuojamą projektą „Anglų–lietuvių–anglų ir prancūzų–lietuvių–prancūzų kalbų mašininio vertimo, paremto statistiniais metodais, sistemos sukūrimas“. Didžiąją projekto darbų dalį atliko UAB *Tilde*. Buvo sukurta MV sistema ALPMAVIS ir visuomenei tapo prieinama vieša internetinė statistinio MV paslauga (https://www.versti.eu/)[14]. 2014 m. atliktų testų duomenimis, bendro pobūdžio tekstų vertimo kokybės įverčiai (BLEU metrikoje) daugiau kaip dvigubai viršijo lietuviško taisyklinio vertimo rezultatus ir buvo praktiškai tolygūs *Google Translate* vertimo sistemos rezultatams. Ypač gerai ALPMAVIS gali versti teisinius tekstus.

3. MAŠININIO VERTIMO Į LIETUVIŲ KALBĄ KOKYBĖ

Atsiradus naujam kalbos reiškiniui – mašinos verstiems lietuviškiems tekstams, pasidarė įdomu įvertinti tokio vertimo kokybę bendrinės kalbos atžvilgiu. Vertinti buvo pasirinktas bendro pobūdžio tekstas, nes jis geriau atspindi bendrinę kalbą, o vertimo kokybei matuoti paimta BLEU metrika.

Vertimo kokybei testuoti buvo pasirinktas internete paskelbtas bendro pobūdžio tekstas[15]. BLEU skaičiavimai atlikti ilgesnio dokumento nei žemiau rodomos jo ištraukos, kadangi BLEU įverčiai sakinių lygyje blogai koreliuoja su žmogiškuoju kokybės vertinimu (Kos, Bojar 2009: 140). Pateikta informacija padės skaitytojui suprasti, kaip BLEU CI (skaičiuojamas nedarant skirtumo tarp didžiųjų ir mažųjų raidžių) procentiniai įverčiai siejasi su mašininio vertimo kokybe. Kita vertus, ir nežiūrėdamas į skaičius, skaitytojas pats gali susidaryti nuomonę, kaip verčia pagal anksčiau aprašytus modelius veikiančios MV sistemos.

Toliau pateikiami to paties originalaus teksto vertimo variantai, gauti taikant įvairius vertimo metodus. Prieš išverstą tekstą nurodomas vertimo metodas, vertimo internete data ir BLEU CI procentinis įvertis.

Originalus tekstas anglų kalba:

Everyone wants a long, happy and healthy life, and scientists are doing their best to make this possible. They are tackling some of the major current health issues as well as emerging threats such as the increasing impact of Alzheimer's disease, diabetes and antibiotic-resistant 'superbugs'. Investment in health research and innovation will help us stay active, develop new, safer and more effective treatments and help keep our health and care systems

[14] ALPMAVIS taip pat pasiekiama per integruotų lietuvių kalbos ir raštijos išteklių informacinę sistemą *Raštija.lt* (www.raštija.lt).

[15] Iš programos HORIZON 2020. Europos Komisijos Mokslinių tyrimų ir inovacijų generalinis direktoratas 2014.

viable. It will give doctors the tools they need for more personalised medicine, and it will step up prevention and treatment of chronic and infectious diseases.

Vertėjo atliktas vertimas:

Visi nori gyventi ilgai ir laimingai, būti sveiki, o mokslininkai kaip galėdami stengiasi, kad tai taptų įmanoma. Jie mėgina išspręsti kai kurias svarbiausias egzistuojančias sveikatos problemas ir kylančias grėsmes sveikatai, tokias kaip vis augančios Alzheimerio ligos, diabeto ir antibiotikams atsparių mikroorganizmų keliamos problemos. Investicijos į sveikatos priežiūros mokslinius tyrimus ir inovacijas padės mums išlikti aktyviems, kurti naujus, saugesnius ir efektyvesnius gydymo būdus bei išlaikyti mūsų sveikatos priežiūros sistemos gyvybingumą. Gydytojams jos suteiks priemonių, kurios jiems būtinos parenkant vaistus pagal individualius pacientų poreikius; be to, tai prisidės prie lėtinių ir užkrečiamų ligų profilaktikos ir gydymo.

VDU taisyklinis mašininis vertimas, 2014 09 19, BLEU – 8,54:

Kiekvienas nori ilgo, laimingo ir sveiko gyvenimo, ir mokslininkai stengiasi, kad padarytų tai galima. Jie sprendžia kai kurias iš pagrindinių einamųjų sveikatos svarstomų problemų taip pat kaip pasirodo grasinimai tokie kaip didėjantis Alzheimerio ligos, diabeto ir antibiotinių atsparių 'superklaidų' poveikis. Investicija į sveikatos tyrinėjimą ir naujovę padės mums likti aktyvi, vystyti naują, saugesnį ir efektyvesnį gydymą, ir pagalba laiko mūsų sveikatą ir rūpinimosi sistemas gyvybingas. Tai duos gydytojams įrankius, kurių jiems reikia labiau įasmenintai medicinai, ir tai padidins prevenciją ir chroniškų ir infekcinių ligų gydymą.

ALPMAVIS statistinis mašininis vertimas, 2014 09 19, BLEU – 17,59:

Visi nori ilgą, laimingą ir sveiką gyvenimą, ir mokslininkai stengiasi, kad tai taptų įmanoma. Jie yra kai kurie iš pagrindinių sveikatos problemų ir iškylančių grėsmių, pvz., didėjantis poveikis Alzheimerio ligos, diabeto ir antibiotikai-atsparūs “superbugs “. Investicijos į sveikatos tyrimai ir inovacijos padės mums išlikti aktyvūs, kurti naujus, saugesnis ir veiksmingesnis ir padėti išlaikyti sveikatos apsaugos ir priežiūros sistemų gyvybingumą. Tai suteiks gydytojai įrankių jiems reikia daugiau suasmeninta medicina, ir tai skatins prevencijai ir gydymui lėtinėmis ir infekcinėmis ligomis.

Google statistinis mašininis vertimas, 2014 09 19, BLEU – 17,10:

Kiekvienas iš mūsų siekia ilgas, sveikas ir laimingas, o mokslininkai deda visas pastangas, kad būtų tai padaryti. Jie kreipiasi į kai kurių pagrindinių sveikatos problemų, taip pat su naujais pavojais, pavyzdžiui, didėjančio poveikio Alzheimerio ligos, diabeto ir "Superbugs" atsparių antibiotikams. Investicijos į mokslinius tyrimus ir inovacijas sveikatos padės mums išlikti aktyvus, sukurti naujų gydymo būdų saugiau ir efektyviau, ir padėti remti mūsų sveikatos priežiūros sistemų tvarumą. Tai suteiks gydytojams, turintiems įrankiai, jie turi praktikuoti labiau atitinkančio vaisto vartojimą ir dėti daugiau pastangų siekiant užkirsti kelią ir gydyti infekcines ir lėtinėmis ligomis.

Microsoft Translator (*Bing*) statistinis mašininis vertimas, 2017 09 27, BLEU –16,68:

Kiekvienas nori ilgą, laimingą ir sveiką gyvenimą, ir mokslininkai daro viską, kad tai įmanoma. Jie yra kova su kai kurių pagrindinių dabartines sveikatos problemas taip pat naujų grėsmių, tokių kaip Alzheimerio liga, diabetu ir antibiotikams atsparių superbugs didėjantį poveikį. Investicijos į sveikatos moksliniai tyrimai ir inovacijos padės mums išlikti aktyviems, kurti naujus, saugesnio ir veiksmingesnio gydymo ir padeda išlaikyti mūsų sveikatai ir sveikatos priežiūros sistemų gyvybinga. Jis suteikia priemones, reikalingas daugiau mediciną, ir tai sustiprins prevencijai ir gydymui lėtinės ir infekcinių ligų gydytojai.

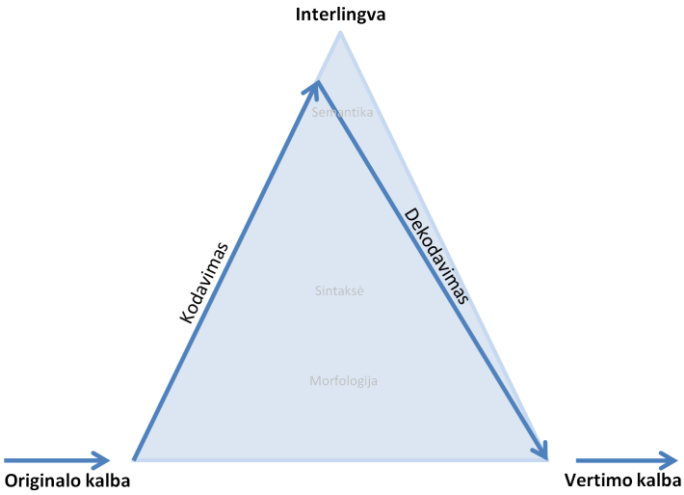
Pateiktos mašininio vertimo ištraukos atspindi tipišką mašininio vertimo kokybę verčiant bendro pobūdžio tekstus. Daug geresni bendro pobūdžio vertimai gaunami tik tuomet, jeigu verčiamas tekstas buvo kada nors panaudotas sistemai mokyti. Objektyvus MV sistemos vertimo kokybės įvertinimas galimas tik naudojant tekstus, kurie niekada nepateko į sistemos mokymus. Tenka pripažinti, jog realiai veikiančios statistinio ir taisyklinio MV sistemos bendro pobūdžio tekstus į lietuvių kalbą kol kas verčia prastai.

4. NEURONINIS MAŠININIS VERTIMAS

2013 m. Nalas Kalchbrenneris ir Philas Blunsomas pasiūlė mašininiam vertimui panaudoti neuroninius tinklus (Kalchbrenner, Blunsom 2013: 1700–1709). Idėja neuroninių tinklų skaičiavimui pasitelkti kompiuterių grafinius procesorius atvėrė galimybes spręsti realius uždavinius, tarp jų – atlikti ir mašininį vertimą. Neuroniniam MV (NMV) realizuoti naudojami milijonai dirbtinių neuronų. NMV koncepcija gana paprasta – koderis visą originalo kalbos sakinį keleto šimtų dimensijų vektorinėje erdvėje paverčia vadinamuoju minties vektoriumi[16], kurį sudaro skaičių seka, tokiu būdu pavaizduojanti sakinio prasmę. Faktiškai šis vektorius yra vektorizuota sakinio mintis. Vektorių, pavaizduojančių gimininguss daiktus ir idėjas, kryptys mažai skiriasi[17]. Dekoderis pagal minties vektorių sugeneruoja vertimą. Tiek koderis, tiek ir dekoderis konstruojami rekurentinių neuroninių tinklų pagrindu (Sutskever ir kt. 2014: 3104–3112; Cho ir kt. 2014: 1724–1734).

[16] Teminą išpopuliarino Geoffrey Hintonas, vienas žymiausių dirbtinio intelekto kūrėjų.

[17] Tuo galima įsitikinti, pvz. eksperimentuojant su vektorizavimo priemone *word2vec* (Mikolov ir kt. 2013: 3111–3119).



3 PAV. Neuroninio mašininio vertimo Vauquois trikampis

3 pav. pateikta neuroninio MV schema pagal Vauquois trikampį[18]. Reikia atkreipti dėmesį, kad automatiškai pasiekiamas aukštas analizės lygis, priartėjantis prie interlingvos. NMV sistema apmokoma naudojant lygiagrečius tekstynus. Apmokymo metu rekurentiniai neuroniniai tinklai automatiškai užprogramuoja morfologinę, sintaksinę ir semantinę informaciją, esančią sakiniuose kartu su įvairaus laipsnio aptiktais dėsningumais. Pagaliau mašininis vertimas tapo neatsiejamas nuo dirbtinio intelekto kūrimo. Neuroninis mašininis vertimas sparčiai tobulėja ir duoda daug geresnius rezultatus nei taisyklinis ar statistinis mašininis vertimas.

Tyrinėjant NMV galimybes greitai buvo susidurta su nauju reiškiniu – neuroniniai tinklai gebėjo versti tarp tokių kalbų porų, kurioms tie tinklai nebuvo tiesiogiai apmokyti. Jeigu NMV sistema buvo apmokyta versti iš kalbos A į kalbą B, taip pat iš kalbos A į kalbą C, tai sistema sugeba versti ir iš kalbos B į C (Johnson 2017: 339–351), tiesa, vertimo kokybė šiek tiek mažesnė. Toks reiškinys gali būti įvardintas kaip ekspromtinis vertimas (angl. *zero-shot translation*). Tai reiškia, jog MV sistemos neuroninis tinklas apmokymo metu susikuria savo interlingvą, kuri suprantama tik jam pačiam.

2016 m. pabaigoje *Google Translate* (GT) pradėtas naudoti neuroninis vertimas. Štai kaip atrodo mūsų iliustracinio pavyzdžio neuroninis vertimas (*Google* neuroninis MV 2017 09 14, BLEU CI – 25,94):

[18] Neuroninio mašininio vertimo pavaizdavimas pagal Vauquois trikampį pateiktas SYSTRAN 2016 m. interneto publikacijoje „Pure Neural Machine Translation“. Prieiga internete: http://www.systransoft.com/download/white-papers/systran-white-paper-PNMT-12-2016_2.pdf, (žiūrėta 2017 11 06).

Visi nori ilgo, laimingo ir sveiko gyvenimo, o mokslininkai daro viską, kad tai būtų įmanoma. Jais sprendžiamos kai kurios svarbiausios dabartinės sveikatos problemos, taip pat kylančios grėsmės, pvz., Didėjantis Alzheimerio ligos, diabeto ir antibiotikams atsparių "superbugs" poveikis. Investicijos į sveikatos mokslinius tyrimus ir inovacijas padės mums išlikti aktyvios, kurti naujus, saugesnius ir veiksmingesnius gydymo būdus, o mūsų sveikatos ir priežiūros sistemos bus gyvybingos. Tai suteiks gydytojams jiems reikalingus įrankius labiau individualizuotam vaistui, taip padidins lėtinių ir infekcinių ligų prevenciją ir gydymą.

Reikia pripažinti, kad mūsų pavyzdį neuroninis tinklas išvertė nepalyginamai kokybiškiau nei taisyklinio ar statistinio mašininio vertimo sistemos. Kol kas ne viską GT verčia taip sėkmingai, pvz., pamėginus išversti *Beauty requires no ornaments,* gauta – *Grožis nereikalingas papuošalams*. Panašiai eksperimentuojant galima nustatyti daugybę atvejų, kai GT verčia visai prastai, taigi dar yra ką tobulinti.

Vilniaus universitetas 2018 m. ruošiasi pradėti įgyvendinti naujos kartos neuroninį mašininį vertimą anglų, lietuvių, lenkų, prancūzų, rusų ir vokiečių kalboms. Į vertimo sistemą bus integruotas automatinis teksto keitimas į šneką ir lietuviškai padiktuoto teksto vertimas.

5. BAIGIAMOSIOS PASTABOS

Mašininio vertimo pažanga šiandien jau neatsiejama nuo pasiekimų kuriant dirbtinį intelektą. Pagal vertimo kokybę visi klasikiniai MV metodai tobulintini, galimybės nėra išsemtos ir nėra aišku, kur jų ribos. Romantinis lengvų pasiekimų laikotarpis baigėsi, euristiniai matematiniai ar inžineriniai metodai ne visada padeda. Vaikai išmoksta kalbą tarsi automatiškai, be pastangų, tačiau vėliau jų kalba ištobulinama tik mokantis gramatikos, studijuojant literatūrą. Panašu, kad be kalbininkų reikšmingesnio įsitraukimo toliau tobulinti mašininį vertimą vargiai įmanoma. Tokį įsitraukimą palengvina atvirojo kodo platformose realizuotos mašininio vertimo priemonės: pvz. *Apertium* taisykliniam vertimui (Forcada ir kt. 2011: 127–144); *Moses* (Koehn ir kt. 2007b: 177–180), LetsMT! (Vasiljevs ir kt. 2012: 43–48) – statistiniam, *TensorFlow*, *OpenNMT* – neuroniniam. Reikia tikėtis, kad Lietuvos kalbininkai suvienys savo pastangas spręsdami modernius kompiuterinės lingvistikos iššūkius, nes kai kurie jų yra egzistenciškai svarbūs mūsų kalbai.

LITERATŪRA

Banerjee S., Lavie A. 2005: METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. – *Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005)*, Ann Arbor, Michigan, 62–72.

Bar-Hillel Y. 1958: Some linguistic obstacles to machine translation. – *Proceedings of the Second International Congress on Cybernetics*, Namur, 197–207.

Bojar O. 2007: English-to-Czech factored machine translation. – *Proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, 232–239.

Brown P. F., Cocke J., Della Pietra S. A., Della Pietra V. J., Jelinek F., Mercer R. L., Roossin P. S. 1988: A statistical approach to language translation. – *Proceedings of COLING 1988: The 12th International Conference on Computational Linguistics*, Budapest, 71–76.

Brown P. F., Della Pietra S. A., Della Pietra V. J., Mercer R. L. 1993: Mathematics of statistical machine translation: Parameter estimation. – *Computational Linguistics* 19(2), 263–311.

Cho K., Bart van Merrienboer, Gulcehre C., Bougares F., Schwenk H., Bengio Y. 2014: Learning phrase representations using RNN encoder-decoder for statistical machine translation. – *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP),* Association for Computational Linguistics, Doha, Qatar, 1724–1734.

Commission of the European Communities 1994: Final Evaluation Of The Results Of Eurotra: A Specific Programme Concerning The Preparation Of The Development Of An Operational Eurotra System For Machine Translation. DIANE Publishing Company, 1995, 194.

Daubarienė A., Ziezytė G. 2013: Machine translation: translated texts in terms of standards of textuality. – *Kalbų studijos* 22, 55–61.

Daudaravičius V., 2006: Pradžia į begalybę. Mašininis vertimas ir lietuvių kalba. – *Darbai ir dienos: Pažangos šuoliai* 45, 7–18.

Daumas M. 1965: Les machines à traduire de Georges Artsrouni. – *Revue d'histoire des sciences et de leurs applications*, tome 18, n°3, 283–302.

De Beaugrande R. A., Dressler W. U. 1981*: Introduction to Text Linguistics*, London: Longman, 270.

Doddington G. 2002: Automatic Evaluation of Machine Translation Quality Using Ngram Co-occurrence Statistics. – *Proceedings of the Second International Conference on Human Language Technology Research,* San Francisco, CA, USA., Morgan Kaufmann Publishers Inc., 138–145.

Forcada M. L, Ginestí-Rosell M., Nordfalk J., O'Regan J., Ortiz-Rojas S., Pérez-Ortiz J. A., Sánchez-Martínez F., Ramírez-Sánchez G., Tyers F. M. 2011: Apertium: a free/open-source platform for rule-based machine translation. – *Machine Translation* 25 (2), Free/Open-Source Machine Translation (June 2011), 127–144.

Hajič J., Hric J., Kuboň V. 2000: Machine Translation of Very Close Languages. – *Proceedings of the 6th Applied Natural Language Processing Conference*, Association for Computational Linguistics, 7–12.

Homola P., Rimkutė E. 2004: Artimų kalbų mašininis vertimas. – *Kalbų studijos* 6, 77–81.

Hutchins J. 2004a: Two precursors of machine translation: Artsrouni and Trojanskij. – *International Journal of Translation* 16(1), 11–31.

Hutchins J. 2004b: The Georgetown-IBM experiment demonstrated in January 1954. – *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas*, AMTA, Washington, DC, 102–114.

Johnson M., Schuster M., Le V. Q., Krikun M., Wu Y., Chen Z., Thorat N., Viégas F. B., Wattenberg M., Corrado G., Hughes M., Dean J. 2017: Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. – *Transactions of the Association of Computational Linguistics* 5 (1), 339–351.

Kalchbrenner N., Blunsom P. 2013: Recurrent Continuous Translation Models. – *Proc. of EMNLP*, October, Association for Computational Linguistics, Seattle, Washington, USA, 1700–1709.

Koehn P., Hoang H. 2007a: Factored translation models. – *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ACL, 868–876.

Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R., Dyer C., Bojar O., Constantin A., Herbst E. 2007b: Moses: Open source toolkit for statistical machine

translation. – *45th Annual Meeting of the Association for Computational Linguistics (ACL)*, Companion Volume, 177–180.

Koehn P., Och F. J., Marcu D. 2003: Statistical phrase-based translation. – *Proceedings of HLT-NAACL*, 127–133.

Kos K., Bojar O. 2009: Evaluation of Machine Translation Metrics for Czech as the Target Language. – *Prague Bull. Math. Linguistics* 92, 135–148.

McCulloch W., Pitts W. 1943: A logical calculus of the ideas immanent in nervous activity. – *Bulletin of Mathematical Biophysics* 5, 115–133.

Mikolov T., Sutskever I., Chen K., Corrado G. S., Dean J. 2013: Distributed Representations of Words and Phrases and their Compositionality – *Proceedings of the 26th International Conference on Neural Information Processing Systems*, Lake Tahoe, Nevada, Vol. 2, 3111–3119.

Mitamura T., Nyberg E., Carbonell J. 1991: An efficient interlingua translation system for multi-lingual document production. – *Proceedings of Machine Translation Summit III*, Washington, DC. Perspausdinta: – *Progress in Machine Translation*, ed. S. Nirenburg, IOS Press, (1993), 105–118.

Papineni K., Roukos S. Ward T., Zhu W. J. 2002: BLEU: a method for automatic evaluation of machine translation. – *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics*, 311–318.

Reifler E. 1960: The solution of MT linguistic problems through lexicography. – *Proceedings of the National Symposium on Machine Translation*, (February 2–5, 1960), ed. H.P. Edmunson, London: Prentice-Hall, (1961), 312–316.

Rimkutė E., Kovalevskaitė J. 2008: Linguistic Evaluation of the First English-Lithuanian Machine Translation System. – *Proceedings of the Third Baltic Conference on Human Language Technologies (2007)*, Kaunas, 257–264.

Skadiņš R., Goba K., Šics V. 2010: Improving SMT for Baltic Languages with Factored Models. – *Proceedings of the Fourth International Conference Baltic HLT, Frontiers in Artificial Intelligence and Applications*, Vol. 219, IOS Press, 125–132.

Snover M., Dorr B., Schwartz R., Micciulla L., Makhoul J. 2006: A Study of Translation Edit Rate with Targeted Human Annotation. – *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, Morristown, NJ, USA, 223–231.

Sutskever I., Vinyals O., V. Le Q. V. 2014: Sequence to sequence learning with neural networks. – *Proceedings of the 27th International Conference on Neural Information Processing Systems*, Montreal, Canada, Vol. 2, MIT Press Cambridge, MA, USA, 3104–3112.

Šveikauskienė D. 2005: Graph Representation of the Syntactic Structure of the Lithuanian Sentence. – *Informatica* 16 (3), 407–418.

Toma P. 1977: Systran as a multilingual machine translation system. – *Proceedings of the Third European Congress on Information Systems and Networks, Overcoming the language barrier*, München, 569–581.

Vasiļjevs A., Skadiņš R., Tiedemann J. 2012: LetsMT!: a cloud-based platform for do-it-yourself machine translation – *Proceedings of the ACL, System Demonstrations*, Jeju Island, Korea, 43–48.

Vauquois B. 1976: Automatic translation – A survey of different approaches. – *COLING-76*, Ottawa. Perspausdinta: – *Readings in machine translation*, eds. S. Nirenburg, H. L. Somers, Y. Wilks, MIT Press, (2003), 333-337.

Weaver W. 1949: Translation. – Perspausdinta: – *Machine Translation of Languages*, MIT Press, Cambridge, MA. (1955), 15–23.

Įteikta 2017 11 13
Priimta 2017 12 20

MACHINE TRANSLATION FOR LITHUANIAN LANGUAGE

Summary

The paper presents a historical overview as well as current state of the art of the machine translation. Translation from another language is always a certain intellectual challenge. In 1949 Warren Weaver suggested using computers to translate texts. The term "machine translation" (MT) appears. Machine translation has been rapidly developing during the first decades in order to gain a strategic advantage in the Cold War. Most popular translated languages were Russian and English. Word-to-word translation and large bilingual dictionaries covering more than 170,000 words prevailed.

In the 1950s and 1960s, machine translation systems appeared which could be called rule-based. They are based on the assumption that a language can be described using a set of rules (including grammatical). It was a rather optimistic period - it was expected to create a perfect machine translation in a few years. But a computer hardly "understands" grammar. Highly inflected languages require tens of thousands thoroughly hand-tuned and mutually consistent rules. Nobody has done this properly yet. The most advanced systems were the SYSTRAN MT system, launched in the European Commission and the Russian PROMT translation system. Subsequently, the rule-based MT progress slowed down. The EUROTRA project (1982-1992), by some estimates costing more than 50,000,000 ECU, fails – even hundreds of recruited specialists failed to create a functioning MT system. This marked a serious rule-based MT crisis. Development simply stalled for many forthcoming years.

The question was raised: if we cannot write so many rules, can it be translated without grammar at all? In 1990 a new breakthrough emerges - the research team at the *IBM Thomas J. Watson Research Center* formulates the basics of statistical machine translation. The translation process has been considered as a transmission of a certain message over a noisy channel. Decoding then has been performed on the basis of the Bayesian theorem. Translation is based on text corpora, especially on large parallel bilingual text corpora. There was a rapid improvement of statistical MT. The *EuroMatrix* project, supported by the European Commission, has created a universal open source machine translation software package MOSES, based on industry-level MT systems. Good results have been obtained - it turns out that you can translate without any dictionary or grammar! This method has greatly facilitated the translation of highly inflected languages too.

The achievements of machine translation today are effectively applied to the Lithuanian language as well. During 2005-2007 Vytautas Magnus University has carried out an EU-funded project "Internet Information Translator". The result was a public online translation service from English to Lithuanian. (http://vertimas.vdu.lt/twsas/). The rule-based translation engine was provided by the Russian company PROMT, while other linguistic resources were prepared in Lithuania. The overall quality of text translation in BLEU metrics (in percent) is about 10. In practice, this means that only every third sentence can be adequately understood. This translation tool still has considerable potential for improving, for example by expanding phrase dictionary. Since September 25, 2008 *Google Translate* also supports Lithuanian. According to the results of the tests (2014), the BLEU translation quality was estimated to be around 17.

2012–2014 Vilnius University implemented the EU-funded project "Creation of English–Lithuanian–English and French–Lithuanian–French machine translation system based on statistical methods". The result was a public online translation service (https://www.versti.eu/). According to the tests carried out in 2014, the BLEU estimates of the translation quality exceeded more than twice the rule-based translation results and were practically equivalent to the Google translation system. When translating the documents of certain domain (such as law) the achieved BLEU score is roughly twice as high as translating general texts and far exceeds Google's results (19-09-2014). However, even the best machine translations often require human intervention and final editing in order to get the perfect translation. So, can machines ultimately translate fine?

The last few years promise new breakthroughs using a neural machine translation. Neural networks themselves construct transformation rules. It is likely that in the near future the neural MT systems will translate better than an average translator. In 2018 Vilnius University is preparing to launch the EU-funded project of a new generation neural machine translation of English, Lithuanian, Polish, French, Russian and German.

Thus, the latest achievements in machine translation apply to the Lithuanian language as well.

KEYWORDS: Lithuanian, machine translation, computational linguistics, history, neural networks, artificial intelligence.

DANIELIUS ALGIRDAS RALYS
Vilniaus universiteto Taikomųjų mokslų institutas
M. K. Čiurlionio g. 29, 03100 Vilnius
danielius.ralys@gmail.com