

General overview

Corpus	Date	Language
hplt-v3-lvs_Latn	9/18/2025	Latvian (lv)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
11,323,409	296,683,054	173,552,923 (58.50 %)	7.5B	44,342,659,474	45 GB

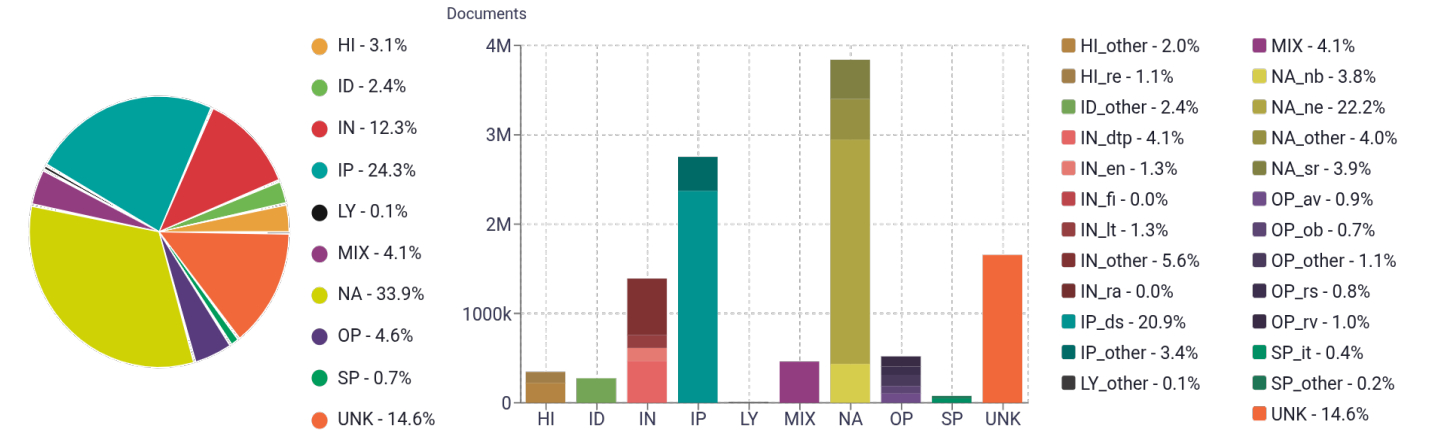
Top 10 domains

Domain	Docs	% of total
lsm.lv	261K	2.30%
delfi.lv	259K	2.29%
tvnet.lv	185K	1.63%
skaties.lv	180K	1.59%
la.lv	114K	1.01%
diena.lv	109K	0.96%
hotels.com	106K	0.94%
wikipedia.org	98K	0.87%
viss.lv	94K	0.83%
nra.lv	85K	0.75%

Top 10 TLDs

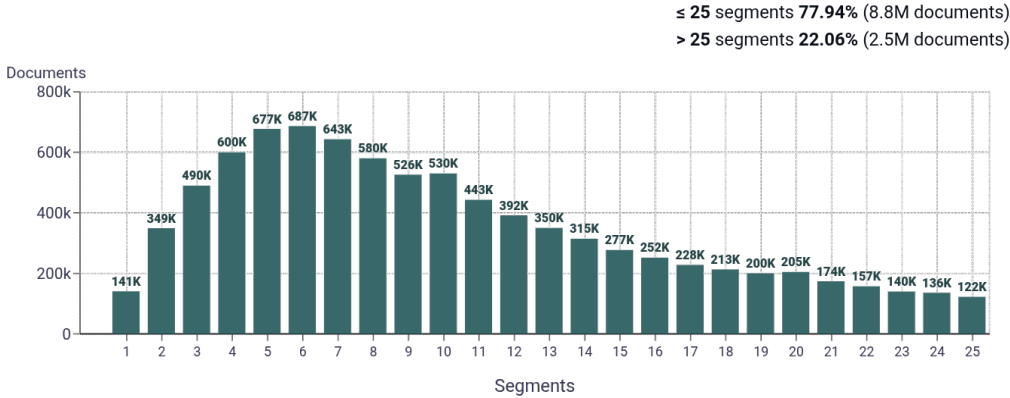
Domain	Docs	% of total
lv	7.9M	69.63%
com	2.1M	18.56%
org	256K	2.26%
eu	229K	2.03%
gov.lv	163K	1.44%
info	94K	0.83%
net	92K	0.81%
pt	39K	0.34%
it	37K	0.33%
ie	29K	0.25%

Register labels

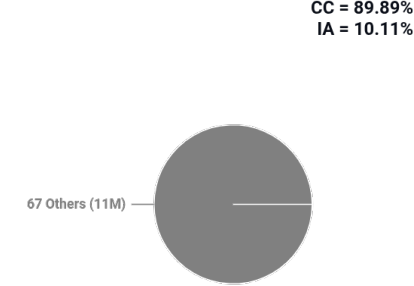


MT:11.7% | 1.3M Documents

Documents size (in segments) ⓘ

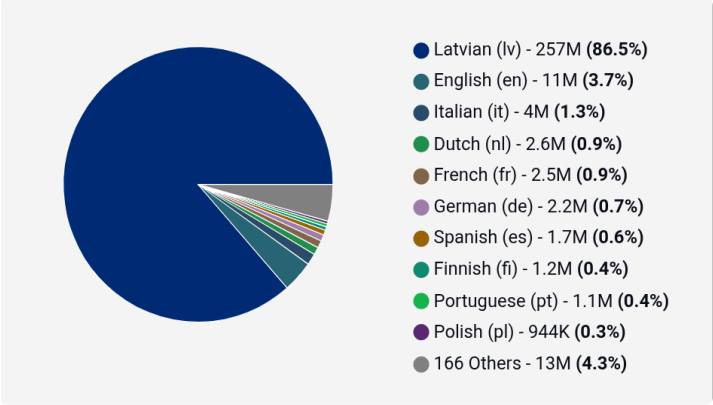


Document collections

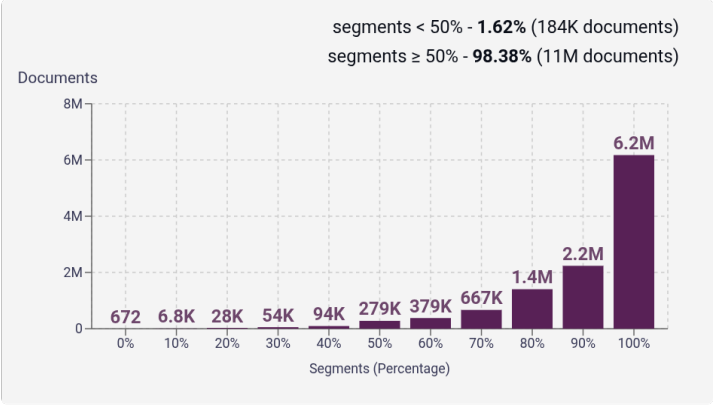


Language Distribution

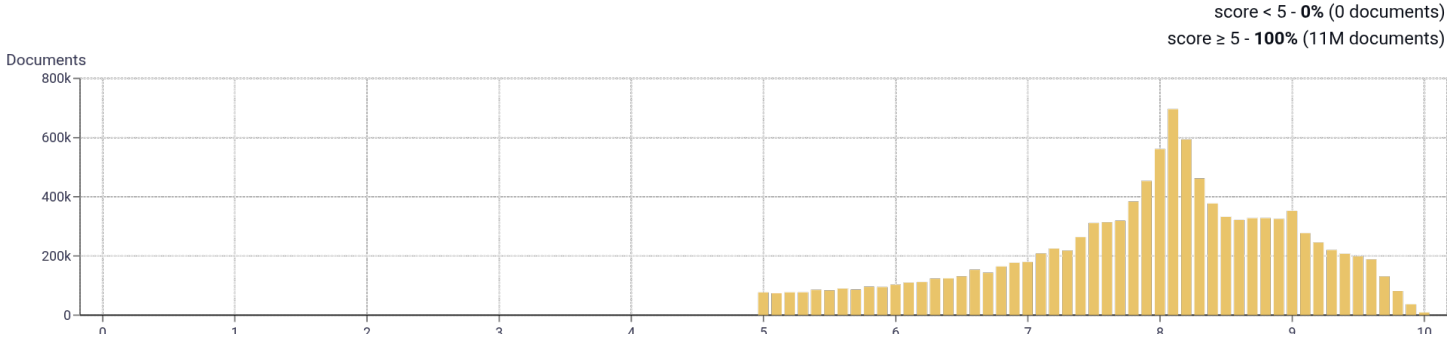
Number of segments in the Latvian (lv) corpus



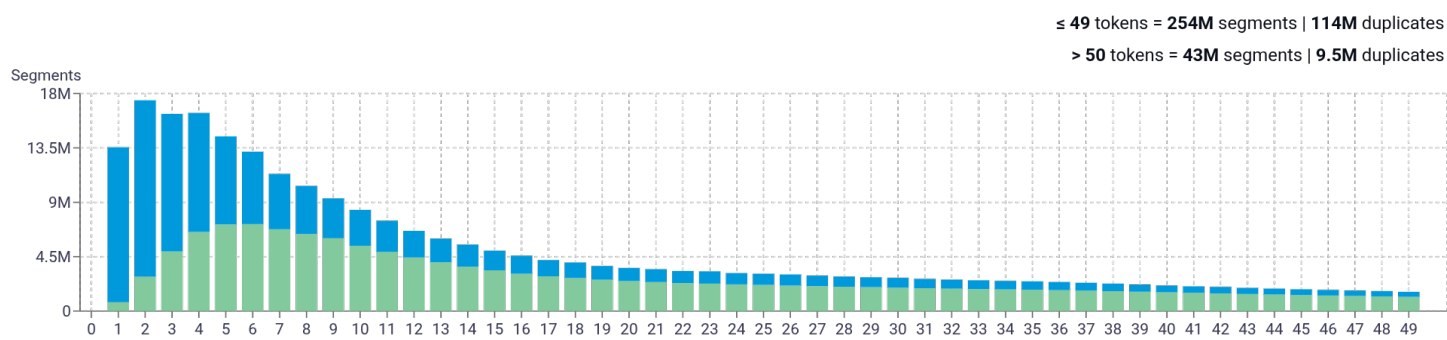
Percentage of segments in Latvian (lv) inside documents



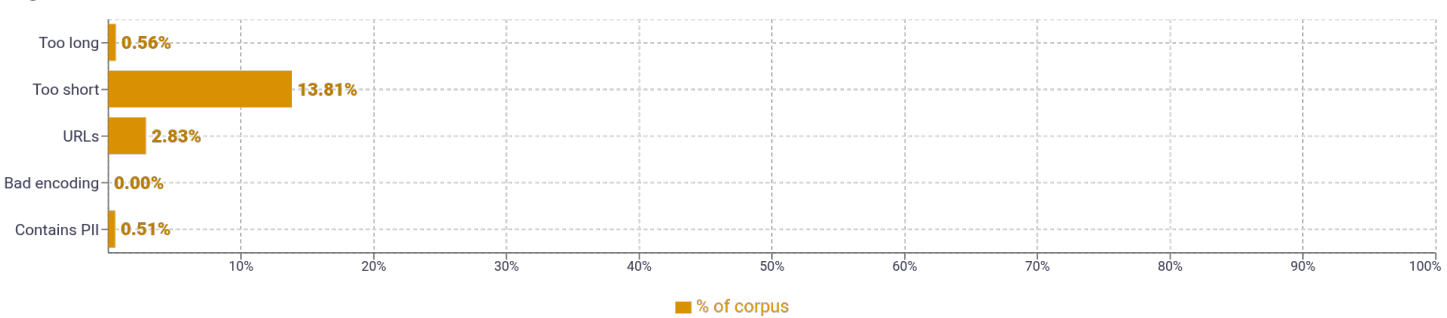
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	to   26,915,826    tas   25,627,246    nav   20,600,209    ko   12,627,152    es   12,149,122	
2	ņemot vērā   1,173,872    šajā gadījumā   959,662    tas nav   886,713    augstas kvalitātes   843,781    tajā pašā   842,340	
3	tajā pašā laikā   606,216    sazinieties ar mums   348,429    neskatoties uz to   262,323    sirds un asinsvadu   260,533    tas ir ļoti   238,934	
4	pavisam jaunu un augstas   129,461    jaunu un augstas kvalitātes   128,101    eiropas parlamenta un padomes   88,890 vides aizsardzības un reģionālās   87,403    aizsardzības un reģionālās attīstības   87,020	
5	pavisam jaunu un augstas kvalitātes   122,208    vides aizsardzības un reģionālās attīstības   85,834    ienāc arī ar savu draugiem.lv   54,118 tas ir saistīts ar faktu   40,841    slimību profilakses un kontroles centra   40,225	

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 of 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				