HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-srp_Cyrl	10/27/2025	Serbian

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
7,081,710	169,842,312	101,707,300 (59.88 %)	40.12%	5B	27,521,554,075	45.96 GB

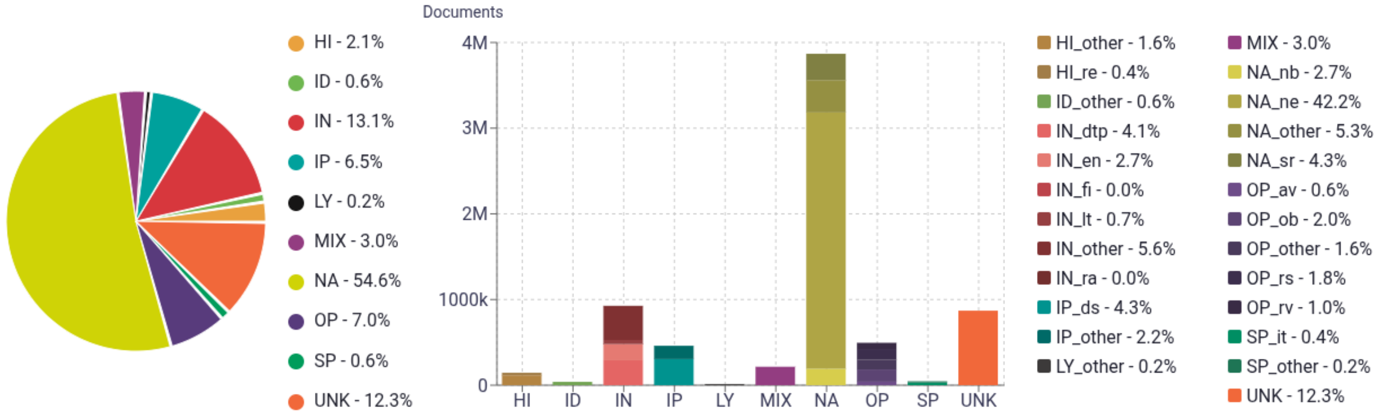
Top 10 domains

Domain	Docs	% of total
novosti.rs	202K	2.85%
wikipedia.org	153K	2.16%
sputniknews.com	139K	1.97%
politika.rs	114K	1.61%
rts.rs	103K	1.46%
juznasrbija.info	81K	1.14%
vostok.rs	79K	1.12%
srbin.info	72K	1.02%
mojenovosti.com	68K	0.96%
rtrs.tv	65K	0.92%

Top 10 TLDs

Domain	Docs	% of total
rs	2.1M	30.01%
com	2.1M	29.01%
org	647K	9.14%
net	377K	5.32%
info	313K	4.42%
org.rs	247K	3.48%
edu.rs	180K	2.54%
gov.rs	156K	2.20%
co.rs	116K	1.64%
ac.rs	80K	1.13%

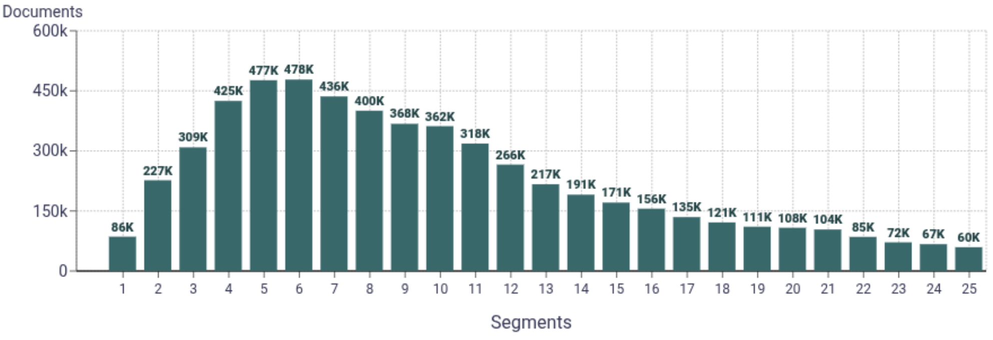
Register labels



MT:8.7% | 617K Documents

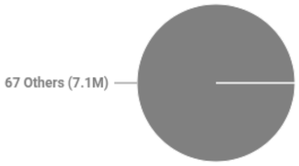
Documents size (in segments)

≤ **25** segments **81.16%** (5.7M documents)
> **25** segments **18.84%** (1.3M documents)



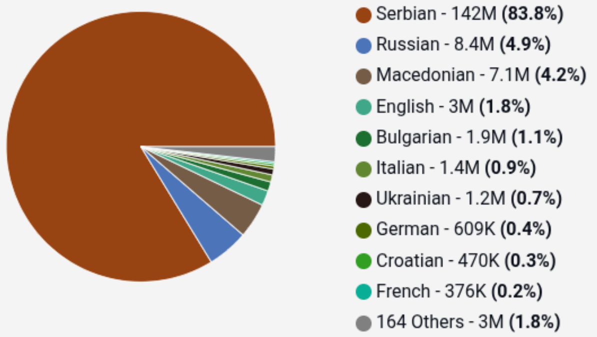
Document collections

CC = 92.15%
IA = 7.85%



Language Distribution

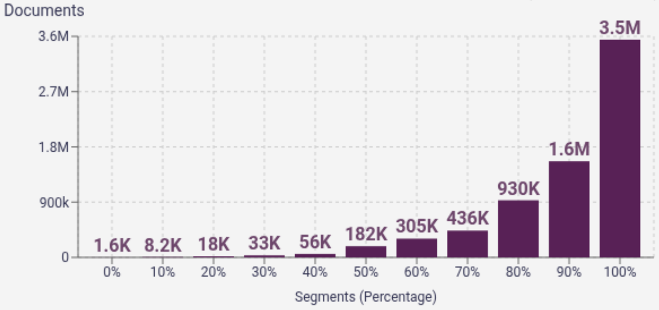
Number of segments in the Serbian corpus



Percentage of segments in Serbian inside documents

segments < 50% - **1.64%** (116K documents)

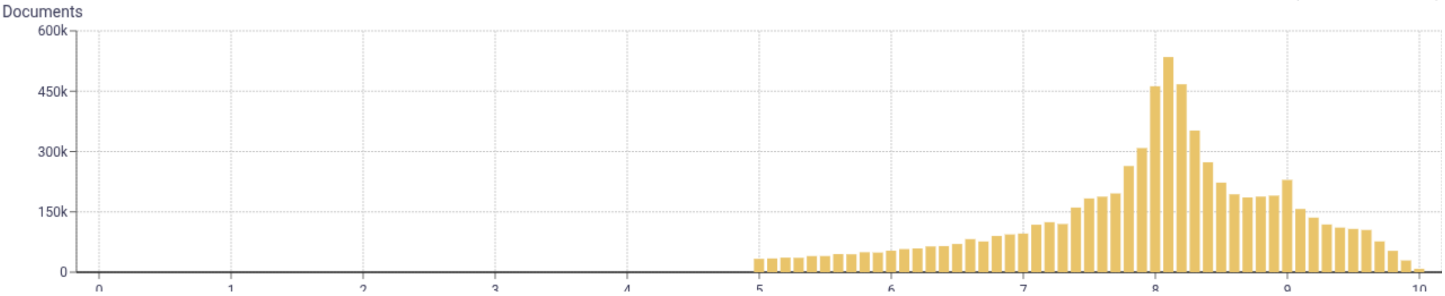
segments ≥ 50% - **98.36%** (7M documents)



Distribution of documents by document score

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

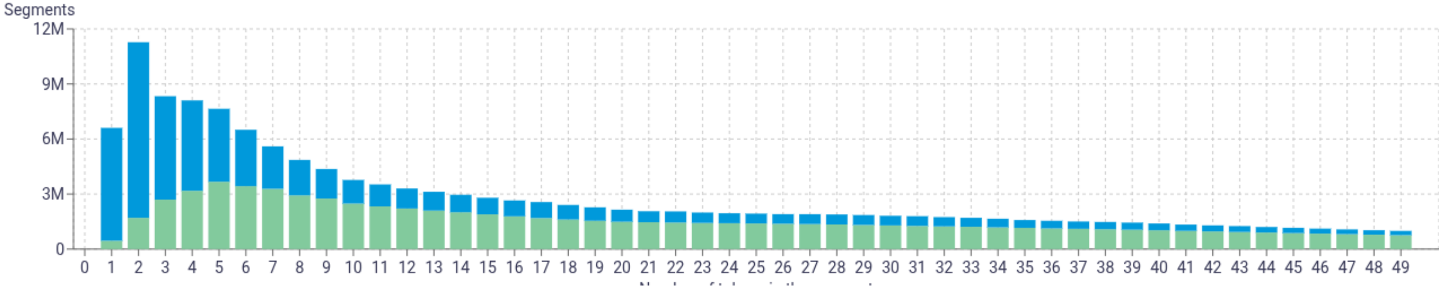
score ≥ 5 - **100%** (7.1M documents)



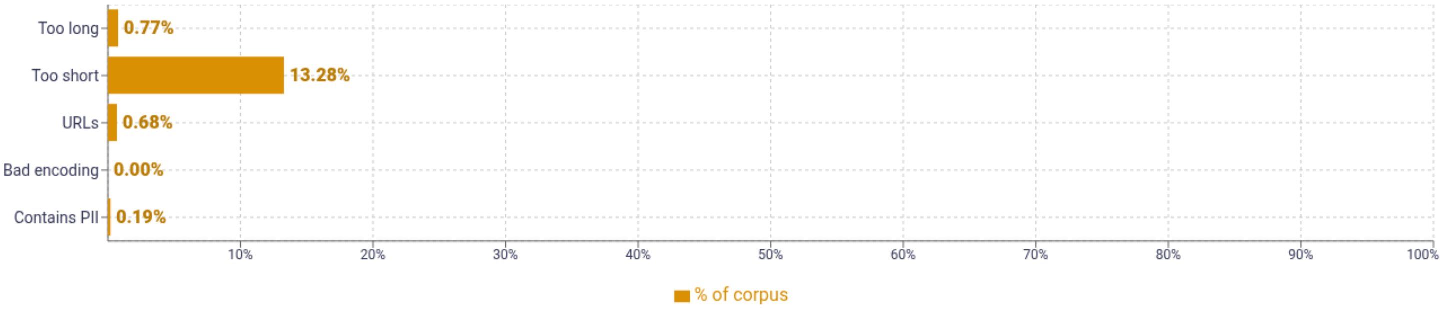
Segment length distribution by token

≤ 49 tokens = **141M** segments | **62M** duplicates

> 50 tokens = **28M** segments | **6.1M** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	године | 9,938,251 како | 9,450,200 није | 9,272,392 све | 8,464,440 која | 8,269,916
2	републике српске | 852,476 због тога | 824,573 погледај још | 702,473 пре него | 650,862 републике србије | 636,443
3	косову и метохији | 194,462 српске православне цркве | 192,980 косова и метохије | 167,529 другог светског рата | 157,730 босне и херцеговине | 153,092
4	наводи се у саопштењу | 84,425 председник србије александар вучић | 75,955 науке и технолошког развоја | 66,369 остале новости из рубрике | 57,569 оставите ваш коментар објавите | 57,569
5	оставите ваш коментар објавите новост | 57,563 agencija za brak i druzenje | 47,810 видљив чим га администратор одобри | 35,890 бити видљив чим га администратор | 35,890 коментар ће бити видљив чим | 35,889

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				