HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-hrv_Latn	9/24/2025	Croatian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
31,156,692	715,111,681	389,573,075 (54.48 %)	19B	108,423,433,937	103.92 GB

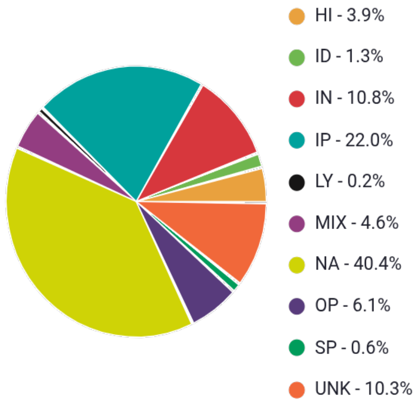
Top 10 domains

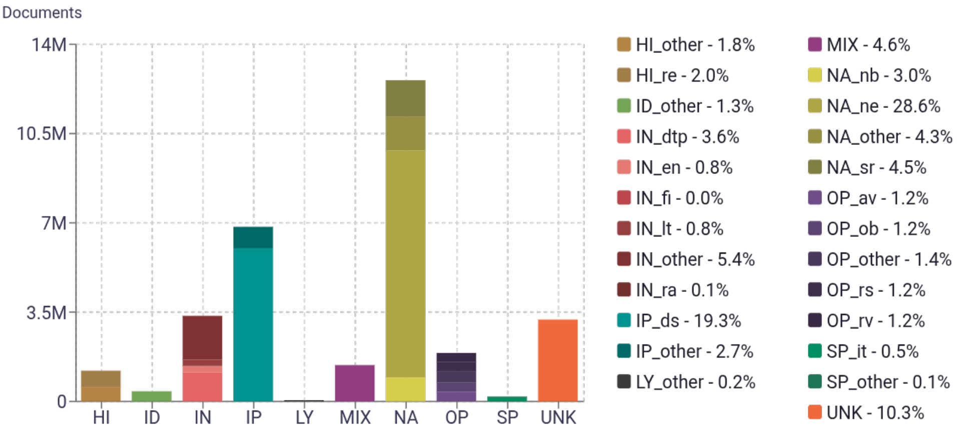
Domain	Docs	% of total
dnevnik.hr	580K	1.86%
tportal.hr	505K	1.62%
index.hr	400K	1.28%
24sata.hr	325K	1.04%
rtl.hr	280K	0.90%
hrt.hr	266K	0.86%
net.hr	257K	0.83%
jutarnji.hr	249K	0.80%
vecernji.hr	246K	0.79%
skole.hr	225K	0.72%

Top 10 TLDs

Domain	Docs	% of total
hr	15M	47.38%
com	8M	25.67%
org	1.4M	4.44%
net	1.3M	4.11%
ba	1.2M	3.72%
info	852K	2.73%
com.hr	642K	2.06%
rs	636K	2.04%
eu	548K	1.76%
news	278K	0.89%

Register labels

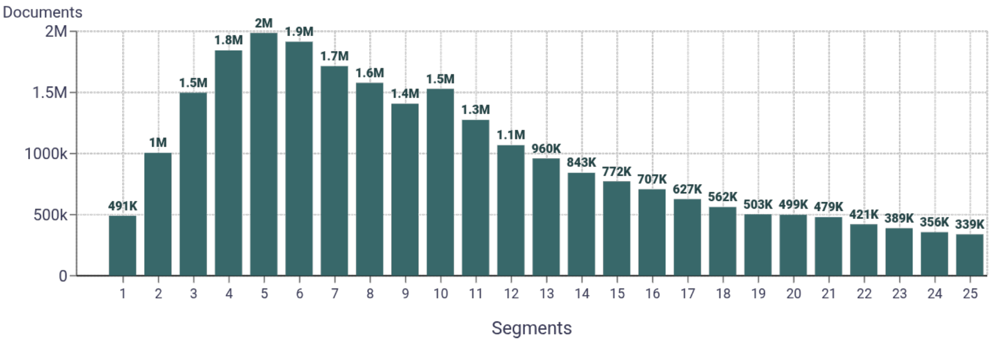




MT:7.9% | 2.5M Documents

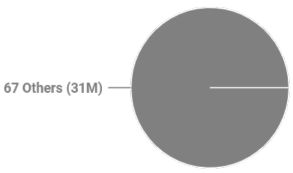
Documents size (in segments)

≤ **25** segments **79.5%** (25M documents)
> **25** segments **20.5%** (6.4M documents)



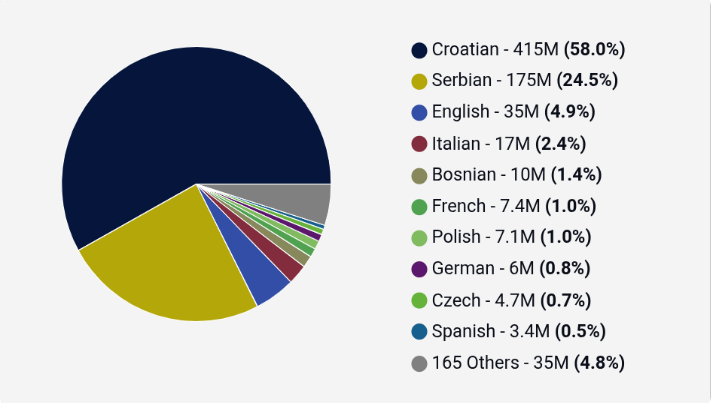
Document collections

CC = 90.76%
IA = 9.24%

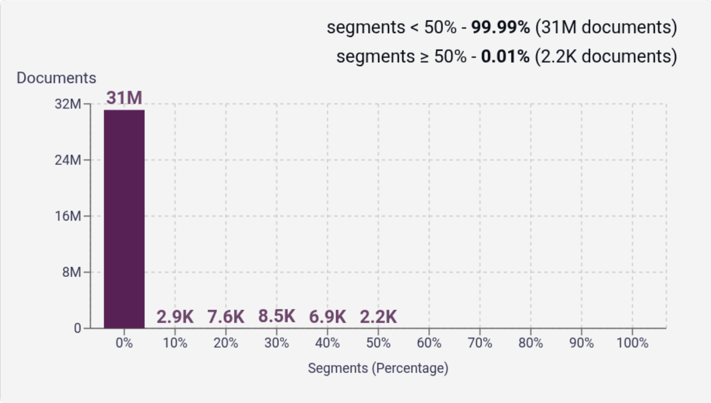


Language Distribution

Number of segments in the Croatian corpus



Percentage of segments in Croatian inside documents

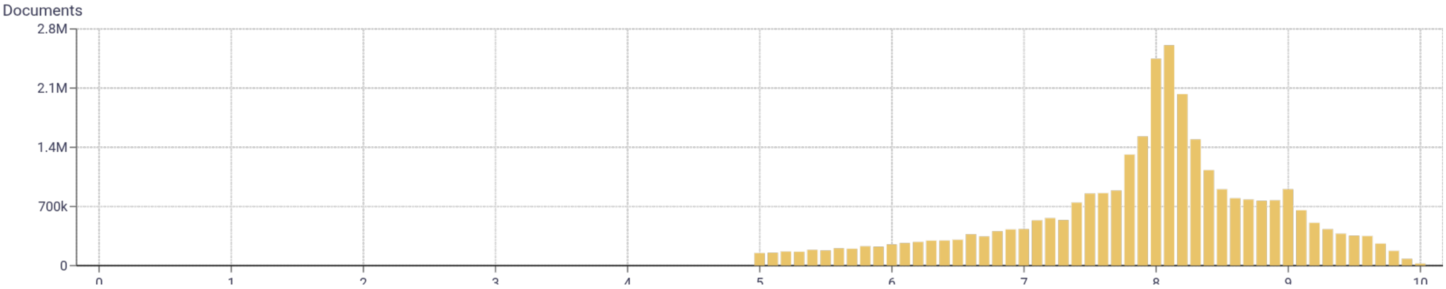

segments < 50% - **99.99%** (31M documents)

segments ≥ 50% - **0.01%** (2.2K documents)

Distribution of documents by document score

score < 5 - **0%** (0 documents)

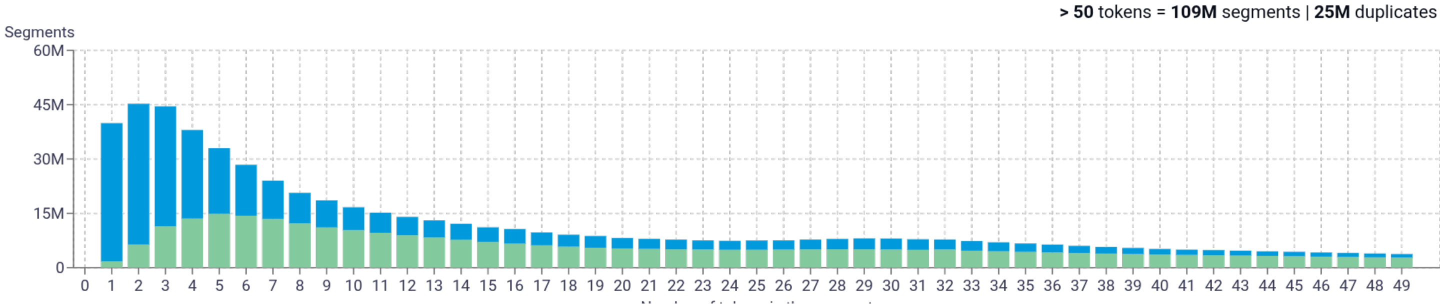
score ≥ 5 - **100%** (31M documents)



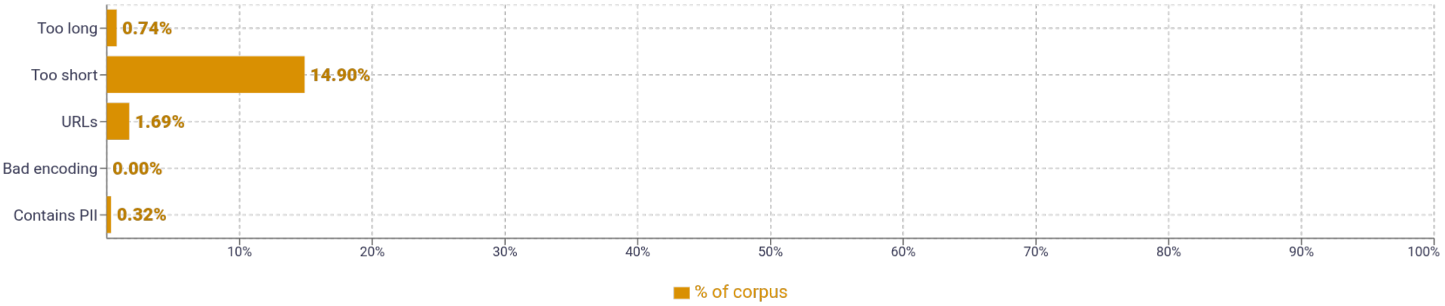
Segment length distribution by token

≤ 49 tokens = **606M** segments | **302M** duplicates

> 50 tokens = **109M** segments | **25M** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>SIZE</th><th>N-GRAMS</th></tr>
<tr><td>1</td><td>godine | 26,546,997
ima | 21,079,219
dana | 19,462,659
bez | 19,298,407
zbog | 18,694,116</td></tr>
<tr><td>2</td><td>ove godine | 3,054,166
osim toga | 2,489,397
bez obzira | 2,138,170
kod kuće | 1,887,456
zbog toga | 1,800,870</td></tr>
<tr><td>3</td><td>članka imaju tag | 950,289
bosne i hercegovine | 595,558
uzeti u obzir | 581,302
imajte na umu | 570,558
prof. dr. sc. | 449,551</td></tr>
<tr><td>4</td><td>imate bilo kakvih pitanja | 261,823
tekst se nastavlja ispod | 238,299
treba imati na umu | 206,687
ovaj komentar kao spam | 198,683
zabrani komentiranje autoru ovog | 198,682</td></tr>
<tr><td>5</td><td>tekst se nastavlja ispod oglasa | 237,401
zabrani komentiranje autoru ovog komentara | 198,682
prijavi ovaj komentar kao spam | 198,682
svojim pametnim telefonima i tabletima | 194,996
pratite nas i na društvenim | 152,077</td></tr>
</table>

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				