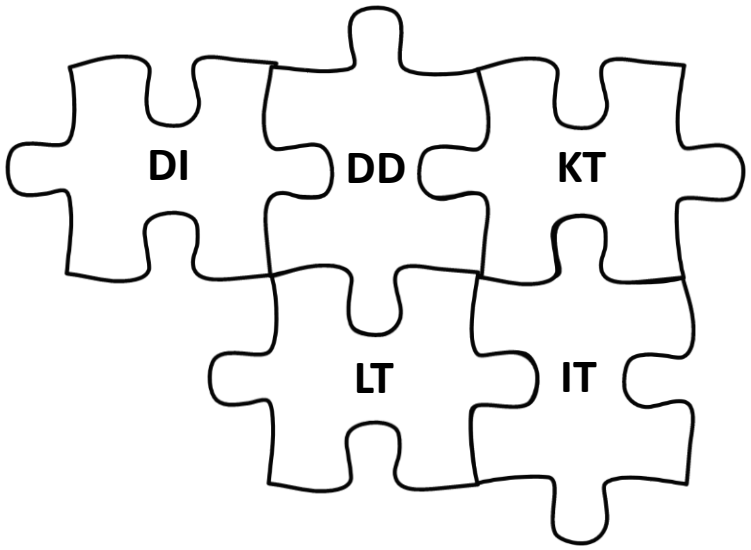
# Kalbinės technologijos Lietuvoje

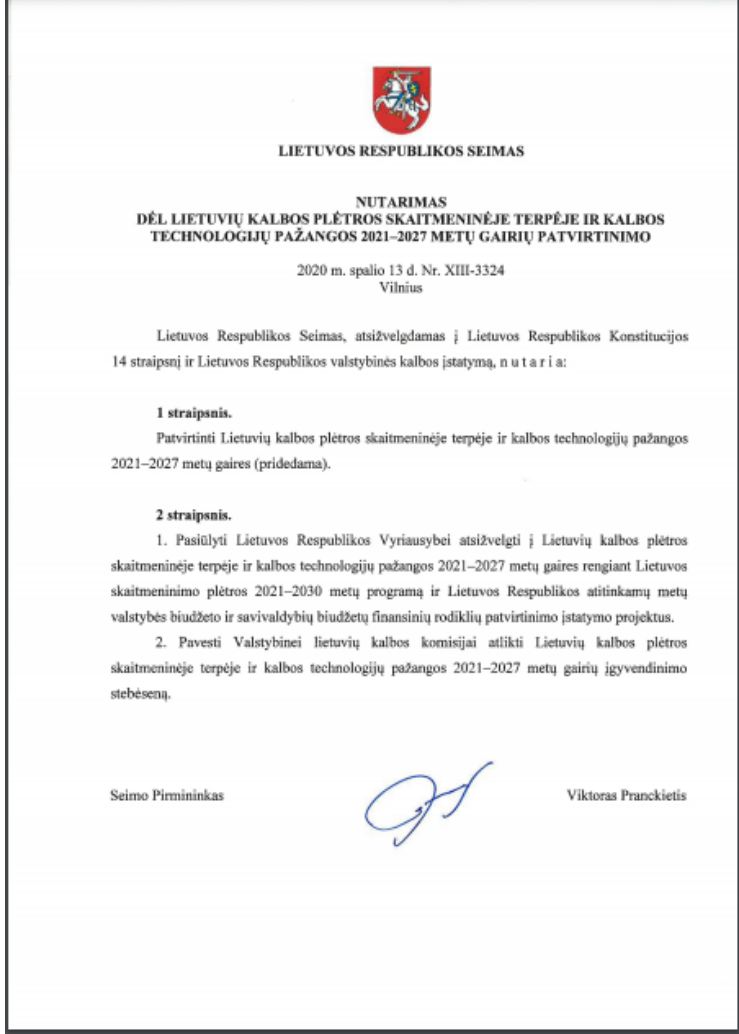
Andrius Utka, VLKK, VDU

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:

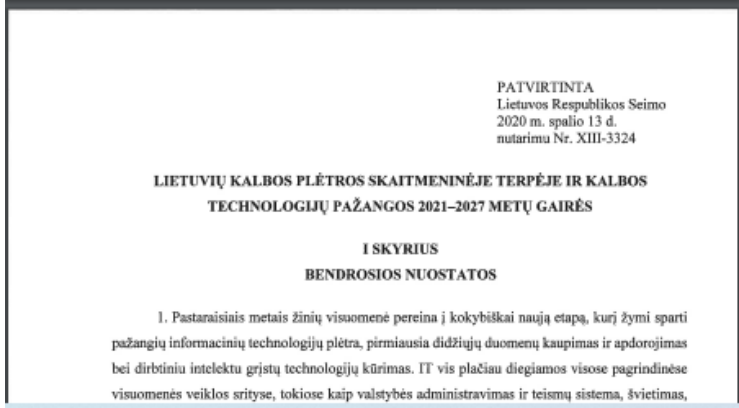
**1 straipsnis.**

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

**2 straipsnis.**

1. Pasiūlyti Lietuvos Respublikos Vyriausybei atsižvelgti į Lietuvių kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021–2027 metų gaires rengiant Lietuvos skaitmeninimo 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 Lietuvių kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021–2027 metų gairių įgyvendinimo stebėseną.

Seimo Pirmininkas Viktoras Pranckietis



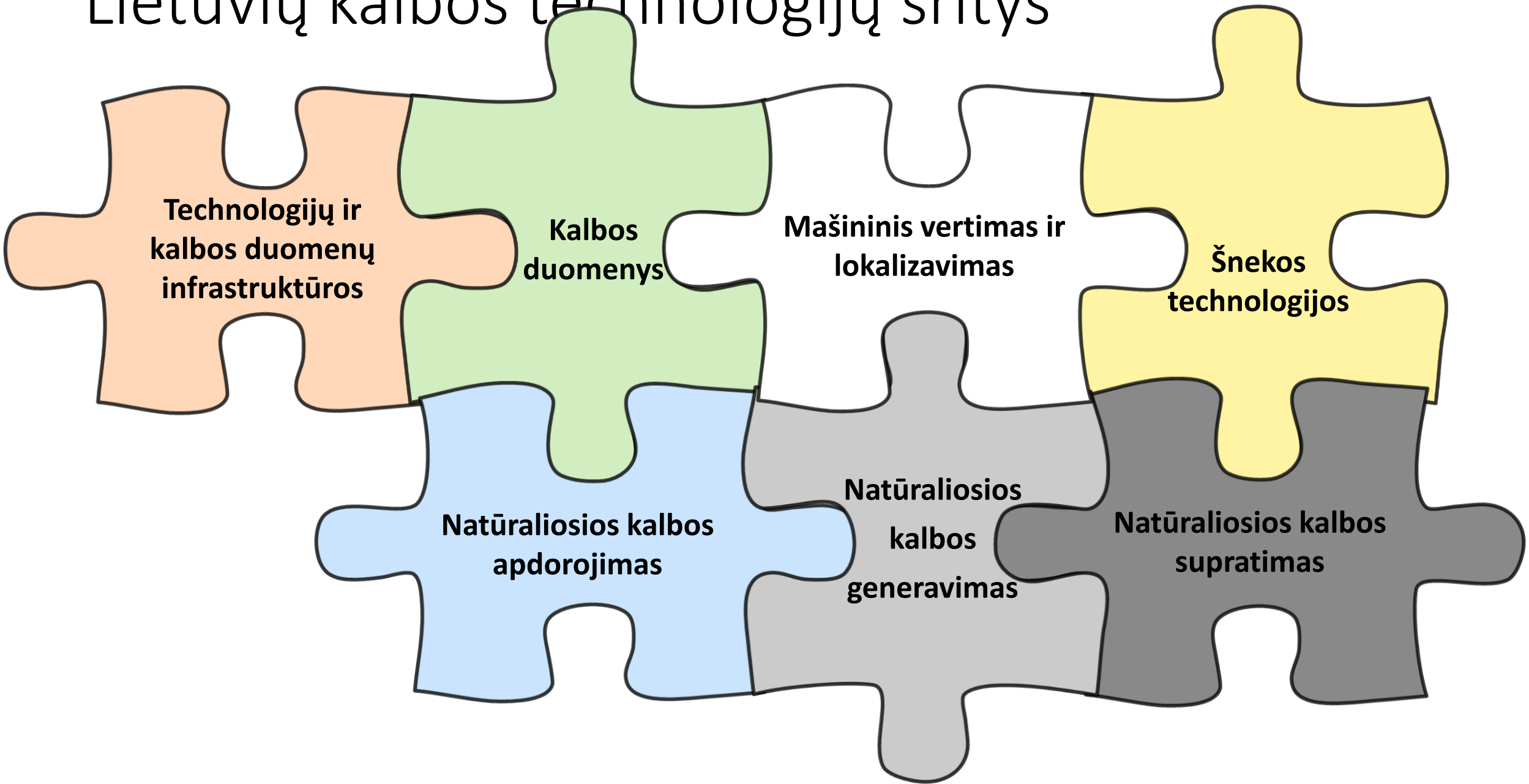
PATVIRTINTA
Lietuvos Respublikos Seimo
2020 m. spalio 13 d.
nutarimu Nr. XIII-3324

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

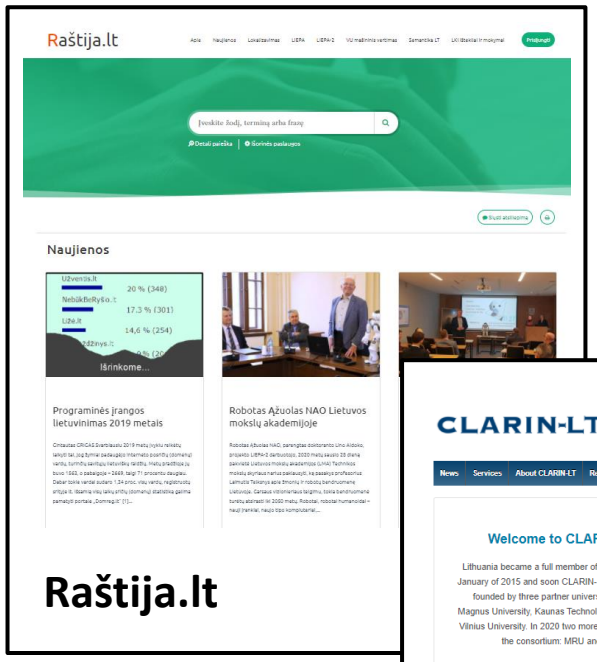
**I SKYRIUS**
**BENDROSIOS NUOSTATOS**

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

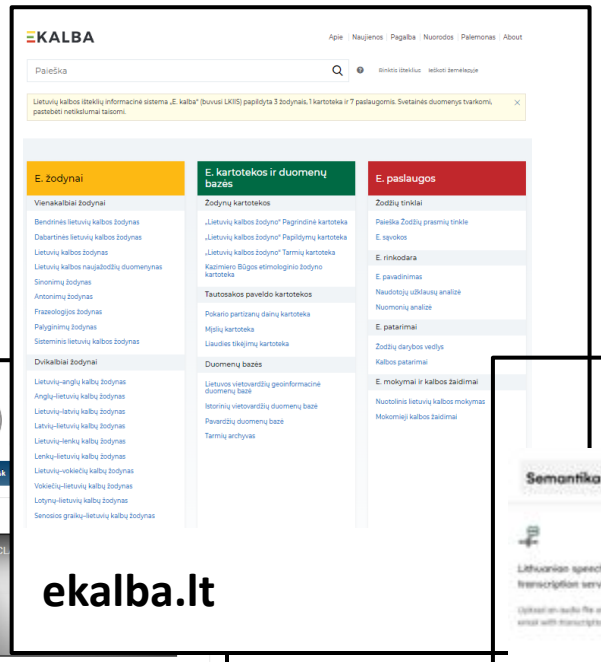
# Lietuvių kalbos technologijų sritys



# Nacionalinės technologijų ir kalbos duomenų infrastruktūros (1)




**Raštija.lt**




**ekalba.lt**

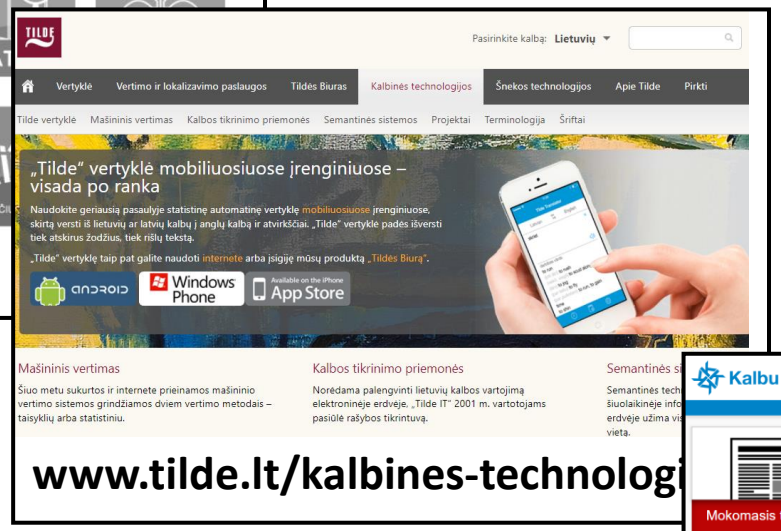

**clarin-lt.lt**


**semantika.lt**

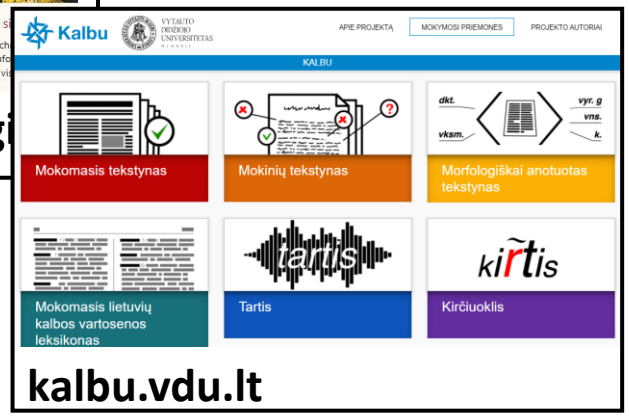
**Technologijų ir kalbos duomenų infrastruktūros**

# Technologijų ir kalbos duomenų infrastruktūros


klc.vdu.lt




www.tilde.lt/kalbines-technologi




kalbu.vdu.lt

**PASTOVU: mwe.lt**

**VLE: www.vle.lt**

**CorALIT: coralit.lt/node/8**

**SKT: sakytinistekstynas.vdu.lt**

**Technologijų ir kalbos duomenų infrastruktūros**

# 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.

EUROPEAN OPEN SCIENCE CLOUD

# Duomenys

**Garsynai**
Liepa-2 garsynas
VDU ir KTU garsynai

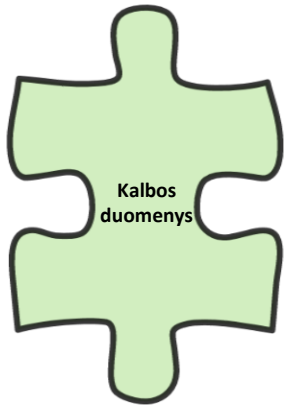
**Ontologijos**
LitWordNet,
SNOMED,
E.sąvokos

**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*

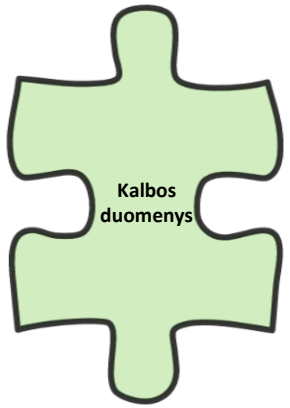
**Įterptinių vektorių kalbos modeliai**
VDU ir *Tokenmill* bandymai

**Geoinformaciniai kalbos duomenys**
E.Kalba geoinformacinė duomenų bazė



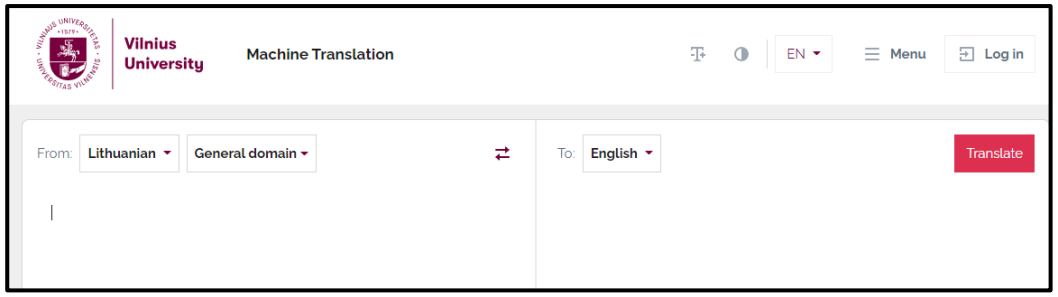
# Duomenų plėtros kryptys

- duomenys yra pagrindinis DI technologijų ir sprendinių šaltinis, todėl reikalingi tiek nestruktūruoti, 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

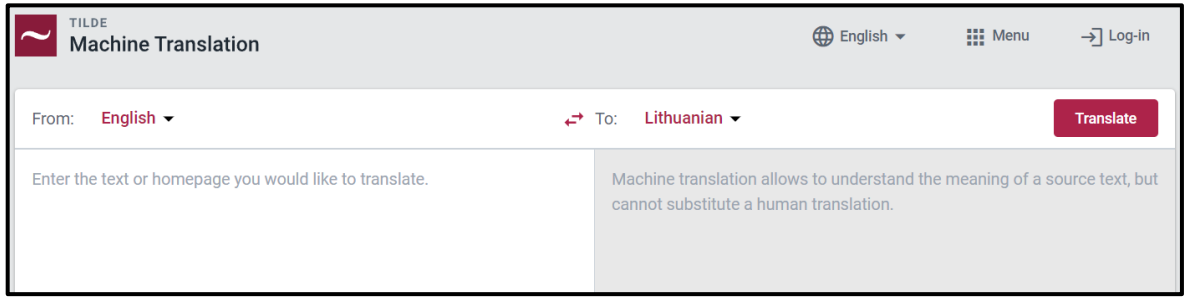


# Mašininis vertimas ir lokalizavimas


**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 Vertėjas

**Google: LT ⇆ 108 kalbos**

Microsoft Bing Translator Text

**MS Bing: LT ⇆ 86 kalbos**

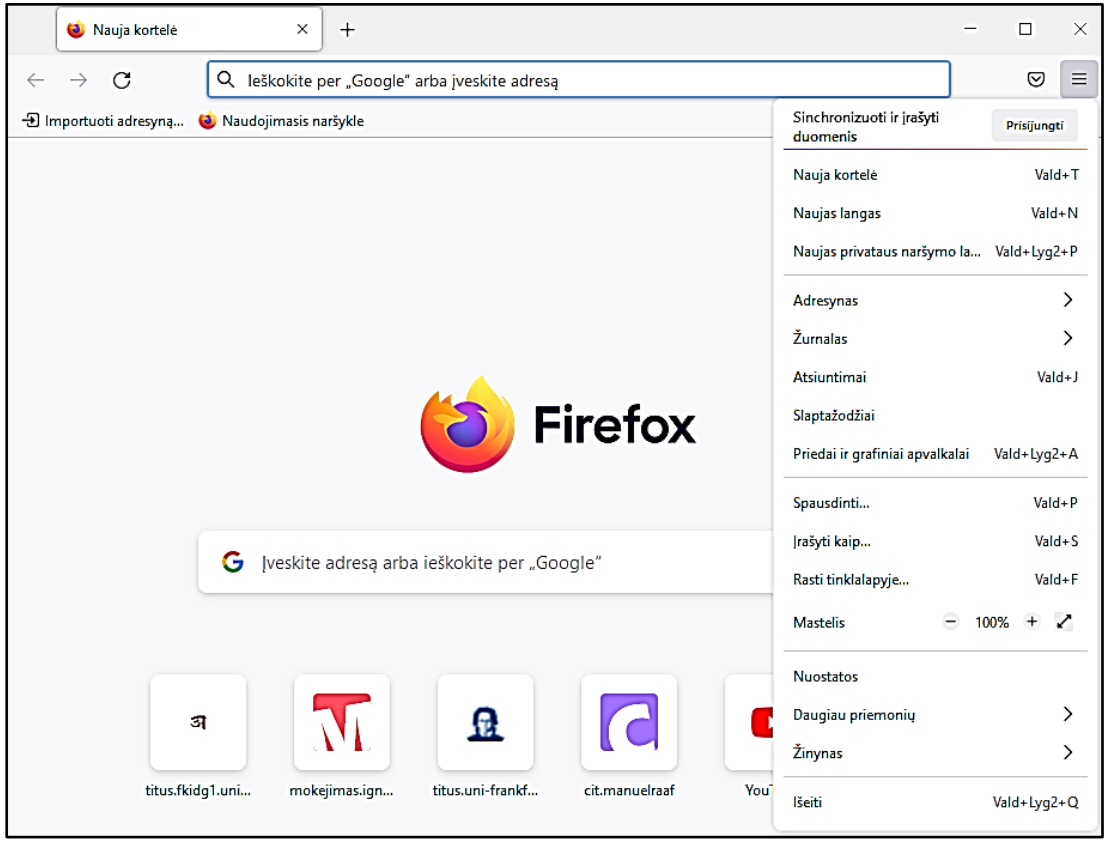
eTranslation

**eTranslation: LT ⇆ 24 ES kalbos**

**Mašininis vertimas ir lokalizavimas**

# 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

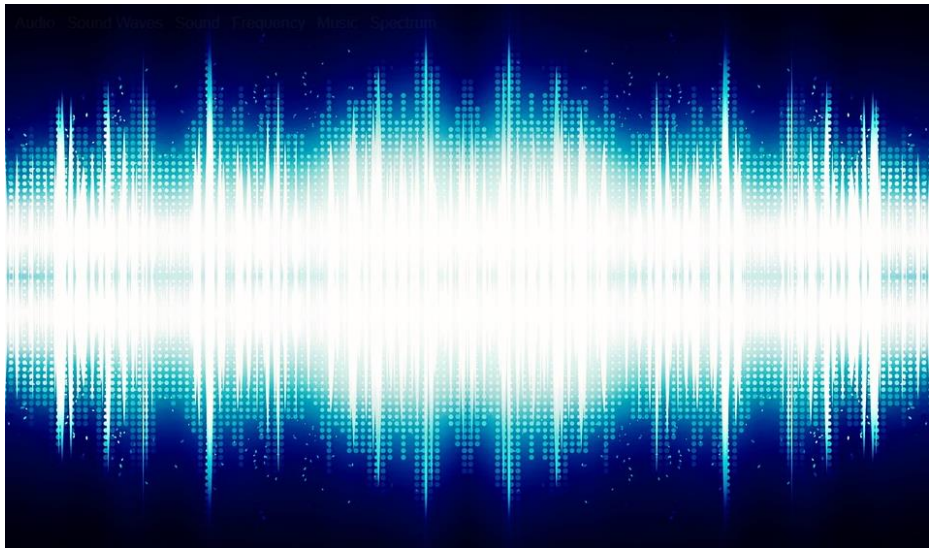
Tildė balso sintezė

Google

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

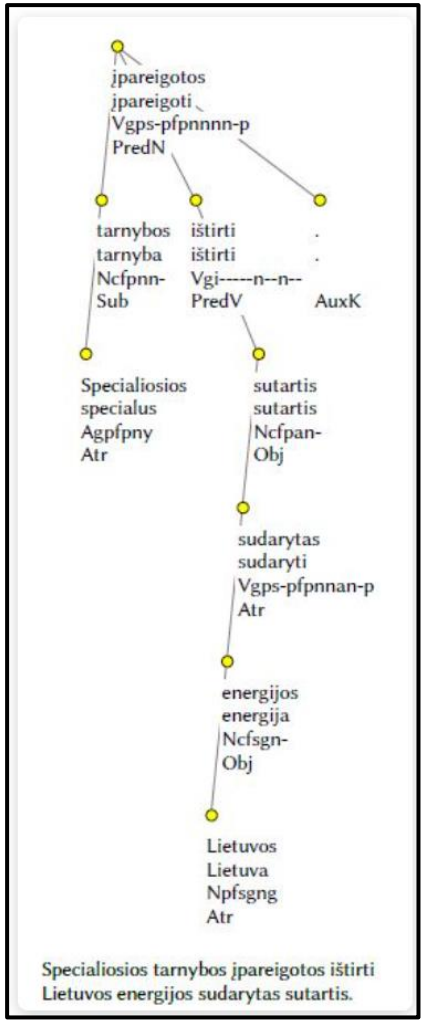
PASTOVU projekte

- 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

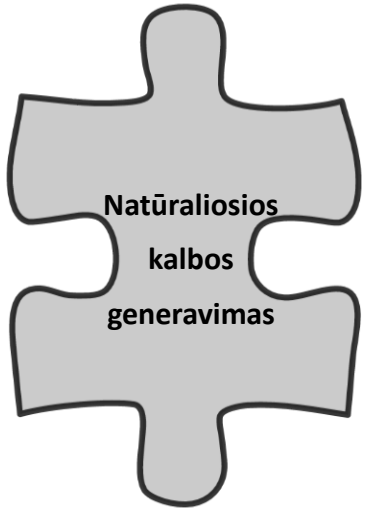
Tobulėjant giliojo mokymosi algoritmams, natūraliosios kalbos supratimo plėtrai reikalingi gausūs, patikimi ir įvairių temų duomenys, parengti mašininiam mokymuisi.

# 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**

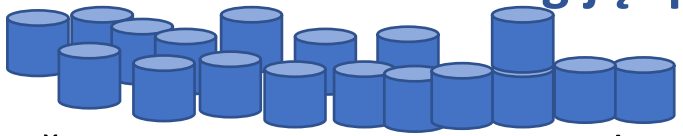
***Finansavimo lėšos***

- ❑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!