HPLT Analytics report

HPLTAnalytics

General overview

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

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

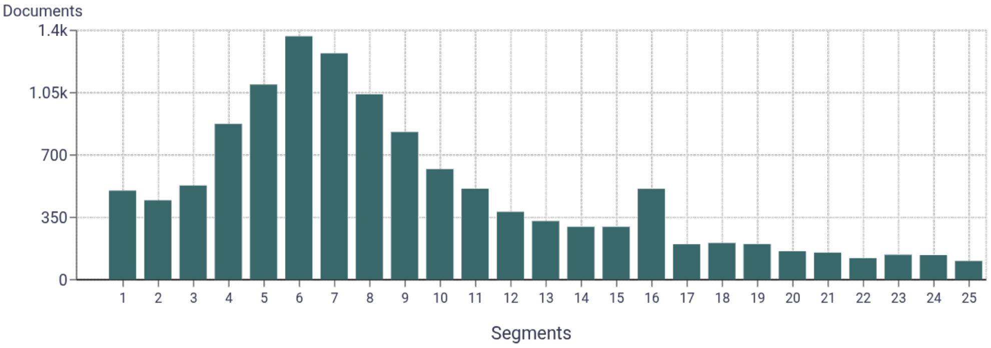
Domain	Docs	% of total
lakuga.lv	7.1K	50.01%
lgsc.lv	775	5.48%
lsm.lv	716	5.06%
ailab.lv	487	3.44%
wikipedia.org	439	3.10%
daugavpilszinas.lv	279	1.97%
cyxob.lv	271	1.92%
naktineica.lv	242	1.71%
bonuks.lv	238	1.68%
lfk.lv	159	1.12%

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%

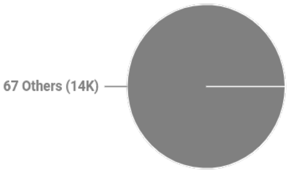
Documents size (in segments) ⓘ

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



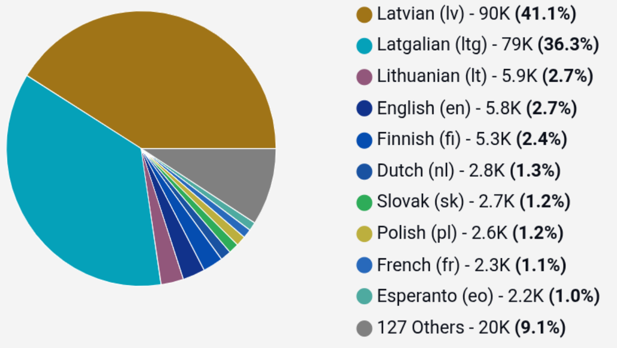
Document collections

CC = 93.63%
IA = 6.37%



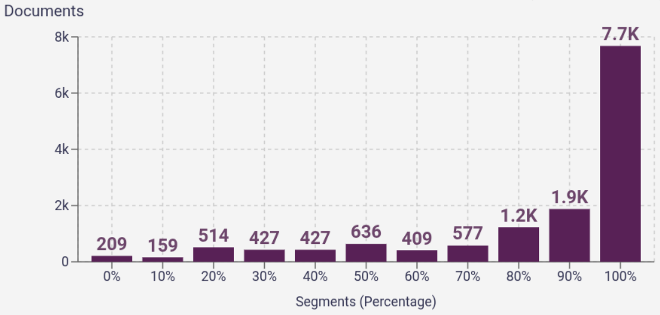
Language Distribution

Number of segments in the Latgalian (ltg) corpus



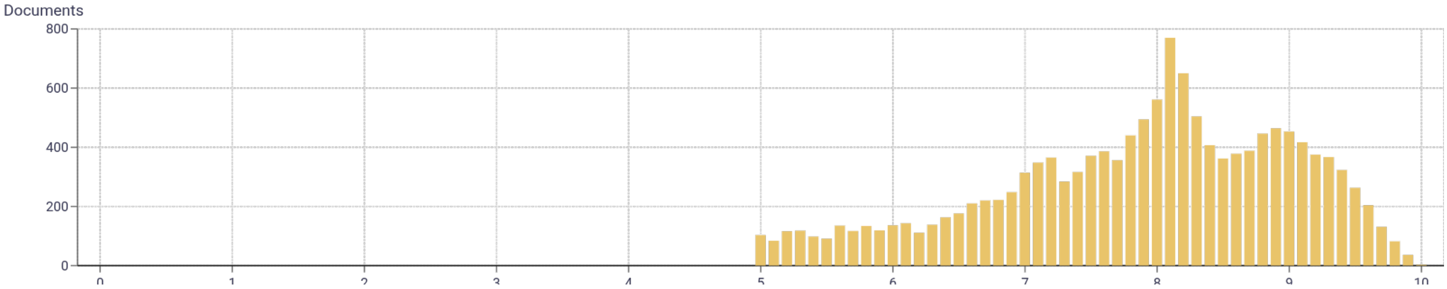
Percentage of segments in Latgalian (ltg) inside documents

segments < 50% - **12.28%** (1.7K documents)
segments ≥ 50% - **87.72%** (12K documents)



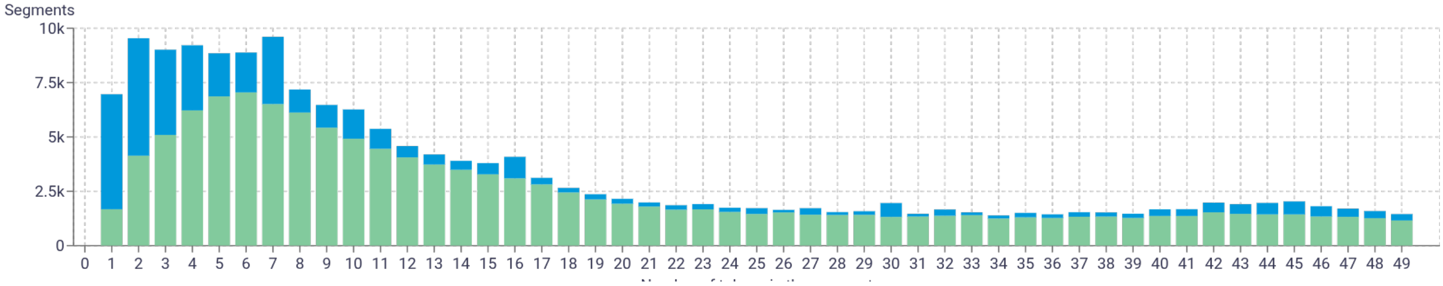
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (14K documents)

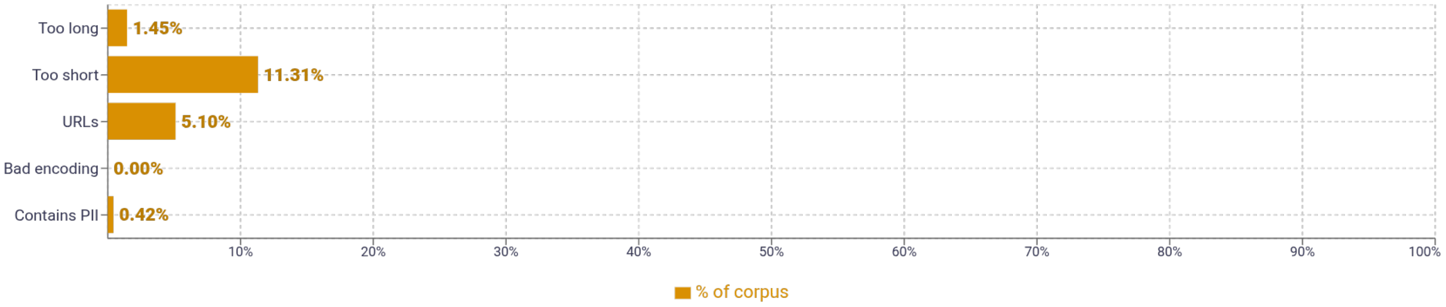


Segment length distribution by token

≤ 49 tokens = **168K** segments | **41K** duplicates
> 50 tokens = **50K** segments | **4.8K** duplicates



Segment noise distribution



Frequent n-grams

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

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, 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 *number or types (uniques)/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>, <ul>, <ol>, 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	dtp
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				