

Līdzsvarotais mūsdienu latviešu valodas tekstu korpus 2022

Kristīne Levāne-Petrova
kristine.levane-petrova@lumii.lv

Darbs veikts ar “Latviešu valodas aģentūras” atbalstu, projekts Nr. 4.6/2019-029.

Clarín darbseminārs
2022. gada 8. decembrī



Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss (LVK)

- Latviešu valodas korpusa koncepcija 2005
- LVK tiek veidots kopš 2007. gada
- 2018. gadā LVK2018 sasniedza 10 miljonus vārdlietojumu

Līdzsvarotā mūsdienu latviešu valodas tekstu korpusa paplašināšana

- LVK2018 → LVK2022 (10 milj. → 100 milj.)
- Tiek veidots, pamatā ievērojot LVK uzbūves principus
- Automatizēta tekstu atlase
- Vienošanās ar izdevējiem (16 līgumi)



Latviešu valodas
aģentūra



LATVIJAS
UNIVERSITĀTE

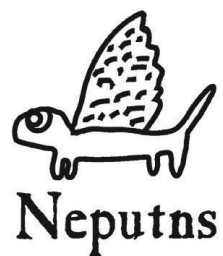


LIEPĀJAS
UNIVERSITĀTE

LAMI

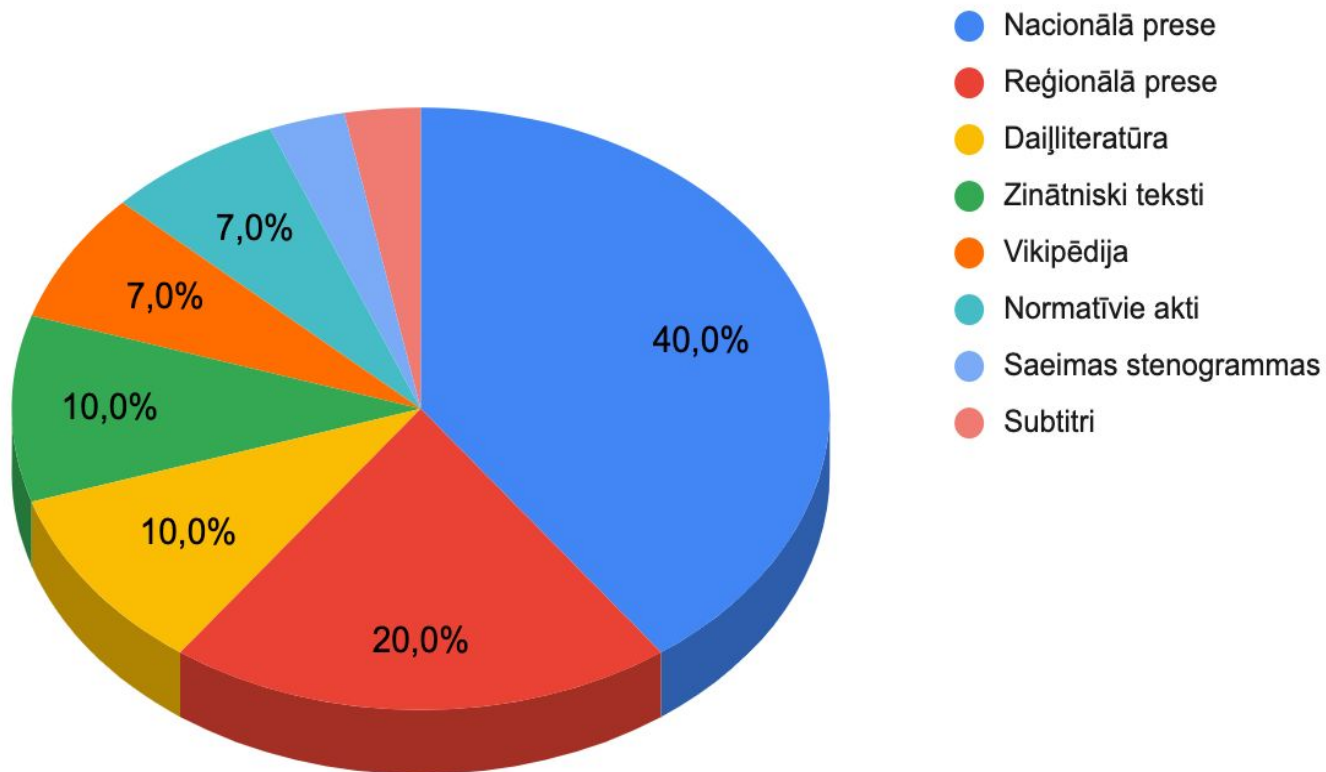
LU LITERATŪRAS,
FOLKLORAS
UN MĀKSLAS
INSTITŪTS





Līdzsvarotais mūsdienu latviešu valodas tekstu korpuss (LVK2022 (alpha))

- Vispārīgs
- Līdzsvarots
- Ar metadatiem
- Automātiski morfoloģiski marķēts



Tekstu atlases kritēriji

- **Laiks**

Teksti, kas publicēti kopš 1991. gada.

- **Orīginālteksti/tulkojumi**

Korpusā tiek iekļauti arī tulkojumi.

- **Autentiski teksti**

Teksti tiek iekļauti korpusā bez labojumiem.

- **Valodas parauga apjoms**

Gan pilni teksti, gan tekstu fragmenti.

LVK2022 nacionālā prese (60%)

- Raksti no dažādiem Latvijas ziņu portāliem (Delfi, Kasjauns)
- No 2000. līdz 2022. gadam (624 datumi)
- No katra mēneša 3-4 datumi

Tematiskais daudzveidīgums

Unikalitāte

LVK2022 reģionālā prese (20%)

Reģionu mediji (Alūksnes un Malienas Ziņas, Bauskas dzīve, Zemgales Ziņas, Dzirkstele, Staburags, Ziemeļlatvija); 506 datumi

IZVĒLIES SAVU LAIKRAKSTU!



**Alūksnes un
Malienas Ziņas**

Alūksnes novads
✉ reklama@aluksniesiem.lv
☎ 26410083



Bauskas dzīve

Bauskas novads
✉ abone@bauskasdzive.lv
☎ 63922963; 26521719



Zemgales Ziņas

Jelgavas novads
✉ reklama@zz.lv
☎ 29522152

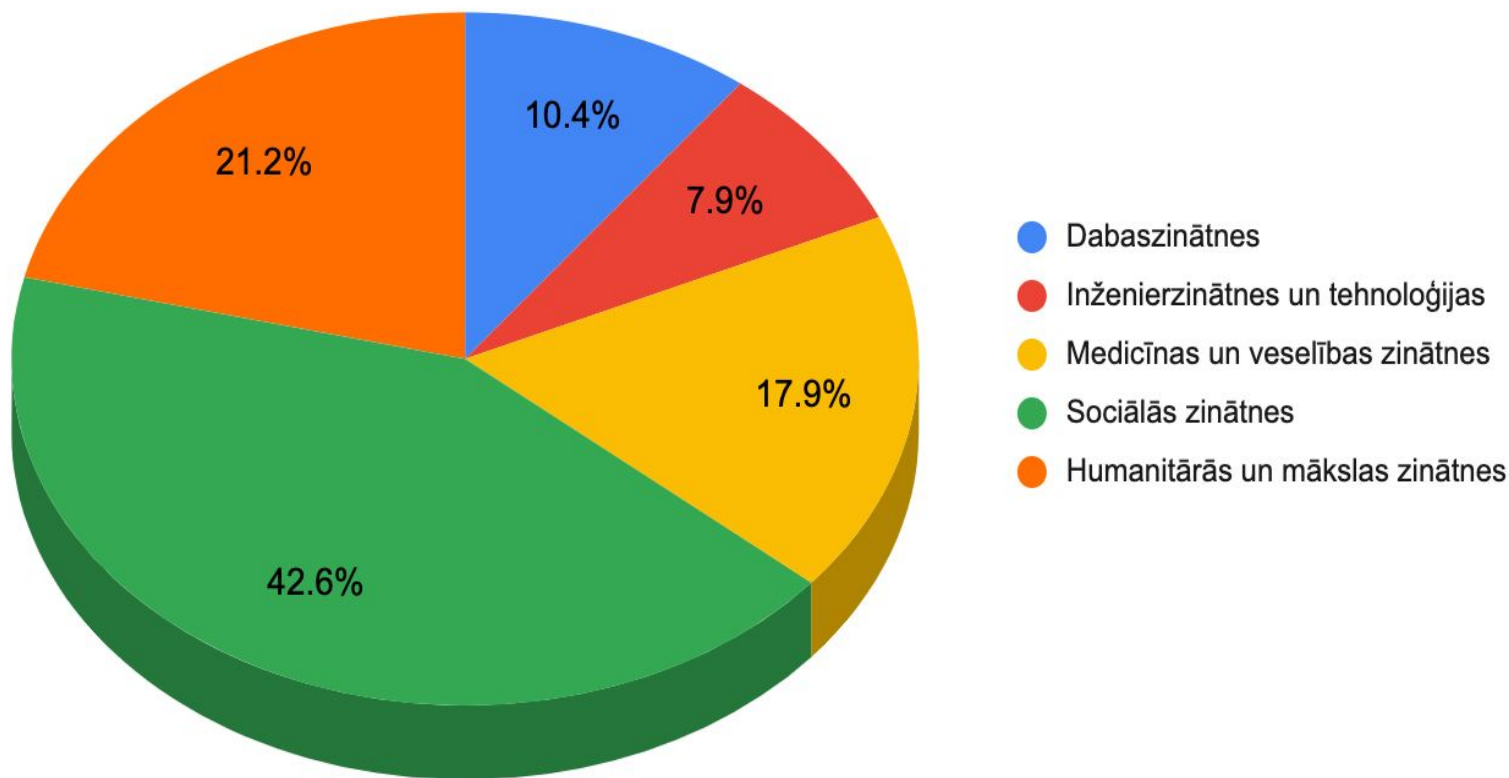


Dzirkstele

Gulbenes novads
✉ reklama@dzirkstele.lv
☎ 26136786

LVK2022 zinātniski teksti (10%)

Kopš 2001. gada; 793 darbi, ne vairāk kā 19 548 vārdlietojumi no katra darba, 506 faili īsāki par limitu



LVK2022 normatīvie akti (7%)

Normatīvie akti no 271 avota (piem., Ministru kabinets, Saeima, Konkurences padome, Smiltenes novada dome u.c.)

19 dokumentu veidi (piem., likums, rīkojums, deklarācija, pavēle, protokols, instrukcija u.c.)

LVK2022 Vikipēdija (7%)

- Raksti, kas garāki par 300 vārdlietojumiem
- Limits – 648 vārdlietojumi
- 4500 no 14105 rakstiem ir vairāk par noteikto limitu

LVK2022 Saeimas stenogrammas (3%), Subtitri (3%)

- Saeimas stenogrammas
- Latvijas sabiedrisko mediju subtitri korpuss (raidījumi, šovi, filmas un seriāli)

Metadati

Katram korpusa avotam tiek pievienoti metadati

☐ Token number	567612	
☒ Document number	96	
☐ Atsauce	Norberts Klaučēns, Nauda, bauda un citas dēkas. Rīga, Madry, 2003.	
☐ ID	d176	
☐ Sadaļa	Daiļliteratūra	

☐ Token number	216885	
☒ Document number	82	
☐ Atsauce	Ilze Boldāne-Zeļeņkova, Etnisko stereotipu veidošanās apstākļi Latvijā: 1850. – 2004. (Vēsture), 2011	
☐ ID	z89	
☐ Sadaļa	Zinātne	

☐ Token number	1208	
☒ Document number	1	
☐ Atsauce	No Saeimas stenogrammām 22.01.2014., Andris Buiķis (Vienotība)	
☐ ID	s27	
☐ Sadaļa	Saeimas stenogrammas	

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 celtnie rātsnama kompleksā bija
d121_0	tramvaja logu lūkdamās uz tukšajiem fabriku	korpusiem	. Kad Millija bija aprakta, Florence
d123_0	, kad sadzirdējām medmāsu kliedzienu 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 kūtīs	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	izpaušmes pilsētas punktus - Spīķerus un VEF	korpusus	, K. Barona ielu un tās apkārtni.