European Language Resource Coordination
Connecting Europe Facility

Europos kalbų išteklių konsorciumo seminaras

Diskusija Kalbos duomenys ir skaitmeniniai ištekliai Lietuvoje

Andrius UTKA
Jolanta ZABARSKAITĖ
Audrius VALOTKA
Audra IVANAUSKIENĖ

European Language Resource Coordination
Connecting Europe Facility

Europos kalbų išteklių konsorciumui aktualūs lietuvių kalbos duomenys ir skaitmeniniai ištekliai

Andrius UTKA

European Language Resource Coordination
Connecting Europe Facility

Diskusiniai klausimai

1. Kokie lietuvių kalbos ištekliai yra aktualūs Europos kalbų konsorciumui?
2. Kokie lietuvių kalbos ištekliai nėra aktualūs?
3. Kaip atsiranda nauji kalbos ištekliai?
4. Kokios Lietuvos institucijos šiais kalbos ištekliais disponuoja?
5. Kaip juos galima gauti?

Kokie lietuvių kalbos ištekliai Europos kalbų konsorciumui yra aktualūs?

1. Lygiagretūs duomenys, (lietuviu kalba su viena iš 24 oficialių Europos Sąjungos kalbų):
   - Lygiagretūs tekstynai,
   - Vertimo atmintys,
   - Žodynai ir leksikonai.
2. Vienkalbiai lietuvių kalbos duomenys:
   - Bendrieji tekstynai,
   - Specialieji tekstynai (tam tikros temos).

Kokie lietuvių kalbos ištekliai nėra aktualūs Europos kalbų konsorciumui?

- Lygiagretūs duomenys su **ne oficialiomis Europos Sąjungos kalbomis**.
- **Nedabartinės** lietuvių kalbos duomenys.
- Įvairūs **vienkalbiai** lietuvių kalbos **žodynai**.
- Mažiau reikalingi ir **grožinės literatūros** tekstynai.

Kaip atsiranda nauji kalbos ištekliai?

Lygiagretūs ištekliai

- Vertimo atmintys kaupiasi verčiant su vertimo atminčių sistemomis (pvz., Trados),
- Lygiagrečius tekstynus kaupiant reikia ne tik gauti tekstus, bet ir juos sulygiuoti naudojant specialius lygiagretinimo įrankius,
- Kuriant žodynus paprastai atliekamas nelengvas leksikografinis darbas.

Vienkalbiai ištekliai

- Juos kaupti lengviau, nei lygiagrečius:

turint tinkamas kopijavimo teises tekstus galima automatiškai siurbti iš interneto arba gauti iš leidyklų.

Kokios Lietuvos institucijos šiais kalbos ištekliais disponuoja?

- Vilniaus universitetas (sukūręs pirmąją statistinio mašininio vertimo sistemą),
  - lietuvių/anglų ir lietuvių/prancūzų lygiagretūs tekstynai,
  - dvikalbiai žodynai.
- Vytauto Didžiojo universitetas (sukūręs pirmąją anglų-lietuvių internetinę taisyklinę mašininio vertimo sistemą)
  - anglų-lietuvių, latvių-lietuvių, vokiečių-lietuvių lygiagretieji tekstynai
  - dabartinės lietuvių kalbos tekstynas (200 mln. žodžių) ir Bendrasis internetinis tekstynas (1 mlrd. žodžių).
- Lietuvių kalbos institutas
  - dvikalbiai žodynai (lietuvių/anglų, latvių, lenkų, vengrų, vokiečių)
- Valstybinė lietuvių kalbos komisija
  - terminų bankas (terminai, kurie turi vertimus)
- Vertimo biurai
  - įvairios vertimo atmintys

Kaip juos galima gauti?

- Tiesiogiai tariantis su institucijomis:
  - nes daugelis išteklių nors ir yra prieinami internete, tačiau duomenų parsisiuntimas yra ribojamas licenciniais susitarimais.
- Per mokslinių tyrimų infrastruktūras (MTI):
  - META-SHARE (metashare.tilde.com),
  - CLARIN (šiuo metu kūrimo fazėje) (www.clarin-lt.lt)

  (kuriose licencinės sutartys sprendžiamos sistemiškai)
- Tariantis su vertimo biurais:
  - turint omeny, kad daugelis vertimo tekstų yra privačių klientų nuosavybė su konfidencialumo apribojimais.

Audrius VALOTKA

VILNIAUS UNIVERSITETO FILOLOGIJOS FAKULTETAS

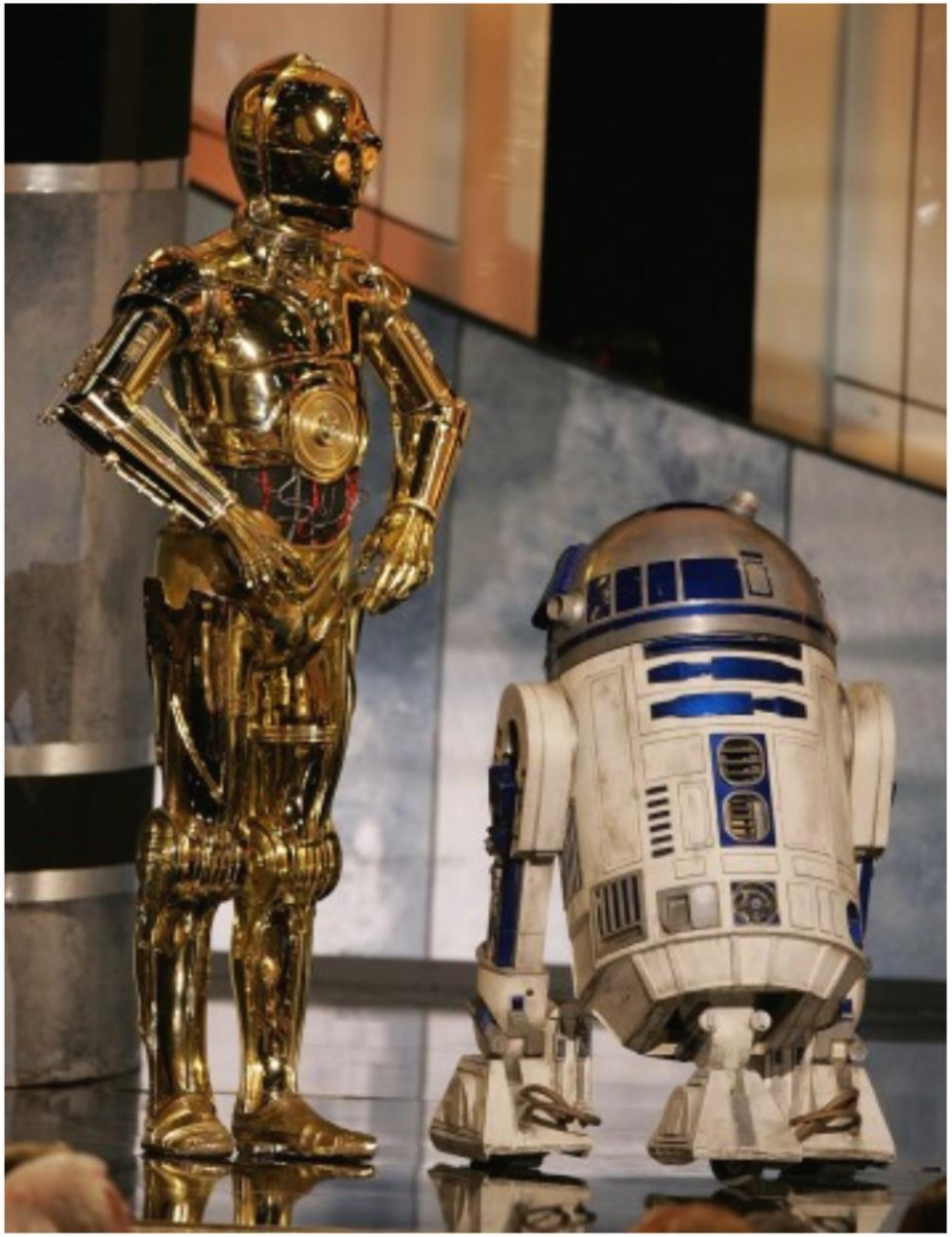
European Language Resource Coordination
Connecting Europe Facility

European Language Resource Coordination
Connecting Europe Facility

Kalba A (tekstas) - Kalba B (tekstas)

Šneka A – Kalba A – Kalba B – Šneka B

European Language Resource Coordination
Connecting Europe Facility



AČIŪ UŽ DĖMESĮ

European Language Resource Coordination
Connecting Europe Facility

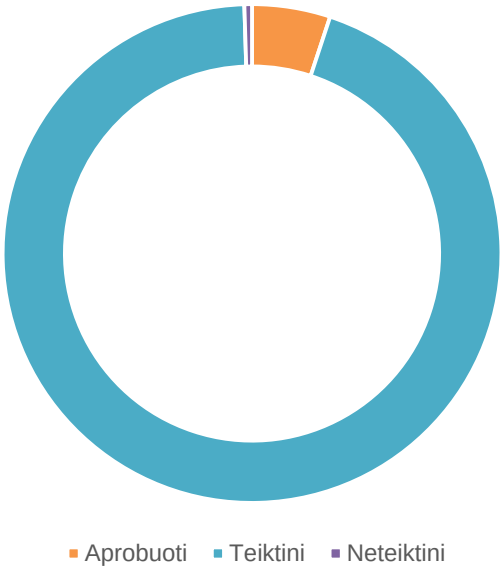
Europos kalbų išteklių konsorciumo seminaras

Diskusija Kalbos duomenys ir skaitmeniniai ištekliai Lietuvoje

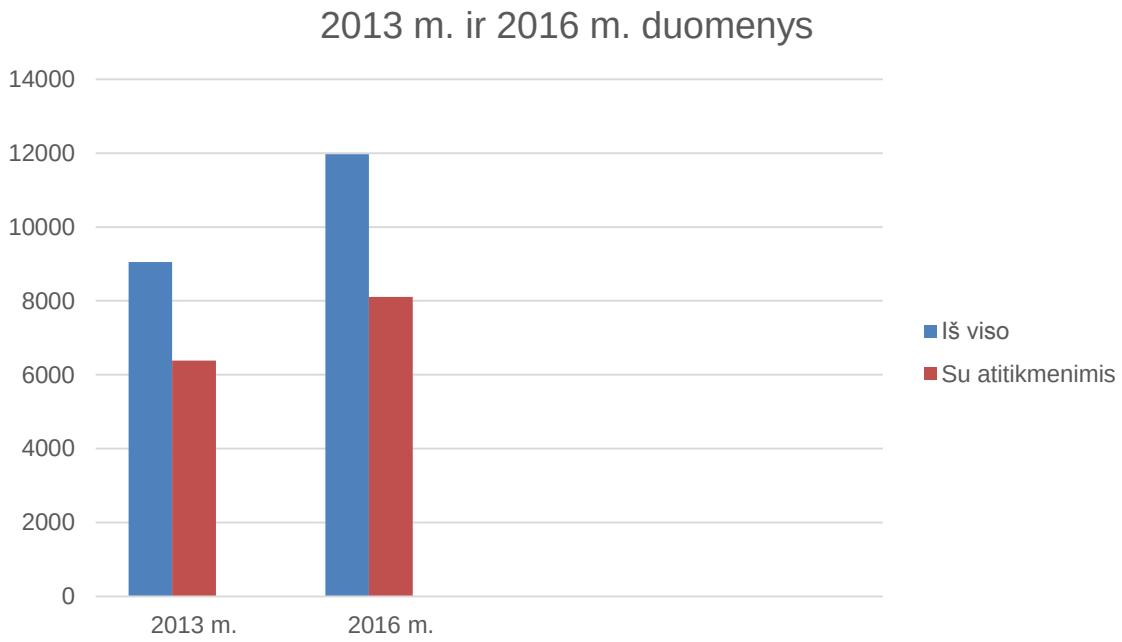
Audra IVANAUSKIENĖ

Lietuvos Respublikos terminų banko (http://terminai.vlkk.lt) statistika

Aprobuota – 11971 termino straipsnis
Teiktina – 222834 terminų straipsniai
Neteiktina – 1160 terminų straipsnių



Aprobuoti LRTB terminų straipsniai, turintys atitikmenų kitomis kalbomis



European Language Resource Coordination
Connecting Europe Facility

Europos kalbų išteklių konsorciumo seminaras

Diskusija Kalbos duomenys ir skaitmeniniai ištekliai Lietuvoje

Jolanta ZABARSKAITĖ

Skaitmeniniai Lietuvių kalbos instituto ištekliai

- LKI valdomų ir kuriamų išteklių tipai
- LKI išteklių ypatybės
- Vartotojai
- Išteklių kūrimo ir tvarkymo strateginės kryptys, pridėtinė vertė ir sklaida
- Ateities perspektyvos

Skaitmeniniai Lietuvių kalbos instituto ištekliai

- LKI mokslinės ir taikomosios veiklos, kaupiant, skaitmeninant ir kuriant skaitmeninius lietuvių kalbos išteklius, metodologinis pagrindas yra požiūris, kad:
- kai postmodernaus pasaulio raidos būtina sąlyga yra žinios ir kūrybiškumas (kurių raiškos yra kalbinės), tinkamai suvokta, organizuota ir naudojama kalba tampa ne tik praktine informacijos keitimosi priemone daugiakalbėje aplinkoje, bet ir kultūrinės, ekonominės bei socialinės vertės kūrimo priemone ir medžiaga.

Skaitmeniniai Lietuvių kalbos instituto ištekliai

LKI kaupiamų skaitmeninių išteklių tipai:

- skaitmenizuoti vienakalbiai žodynai (DŽ, LKŽ, BŽ, sinonimų, antonimų, palyginimų, frazeologijos žodynai ir t. t.),
- skaitmenizuoti dvikalbiai žodynai (lietuvių-vengrų, baltarusių-lietuvių),
- specializuotos duomenų bazės (pavardžių, vietovardžių, naujažodžių, naujųjų skolinių ir kt.),
- specializuoti tarminiai ištekliai (geolingvistinių duomenų daugiafunkcinė integruota bazė, tarmių archyvo informacinė duomenų bazė ir kt.),
- specializuoti tekstynai (sakytinės žiniasklaidos tekstynas 1960–2010, senųjų raštų anotuotas tekstynas SLIEKKAS, senųjų raštų duomenų bazė ir kt.),
- geoinformacinė Lietuvos vietovardžių duomenų bazė,
- naujažodžių tartuvas it t.t.

Skaitmeniniai Lietuvių kalbos instituto ištekliai

- LKI skaitmeninių išteklių ypatybės:
- struktūrizuoti,
- su labai detalia atributine paieška,
- integruojantys įvairaus formato medijas (įvairialypis turinys),
- nuolat atnaujinami atsižvelgiant į naujausius kalbos pokyčius,
- integruoti.

Skaitmeniniai Lietuvių kalbos instituto ištekliai

- http://www.prusija.lki.lt/
- http://lkiis.lki.lt/
- http://titus.uni-frankfurt.de/sliekkas/

Skaitmeniniai Lietuvių kalbos instituto ištekliai

Plėtros strategija:

- tolimesnis visų LKI išteklių integravimas į bendrą infrastruktūrą,
- prasminių sąsajų sistemos sukūrimas,
- paieškų galimybių didinimas,
- paieškų vizualizavimas,
- paslaugų vartotojui plėtra ir įvairovė,
- integracija į viešojo administravimo paslaugas,
- integracija į daugiakalbes infrasruktūras.