Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss 2018

Kristīne Levāne-Petrova, Roberts Darģis

Praktiskas ievirzes pētījumu projekts "Daudzslāņu valodas resursu kopa teksta semantiskai analīzei un sintēzei latviešu valodā" nr.1.1.1.1/16/A/219

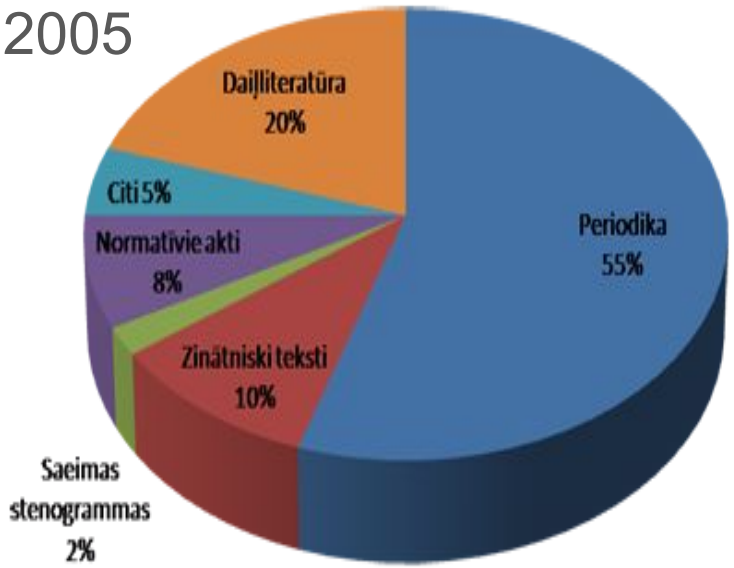
LU 76. Zinātniskās konferences
Datorlingvistikas sekcija 2018. gada 1. martā

NACIONĀLAIS ATTĪSTĪBAS PLĀNS 2020
EIROPAS SAVIENĪBA
Eiropas Reģionālās attīstības fonds
IEGULDĪJUMS TAVĀ NĀKOTNĒ

Mākslīgā intelekta laboratorija
LU MII

Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss (LVK)

- Latviešu valodas korpusa koncepcija 2005
- LVK tiek veidots kopš 2007. gada
- 2013. gadā LVK 2013 sasniedza 4,5 miljonus vārdlietojumu (ar LVA atbalstu)

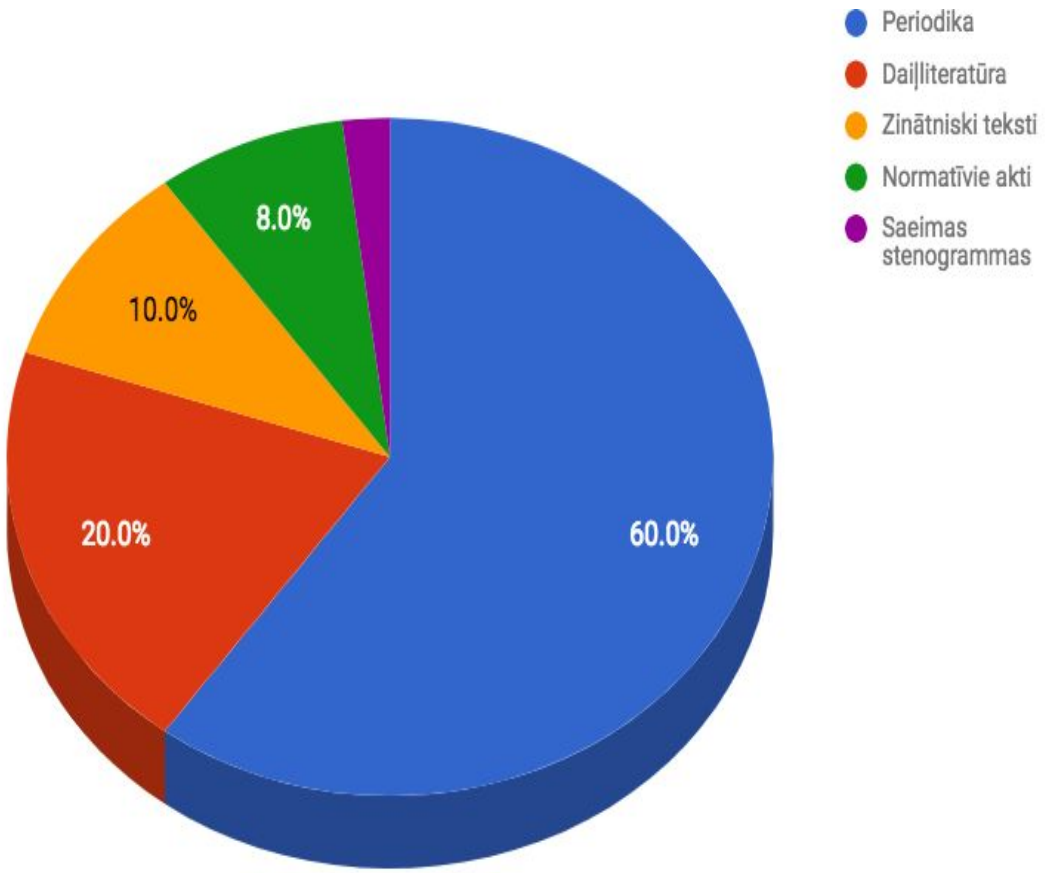


Līdzsvarotā mūsdienu latviešu valodas tekstu korpusa paplašināšana

- LVK2013 → LVK2018 (4,5 milj. → 10 milj.)
- Tiek veidots, pamatā ievērojot LVK2013 uzbūves principus
- Precizēti iekļaujamie metadati
- Automatizēta tekstu atlase

Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss (LVK2018)

- Vispārīgs
- Līdzsvarots
- Papildināms
- Pievienoti metadati
- Automātiski morfoloģiski marķēts



Tekstu atlases kritēriji (1)

- **Valodas paveids**

Daiļliteratūra, periodika, zinātniski teksti utt.

- **Laiks**

Teksti, kas publicēti kopš 1991.gada.

- **Oriģināllteksti**

Korpusā netiek iekļauti tulkojumi.

- **Publikācijas veids**

Teksti, kas publicēti gan elektroniskā formātā, gan drukātā formātā.

Tekstu atlases kritēriji (2)

- **Valodas parauga apjoms**

Gan pilni teksti, gan tekstu fragmenti.

Valodas parauga izmērs nedrīkst pārsniegt 5% no korpusa sadaļas apjoma.

- **Autentiski teksti**

Teksti tiek iekļauti korpusā tādi, kādi tie ir — bez labojumiem.

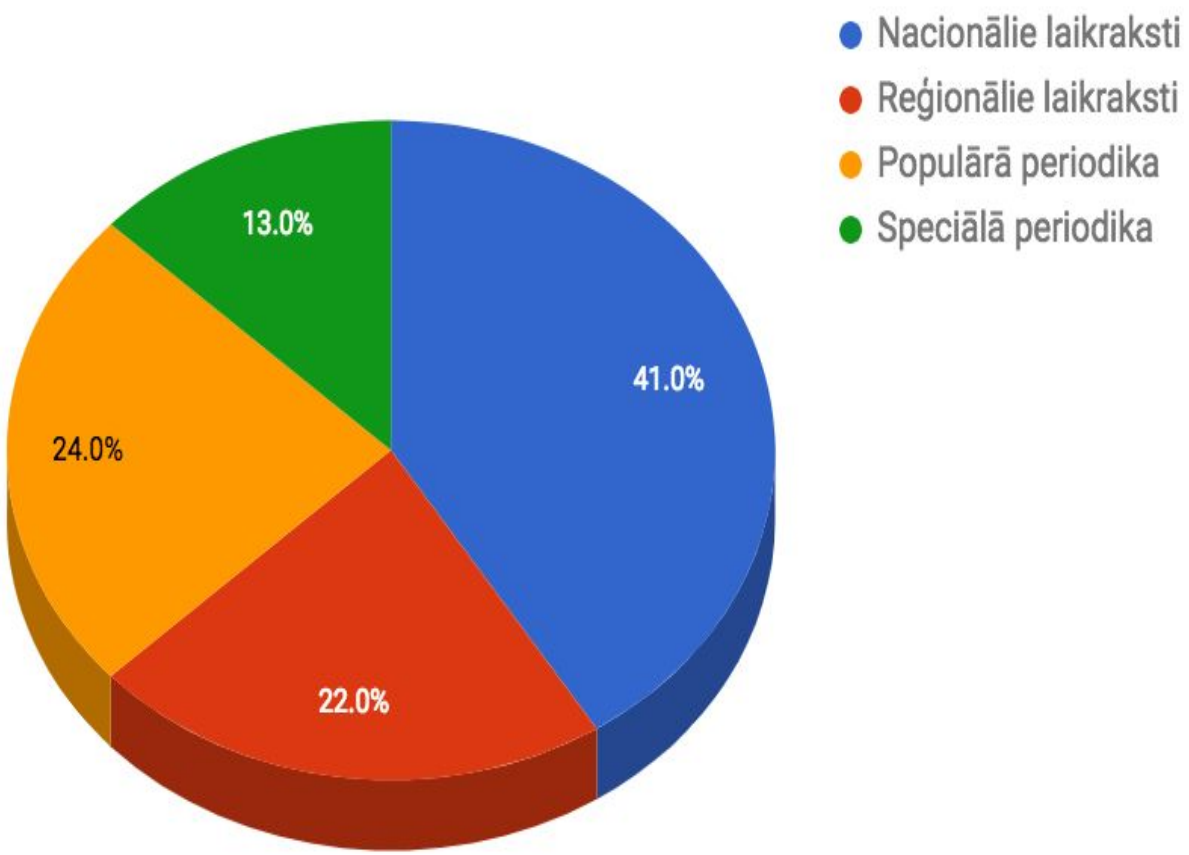
- **Pieejamība, subjektīvais, nejaušības princips, formālas atbilstības princips u.c.**

LVK2018 periodika (1)

Tematiskais daudzveidīgums

Ģeogrāfija

Unikalitāte



LVK2018 periodika (2)

Reģionālā periodika

(rekurzeme.lv, ezerzeme.lv, edruva.lv u.c.)

Speciālā periodika

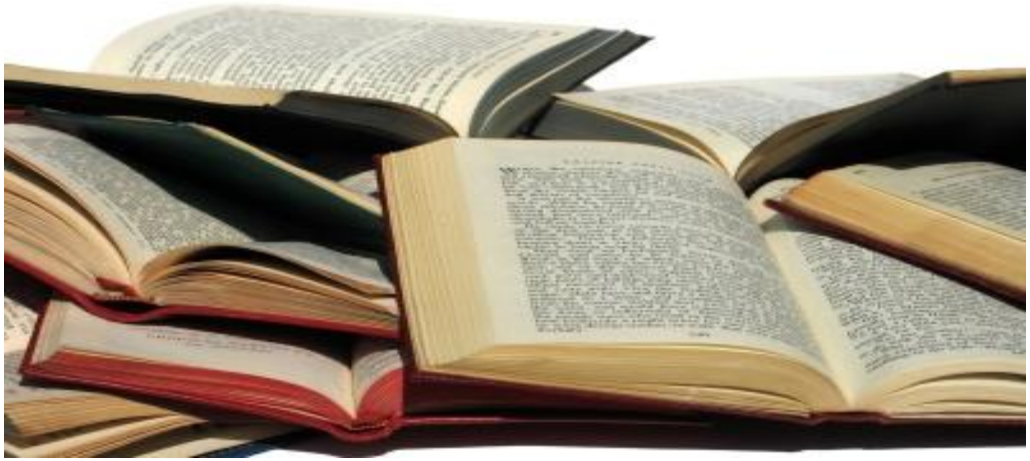
(kroders.lv, cetrassezonas.lv, mammamuntetiem.lv, garaza.lv u.c.)

Populārā periodika

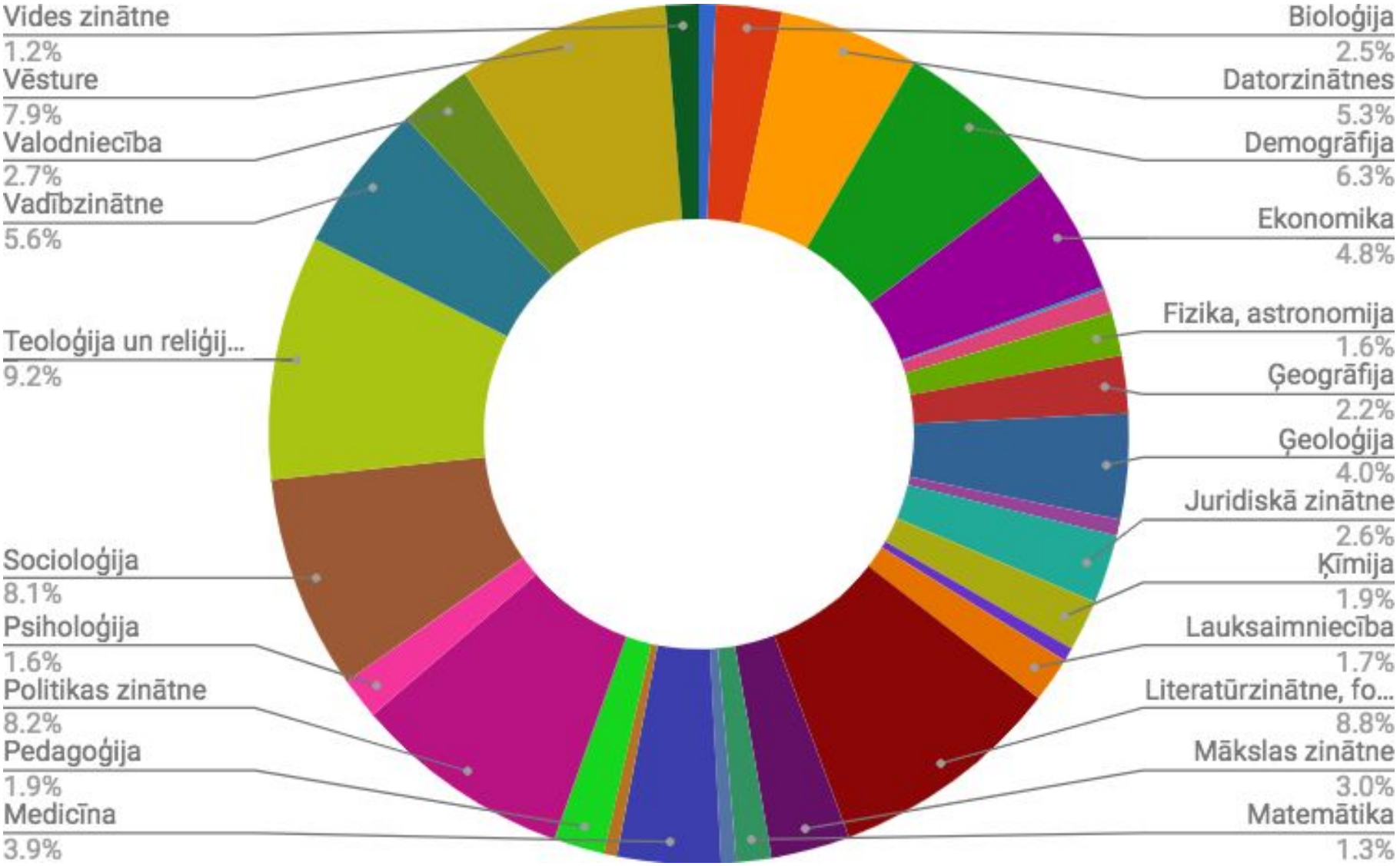
(cosmo.lv, slavenibas.lv u.c.)

LVK2018 daiļliteratūra

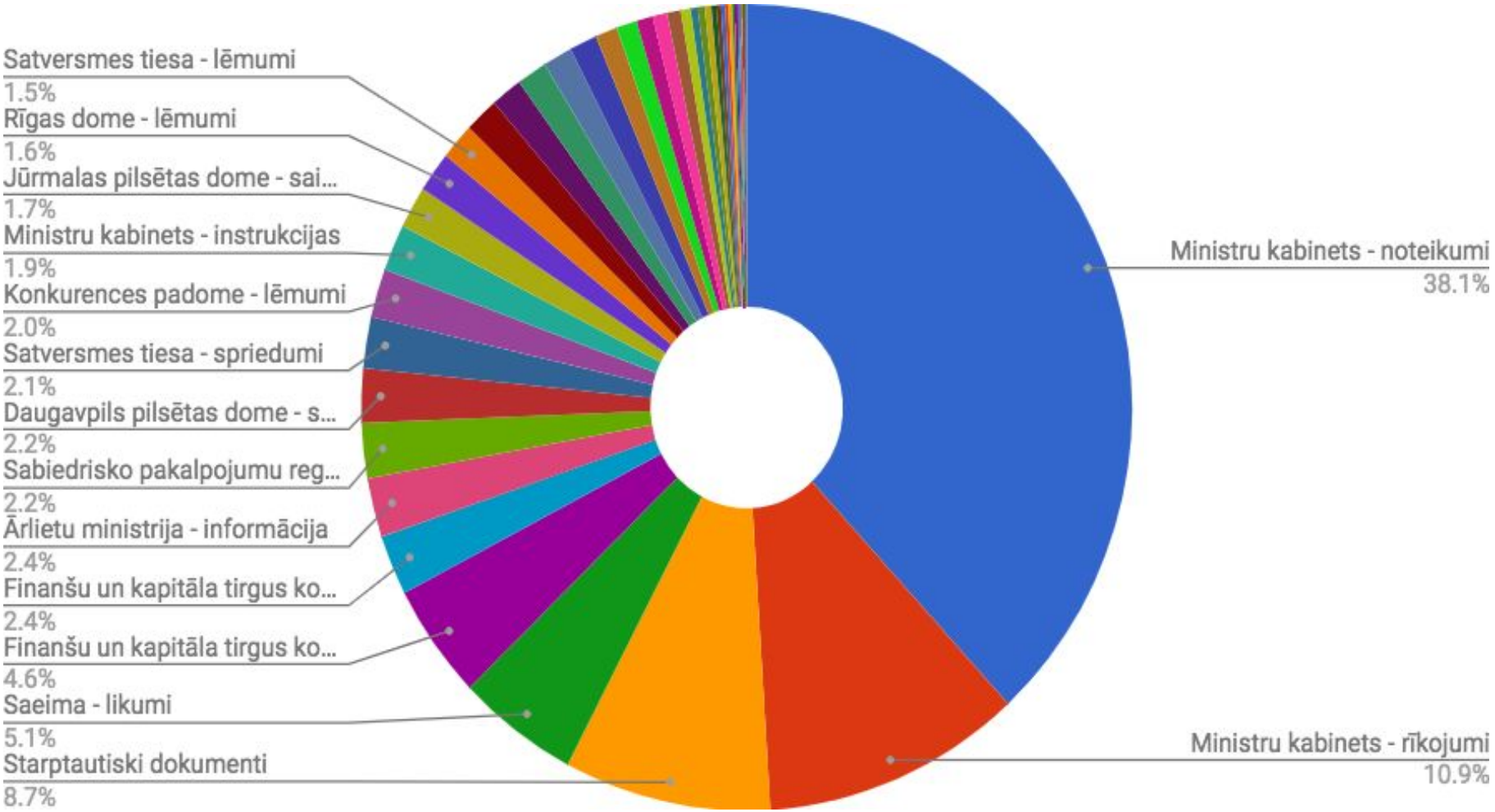
Atzītu autoru darbi (Nora Ikstena, Pauls Bankovskis, Dace Rukšāne u.c.)



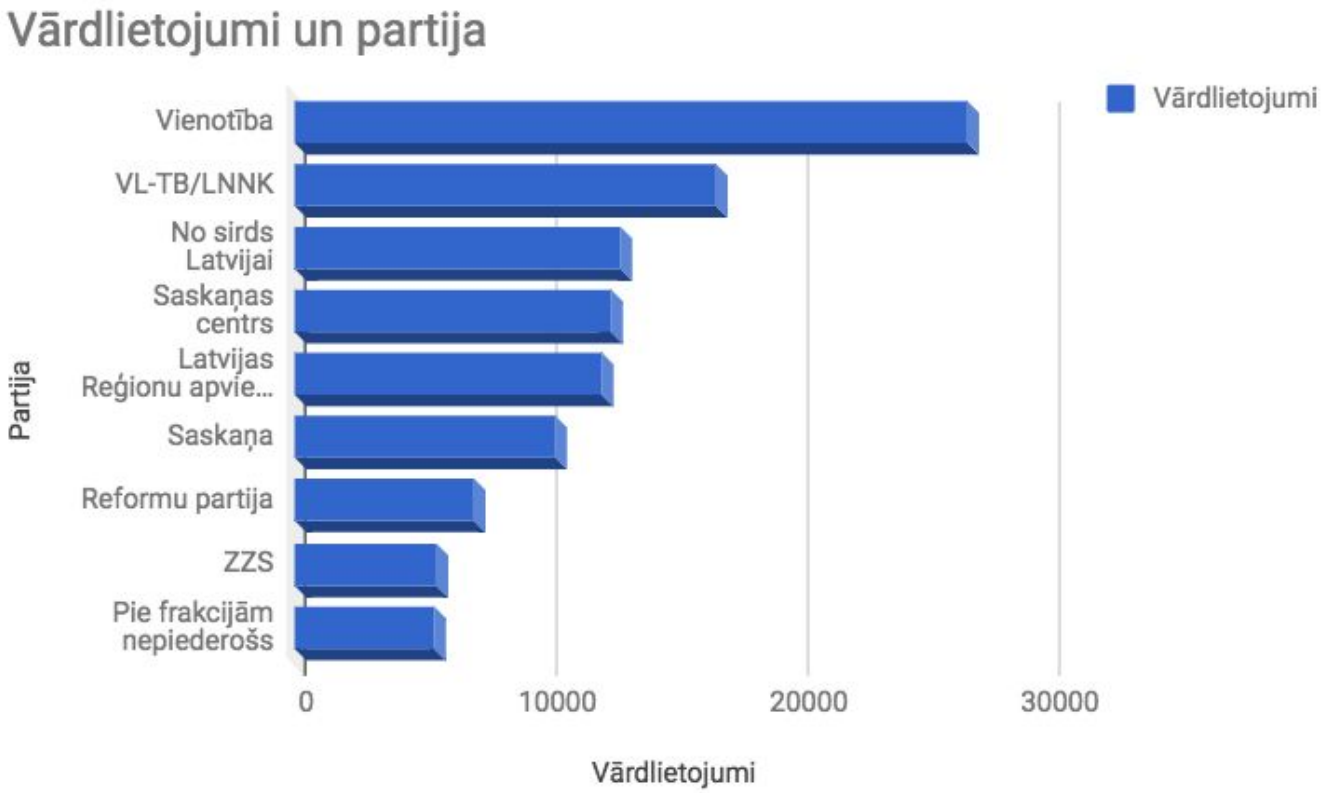
LVK2018 zinātniski teksti



LVK2018 normatīvie akti



LVK2018 Saeimas stenogrammas



Metadati

Katram korpusa avotam tiek pievienoti metadati, piemēram,

- Avota autors
- Avota nosaukums
- Izdevējs
- Izdošanas gads
- Vietrādis URL

Kā atsaukties uz korpusu?

Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss

LVK2013

LVK2018

Paldies par uzmanību!

www.korpuss.lv

268235	burtu " k" un defisi, aiz defises norādot	korpusa	numuru (20.2. apakšpunkts). [2.9
z71_0	sagatavot pētījumam tekstu (arī audio)	korpusu	, tas ir, apkopot, sistematizēt un apstrādāt
z72_0	ģenerālgubernatora pils jaunais rietumu	korpuss	; nozīmīga celtne rātsnama kompleksā bija
d121_0	tramvaja logu lūkodamās uz tukšajiem fabriku	korpusiem	. Kad Millija bija aprakta, Florence
d123_0	, kad sadzirdējām medmāsu kliedzienus no	korpusa	, tad priecīgi saskatījāmies, jo zinājām
d131_0	spokoties pamestu rūpnīcu un neražošanas	korpusu	, nedaceltu būvju, un tas viss kādam piederēja
d165_0	griežas apkārt, cenšoties ieraudzīt rūpnīcas	korpusus	. Taču kārklu un nezāļu puduri ir lielāki
d71_0	sētu, bet caurspīdīgu, vienā pusē slimnīcas	korpuss	, otrā — mežs. Bet blakus placī pastaigājās
p107_0	pielietojums to senlaicīgā un greznā koka	korpusa	dēļ. Tolaik radio korpuss bija no riekstkoka
p107_0	un greznā koka korpusa dēļ. Tolaik radio	korpuss	bija no riekstkoka, vēlāk padomju laikos
d57_0	raibās kleitas. Bodītes lillā krāsotais	korpuss	izrādījās vecs kuģa bunkurs no kniedēta
p1459_0	turpmāk jāpieliek roka — slimnīcas piecstāvu	korpusam	un ambulatorajai daļai nepieciešama siltināšana
p1461_0	Rāviņi visus šos 13 gadus mitinājušies kūts	korpusā	un arī tagad turpina to darīt, tā viņiem
p1482_0	pakalpojumu uzlabošana, slimnīcas trīsstāvu	korpusa	celtniecības uzsākšana. Atbalsts vēlams
p1515_0	Rudzātu speciālās internātskolas mācību	korpusa	atjaunošana, Salenieku pansionātam tiks
p180_0	visticamāk tapis ekstrēmā veidā. " Gar (rūpnīcas	korpusa	nogruvušās sienas) ārmalu ved kāpnes līdz
p1893_0	iekārtots jaunā, plašā, gaišā telpā jaunā	korpusa	pirmajā stāvā. Centra vadītāja Mirdza
p2171_0	gadījumā iespējams salocīt gar automašīnas	korpusa	sāniem. Konstruktors apgalvoja, ka
p2510_0	visas sev vajadzīgās paroles. Arī atslēgas	korpuss	ir drošs pret mehāniskiem bojājumiem, jo
p2252_0	izpausmes pilsētas punktus - Spīķerus un VEF	korpusus	, K. Barona ielu un tās apkārtni.