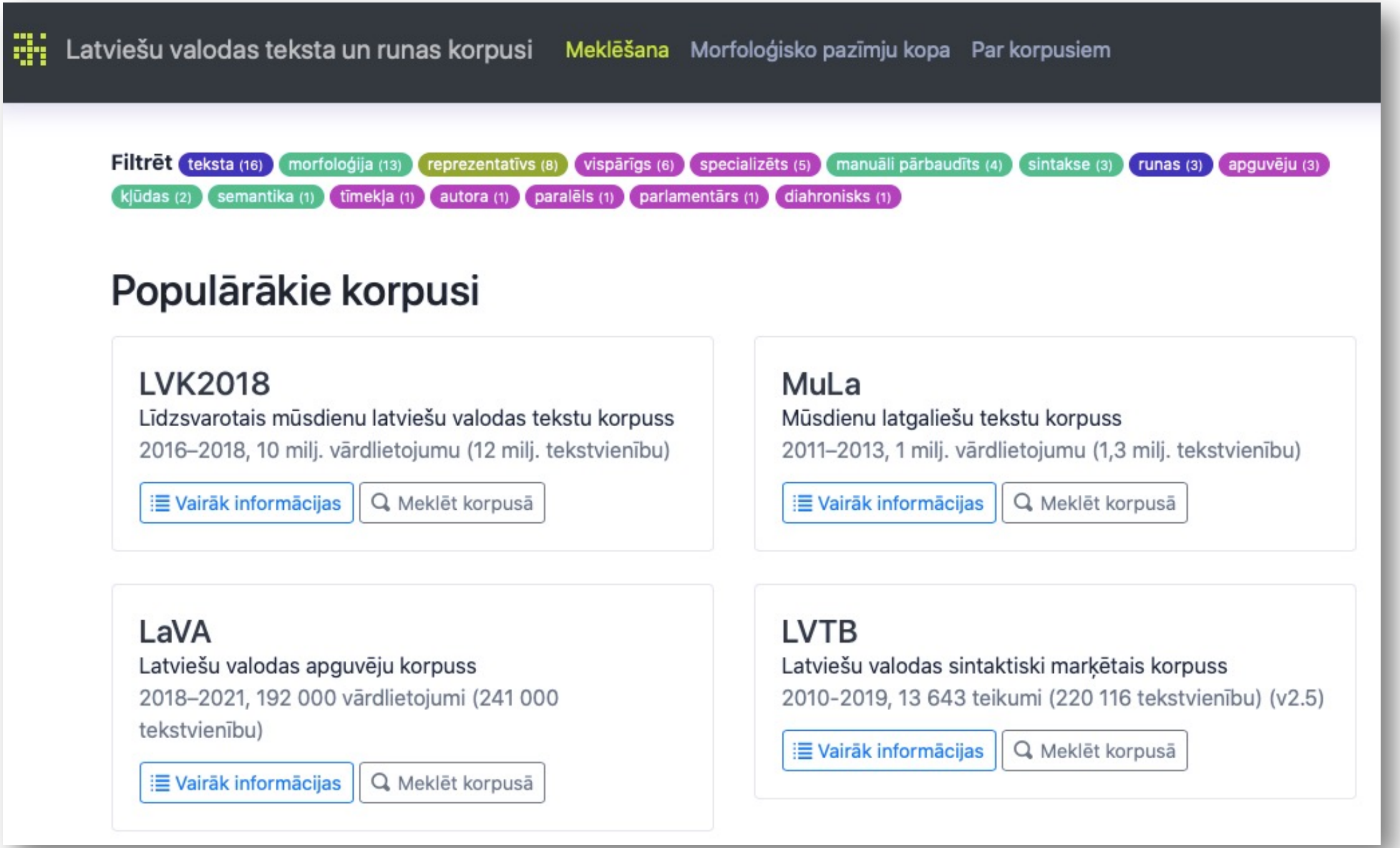
Latvijas Universitātes
Matemātikas un informātikas institūts

Latviešu valodas teksta un runas korpusi vietnē *korpuss.lv*

Baiba Saulīte, Ilze Auziņa

Praktiskais seminārs
2022. gada 12. janvāris


Latviešu valodas teksta un runas korpusi
Meklēšana
Morfoloģisko pazīmju kopa
Par korpusiem
Filtrēt
teksta (16)
morfoloģija (13)
reprezentatīvs (8)
vispārīgs (6)
specializēts (5)
manuāli pārbaudīts (4)
sintakse (3)
runas (3)
apguvēju (3)
kļūdas (2)
semantika (1)
tīmekļa (1)
autora (1)
paralēls (1)
parlamentārs (1)
diahronisks (1)
Populārākie korpusi
LVK2018
Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss
2016–2018, 10 milj. vārdlietojumu (12 milj. tekstvienību)
Vairāk informācijas
Meklēt korpusā
MuLa
Mūsdienu latgaliešu tekstu korpuss
2011–2013, 1 milj. vārdlietojumu (1,3 milj. tekstvienību)
Vairāk informācijas
Meklēt korpusā
LaVA
Latviešu valodas apguvēju korpuss
2018–2021, 192 000 vārdlietojumi (241 000 tekstvienību)
Vairāk informācijas
Meklēt korpusā
LVTB
Latviešu valodas sintaktiski marķētais korpuss
2010-2019, 13 643 teikumi (220 116 tekstvienību) (v2.5)
Vairāk informācijas
Meklēt korpusā

Korpusa vizītkarte

LaVA | Meklēt korpusā | Skatīt CLARIN-LV

Latviešu valodas apguvēju korpuss

Korpusā iekļauti to Latvijas augstākajās mācību iestādes studējošo ārvalstnieku darbi, kuri latviešu valodu apgūst kā svešvalodu pirmo vai otro semestri. Teksti ir automātiski morfoloģiski marķēti, un tajos ir manuāli marķētas valodas apguvēju kļūdas.

Publikācija, uz kuru atsaukties:
R. Dargis and I. Auzina and K. Levane-Petrova and I. Kaija
Quality Focused Approach to a Learner Corpus Development
Proceedings of The 12th Language Resources and Evaluation Conference (LREC), 392-396, 2020
PDF

teksta (16) | apguvēju (3) | morfoloģija (13) | kļūdas (2) | manuāli pārbaudīts (4)

Korpusa apjoms	192 000 vārdlietojumi (241 000 tekstvienību)
Izstrādes periods	2018–2021
Izstrādātājs	LU MII
Finansējuma avots	LZP Fundamentālo un lietišķo pētījumu programas projekts Nr. lzp-2018/1-0527
Citas publikācijas	I. Kaija and I. Auziņa **Data collection for learner corpus of Latvian: copyright and personal data protection** *Selected papers from the CLARIN Annual Conference 2019, 41-47, 2020* PDF DOI I. Auzina and I. Kaija and K. Levane-Petrova **Mērķhipotēžu izvirzīšana latviešu valodas apguvēju korpusā** *Valoda: nozīme un forma, 11, 7-26, 2020* PDF DOI

Korpusu iedalījums

Dati: teksta vs. runas

Struktūra: reprezentatīvs vs. nav reprezentatīvs

Tips: vispārīgs vs. specializēts

Marķējums: marķēts vs. nemarķēts

Korpusu iedalījums: dati

Teksta korpuss – liels un strukturēts mašīnlasāmu tekstu kopums, kas paredzēts rakstītās valodas analīzei.

teksta (16)

LVK2018, LVK2013, LVTB, UDLV, FullStack-LV, LiLa, MuLa, LaVA, VVPP, Pārspriedumi, Saeima, Senie, Rainis, Emuāri, Barometrs, Tīmeklis 2007

Runas korpuss – audio ierakstu un to transkripciju kopums, kas paredzēts runātās valodas izpētei.

runas (3)

LVR100, LaRKo, LAMBA, Subtitri

Korpusu iedalījums: struktūra

Reprezentatīvā korpusā ir ietverti daudzveidīgi dati noteiktās proporcijās.

Vispārīgā un specializētā korpusa reprezentativitāte tiek sasniegta un mērīta dažādos veidos.

reprezentatīvs (8)

LVK2018, LVK2013, LVTB, UDLV, FullStack-LV, LiLa, MuLa
LVR100

Korpusu iedalījums: tips

Vispārīgais korpuss nav ierobežots tematiski vai pēc kādas citas pazīmes.

vispārīgs (6)

LVK2018, LVK2013, LVTB, UDLV, FullStack-LV, LRK100

Specializētais korpuss ir ierobežots tematiski, ģeogrāfiski, pēc runātāju vecuma, laikā vai pēc kādas citas pazīmes. Kopā 13 tādu korpusu.

specializēts (5)

LAMBA, LaRKo, MuLa, Emuāri, Barometrs

Saeima, LaVa, VVPP, Pārspriedumi, Senie, Tīmeklis 2007, LiLa, Rainis

parlamentārs (1) apguvēju (3) diahronisks (1) tīmekļa (1) paralēls (1) autora (1)

Korpusu iedalījums: marķējums

morfoloģija (13) – morfoloģiski marķēts

Emuāri, FullStack-LV, LAMBA, LaVA, LVK2013, LVK2018, LVTB, LRK100, VVPP, Pārspriedumi, Rainis, Saeima, Subtitri, Tīmeklis2007, UDLV

sintakse (3) – sintaktiski marķēts FullStack-LV, LVTB, UDLV,

semantika (1) – semantiski marķēts (*FrameNet* un *PropBank*) FullStack-LV

kļūdas (2) – marķētas valodas apguvēju kļūdas LaVA, VVPP

manuāli pārbaudīts (4) – manuāli labotas automātiskā marķējuma kļūdas

LVTB, UDLV, FullStack-LV, LaVA

Piekļuve korpusiem

Konkrētā korpusa mājaslapā (korpuss vai plašāka informācija par to):

Barometrs	http://barometrs.korpuss.lv/
LAMBA	runa.lamba.lv
LaRKo	larko.ailab.lv
LaVA	lava.korpuss.lv/search
Saeima	saeima.korpuss.lv
Senie	senie.korpuss.lv
LVTB, LVUD	http://sintakse.korpuss.lv/
LRK100	http://runa.korpuss.lv/

Daži korpusi nav publiski pieejami (VVPP) vai ir pieejami tikai daļēji (LRK100, Subtitri)

Piekļuve korpusiem *NoSketchEngine* platformā

SketchEngine

- korpusu pārlūks un teksta analīzes rīks
- ar sarežģītiem vaicājumiem ļauj pētniekiem meklēt valodas parādības apjomīgās tekstu kolekcijās
- korpusi vairāk nekā 90 valodās

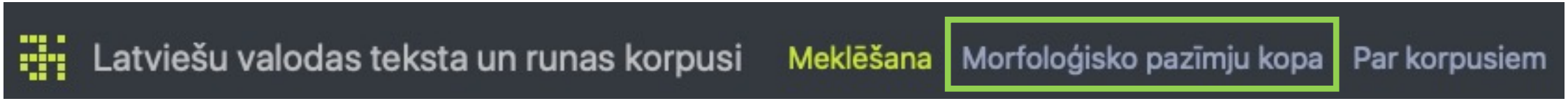
NoSketchEngine – *SketchEngine* bezmaksas versija

AiLab korpusi (14)

Emuāri, LaVA, LiLa, LVK2013, LRK100, LVK2018, Pārspriedumi, Rainis, Saeima, Subtitri, Tīmeklis2007, UDLV

MuLa, Senie

Morfoloģisko pazīmju kopa

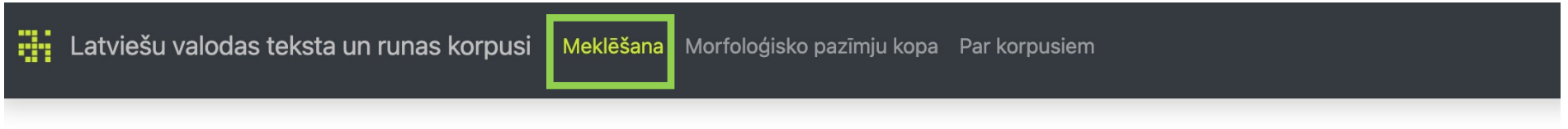


Morfoloģiskajā marķēšanā izmantotā pazīmju kopa, kurā uzskaitītas vārdšķiras (11) un citas tekstvienību grupas (3), katrai no tām:

- morfoloģisko pazīmju un to apzīmējumu (tagu) kopums (1–13)
- katras pazīmes vērtība
- skaidrojums
- piemēri

Kur nepieciešams, pie vārdšķiras dotas īsas piezīmes, piemēram, par pamatformas izvēli vai par atšķirībām no tradicionālās latviešu valodas gramatikas.

Meklēšana vairākos korpusos vienlaikus



??kūkot ? Meklēt

Meklēšanas instrukcija

Ierakstot pamatformu, tiks atrastas visas vārdformas:
iet → eju, iet, ejam, gāja utt.

Ierakstot konkrētu vārdformu, tiks meklēti lietojumi tikai ar šo vārdformu:
gāja → gāja, ziemā → ziemā, kūku → kūku u. c.

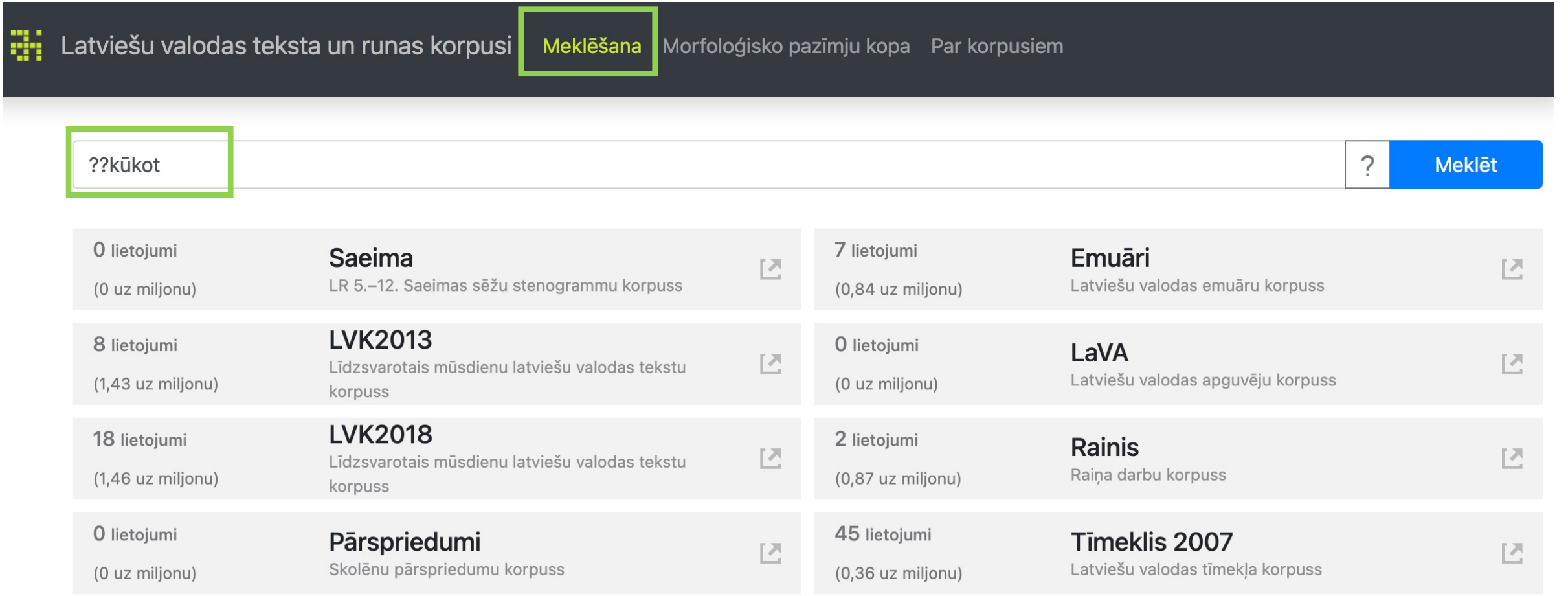
Ierakstot divu vai vairāku vārdu pamatformas, tiks atrastas konkordances ar šiem vārdiem dažādos locījumos:
lasīt grāmata → lasu grāmatu, jālasa grāmata, jālasa grāmatu, lasīju grāmatu u. c.

Vaicājumā var izmantot zvaigznīti (*) nenoteikta rakstzīmju skaita norādīšanai, jautājuma zīmi (?) tieši vienas rakstzīmes apzīmēšanai, vertikālu līniju (|) vairāku vārdu vai frāžu meklēšanai.

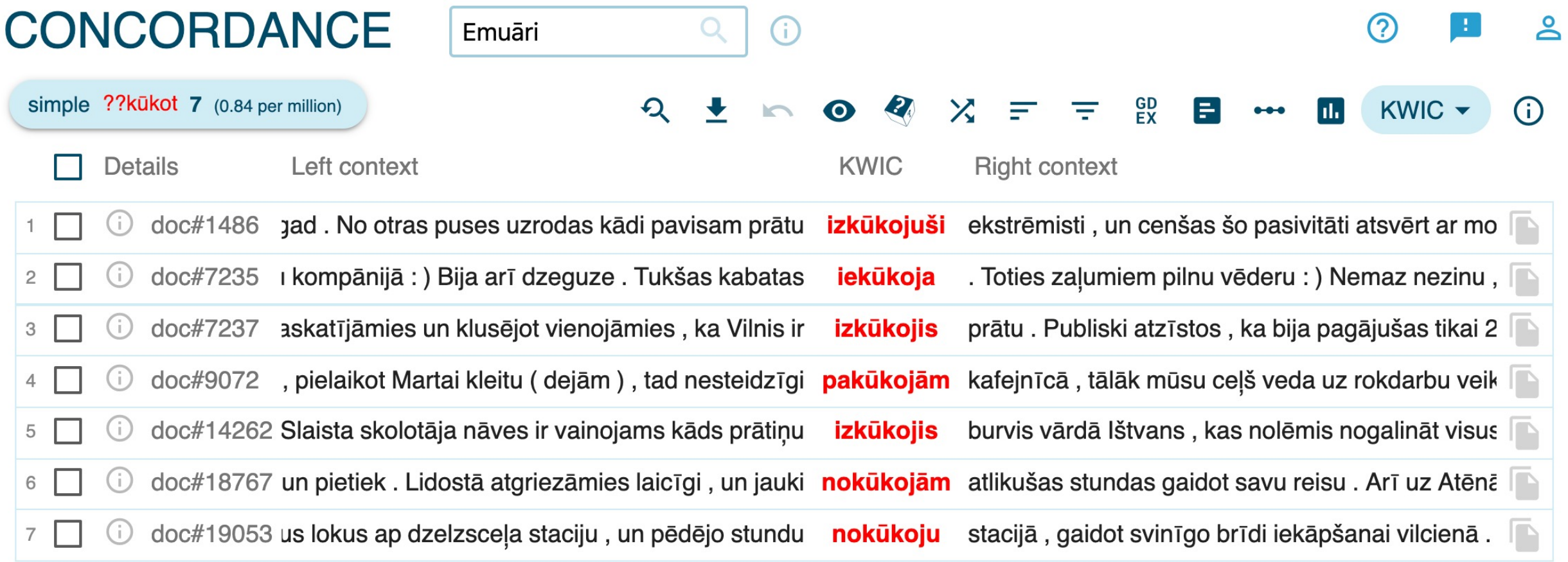
Vaicājumu piemēri

Vaicājums	Atrastie lietojumi	Komentāri
`inic*`	*iniciators, iniciatīva, iniciēt, iniciālis* utt.	Vārdi, kas sākas ar *inic-*.
`*šana`	*dibināšana, piekrišana, ražošanas, pagatavošanai, zināšanu* utt.	Vārdi, kas beidzas ar *-šana*, un to formas.
`??kūkot`	*izkūkot, pakūkot, nokūkot, iekūkot* utt.	Verbs *kūkot* ar divburtu priedēkli dažādos locījumos.
`finansu|finanšu`	*finansu, finanšu*	Atrod abas vārdformas.

Meklēšana vairākos korpusos: piemērs



Meklēšana vairākos korpusos: piemērs



korpuss.lv nākotne

Paplašināt meklēšanu visos korpusos – pievienot meklēšanu arī citu iestāžu korpusos.

Pilnveidot esošos korpusus un pievienot jaunus.

Regulāri ievietot jaunākās versijas.

Pēc iespējas pievienot meklēšanas instrukcijas.

Ieskaties: http://ailab.lv

Seko mums: @AiLab.lv

Jautājumi, ieteikumi: korpuss@ailab.lv

Projekti

Nr. VPP-IZM-DH-2020/1-0001
Nr. VPP-Letonika-2021/1-0006

Nr. lzp-2018/1-0527 (2018-2021)
Nr. lzp-2019/1-0464 (2020–2022)

Nr. VPP-IZM-2018/2-0002 (2018–2021)