

General overview

Corpus	Date	Language
hplt-v3-ltg_Latn	9/18/2025	Latgalian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
14,140	218,599	172,971 (79.13 %)	8.1M	45,139,834	46.49 MB

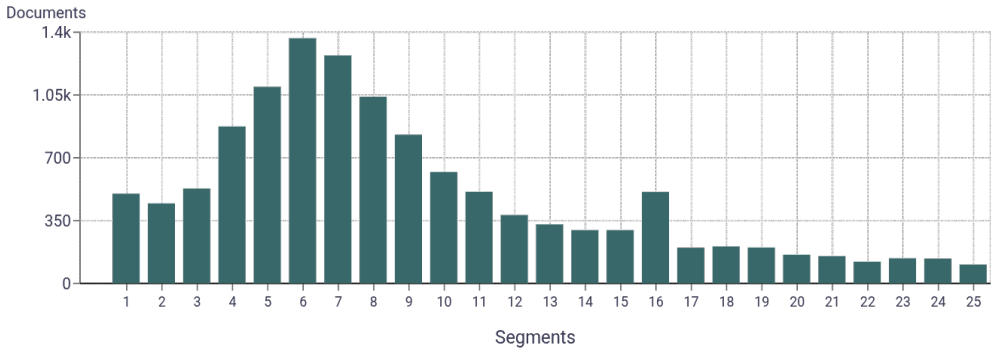
Top 10 domains

Domain	Docs	% of total
lakuga.lv	7.1K	50.01%
lgsc.lv	775	5.48%
ism.lv	716	5.06%
ailab.lv	487	3.44%
wikipedia.org	439	3.10%
daugavpilszinas.lv	279	1.97%
cyxob.lv	271	1.92%
naktineica.lv	242	1.71%
bonuks.lv	238	1.68%
lfk.lv	159	1.12%

Top 10 TLDs

Domain	Docs	% of total
lv	13K	89.44%
org	533	3.77%
cz	314	2.22%
com	228	1.61%
eu	197	1.39%
gov.lv	65	0.46%
ru	45	0.32%
net	31	0.22%
lt	21	0.15%
edu.lv	13	0.09%

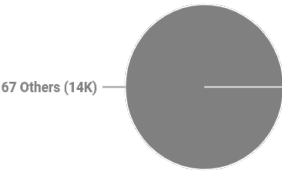
Documents size (in segments) ⓘ



≤ 25 segments **87.14%** (12K documents)
> 25 segments **12.86%** (1.8K documents)

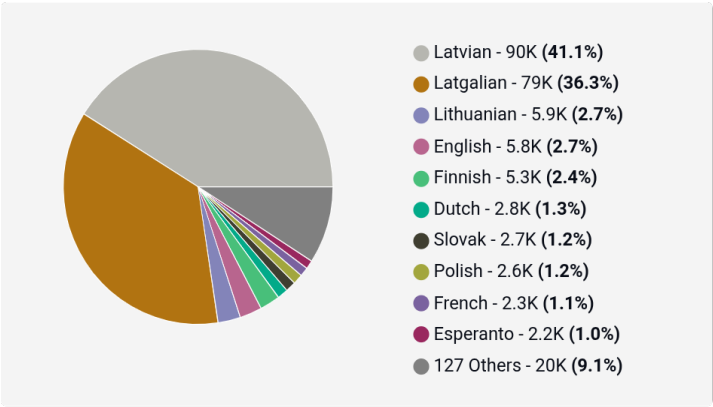
Document collections

CC = **93.63%**
IA = **6.37%**

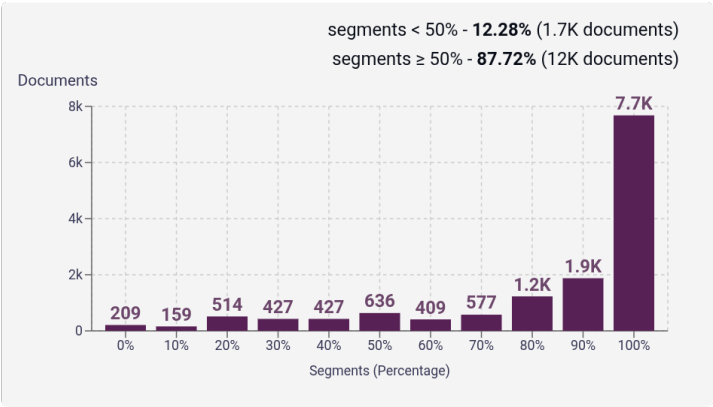


Language Distribution

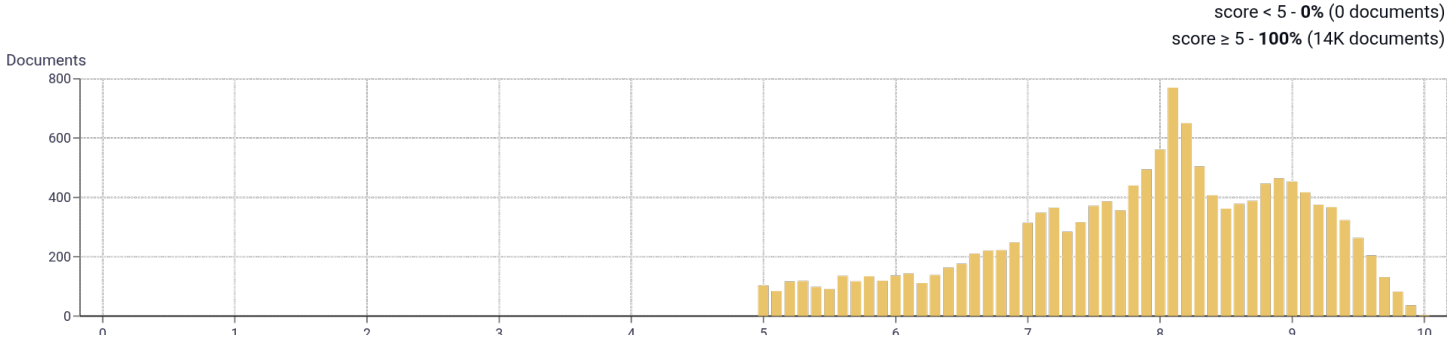
Number of segments in the Latgalian corpus



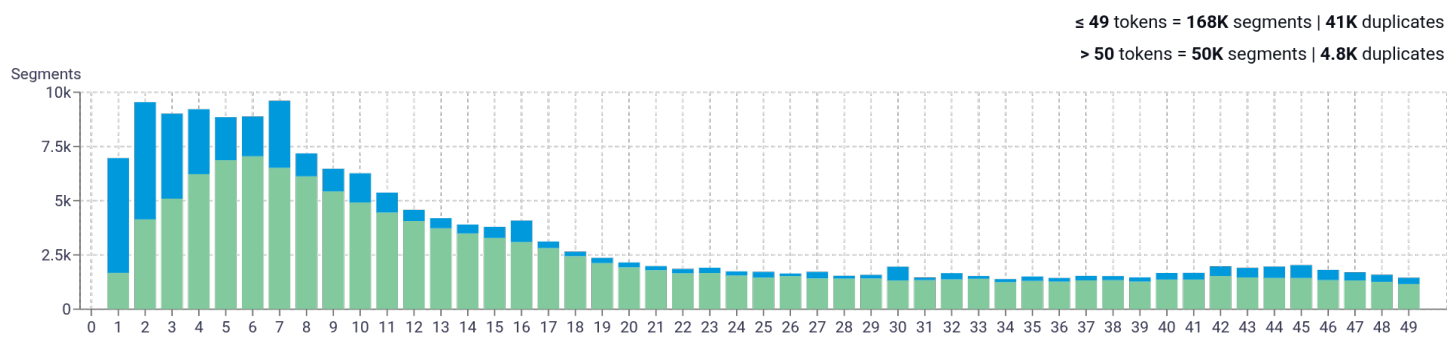
Percentage of segments in Latgalian inside documents



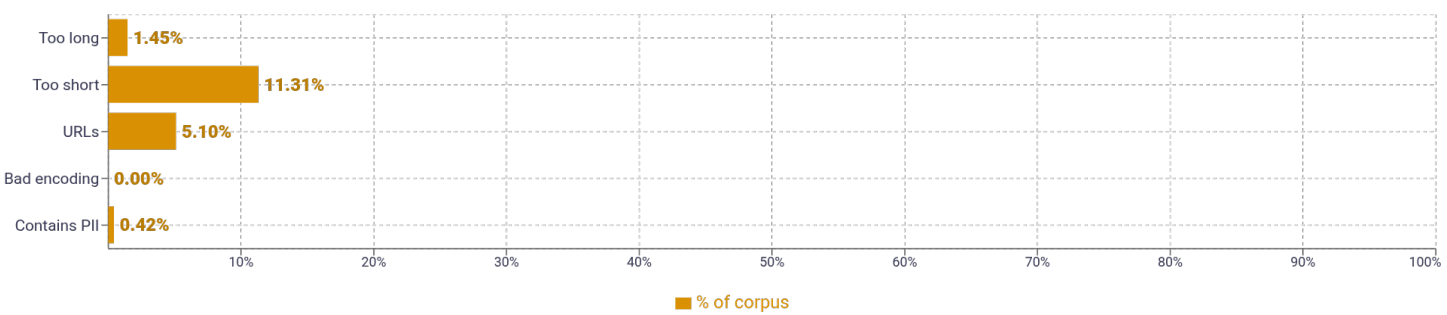
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>i 184,653</div> <div>ir 96,254</div> <div>par 65,376</div> <div>nu 61,608</div> <div>kai 46,514</div>	
2	<div>skaiteit vaira 6,257</div> <div>tys ir 4,330</div> <div>par tū 4,315</div> <div>portals lakuga 3,414</div> <div>latgališu kulturs 3,311</div>	
3	<div>latgolys studentu centrs 1,411</div> <div>latgališu rokstu volūdys 1,246</div> <div>latgališu kulturs goda 1,016</div> <div>sacky do vysavace 896</div> <div>pi myusim latgolā 757</div>	
4	<div>ienāc arī ar savu 622</div> <div>arī ar savu draugiem 622</div> <div>latgališu kulturs goda bolvys 515</div> <div>facebook vai twitter profilu 472</div> <div>viedokli par raidījumā dzirdēto 462</div>	
5	<div>ienāc arī ar savu draugiem 622</div> <div>tomēr patur tiesības dzēst komentārus 462</div> <div>savu viedokli par raidījumā dzirdēto 462</div> <div>raidījumā dzirdēto un atbalsta diskusijas 462</div> <div>radio aicina izteikt savu viedokli 462</div>	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , etc.) replaced by newlines.

Volumes - Tokens

Tokenized with <https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md>

Type-Token Ratio

Lexical variety computed as $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, after removing punctuation (<https://www.sltinfo.com/wp-content/uploads/2014/01/type-token-ratio.pdf>).

Document size (in segments)

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , etc.) replaced by newlines.

Language distribution

Language identified with FastSpell (<https://github.com/mbanon/fastspell>).

Distribution of segments by fluency score

Obtained with Monocleaner (<https://github.com/bitextor/monocleaner>).

Distribution of documents by average fluency score

Obtained with Monocleaner (<https://github.com/bitextor/monocleaner>).

Distribution of documents by document score

Obtained with Web Docs Scorer (<https://github.com/pablop16n/web-docs-scorer/>).

Segment length distribution by token

Tokenized with <https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md>

Segment noise distribution

Obtained with Bicleaner Hardrules (<https://github.com/bitextor/bicleaner-hardrules/>).

Frequent n-grams

Tokenized with <https://github.com/hplt-project/data-analytics-tool/blob/main/tokenizers-info.md>, after removing n-grams starting or ending in a stopword. Stopwords from <https://github.com/hplt-project/data-analytics-tool/blob/main/scripts/resources/README.txt>

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	How-to or instructions	HI	Description of a thing or person	ntp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	Informational persuasion	IP	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	Opinion	OP
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	Informational description	IN	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				