HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
ltg_Latn.jsonl.tsv	12/6/2024	Latgalian (ltg)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
9,209	151,382	77,506 (51.20 %)	4.8M	26,735,255	27.41 MB

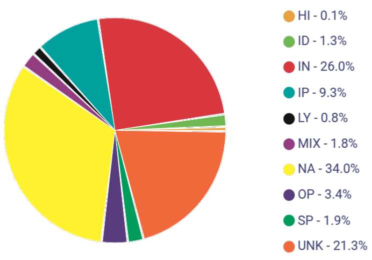
Top 10 domains

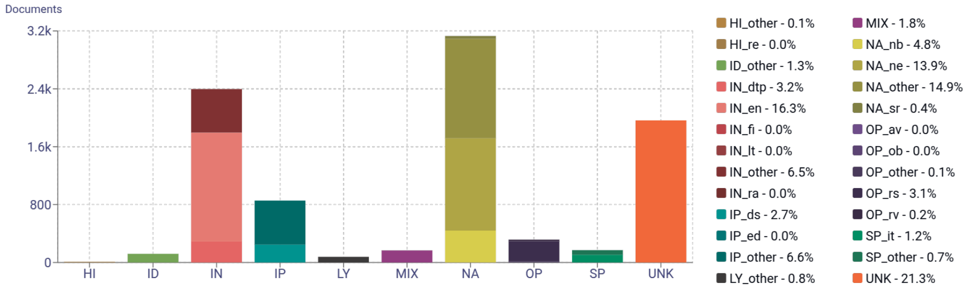
Domain	Docs	% of total
lakuga.lv	3.3K	36.23%
wikipedia.org	1.8K	19.43%
lgsc.lv	1.3K	14.27%
naktineica.lv	352	3.82%
bonuks.lv	295	3.20%
cyxob.lv	268	2.91%
lsm.lv	191	2.07%
rezeknesbibliot...	96	1.04%
sciencegraph.net	89	0.97%
jw.org	80	0.87%

Top 10 TLDs

Domain	Docs	% of total
lv	6.7K	72.84%
org	1.9K	20.76%
com	148	1.61%
eu	143	1.55%
net	108	1.17%
cz	86	0.93%
ru	35	0.38%
gov.lv	16	0.17%
in	12	0.13%
info	7	0.08%

Register labels

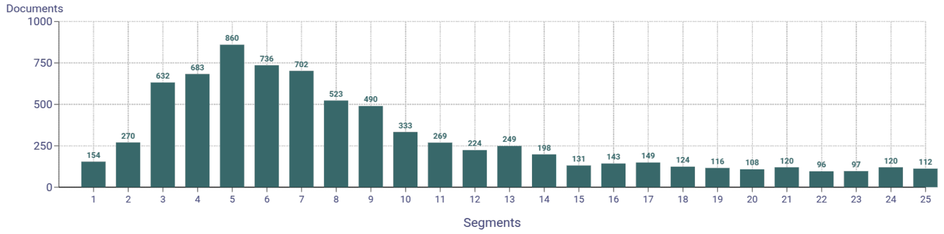




🤖 **MT**:1.5% | 136 Documents

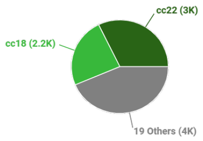
Documents size (in segments)

<= 25 segments **82.95%** (7.6K documents)
> 25 segments **17.05%** (1.6K documents)



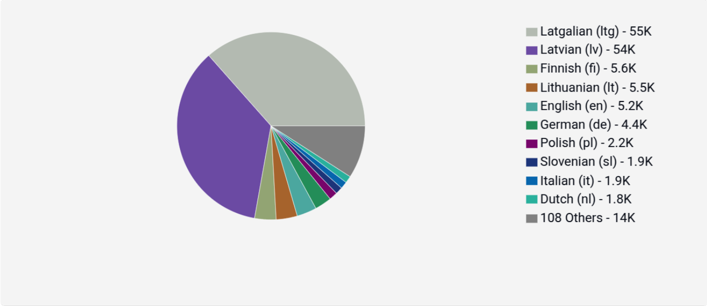
Documents by collection

CC = 75.07%
IA = 24.93%

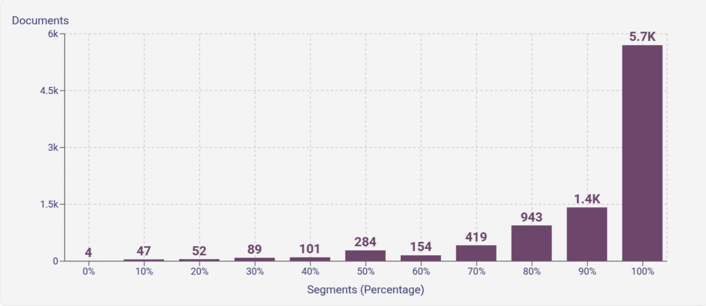


Language Distribution

Number of segments in the Latgalian (ltg) corpus

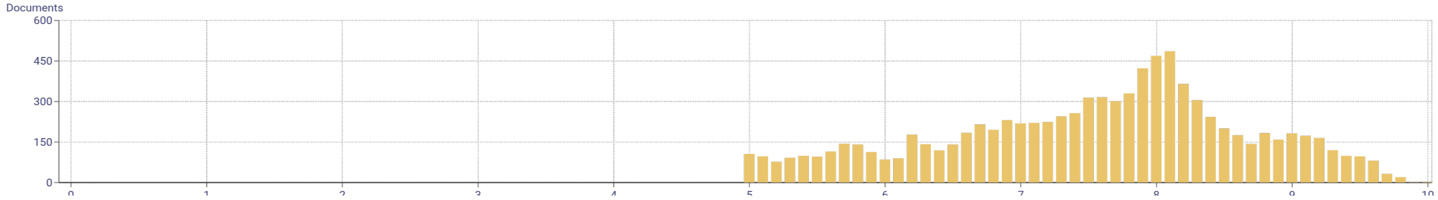


Percentage of segments in Latgalian (ltg) inside documents



Distribution of documents by document score

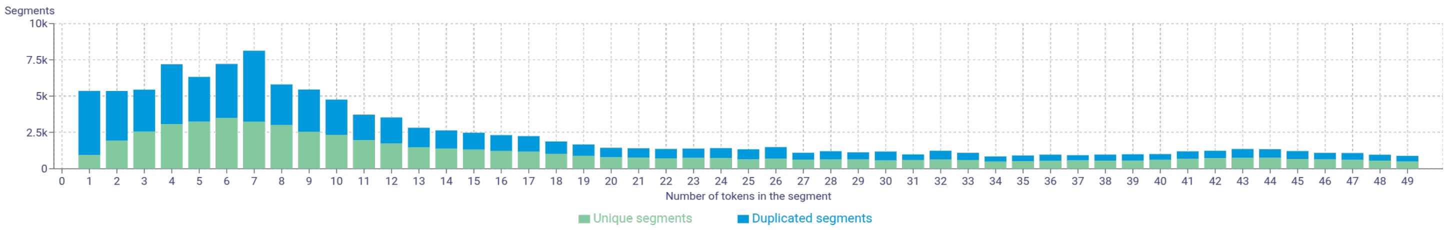
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (9.2K documents)



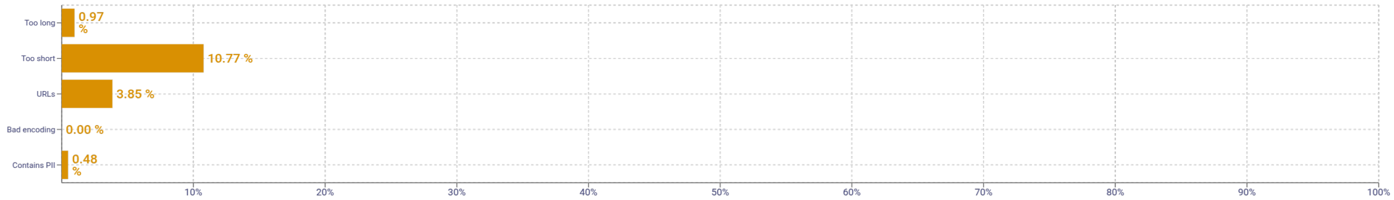
Segment length distribution by token

≤ **49** tokens = **58K** segments | **61K** duplicates

> **50** tokens = **32K** segments | **13K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	i | 118799 ir | 55420 nu | 38709 par | 38067 kai | 28443
2	labot pirmkodu | 2730 tys ir | 2658 latgolys studentu | 2636 par tū | 2421 latgalīšu rokstu | 2093
3	latgolys studentu centrs | 1875 latgalīšu rokstu volūdys | 1133 latgalīšu kulturys goda | 706 pi myusim latgolā | 637 fikys fikys fikys | 579
4	fikys fikys fikys fikys | 578 latgalīšu kulturys goda bolvys | 356 nūvoda teritoriskais padalīņs latgolā | 235 kyskys kys kyskys kys | 232 kys kyskys kys kyskys | 232
5	fikys fikys fikys fikys fikys | 577 kys kyskys kys kyskys kys | 232 kyskys kys kyskys kys kyskys | 224 latgalīšu kulturys ziņu portals lakuga | 206 godā solu reorganizej kai pogostu | 184

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av