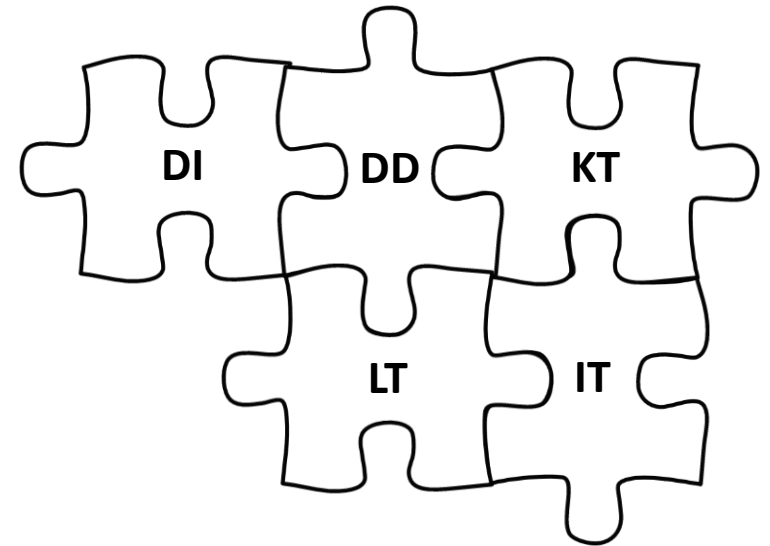


Kalbinės technologijos Lietuvoje

Andrius Utkas, VLKK, VDU

2021-12-01



Lietuvių kalbos plėtros skaitmeninėje terpėje ir kalbos pažangos 2021–2027 Gairės

- 2020 m. spalio 13 d. Seimo nutarimu buvo patvirtintos „Lietuvių kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021–2027 metų gairės“.
- Gairių autoriai – ekspertų komanda iš mokslo ir verslo institucijų: Baltijos pažangių technologijų instituto, KTU, LKI, UAB *Tildė*, VDU, VLKK, VU. Buvo konsultuojamasi su EIMIN ir KM.



LIETUVOS RESPUBLIKOS SEIMAS

NUTARIMAS DĖL LIETUVIŲ KALBOS PLĖTROS SKAITMENINĖJE TERPĖJE IR KALBOS TECHNOLOGIJŲ PAŽANGOS 2021–2027 METŲ GAIRIŲ PATVIRTINIMO

2020 m. spalio 13 d. Nr. XIII-3324
Vilnius

Lietuvos Respublikos Seimas, atsižvelgdamas į Lietuvos Respublikos Konstitucijos 14 straipsnį ir Lietuvos Respublikos valstybinės kalbos įstatymą, n u t a r i a:

1 straipsnis.

Patvirtinti Lietuvos kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021–2027 metų gairės (pridedama).

2 straipsnis.

1. Pasiūlyti Lietuvos Respublikos Vyriausybei atsižvelgti į Lietuvos kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021–2027 metų gaires rengiant Lietuvos skaitmeninio plėtros 2021–2030 metų programą ir Lietuvos Respublikos atitinkamų metų valstybės biudžeto ir savivaldybių biudžetų finansinių rodiklių patvirtinimo įstatymo projektus.

2. Pavesti Valstybinei lietuvių kalbos komisijai atlikti Lietuvos kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021–2027 metų gairių įgyvendinimo stebėseną.

Seimo Pirmininkas

Viktoras Pranckietis

PATVIRTINTA
Lietuvos Respublikos Seimo
2020 m. spalio 13 d.
nutarimu Nr. XIII-3324

LIETUVIŲ KALBOS PLĖTROS SKAITMENINĖJE TERPĖJE IR KALBOS TECHNOLOGIJŲ PAŽANGOS 2021–2027 METŲ GAIRĖS

I SKYRIUS BENDROSIOS NUOSTATOS

1. Pastaraisiais metais žinių visuomenė pereina į kokybiškai naują etapą, kurį žymi sparti pažangių informacinių technologijų plėtra, pirmiausia didžiųjų duomenų kaupimas ir apdorojimas bei dirbtiniu intelektu grįstų technologijų kūrimas. IT vis plačiau diegiamos visose pagrindinėse visuomenės veiklos srityse, tokiose kaip valstybės administravimas ir teismų sistema, švietimas,

Lietuvių kalbos technologijų sritys

Technologijų ir kalbos duomenų infrastruktūros

Kalbos duomenys

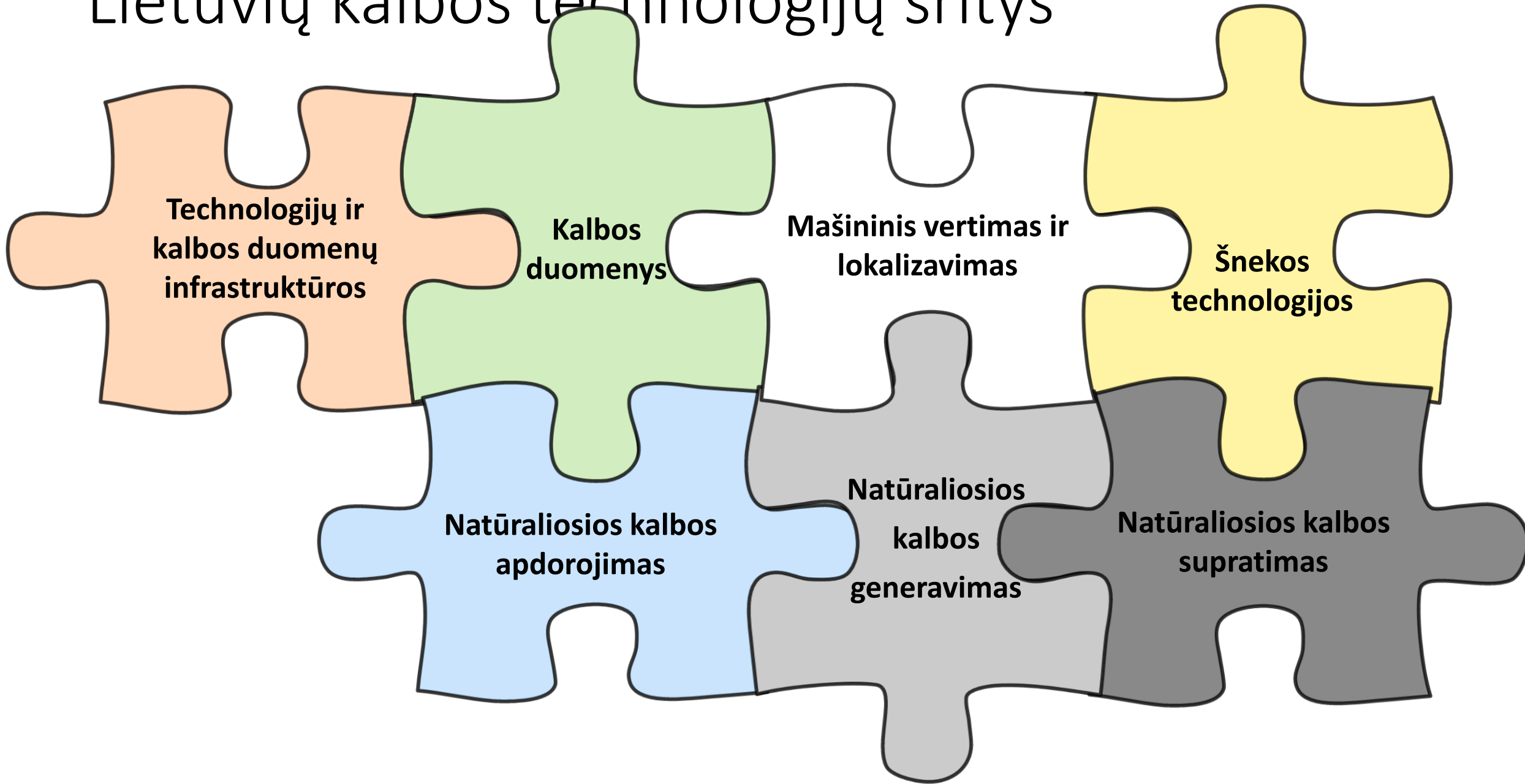
Mašininis vertimas ir lokalizavimas

Šnekos technologijos

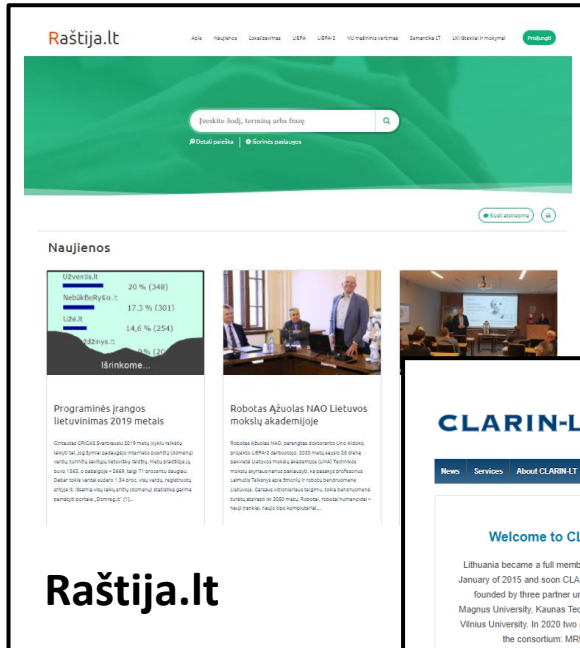
Natūraliosios kalbos apdorojimas

Natūraliosios kalbos generavimas

Natūraliosios kalbos supratimas



Nacionalinės technologijų ir kalbos duomenų infrastruktūros (1)



Raštija.lt

Apie Naujienos Lietuviškosis UŽTA UŽTA-1 Vokėtinis veiksmai Semantika LT-1 Būkime ir mokymai Tiesiogiai

Įrašyti žodį, terminą arba frazą

#Gausi paroda | @Dorėta Paulauskaitė

Facebook Twitter

Naujienos

Užvenkite 20% (348)
Neakademiško 17,3% (301)
UŽTA.R 14,6% (254)
Kitos 1,1% (20)

Išrinkite...

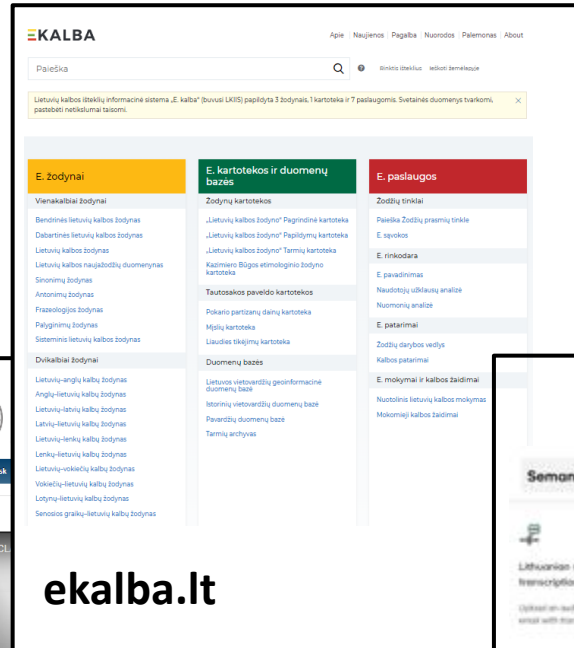
Programinės įrangos lietuvinimas 2019 metais

Ortautas OFCLA ir partneriai 2019 metais įvykdė milijonų eurų vertės programines įrangas lietuvininti. Šiandien Lietuvoje yra 1600 lietuvinintų programinių įrangų. Šiandien Lietuvoje yra 1600 lietuvinintų programinių įrangų. Šiandien Lietuvoje yra 1600 lietuvinintų programinių įrangų.

Robotas Ažulais MAO Lietuvos mokyklų akademijoje

Robotas Ažulais MAO, parengtas partnerėms UAB Ažulais, pradėjo veiklą Lietuvos mokyklų akademijoje. Robotas Ažulais MAO, parengtas partnerėms UAB Ažulais, pradėjo veiklą Lietuvos mokyklų akademijoje.

Raštija.lt



ekalba.lt

Apie Naujienos Pagalba Nuorodos Pateiktos About

Paieška

Lietuvių kalbos (išskirti) informacinė sistema „E. kalba“ (buvusi LKIS) papildyta 3 žodynais, 1 kartoteka ir 7 paslaugomis. Svetainės duomenys tvarkomi, pastebėti netikslumai rašomi.

E. žodynai	E. kartotekos ir duomenų bazės	E. paslaugos
Vienakabiai žodynai <ul style="list-style-type: none">Bendrinės lietuvių kalbos žodynasDabartinės lietuvių kalbos žodynasLietuvių kalbos žodynasLietuvių kalbos naujų žodžių duomenynasSporinčių žodynasAsponimų žodynasFrazologijos žodynasPalyginimų žodynasSisteminių lietuvių kalbos žodynas	Žodynių kartotekos <ul style="list-style-type: none">„Lietuvių kalbos žodyno“ Pagrindinė kartoteka„Lietuvių kalbos žodyno“ Papildomi kartoteka„Lietuvių kalbos žodyno“ Tarmių kartotekaAkademių Būgys etimologinis žodyno kartoteka Tautosakos paveldo kartotekos <ul style="list-style-type: none">Dobruojo partonų dainų kartotekaMėgų kartotekaLiaudies tikėjimų kartoteka Duomenų bazės <ul style="list-style-type: none">Lietuvių-anglų kalbų žodynasAnglų-lietuvių kalbų žodynasLietuvių-lietuvių kalbų žodynasLietuvių-lietuvių kalbų žodynasLietuvių-lietuvių kalbų žodynasLietuvių-ukrainų kalbų žodynasVokiečių-lietuvių kalbų žodynasLietuvių-lietuvių kalbų žodynasSenosios graikų-lietuvių kalbų žodynas	Žodžių trinkiai <ul style="list-style-type: none">Paieška Žodžių prasmių trinkleE. sąvokosE. rinkodaraE. pavadinimasNaudotųjų vikiščių analizėNuomonių analizėE. patarimaiŽodžių darybos vedlysKalbos patarimaiE. mokymai ir kalbos žaidimaiNuostolius lietuvių kalbos mokymasMokomųjų kalbos žaidimai

ekalba.lt



CLARIN-LT

News Services About CLARIN-LT Repository Partners Contact Help Desk

Welcome to CLARIN-LT!

Lithuania became a full member of CLARIN ERIC in January of 2015 and soon CLARIN-LT consortium was founded by three partner universities: Vytautas Magnus University, Kaunas Technology University and Vilnius University. In 2020 two more institutions joined the consortium: MRU and SPTI.

CLARIN NEWSFLASH	CLARIN TWITTER	CLARIN VIDEOLECTURES
-------------------------	-----------------------	-----------------------------

News Services About CLARIN-LT Repository Partners Contact Help Desk Events

clarin-lt.lt



Semantika

Lithuanian speech-to-text transcription service

Automatic document summary service for Lithuanian documents

Desired / appropriate university study field selection service

Analysis and correction of Lithuanian text

Analysis and search in Lithuanian online media corpus

semantika.lt



Technologijų ir kalbos duomenų infrastruktūros

Vytauto Didžiojo Universitetas
Kompiuterinės lingvistikos centras
CLARIN-LT

NAUJIENOS TEKSTYNAI IRANKIAI DUOMENYNAI BIBLIOGRAFIJA PROJEKTAI APIE

DABARTINIS LIETUVIŲ KALBOS TEKSTYNAS
MORFOLOGIŠKAI ANOTUOTAS DLKT
LYGIAGRETUSIS TEKSTYNAS
LILA

klc.vdu.lt

TILDE

Pasirinkite kalbą: Lietuvių

Vertyklė Vertimo ir lokalizavimo paslaugos Tildės Biuras Kalbines technologijos Šnekos technologijos Apie Tildę Pirkti

Tilde vertyklė Mašininis vertimas Kalbos tikrinimo priemonės Semantinės sistemos Projektai Terminologija Šriftai

„Tilde“ vertyklė mobiliuosiuose įrenginiuose – visada po ranka

Naudodite geriausias pasaulijje statistine automatinę vertyklę mobiliuosiuose įrenginiuose, skirta versti iš lietuvių ar latvių kalbų į anglų kalbą ir atvirkščiai. „Tilde“ vertyklė padės išversti tiek atskirus žodžius, tiek rišli tekstą.

„Tilde“ vertyklę taip pat galite naudoti internete arba įsigyje mūsų produkty „Tildės Biuras“.

Android Windows Phone App Store

Mašininis vertimas: Šiuo metu sukurtos ir internete prieinamos mašininio vertimo sistemos grindžiamos dviem vertimo metodais – taisyklių arba statistiniu.

Kalbos tikrinimo priemonės: Noredama palengvinti lietuvių kalbos vartojimą elektroninėje erdvėje, „Tilde IT“ 2001 m. vartotojams pasiūlė rašybos tikrintuvą.

Semantinės sistemos: Semantinės sistemos šiuolaikinėje informacijoje užima vieni iš svarbiausių vietų.

www.tilde.lt/kalbines-technologijos

- PASTOVU:** mwe.lt
- VLE:** www.vle.lt
- CorALIT:** coralit.lt/node/8
- SKT:** sakytinistekstynas.vdu.lt

Kalbu

KALBU

Mokomasis tekstynas
Mokinių tekstynas
Morfologiškai anotuotas tekstynas
Mokomasis lietuvių kalbos vartosenos leksikonas
Tartis
Kirčiuoklis

kalbu.vdu.lt



Duomenų infrastruktūrų plėtros kryptys

- pildyti infrastruktūras reikalingais skaitmeniniais ištekliais;
- atnaujinti infrastruktūrų techninę įrangą bei užtikrinti jos palaikymą;
- integruoti infrastruktūras į stambesnes nacionalines, Europos ir tarptautines kalbų išteklių sistemas;
- užtikrinti infrastruktūrose saugomų technologijų ir duomenų atvirumą.
- skatinti technologijų ir duomenų infrastruktūrų kūrėjų bendradarbiavimą ir specializavimąsi tam tikroje srityje.



Duomenys

Garsynai

Liepa-2 garsynas
VDU ir KTU garsynai

Ontologijos

LitWordNet,
SNOMED,
E.sąvokos

Skaitmeniniai žodynai ir vertimo atmintys

E.Kalba žodynai
Raštijos.lt
vertimo atmintys
VLE

Tekstynai

DLKT
ALKSNIS
LITIS
MATAS
BIT
CorALIT
SKT
Tarmių tekstynas
Senosios lietuvių k.
Senųjų raštų

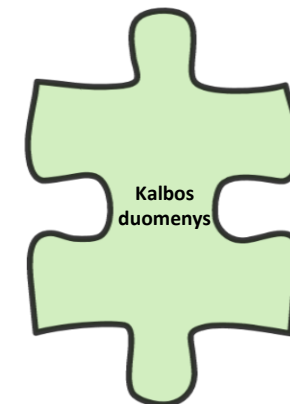
ELRC tekstynai
Sketch Engine tekstynai
CLARIN-PL tekstynai

Įterptinių vektorių kalbos modeliai

VDU ir *Tokenmill*
bandymai

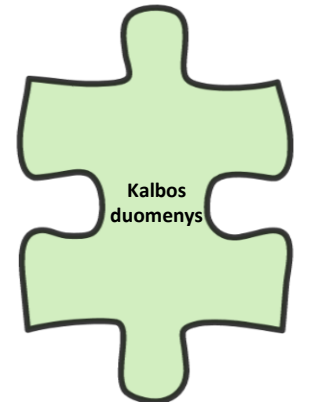
Geoinformaciniai kalbos duomenys

E.Kalba geoinformacinė
duomenų bazė



Duomenų plėtros kryptys

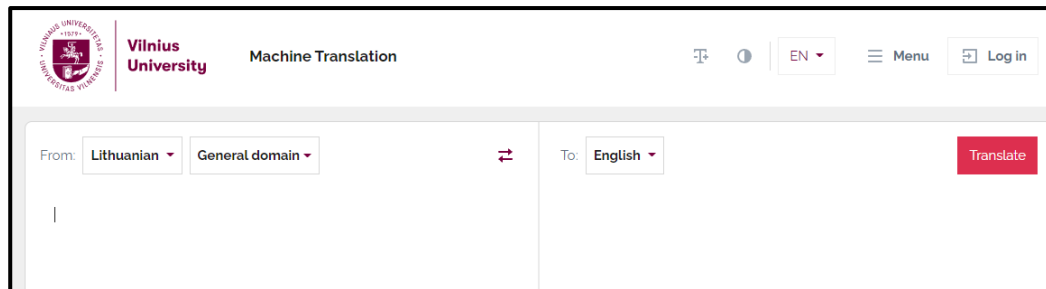
- duomenys yra pagrindinis DI technologijų ir sprendinių šaltinis, todėl reikalingi tiek nestructūruoti, tiek sudėtingi struktūruoti duomenys.
- duomenys turi būti nuolat gausinami ir atnaujinami, kad atspindėtų kuo įvairesnes kalbos vartojimo sritis bei kalbos pokyčius;
- svarbu užtikrinti ne tik didelę apimtį, bet ir kokybę, privačių duomenų apsaugą bei prieinamumą



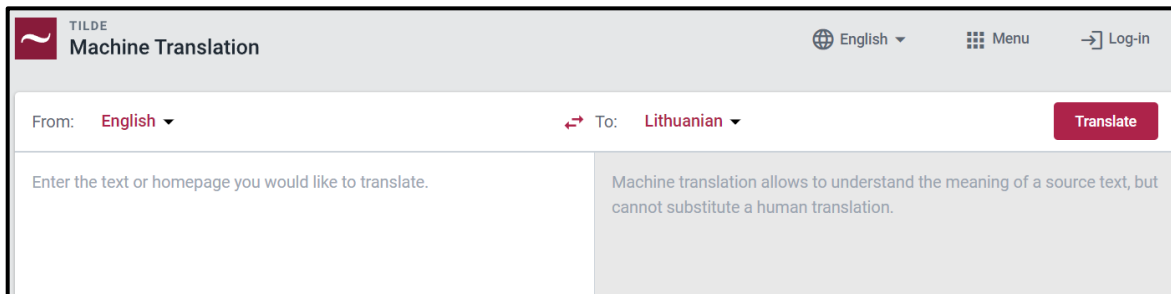
Mašininis vertimas ir lokalizavimas



VDU (2008) : EN-LT



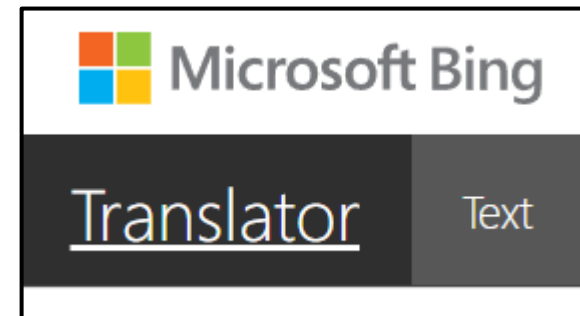
VU (2020): LT ↔ DE, EN, FR, PL, RU



Tildės vertyklė (2021): LT ↔ DE, EN, RU, PL
EN ↔ EE, LV, FI, DE, BG, NO, PL, CR, RO, ES, LT



Google: LT ↔ 108 kalbos



MS Bing: LT ↔ 86 kalbos

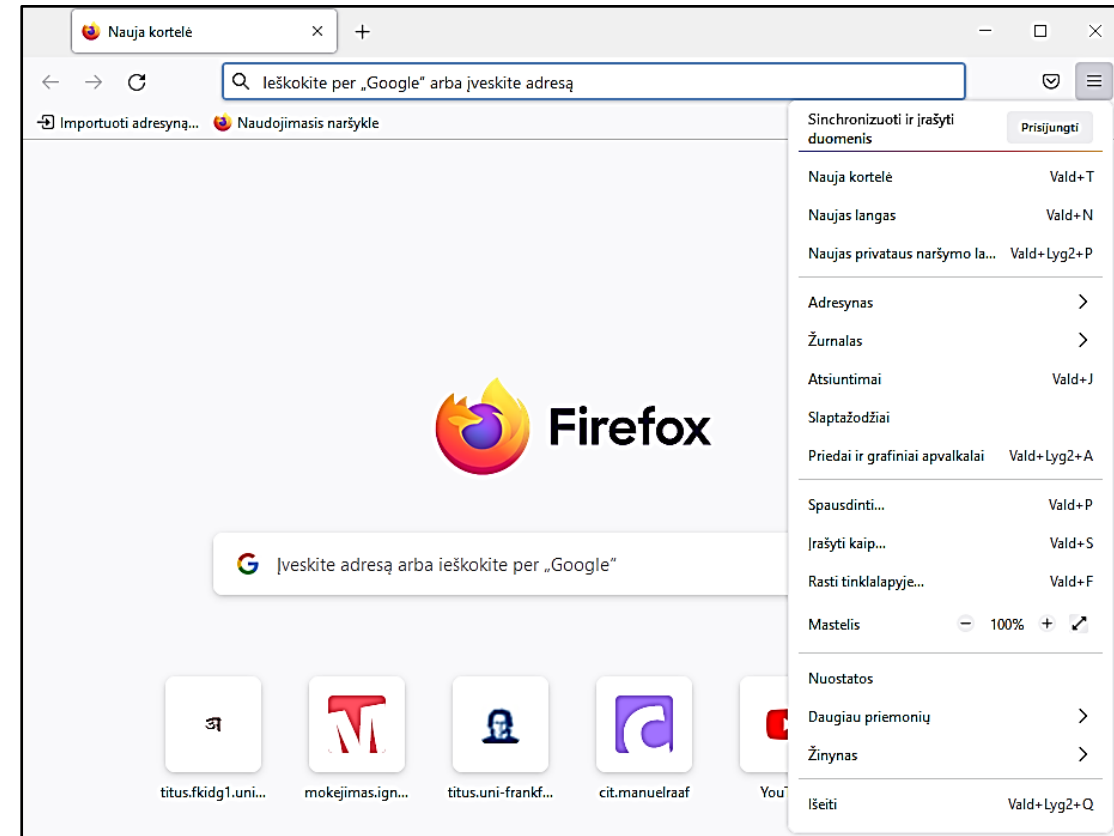


eTranslation: LT ↔ 24 ES kalbos



Mašininio vertimo ir lokalizavimo plėtros kryptys

- Tolesniam mašininio vertimo sistemų tobulinimui trūksta atvirų daugiakalbių duomenų bazių (dvikalbių lygiagrečiųjų tekstynų), taip pat specializuotų tekstų duomenų.
- Reikia toliau lokalizuoti naujausią programinę įrangą.
- Taip pat svarbu pridėti naujas aktualias kalbų poras.



Šnekos technologijos

Šnekos atpažinimas

LIEPA ir LIEPA-2 valdymo balsu sistemos

Semantika-2 automatinė transkripcija

TILDĖS šnekos atpažinimas

Šnekos sintezavimas

LIEPA ir LIEPA-2 Tartuvas ir sintezatoriai akliesiems

VDU sintezatoriai

Tildė balsų sintezė



Transkripcijos paslauga



Šnekos technologijų plėtros kryptys

- Kadangi šiuolaikinės šnekos atpažinimo ir sintezės pagrindą sudaro anotuoti garsynai, tai svarbiausia ir pagrindinė šnekos technologijų plėtros kryptis yra naujų atvirų įvairių sričių, dialektų, amžiaus grupių, foninės aplinkos garsynų kūrimas.



Nuotrauka iš Max Pixel (maxpixel.net)



Natūraliosios kalbos apdorojimas

Baziniai NKA sprendiniai

Sukurti SEMANTIKA-2 projekte (prieinami SEMANTIKA-2 portale)

Segmentatorius

Lemuoklis

Morfologijos analizatorius

Kalbos dalių atpažintuvas

Rašybos klaidų tikrintuvas

Teksto normalizatorius

PASTOVU projekte

Pastoviųjų junginių atpažintuvas

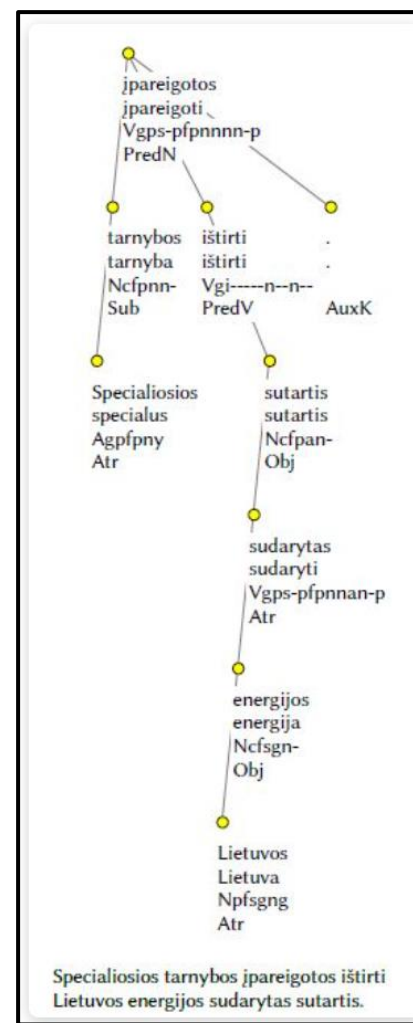
Ir kt.

SpaCy Python bibliotekos sprendimai lietuvių kalbai sukurti „Tokenmill“ kompanijos.



Natūraliosios kalbos apdorojimo plėtra

- Tobulėjant giliojo mokymosi algoritmams tolesnei natūraliosios kalbos apdorojimo plėtrai reikalingi gausūs, patikimi ir įvairių temų duomenys, parengti mašininiam mokymuisi.



Natūraliosios kalbos supratimas

SEMANTIKA ir SEMANTIKA-2 sukurti sprendimai:

- Sentimentų (nuomonių) analizatorius

- Neapykantos / įžeidžiančios kalbos atpažintuvas

- Santraukų sudarymas

- Įvardintųjų esybių atpažintuvas

Natūraliosios kalbos supratimo plėtros kryptys

Tobulėjant giliojo mokymosi algoritmams, natūraliosios kalbos supratimo plėtrai reikalingi gausūs, patikimi ir įvairių temų duomenys, parengti mašininiam mokymuisi.

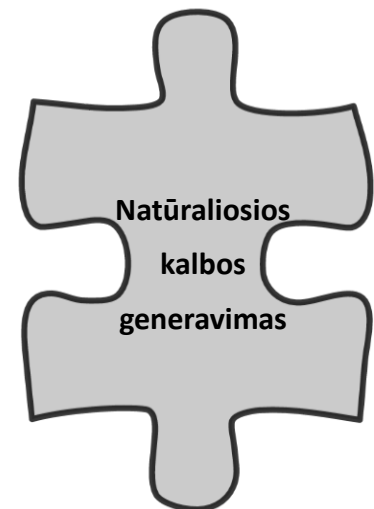


Natūraliosios kalbos generavimas

- Reikia pripažinti, kad šioje kalbos technologijų srityje Lietuvoje žengiami tik pirmieji žingsniai;
- Čia galima paminėti UAB „Tokenmill“, kuri sukūrė ne vieną natūraliosios kalbos generavimo sprendinį verslo reikmėms.
- GPT-3 (*Generative Pre-trained Transformer*) gali generuoti tekstus lietuvių kalba.

Natūraliosios kalbos generavimo plėtra

Svarbiausia natūralios kalbos generavimo sąlyga – didžiuliai atviri rišlios kalbos duomenys.



Kas yra daroma ir kokie tolimesni žingsniai?

2021-2030 Nacionalinis pažangos planas (NPP) *(1.7 Skatinti valstybės skaitmeninimą)*

2021 m. lapkričio 17 d. LR Vyriausybė patvirtino

2021–2030 m. Lietuvos Respublikos ekonomikos ir inovacijų ministerijos valstybės skaitmeninimo plėtros programą

Finansavimo lėšos

- Lietuvos 2021–2027 m. ES struktūrinių fondų investicijos
- Ekonomikos gaivinimo ir atsparumo didinimo priemonės (Naujos kartos Lietuva, NKL) (Economic recovery and resilience facility (RRF))
- Valstybės biudžeto lėšos

Lietuvių kalbos ištekliai, suskaitmeninti siekiant vystyti dirbtinį intelektą ir inovatyvias technologijas

1 kryptis: Kalbinių išteklių dirbtinio intelekto technologijų sprendimų poreikiams plėtra.

18 priemonių, 27,38 mln. €



Efektyviai šnekos, teksto ir mašininio vertimo DI sistemų plėtrai užtikrinti bus kuriamos šios priemonės: atviri garsynai, tekstynai, vektorizuoti modeliai bei daugiakalbiai tekstynai.

2 kryptis: Lietuvių k. išsaugojimą ir gyvybingumą palaikančių skaitmeninių išteklių plėtra

3 priemonės, 6,474 mln. €



1. Lietuvių kalbos paveldo ir šiuolaikinius išteklių modernizavimas bei pritaikymas mokslo, kultūros ir švietimo poreikiams.
2. geoinformacinių sistemų (GIS) pasitelkimas.

3 kryptis: Specializuotų žinių išteklių kūrimas

2 priemonės, 0,94. mln. €



Specializuotų išteklių (ontologijų) kūrimas medicinai ir kitoms sritims.

Ačiū už dėmesį!

