HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
lvs_Latn.jsonl.tsv	6/4/2025	Latvian (lvs)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
6,771,633	173,791,676	79,908,770 (45.98 %)	4.2B	25,015,494,704	25.39 GB

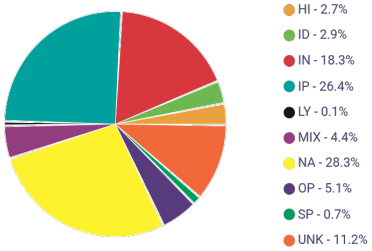
Top 10 domains

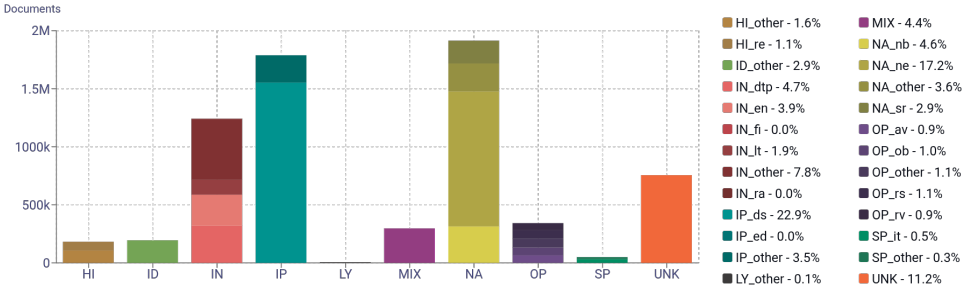
Domain	Docs	% of total
wikipedia.org	253K	3.74%
atlants.lv	199K	2.93%
tvnet.lv	122K	1.80%
skaties.lv	93K	1.37%
hotels.com	89K	1.32%
viss.lv	85K	1.26%
delfi.lv	74K	1.09%
lsm.lv	72K	1.06%
agoda.com	61K	0.90%
blogspot.com	54K	0.79%

Top 10 TLDs

Domain	Docs	% of total
lv	4.8M	70.33%
com	1.1M	16.32%
org	327K	4.82%
gov.lv	126K	1.86%
eu	122K	1.80%
info	57K	0.84%
net	45K	0.66%
ie	19K	0.27%
edu.lv	17K	0.24%
lt	16K	0.23%

Register labels

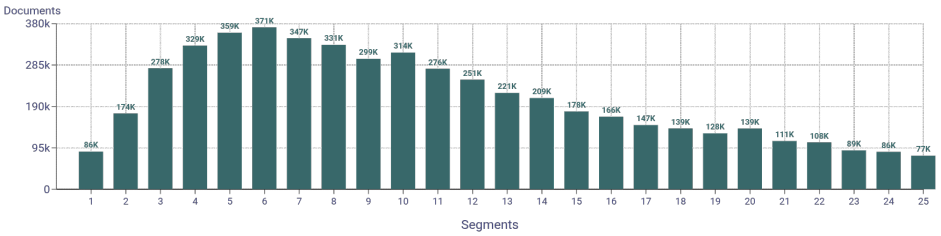




MT:8.6% | 584K Documents

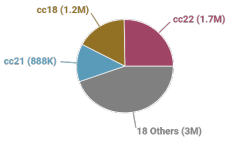
Documents size (in segments)

<= 25 segments **77.02%** (5.2M documents)
> 25 segments **22.98%** (1.6M documents)



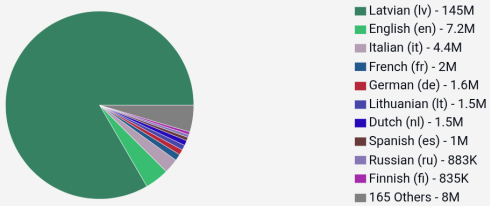
Documents by collection

CC = 68.26%
IA = 31.74%



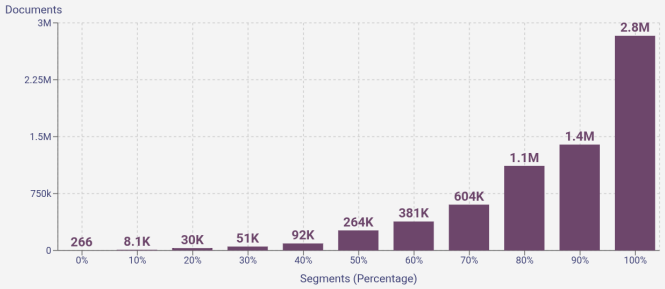
Language Distribution

Number of segments in the Latvian (lvs) corpus



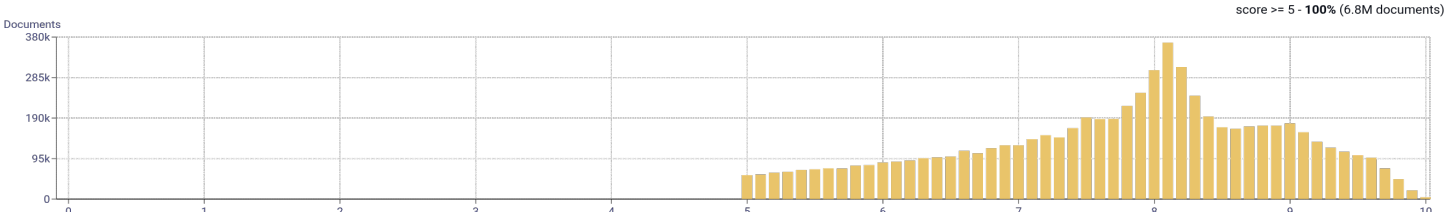
*Latvian (lvs) identification might be inaccurate because it is not supported by FastSpell

Percentage of segments in Latvian (lvs) inside documents



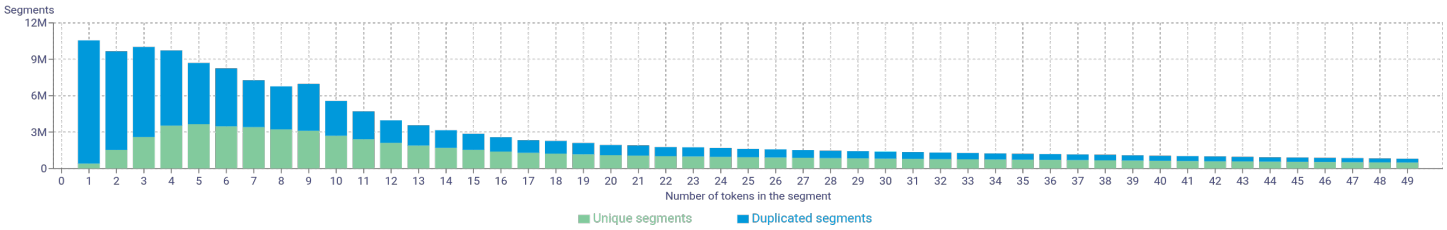
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (6.8M documents)

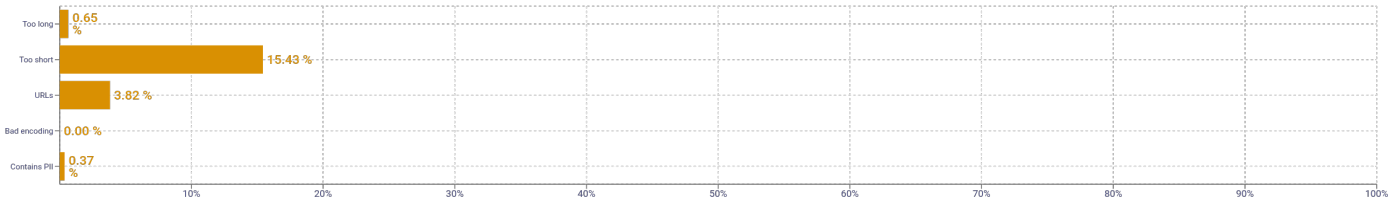


Segment length distribution by token

≤ 49 tokens = **65M** segments | **84M** duplicates
> 50 tokens = **25M** segments | **9.5M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	kas | 23005193 · to | 15137881 · tas | 12521434 · nav | 11485252 · latvijas | 7400622
2	šo uzņēmumu | 1034889 · šo sadaļu | 753720 · labot pirmkodu | 681711 · ņemot vērā | 595807 · labot šo | 585972
3	labot šo sadaļu | 583303 · pēdējās stundas laikā | 258318 · stundas laikā šo | 258117 · viesnīcu ir skatījušas | 258109 · laikā šo viesnīcu | 258109
4	šo viesnīcu ir skatījušas | 258109 · stundas laikā šo viesnīcu | 258109 · pēdējās stundas laikā šo | 258109 · kuru nosūtīt darba saiti | 184171 · eiropas parlamenta un padomes | 70347
5	pēdējās stundas laikā šo viesnīcu | 258109 · laikā šo viesnīcu ir skatījušas | 258109 · vides aizsardzības un reģionālās attīstības | 56856 · rakstu pārpublicēšanas noteikumiem lūdzam kontaktēties | 43110 · noteikumiem lūdzam kontaktēties ar travelnews.lv | 43101

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