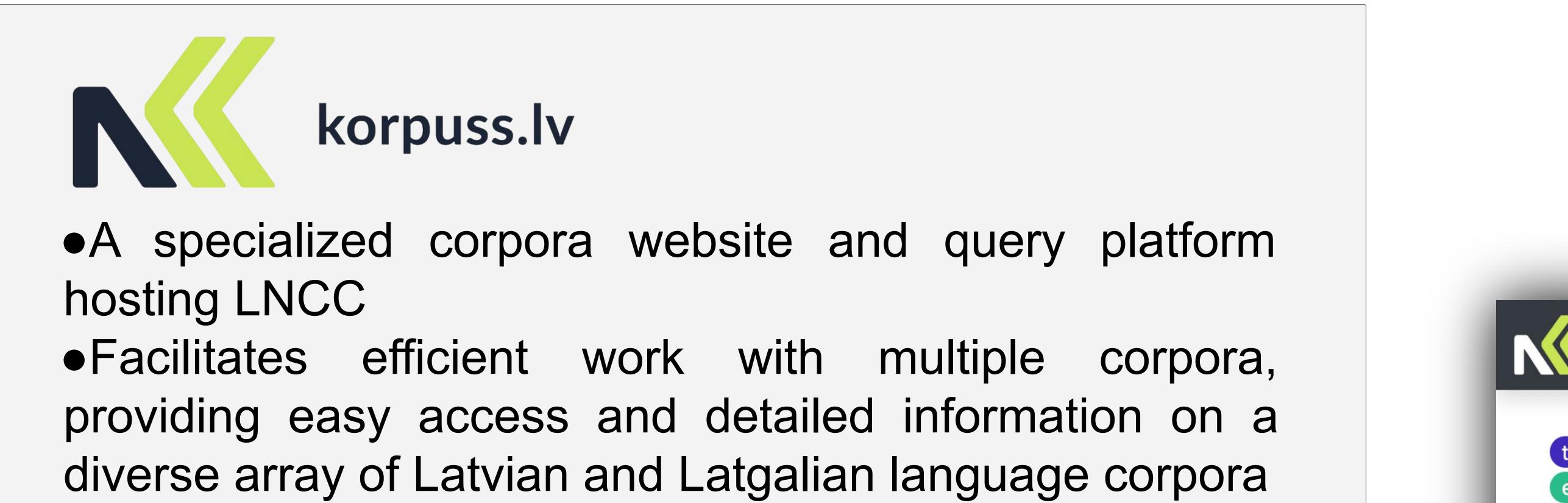
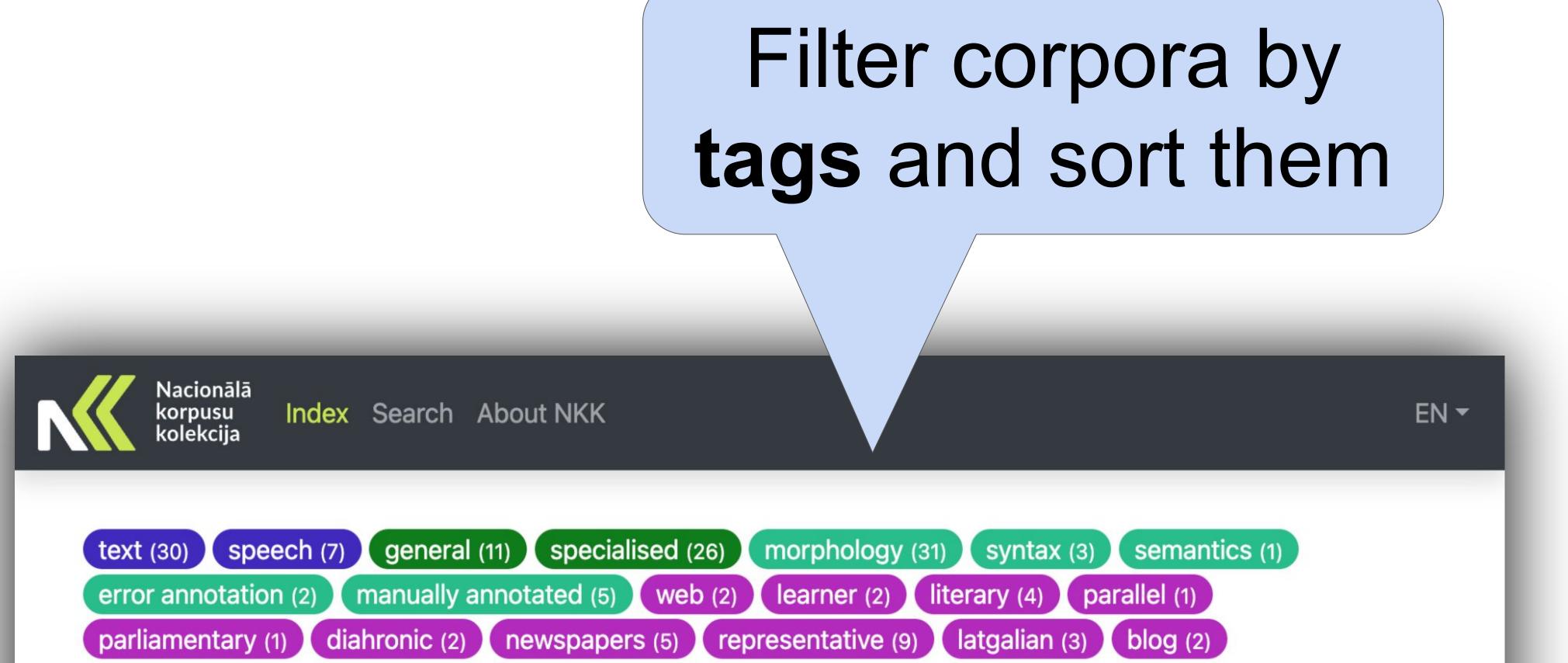
Korpuss.lv – a versatile platform for digital humanities Roberts Darģis, Baiba Saulīte AlLab, IMCS, University of Latvia

Latvian National Corpora Collection (LNCC) is an extensive and diverse collection comprising over 35 text and spoken corpora (totaling 2.8 billion tokens), from two corpus indexer instances (endpoints) maintained by Institute of Mathematics and Computer Science, University of Latvia, and National Library of Latvia. This collection represents a wide range of text types, such as news, blogs, scientific texts, parliamentary debates, colloquial texts, essays and early texts. Notably, almost all corpora in LNCC are re-annotated with a uniform morpho-syntactic annotation scheme, enabling federated search and consistent linguistic analysis across different types and genres.





Information about the corpus and more:

Extensive metadata
References for citation

Relevant publications

Find relevant corpora, term relative and absolute frequency by searching in all corpora at once, using Federated Content Search (FCS)

Nacionālā korpusu kolekcija Index Search About NKK			E				
sirds			Search				
Query <i>sirds</i> returned 326 883 results in 23 of 32 corpora							
Corpus	Relative frequency per 1 million	Absolute frequency	About the corpus				
Pārspriedumi Corpus of Students' Essays	2208	499	More: korpuss.lv/id/Pārspriedumi Developers: IMCS UL, LiepU, RAT Node: nosketch.korpuss.lv				
Senie Corpus of Early Written Latvian Texts	766	2166	More: korpuss.lv/id/Senie Developers: LLI UL, FH UL, IMCS UL Node: nosketch.korpuss.lv				
Rainis Corpus of Texts Written by Rainis	755	1737	More: korpuss.lv/id/Rainis Developers: IMCS UL Node: nosketch.korpuss.lv				
LVMED Latvian Radiology Speech Corpus	616	97	More: korpuss.lv/id/LVMED Developers: IMCS UL, REUH Node: nosketch.korpuss.lv				
BalsuTalka Balsutalka.lv Speech Corpus (Common Voice	397	591	More: korpuss.lv/id/BalsuTalka Developers: IMCS, UL, ILFA UL, LATA Node: nosketch.korpuss.lv				

Corpora with tag specialised (26) Order by: Size -Zinas Barometrs Corpus of News Portal Comments Articles from Latvian news portals 2022, 357.2M words (513.5M tokens) 2011-2022, 26M comments (642M tokens) Developers: RSU, IMCS UL Developers: IMCS UL More Q Search More Search Cīņa Jaunatne "Cīņa" (1904–1991) "Padomju Jaunatne" (1944–1989) 2024, 185M words (231M tokens) 2024, 138M words (176M tokens) **Developers: NLL** Developers: NLL More Q Search More Q Search Likumi LitMāksla Corpus of Legal Acts of the Republic of Latvia "Literatūra un Māksla" 2022, 52.7M words (65.8M tokens) 2022, 73.9M words (116.2M tokens) Developers: IMCS UL Developers: NLL More Q Search More Q Search Most of the corpora can be downloaded from corpus homepage or CLARIN repository

Nacionālā korpusu Index Search About NKK kolekcija EN 🕶 LaVA Q Search Latvian Language Learner Corpus The corpus includes more than 1000 texts created by foreign Latvian language learners studying at Latvian higher education institutions for the first or second semester. The morphologically annotated texts have been checked manually; the language learners' errors have been manually annotated. Citation Publication R. Dargis, I. Auzina, K. Levane-Petrova, I. Kaija Quality Focused Approach to a Learner Corpus Development 2020 PDF Data I. Auzina, I. Kaija, K. Levāne-Petrova, K. Pokratniece, R. Darģis Latvian Language Learner Corpus (LaVA) CLARIN-LV digital library, 2021 http://hdl.handle.net/20.500.12574/42 text (30) specialised (26) learner (2) morphology (31) error annotation (2) manually annotated (5) 192k words (241k tokens) Corpus size Development period 2018–2021 Institute of Mathematics and Computer Science UL Developers Latvian Council of Science, "Development of Learner corpus of Latvian: methods, tools and Funding applications" (lzp-2018/1-0527) http://lava.korpuss.lv/lv/ Homepage http://hdl.handle.net/20.500.12574/42 CLARIN I. Kaija and I. Auzina Other publications Data collection for learner corpus of Latvian: copyright and personal data protection Selected papers from the CLARIN Annual Conference 2019, 41-47, 2020 PDF

N SKETCH ENGINE

- •A corpus management platform facilitating extensive corpus analysis: from frequency lists to condordanes and timelines
- Almost all corpora are morphologically annotated

1,702,347

327,010 -

269,902 -

84,461

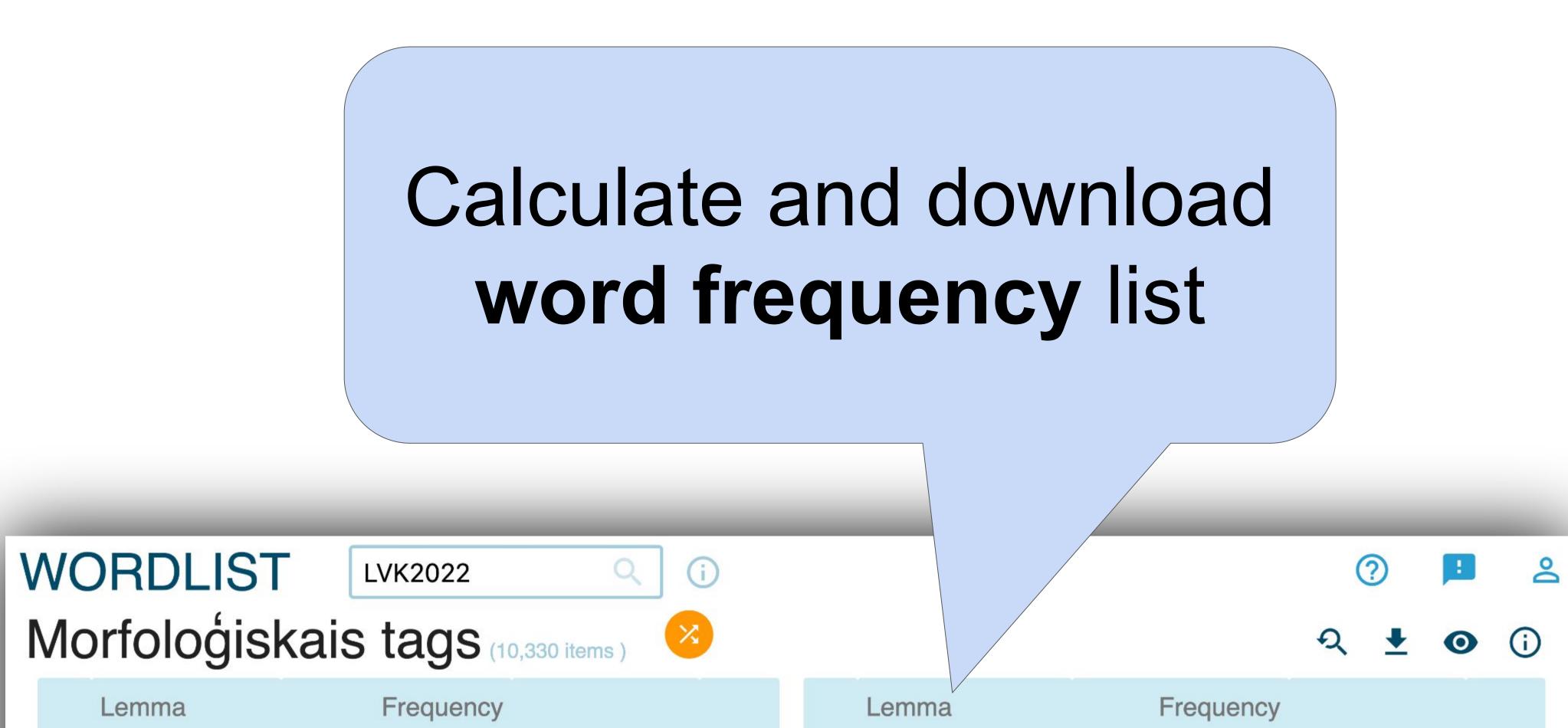
Valsts petijumi

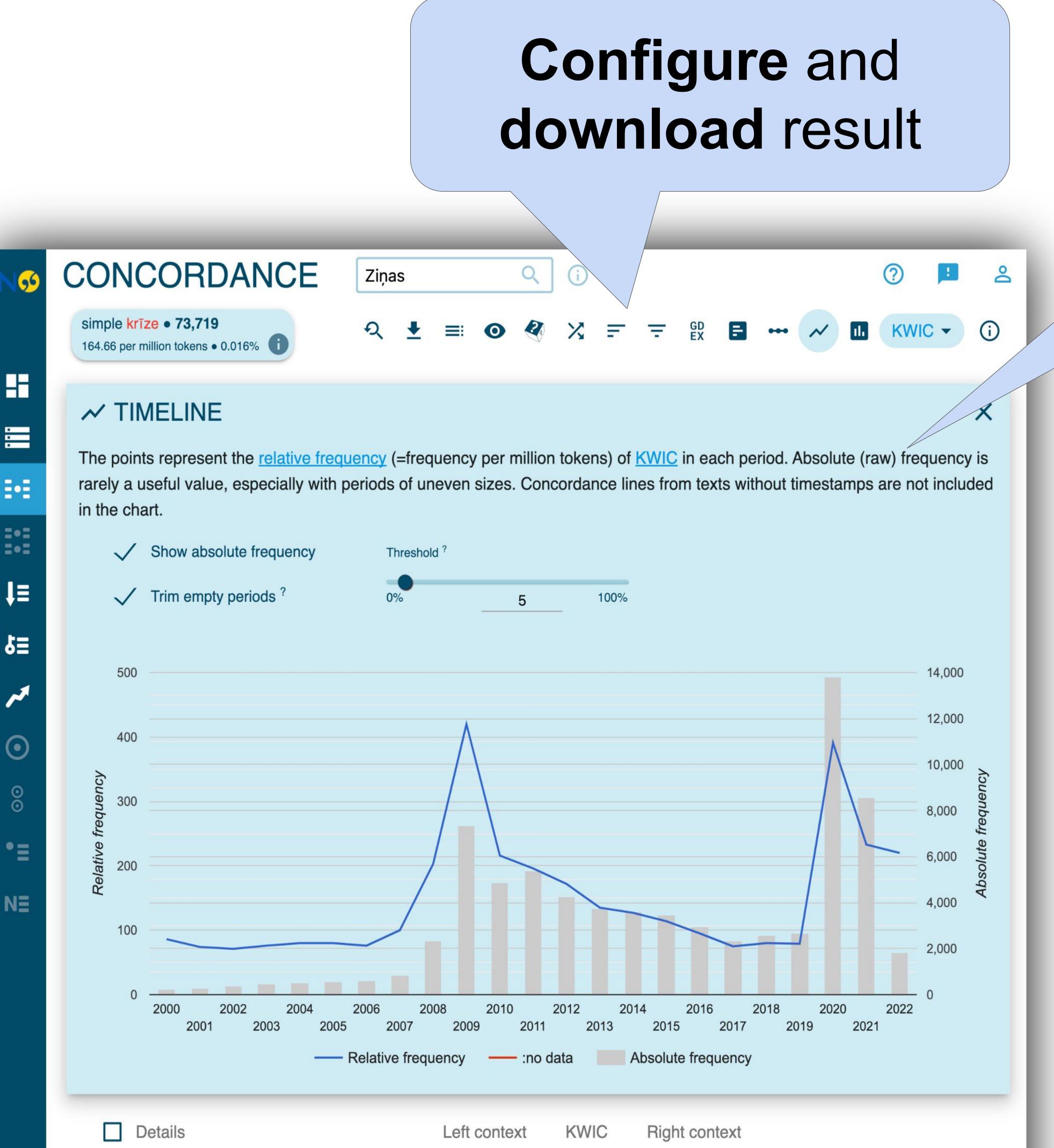
būt

tikt

varēt

norādī





Timeline of relative and absolute frequencies over years for supported corpora

Display and count metadata							
Select the metadata to be displayed in the concordance. Click \Xi to calculate statistics.							
Display above lines ? \checkmark Shorten to <u>15</u> characters							
) . () Tomēr barību nedz cūkas , nedz suņi ar savu darbu nesagādāja ; un viņu bija 个 🗸 🗸							
Document number 🔍 = (1)							
Token number	14663874						
V Document number	156372						
Datums	2007+2007l10	F					
doc.wordcount	269	F					
Dzimums	Male	F					
	2007_10_23_a.txt_seq16	F					

5 veikt	83,897	•••	15 minēt	49,400	•••
6 izmantot	82,580	•••	16 nodrošināt	48,945	•••
7 notikt	80,287	•••	17 uzskatīt	47,600	•••
8 noteikt	75,436	•••	18 sākt	46,320	•••
9 saņemt	58,872	•••	19 kļūt	46,062	•••
10 saistīt	53,649	•••	20 sacīt	44,010	•••

paredzēt

zināt

iegūt

53,446

51,022

50,635

49,808

National Research Programme "Digital Resources of the Humanities"

(VPP-IZM-DH-2020/1-0001; 2020–2022)

...

...

			Left context		riight context		
1	(i) doc#40	par aiziešanu , iņot spsa/par ncfsa4/aiziešana zc/, vm	novērojamas npdfpnppnpn/novērot	krīzes ncfsg5/krīze	pazīmes .	Pēc Birkava d spsg/pēc ncmsg1/birkavs ncf	
2	(i) doc#40	isija nekādā gadījumā emisija pi0msly/nekāds ncmsl1/gadīju		krīzi ncfsa5/krīze		Vienīgais pozitīva afmsnyp/vienīgs afmsnyp/pozi	
3	(i) doc#57	a ministrs .	Atrisinājums ncmsn1/atrisinājums	krīzei ncfsd5/krīze	prokuratūrā eso ncfsl4/prokuratūra vcnrp_i00a	t jāatrod ma n/būt vmnd0t100an/atrast rp_/i	
4	(i) doc#79	konkrēta plāna , afmsgnp/konkrēts ncmsg1/plāns zer	pārdzīvos 230an/pārdzīvot	krīzi ncfsa5/krīze	0	adā tie kļū sl1/gads pd3mpnn/tas vtnifi130a	

Concordances with lemmas, morphological and syntactic annotations if available

Kategorija	09. Saeima+09. SaeimalJaunais laiks	E
Vārds	Ainars Latkovskis	E
Vecums runāšanas brīdī	40	E

Reference and other metadata about each concordance



Funding for the development of the corpus conception, Balanced Corpus of Modern Latvian, etc. (2005–2022)