HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hrv_Latn.jsonl.tsv	6/9/2025	Croatian (hr)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
12,303,799	297,020,352	134,163,584 (45.17 %)	8.4B	47,710,541,358	45.7 GB

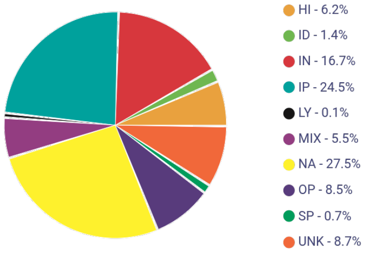
Top 10 domains

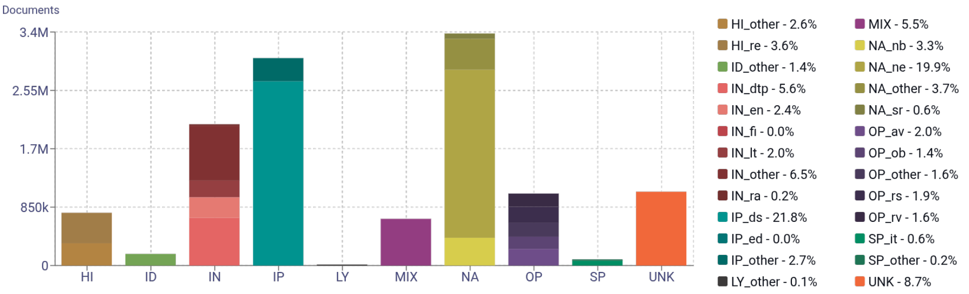
Domain	Docs	% of total
wikipedia.org	264K	2.15%
dnevnik.hr	170K	1.38%
blogspot.com	138K	1.12%
tportal.hr	95K	0.77%
skole.hr	88K	0.72%
jutarnji.hr	84K	0.68%
index.hr	82K	0.67%
metro-portal.hr	73K	0.59%
24sata.hr	66K	0.54%
hrt.hr	59K	0.48%

Top 10 TLDs

Domain	Docs	% of total
hr	5.5M	45.05%
com	3.4M	27.95%
org	592K	4.81%
net	551K	4.48%
ba	421K	3.43%
rs	340K	2.76%
info	288K	2.34%
com.hr	247K	2.01%
eu	219K	1.78%
news	140K	1.14%

Register labels

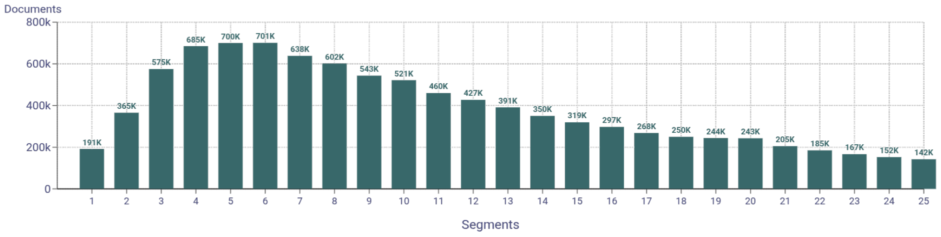




🤖 **MT:**6.0% | 738K Documents

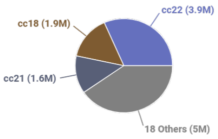
Documents size (in segments)

<= 25 segments **78.21%** (9.6M documents)
> 25 segments **21.79%** (2.7M documents)



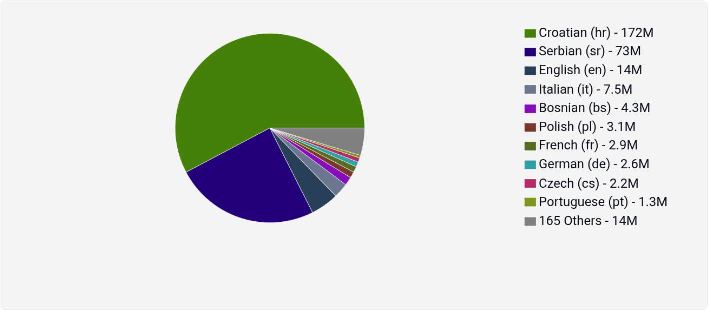
Documents by collection

CC = 69.33%
IA = 30.67%

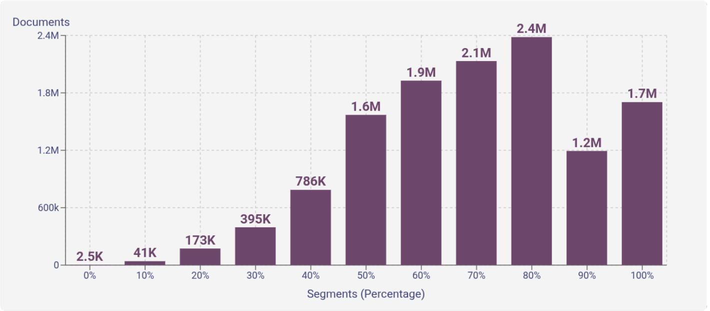


Language Distribution

Number of segments in the Croatian (hr) corpus

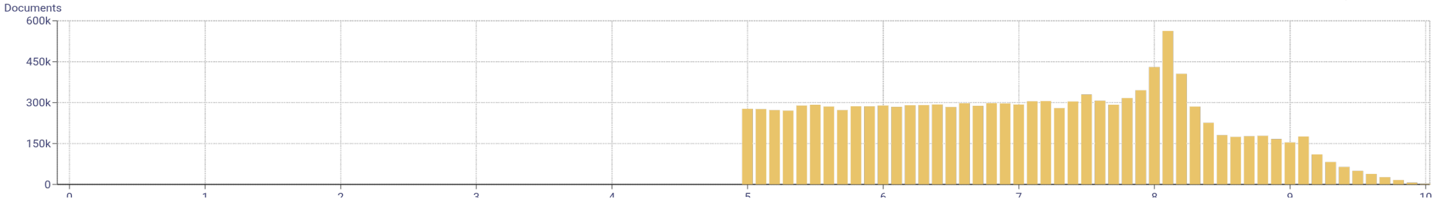


Percentage of segments in Croatian (hr) inside documents



Distribution of documents by document score

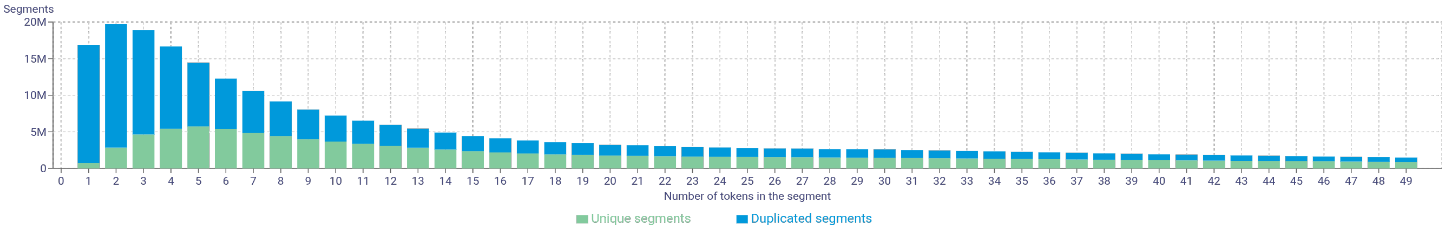
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (12M documents)



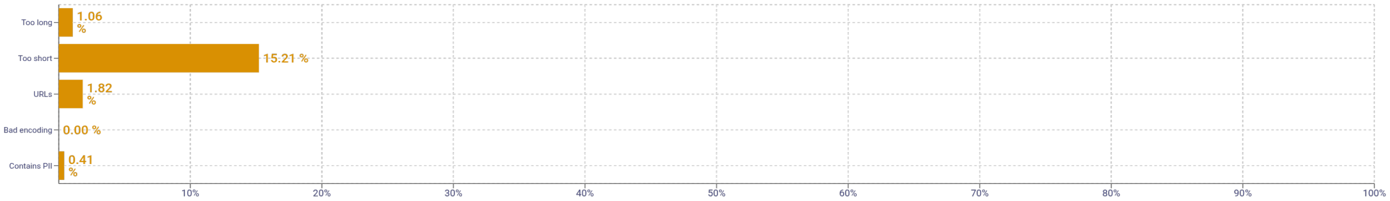
Segment length distribution by token

≤ **49** tokens = **103M** segments | **142M** duplicates

> 50 tokens = **52M** segments | **21M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	godine | 11286744 ima | 9006942 dana | 8338453 bez | 8247403 prema | 8048470
2	republike hrvatske | 1076977 ove godine | 1058244 bez obzira | 973230 zbog toga | 929285 osim toga | 911780
3	dana od dana | 270107 bosne i hercegovine | 220014 prof. dr. sc. | 216626 imajte na umu | 210093 članka imaju tag | 206360
4	imate bilo kakvih pitanja | 91281 stupanja na snagu ovoga | 75062 treba imati na umu | 71039 mijenja se i glasi | 69479 e-mail adresa je zaštićena | 66209
5	stupanja na snagu ovoga zakona | 66212 adresa je zaštićena od spambota | 55837 omogućiti javascript da je vidite | 55755 preporučite ga prijateljima putem ovih | 45588 svojim pametnim telefonima i tabletima | 41990

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