

Mašininis vertimas: Kaip jis veikia?



Pranešėjai
dr. Arūnas Samuilis
Virginijus Dadurkevičius



Apžvalga:

- Kodėl reikalingas mašininis vertimas (MV)?
- Kodėl MV toks sudėtingas?
- MV istorinė apžvalga
- Kaip veikia modernus statistinis MV?
- Svarbiausia - duomenys!
- Kas jau padaryta Lietuvoje?

- Europa = Daugiakalbiškumas
- 24 oficialios kalbos
- Tiek daug informacijos išversti!
2015 metais ~ 2 mln. puslapių, puslapyje ~ 1500 ženklų!
- Vertimo sąnaudos!
2015 metais ~ 330 000 000 €, ~ 1 % viso EU biudžeto
- Vertimo kokybė?
Dažniausiai gera, bet gali trūkti nuoseklumo ir darnos
- Ar gali MV padėti?
Žinoma! Geras rezultatas pigiau ir greičiau ! Puikus pradinis vertimo variantas, nusistovėjusių terminų ir sąvokų vienodas vertimas.



Babelio bokštas, Piteris Breigelis Vyresnysis, 1563

Kodėl MV toks sudėtingas?



- Kalba yra sudėtingas, ne iki galo suprastas ir formalizuotas reiškinys
- Vienas atskiras žodis ar sakinyš gali turėti daug reikšmių
- Yra daug būdų pasakyti tą pačią mintį
- Reikšmė priklauso nuo konteksto
- Tiesioginė ir perkeltinė kalba

- Žodžių tvarka
- Morfologija
- Anaforos, metaforos ir t. t., ir pan.

- Žmogus kalbą supranta „be jokių mokslų“, net maži vaikai. Kompiuteriui viskas iki detalių turi būti formalizuota.

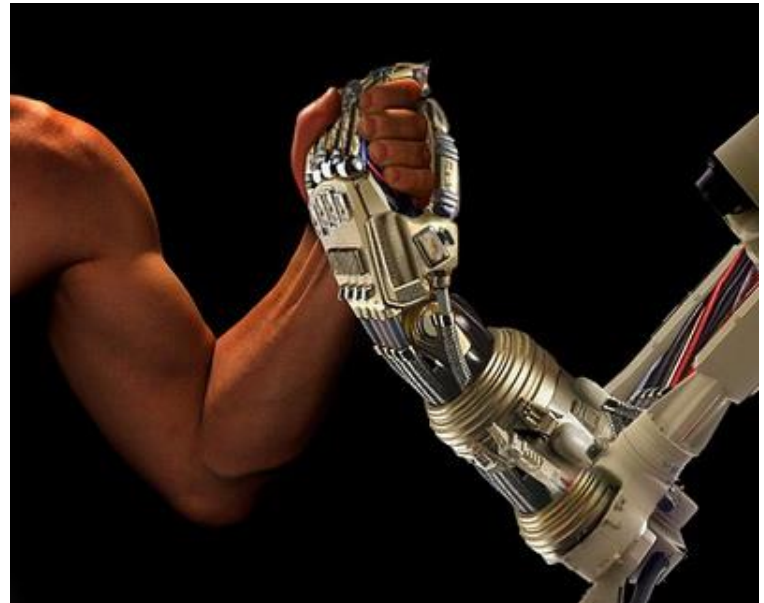


Image: <http://workingtropes.lmc.gatech.edu/wiki/index.php/File:Man-vs-machine.jpg>
License: CC BY-NC-SA 3.0

- 1943 metais pradėtas konstruoti pirmasis kompiuteris ENIAC. Jo tikslas – artilerijos sviedinių trajektorijų skaičiavimas.
- 1947 m. Warren Weaver pasiūlė panaudoti kompiuterius tekstų vertimui. Atsiranda terminas – mašininis vertimas.
- MV imamas sparčiai vystyti JAV ir SSRS, siekiant įgyti strateginį pranašumą šaltajame kare.
- Populiariausios verčiamos kalbos – rusų ir anglų.
- Vyrauja pažodinis vertimas, sukuriami dideli kompiuteriniai dvikalbiai žodynai, apimantys virš 200 000 žodžių.

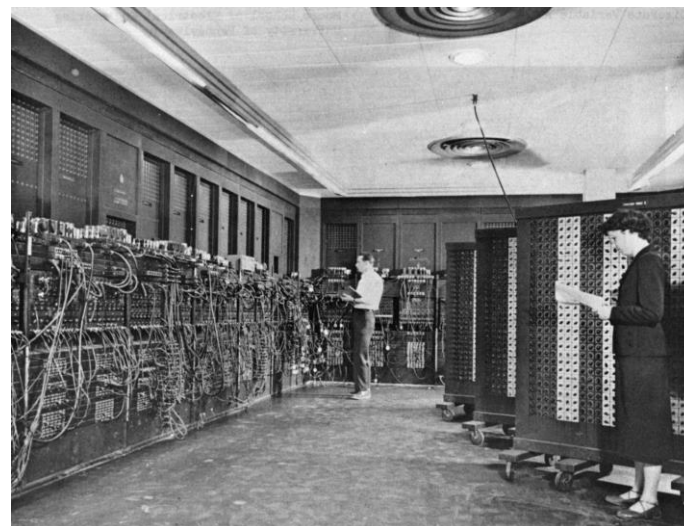


Image: <https://en.wikipedia.org/wiki/ENIAC#/media/File:Eniac.jpg>
License: public domain



- 1950 – 1960 metais atsiranda mašininio (kompiuterinio) vertimo sistemos, kurias galima pavadinti taisyklinėmis (*rule-based*).
- Jos kuriamos laikantis požiūrio, jog kalbą galima aprašyti naudojant tam tikrų taisyklių (taip pat ir gramatinių) sistemą.
- Optimistinis laikotarpis – tikėtasi per keletą metų sukurti tobulą mašininį vertimą.

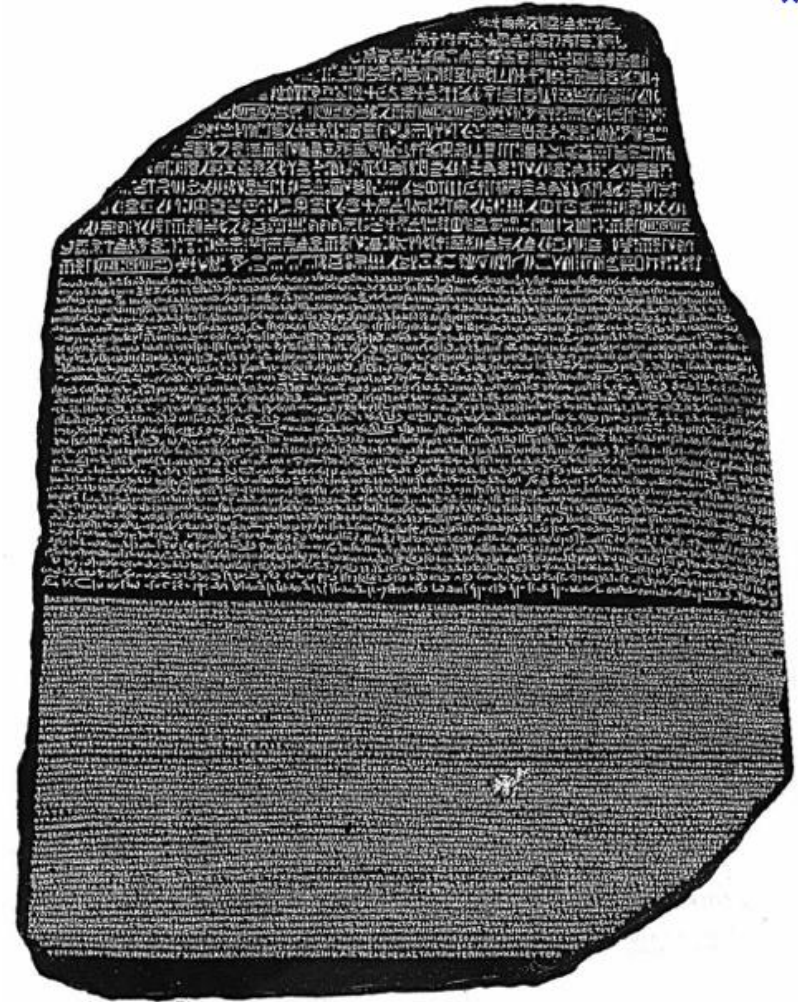


- Vyravo optimistinės MV perspektyvos, tačiau pasiekti prasti praktiniai rezultatai.
- 1966 m. JAV įkurtas ALPAC (Automatic Language Processing Advisory Committee) komitetas nusprendžia, jog MV artimiausiu metu neturi perspektyvų.
- MV projektų finansavimas JAV nutraukiamas dvidešimčiai metų, jis sumenksta ir kitose šalyse.
- Europinis EUROTRA projektas (1982 – 1992 m. m.), kainavęs apie 50 000 000 ECU, baigiasi nesėkme – šimtai specialistų taip ir nesukūrė veikiančios MV sistemos.
- Dar ir šiandien taisyklinio vertimo lyderiai – vis tas pats SYSTRAN bei kelių dešimtmečių senumo rusiška PROMT vertimo sistema.



- 1990 m. įvyksta naujas proveržis - IBM tyrėjų grupė suformuluoja statistinio mašininio vertimo pagrindus (P. Brown ir kt.).
- Vertimo procesas prilyginamas tam tikro pranešimo perdavimui triukšmingu kanalu.
- Dekoduojama remiantis Bajeso teorema.
- Vertimas remiasi tekstynais, vertimui ypač svarbūs lygiagretūs dvikalbiai tekstynai.
- Netikėtai geri rezultatai – pasirodo, galima versti neturint nei žodyno, nei jokio supratimo apie gramatiką!

**Rozetės akmuo – pirmasis
lygiagretus tekstynas, o taip pat
ir statistinio vertimo objektas**

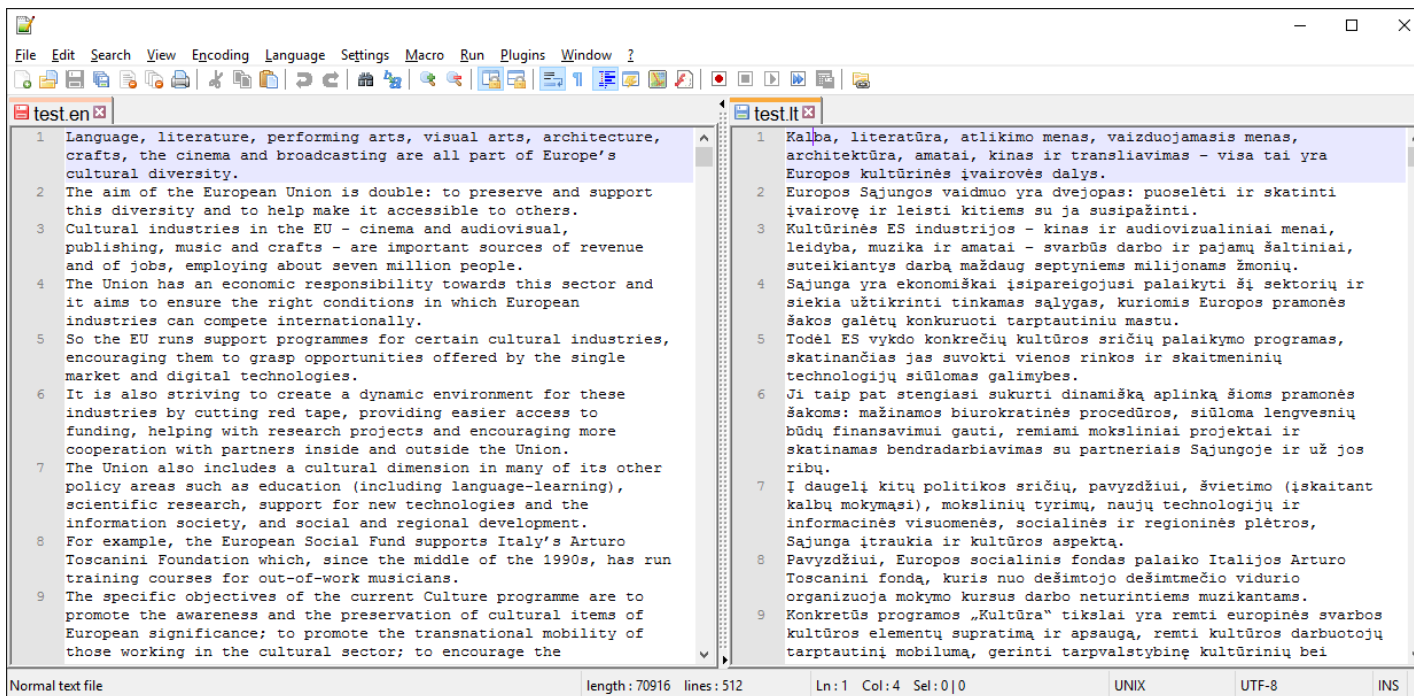




- Žodžiai turi daug prasmių. Daugiaprasmiškumas buvo ir išlieka svarbiausia kompiuterinio mašininio vertimo problema.
- Sunki problema – kaip versti įvardžius (anaforos atpažinimas).
The soldiers killed ten women. They have been buried next day.
Kas buvo palaidoti, **jie** ar **jos**, kareiviai ar moterys?
- Sintaksinių struktūrų nustatymas šių vertimo problemų neišspręs.
- Ieškoma išsigelbėjimo semantikoje bei kuriant įvairias ontologijas.
- MV problemos stimuliuoja pažangą dirbtinio intelekto kūrimo srityje.
- Populiarėja mišrios (hibridinės) vertimo sistemos, apimančios tiek taisyklinį, tiek ir statistinį MV.
- Nuo 2010 m. SYSTRAN (Systran Server 7) inkorporavo ir statistinį vertimą į savo sistemą.
- Europos Komisijos remto projekto *EuroMatrix* metu sukurtas universalus atviro kodo statistinio mašininio vertimo programinės įrangos paketas [MOSES](#).

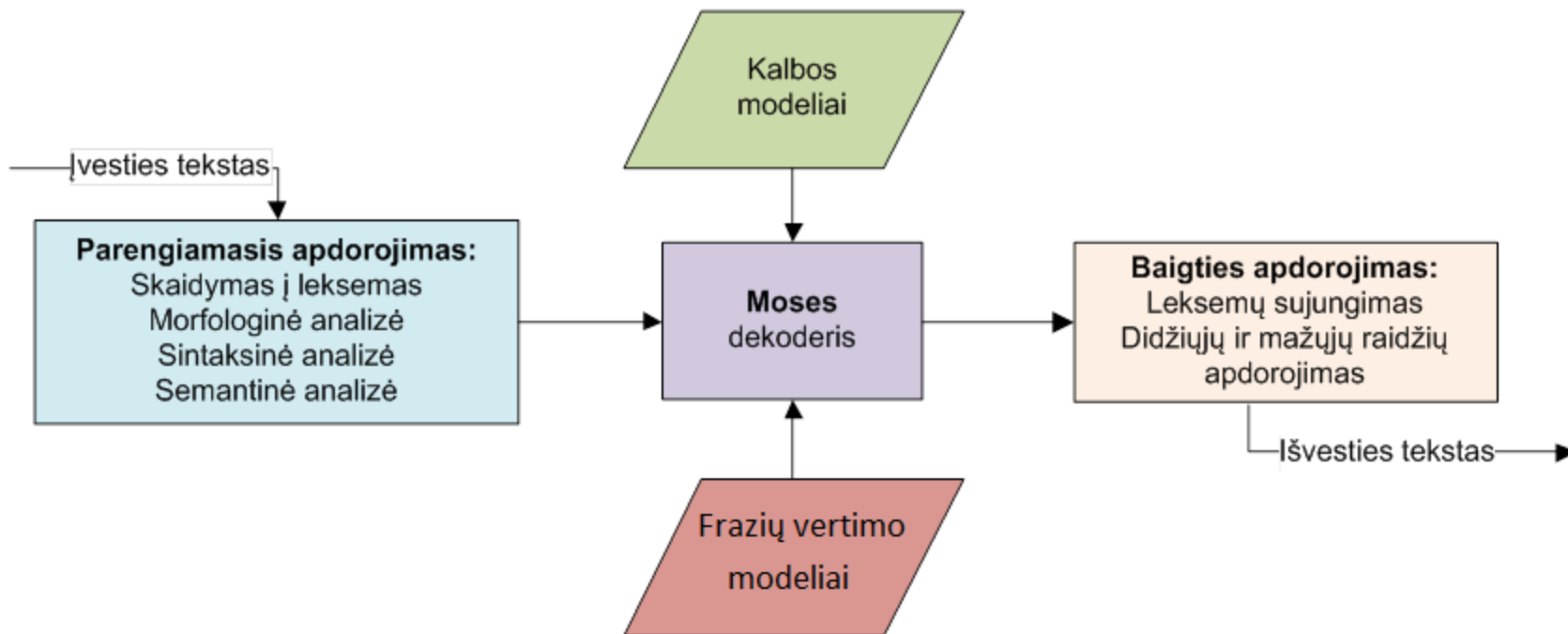
Statistinis MV apmokomas dviejų tipų duomenimis:

- Žmogaus išverstais lygiagrečiais dvikalbiais tekstais
- Vienakalbiu tekstu ta kalba, į kurią yra verčiama
- Kuo daugiau teksto – tuo geriau!
- Taip pat svarbu tekstų kokybė (rašybos klaidos ir pan.), jų srities tinkamumas ir lygiagretinimo tikslumas!



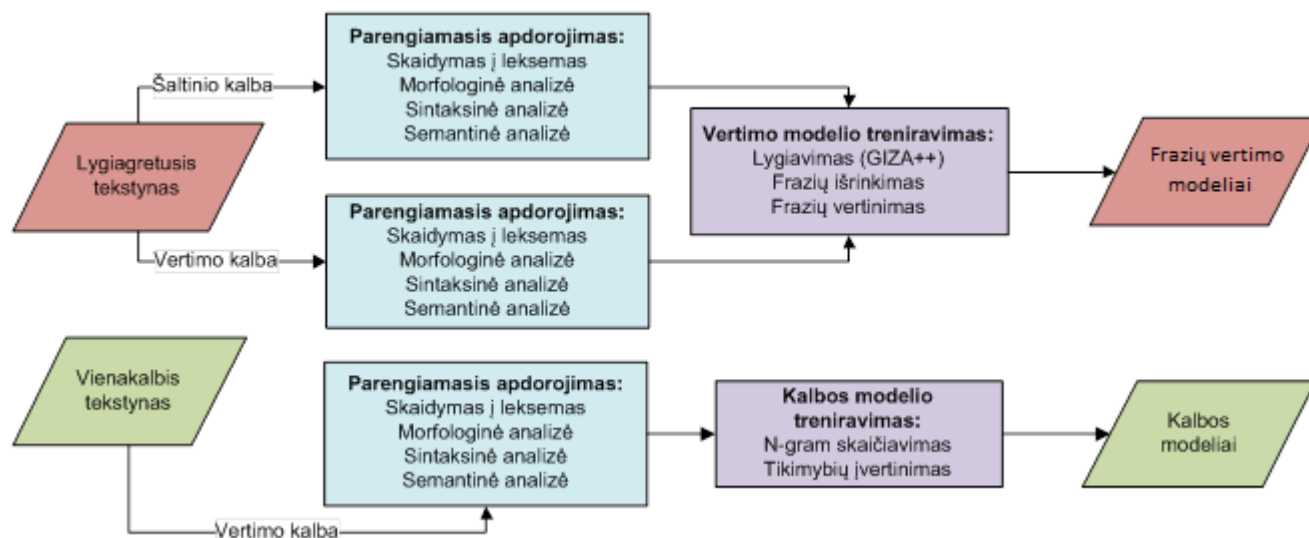
```
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
test.en test.lt
1 Language, literature, performing arts, visual arts, architecture,
crafts, the cinema and broadcasting are all part of Europe's
cultural diversity.
2 The aim of the European Union is double: to preserve and support
this diversity and to help make it accessible to others.
3 Cultural industries in the EU - cinema and audiovisual,
publishing, music and crafts - are important sources of revenue
and of jobs, employing about seven million people.
4 The Union has an economic responsibility towards this sector and
it aims to ensure the right conditions in which European
industries can compete internationally.
5 So the EU runs support programmes for certain cultural industries,
encouraging them to grasp opportunities offered by the single
market and digital technologies.
6 It is also striving to create a dynamic environment for these
industries by cutting red tape, providing easier access to
funding, helping with research projects and encouraging more
cooperation with partners inside and outside the Union.
7 The Union also includes a cultural dimension in many of its other
policy areas such as education (including language-learning),
scientific research, support for new technologies and the
information society, and social and regional development.
8 For example, the European Social Fund supports Italy's Arturo
Toscanini Foundation which, since the middle of the 1990s, has run
training courses for out-of-work musicians.
9 The specific objectives of the current Culture programme are to
promote the awareness and the preservation of cultural items of
European significance; to promote the transnational mobility of
those working in the cultural sector; to encourage the
1 Kalba, literatūra, atlikimo menas, vaizduojamasis menas,
architektūra, amatai, kinas ir transliavimas - visa tai yra
Europos kultūrinės įvairovės dalys.
2 Europos Sąjungos vaidmuo yra dvejopas: puoselėti ir skatinti
įvairovę ir leisti kitiems su ja susipažinti.
3 Kultūrinės ES industrijos - kinas ir audiovizualiniai menai,
leidyba, muzika ir amatai - svarbūs darbo ir pajamų šaltiniai,
suteikiantys darbą maždaug septyniems milijonams žmonių.
4 Sąjunga yra ekonomiškai išpareigojusi palaikyti šį sektorių ir
siekia užtikrinti tinkamas sąlygas, kuriomis Europos pramonės
šakos galėtų konkuruoti tarptautiniu mastu.
5 Todėl ES vykdo konkrečių kultūros sričių palaikymo programas,
skatinančias jas suvokti vienos rinkos ir skaitmeninių
technologijų siūlomas galimybes.
6 Ji taip pat stengiasi sukurti dinamišką aplinką šioms pramonės
šakoms: mažinamos biurokratinės procedūros, siūloma lengvesnių
būdų finansavimui gauti, remiami moksliniai projektai ir
skatinamas bendradarbiavimas su partneriais Sąjungoje ir už jos
ribų.
7 Į daugelį kitų politikos sričių, pavyzdžiui, švietimo (įskaitant
kalbų mokymąsi), mokslinių tyrimų, naujų technologijų ir
informacinės visuomenės, socialinės ir regioninės plėtros,
Sąjunga įtraukia ir kultūros aspektą.
8 Pavyzdžiui, Europos socialinis fondas palaiko Italijos Arturo
Toscanini fondą, kuris nuo dešimtojo dešimtmčio vidurio
organizuoja mokymo kursus darbo neturintiems muzikantams.
9 Konkretūs programos „Kultūra“ tikslai yra remti europinės svarbos
kultūros elementų supratimą ir apsaugą, remti kultūros darbuotojų
tarptautinį mobilumą, gerinti tarpvalstybinę kultūrinių bei
```

Normal text file length: 70916 lines: 512 Ln: 1 Col: 4 Sel: 0|0 UNIX UTF-8 INS



Ką MV išmoksta iš duomenų?

- Atpažinti kurį sakinį versti į kurį: lygiavimas sakinio lygmenyje
- Atpažinti kurį žodį išversti į kurį: žodžių lygiavimas + vertimo tikimybės
- Kaip turi atrodyti išverstas sakinytis (sakinio struktūra): kalbos modelis





- Statistinio MV pagrindą sudaro – DUOMENYS
- Statistinis MV iš duomenų išmoksta VERSTI
- Duomenys
 - Lygiagretūs tekstynai (vertimai)
 - Vienakalbiai tekstynai (tos kalbos tekstai į kurią verčiama)
 - Žodynai, terminologijos, ontologijos, įvardintų esybių duomenų bazės
- Statistinis MV geriausiai verčia tos srities tekstus, kuriais jis apmokytas.



- 2005 - 2007 m. Vytauto Didžiojo universitetas vykdė Europos Sąjungos Struktūrinių fondų finansuojamą projektą „Internetinė informacijos vertimo priemonė“ . Rezultatas – vieša internetinė vertimo iš anglų į lietuvių k. paslauga. Vertimo variklį pateikė rusų kompanija PROMT. Nėra aišku, kiek laiko dar bus teikiama ši paslauga (<http://vertimas.vdu.lt/twsas/>).
- Nuo 2008 m. rugsėjo 25 d. Google Translate palaiko ir lietuvių kalbą. Tačiau to neleidžiama naudoti komerciniams tikslams!
- 2015 m. gegužės 15 d. Vilniaus universitetas pabaigė projektą „Anglų-lietuvių-anglų ir prancūzų-lietuvių-prancūzų kalbų mašininio vertimo, paremto statistiniais metodais, sistemos sukūrimas“ ir pristatė visuomenei MV paslaugų svetainę www.versti.eu

Kalbos:

- lietuvių – anglų
- anglų – lietuvių
- lietuvių – prancūzų
- prancūzų – lietuvių

Sritys:

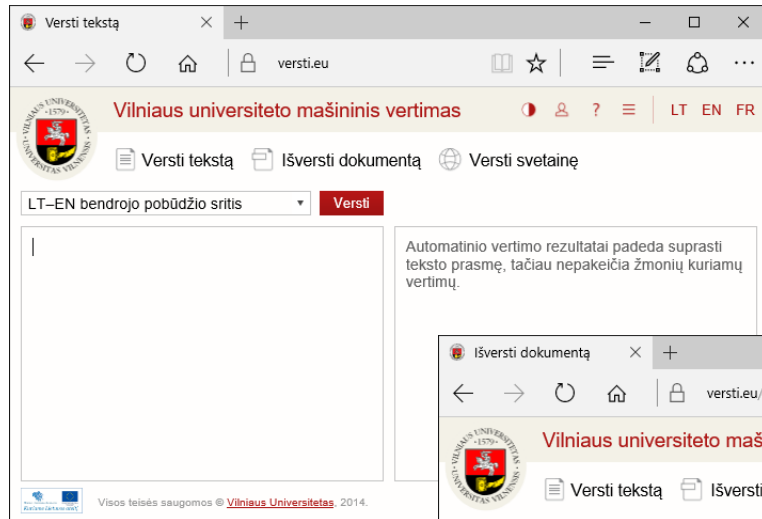
- lietuvių – anglų IT
- lietuvių – anglų teisinė
- anglų – lietuvių IT
- anglų – lietuvių teisinė
- lietuvių – prancūzų teisinė
- prancūzų – lietuvių teisinė

Vertimo kokybė:

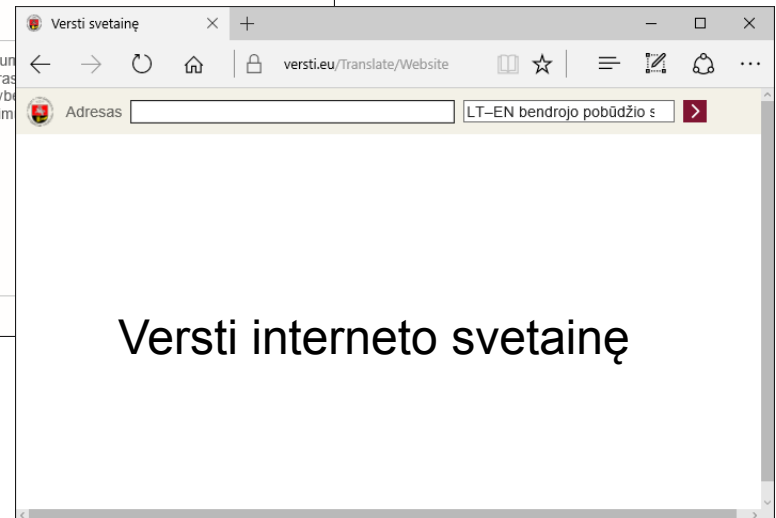
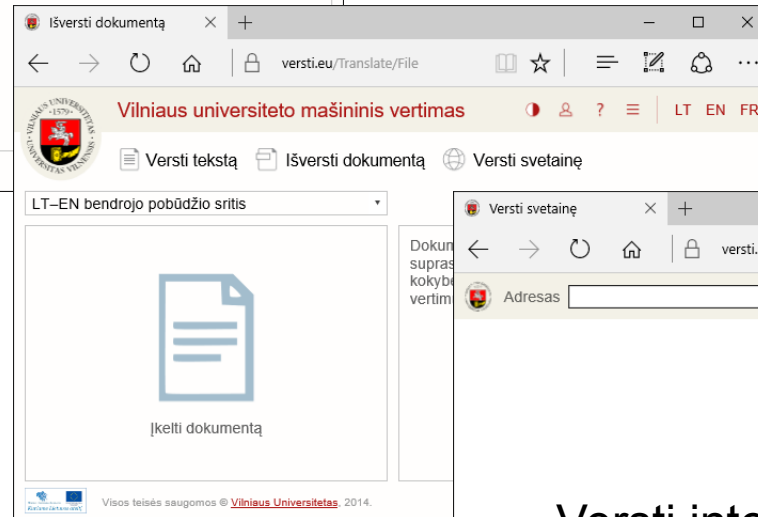
Kalbų pora	Domenas	BLEU „VU“	BLEU, „Google“	BLEU, „Microsoft“
Anglų-lietuvių	Bendrasis	37.78	19.02	16.82
	Teisės aktų	56.04	43.98	45.11
	IT	60.42	22.68	23.89
Lietuvių-anglų	Bendrasis	43.94	32.56	33.36
	Teisės aktų	67.85	40.28	56.19
	IT	76.57	37.03	36.34
Prancūzų-lietuvių	Bendrasis	37.57	17.45	14.33
	Teisės aktų	51.67	32.19	31.22
Lietuvių-prancūzų	Bendrasis	37.19	17.95	18.03
	Teisės aktų	57.90	28.87	39.46

Lingvistiniai resursai:

Tekstynai	Iš viso
Bendrinė sritis	
Vienkalbiai	
Lietuvių	856,7 M žodžių
Anglų	1 928,9 M žodžių
Prancūzų	1 163,3 M žodžių
Lygiagretieji	
Anglų-lietuvių	8,1 M sakinių
Prancūzų-lietuvių	7,0 M sakinių
Teisinis	
Vienkalbiai	
Lietuvių	207,8 M žodžių
Anglų	582,3 M žodžių
Prancūzų	586,3 M žodžių
Lygiagretieji	
Anglų-lietuvių	6,3 M sakinių
Prancūzų-lietuvių	5,7 M sakinių
IT	
Vienkalbiai	
Lietuvių	254,1 M žodžių
Anglų	262,1 M žodžių
Lygiagretieji	
Anglų-lietuvių	5,0 M sakinių



Versti
tekstą



Versti dokumentą, išlaikant
formatavimą (txt, docx,
doc, xlsx, pptx, odf, TMX,
XLIFF)

Galimybė verčiant naudoti
jau sukurtą terminiją ar
prisidėti savo

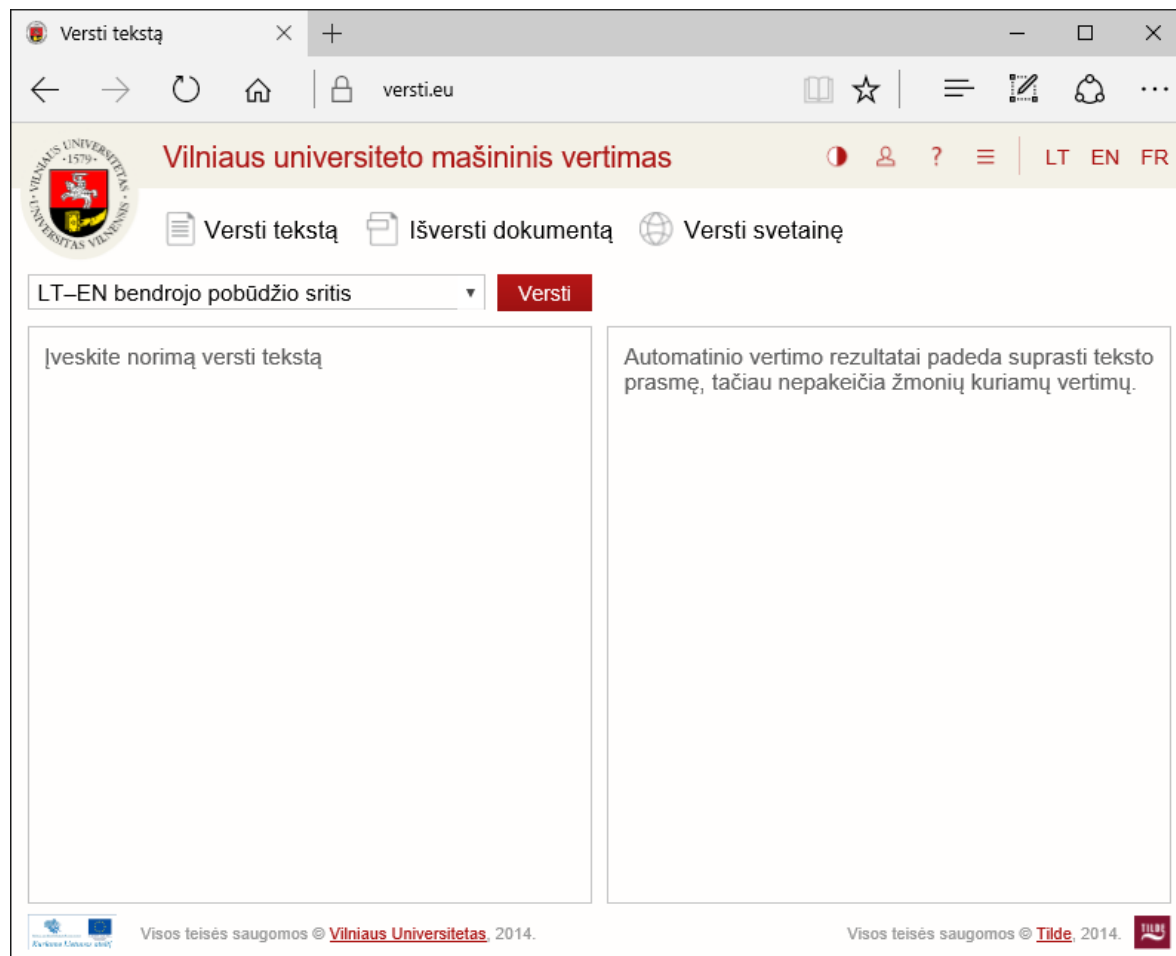
Versti interneto svetainę

Kaip versti?

Tiesiai

www.versti.eu puslapyje

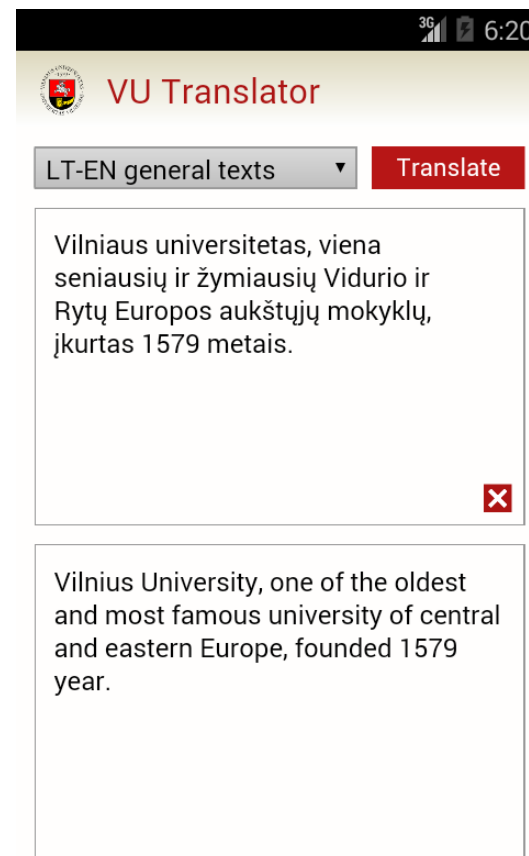
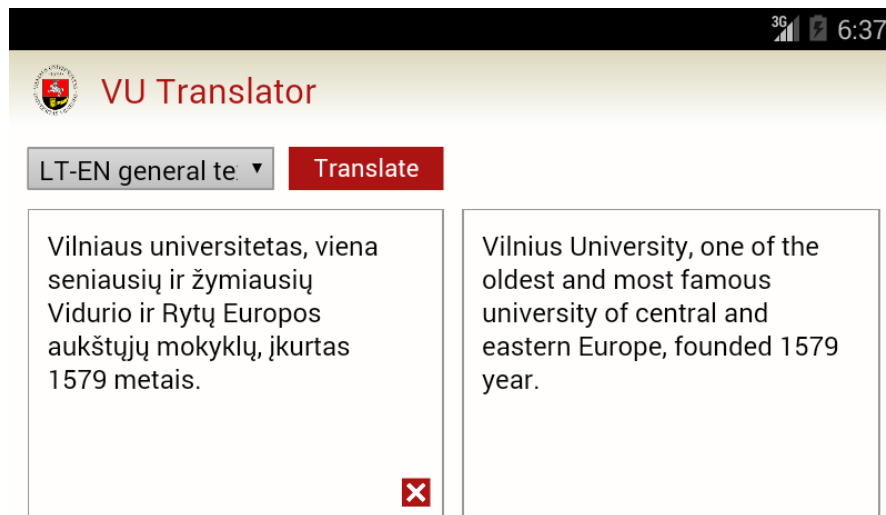
arba www.raštija.lt



The screenshot shows a web browser window with the URL versti.eu. The page title is "Vilniaus universiteto mašininis vertimas". The interface includes a navigation menu with "Versti tekstą", "Išversti dokumentą", and "Versti svetainę". A language selection dropdown is set to "LT-EN bendrojo pobūdžio sritis", and a red "Versti" button is visible. The main content area is split into two columns: the left column contains the placeholder text "[veskite norimą versti tekstą]", and the right column contains the text "Automatinio vertimo rezultatai padeda suprasti teksto prasmę, tačiau nepakeičia žmonių kuriamų vertimų." The footer contains logos for the European Union and Vilniaus Universitetas, along with copyright information: "Visos teisės saugomos © Vilniaus Universitetas, 2014." and "Visos teisės saugomos © Tilde, 2014."



Kaip versti?
Mobiliose programėlėse:





Kaip versti?

- Įsidięgti naršyklės vertimo įskiepi („Chrome“, „Firefox“, „Internet Explorer“, „Safari“ ir „Opera“)
- Naudodami VU mašininio vertimo atvirojo kodo API, teikite vertimų užklausas SMT sistemoms ir integruokite VU mašininį vertimą į savo sprendimą.



- Naujos kalbų poros: vokiečių-lietuvių, lenkų-lietuvių, rusų-lietuvių.
- Geresnė vertimo kokybė.
- Integravimas su garso technologijomis.