HPLT Analytics report

General overview

Corpus	Date	SL	TL
hplt-v2-en-bs.tsv	1/24/2025	English (en)	Bosnian (bs)

Volumes

Segments	SL tokens	SL characters	SL size
4,559,328	119M	629,495,311	603.83 MB

TL tokens	TL characters	TL size
107M	606,814,485	594.95 MB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
wikipedia.org	12.4%	wikipedia.org	11.4%
booking.com	3.9%	nikana.gr	2.3%
nikana.gr	2.4%	booking.com	1.9%
voanews.com	1.6%	tripadvisor.rs	1.5%
airwise.com	1.5%	voanews.com	1.5%
agoda.com	1.5%	kosovotwopointzero.com	1.4%
educationbro.com	1.5%	agoda.com	1.1%
kosovotwopointzero.com	1.4%	cin.ba	1.1%
cin.ba	1.1%	slobodnaevropa.org	0.8%
aljazeera.com	1.0%	setimes.com	0.8%

Dataset top 10 TLDs

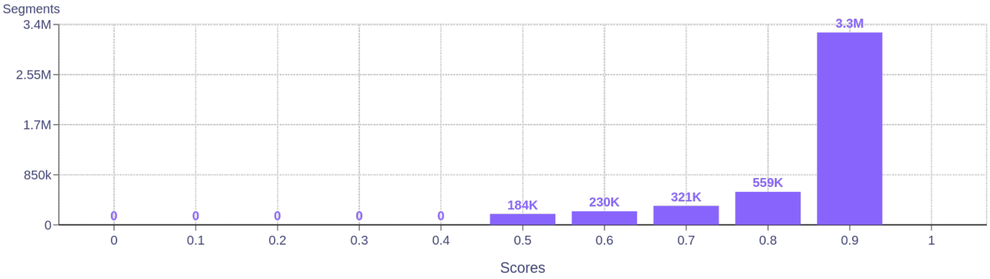
SL domain	Segments	TL domain	Segments
com	59.5%	com	41.0%
org	32.3%	org	27.9%
rs	11.4%	rs	19.6%
net	5.7%	ba	7.1%
ba	5.4%	net	6.5%
gr	2.6%	gr	2.4%
me	2.0%	me	2.4%
org.rs	1.8%	org.rs	1.9%
eu	1.8%	eu	1.3%
gov.rs	1.5%	gov.rs	1.3%

Translation likelihood

≥ 5 = 4.6M segments | **100.0%**

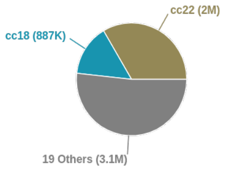
≥ 8 = 3.8M segments | **83.9%**

< 5 = 0 segments | **0.0%**



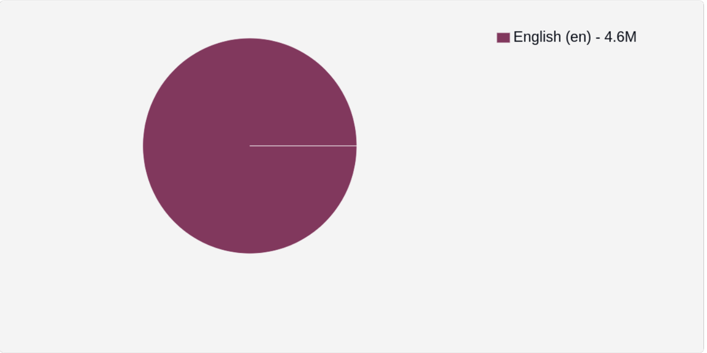
Collections

CC = 69.93%
IA = 30.07%

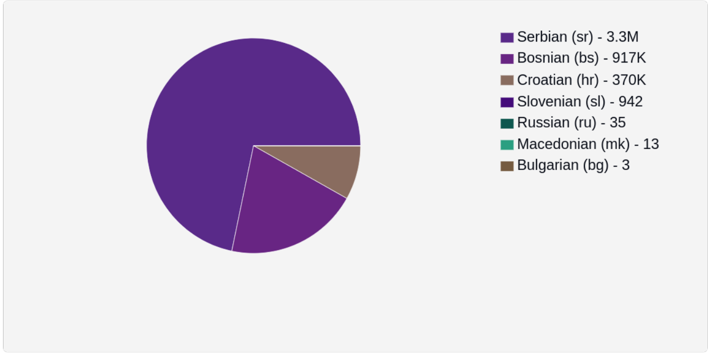


Language Distribution

Source



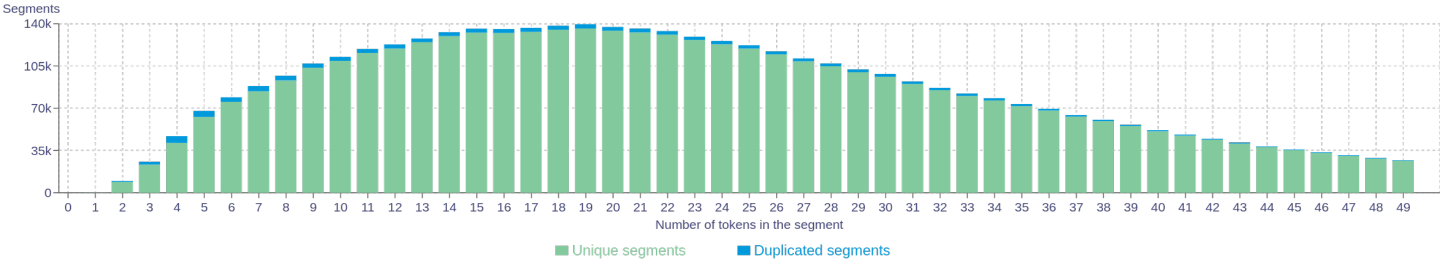
Target



Source segment length distribution by token

<= 49 tokens = **4.1M** segments | **113K** duplicates

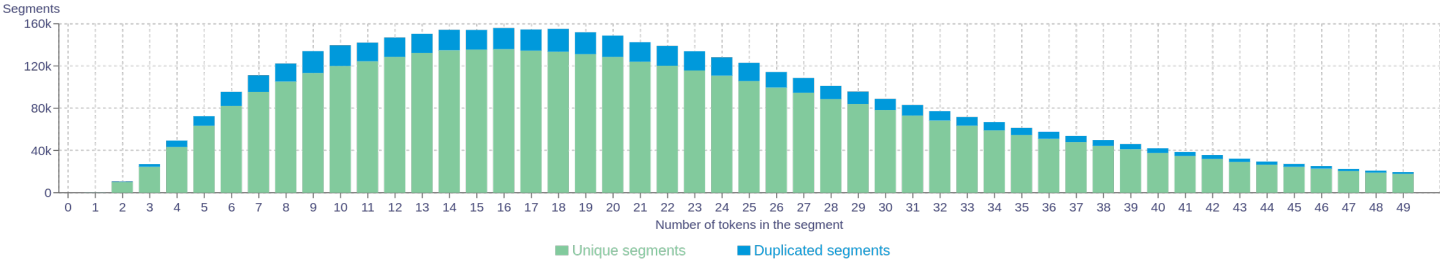
> 50 tokens = **372K** segments | **6.3K** duplicates



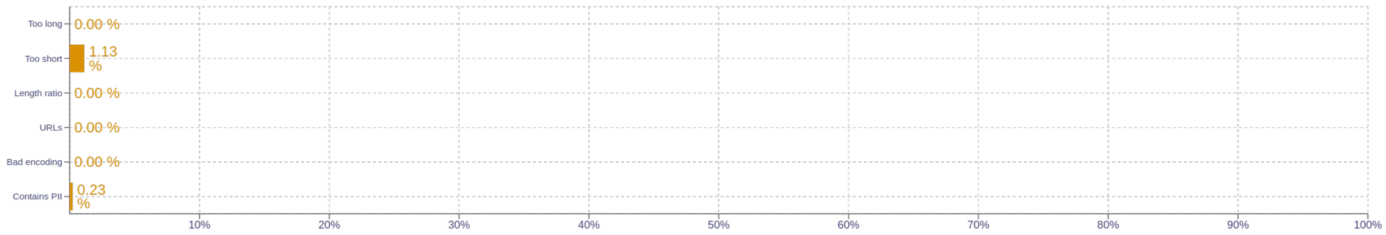
Target segment length distribution by token

<= 49 tokens = **3.8M** segments | **546K** duplicates

> 50 tokens = **247K** segments | **24K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	one | 237109 also | 214524 new | 174740 people | 164216 first | 162292
2	united states | 27078 human rights | 26544 novi sad | 21074 prime minister | 19752 european union | 19694
3	bosnia and herzegovina | 46942 proud to partner | 22254 tripadvisor is proud | 22237 reservations with confidence | 22163 republic of serbia | 22152
4	address is being protected | 12248 international transportation and spedition | 7061 freıght exchange- international transportation | 7061 game with your best | 6801 president of the republic | 5602
5	tripadvisor is proud to partner | 22237 email address is being protected | 11406 proud to partner with booking.com | 8535 exchange- international transportation and spedition | 7061 forget to rate this game | 6811

Target n-grams

Size	n-grams
1	godine | 334611 godina | 107634 hotel | 98037 svoje | 87765 možete | 85546
2	pravite bezbedne | 22254 bezbedne rezervacije | 22254 republike srbije | 18021 uredi izvor | 17200 prvi put | 17087
3	možete da pravite | 22265 ponosan na partnerstvo | 22255 tripadvisor je ponosan | 22254 pravite bezbedne rezervacije | 22254 bosne i hercegovine | 19184
4	možete da pravite bezbedne | 22252 učestvuje u našem programu | 8751 ponuda kamiona za međunarodni | 7133 ovu igru sa svojim | 6815 igru sa svojim najboljim | 6815
5	tripadvisor je ponosan na partnerstvo | 22254 možete da pravite bezbedne rezervacije | 22252 pošte je zaštićena od spambotova | 10638 ponosan na partnerstvo sa booking.com | 8626 ponosan na partnerstvo sa expedia | 7618

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