

General overview

Corpus	Date	Language
hplt-v3-bos_Latn	9/24/2025	Bosnian (bs)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
37,078,307	641,274,904	369,534,314 (57.62 %)	18B	98,661,298,393	94.65 GB

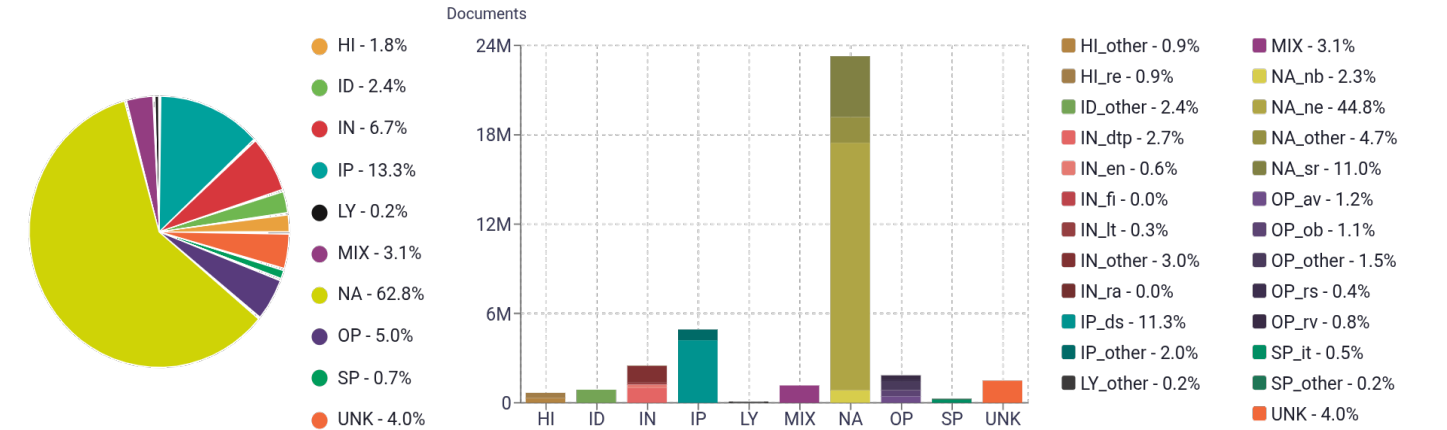
Top 10 domains

Domain	Docs	% of total
klix.ba	739K	1.99%
blic.rs	499K	1.34%
mondo.rs	485K	1.31%
krstarica.com	361K	0.97%
vesti.rs	309K	0.83%
novosti.rs	260K	0.70%
blogspot.com	257K	0.69%
b92.net	256K	0.69%
republika.rs	240K	0.65%
vijesti.me	238K	0.64%

Top 10 TLDs

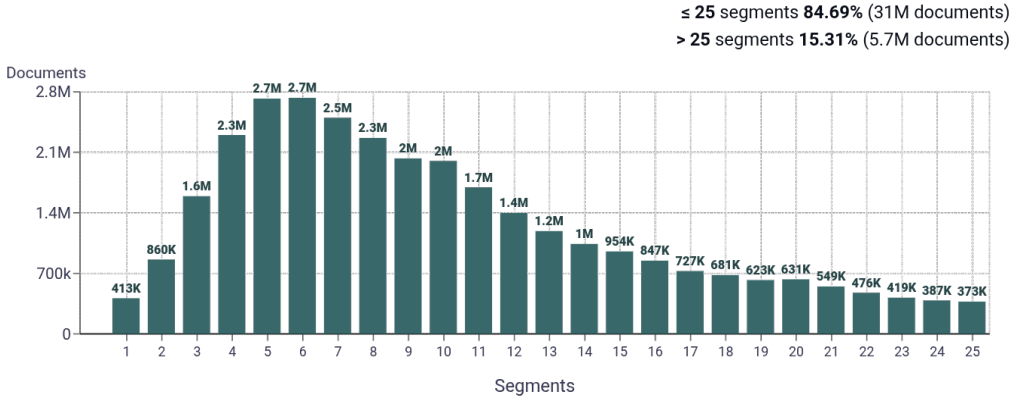
Domain	Docs	% of total
rs	12M	32.05%
com	10M	27.69%
ba	4.8M	12.90%
net	2.7M	7.17%
me	1.5M	3.93%
org	1.3M	3.61%
info	1.3M	3.49%
hr	768K	2.07%
co.rs	373K	1.01%
org.rs	285K	0.77%

Register labels

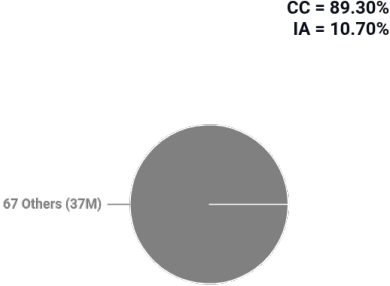


MT:1.2% | 452K Documents

Documents size (in segments) ⓘ

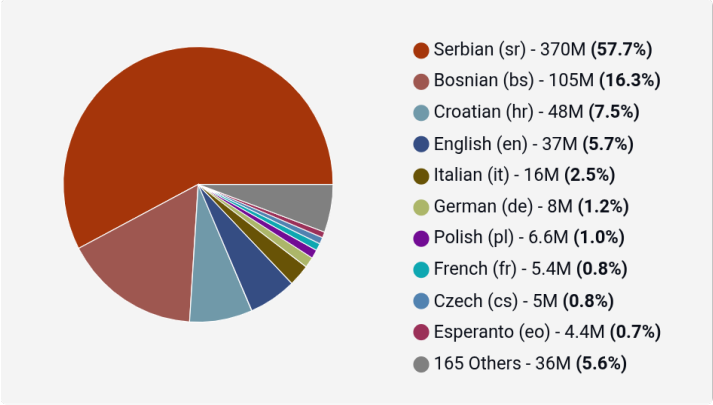


Document collections

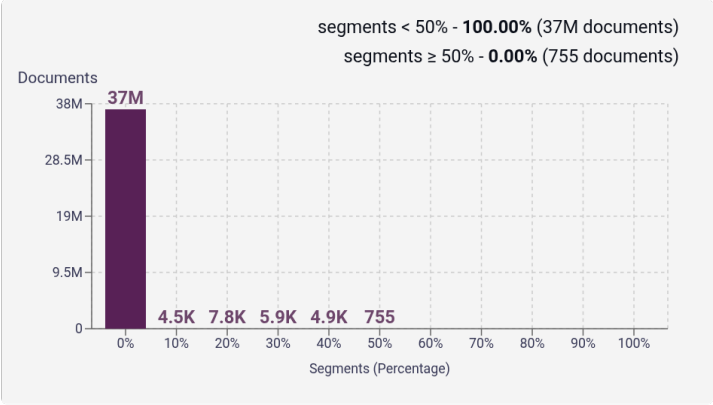


Language Distribution

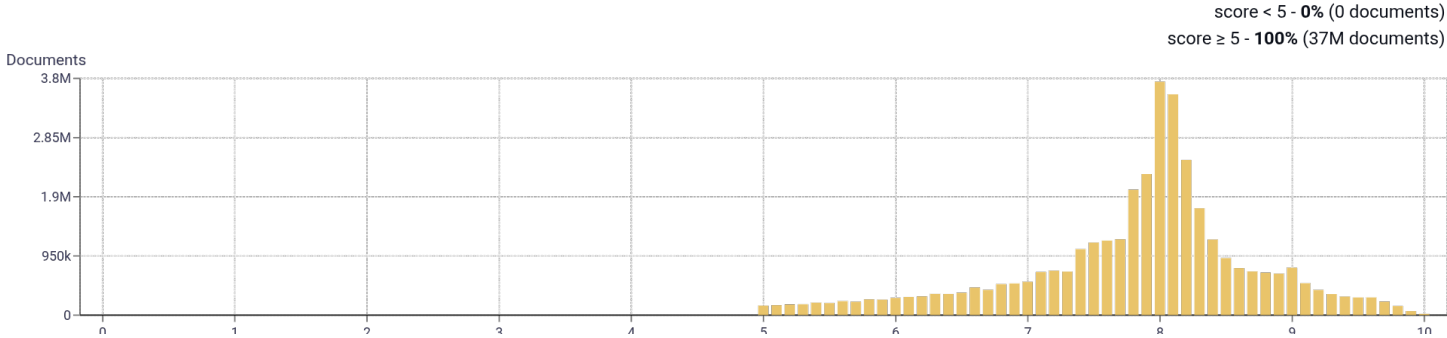
Number of segments in the Bosnian (bs) corpus



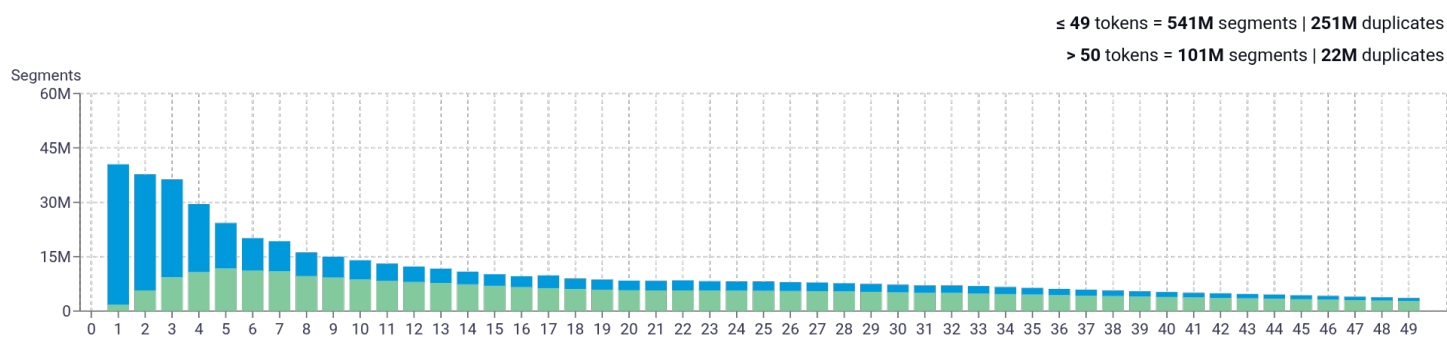
Percentage of segments in Bosnian (bs) inside documents



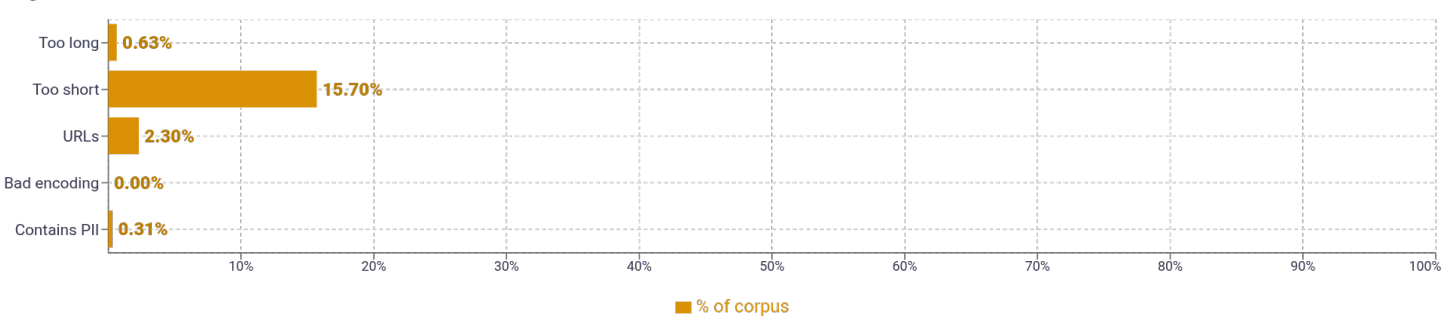
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	godine   35,797,844   dana   18,453,655   godina   16,314,194   dok   15,547,506   kod   14,392,517	
2	prvi put   2,499,949   prošle godine   2,134,707   republike srspe   2,013,416   crnoj gori   1,797,398   druge strane   1,712,132	
3	bosne i hercegovine   2,473,183   bosni i hercegovini   1,271,842   bosna i hercegovina   506,641   srbije aleksandar vučić   443,769 imajući u vidu   391,134	
4	navodi se u saopštenju   381,258   predsednik srbije aleksandar vučić   302,115   nalazi se u ulici   234,732 odražavaju stavove njihovih autora   210,010   komentari odražavaju stavove njihovih   209,719	
5	komentari odražavaju stavove njihovih autora   209,709   komentarima su privatno mišljenje autora   182,893   autora komentara i ne odražavaju   170,254 komentara i ne odražavaju stavove   170,250   iznešena u komentarima su privatno   165,640	

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				