HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-sr.tsv	1/25/2025	English (en)	Serbian (sr)

Volumes

Segments	SL tokens	SL characters	SL size
5,291,686	118M	607,110,021	581.59 MB

TL tokens	TL characters	TL size
105M	576,427,030	991.74 MB

Dataset top 10 domains

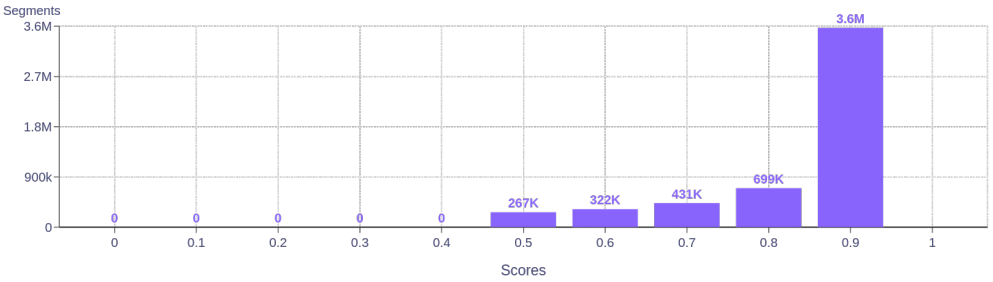
SL domain	Segments	TL domain	Segments
wikipedia.org	15.9%	wikipedia.org	14.0%
google.com	12.0%	google.com	5.6%
biblegateway.com	1.9%	tripadvisor.rs	4.5%
erieaquariumsociety.com	1.9%	rbth.com	1.6%
rbth.com	1.7%	biblegateway.com	1.6%
jw.org	1.5%	erieaquariumsociety.com	1.5%
academiccourses.com	1.4%	jw.org	1.4%
bachelorstudies.com	1.3%	stealthsettings.com	1.2%
stealthsettings.com	1.3%	vsaduidoma.com	1.2%
educationbro.com	1.2%	stroifaq.com	0.9%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	88.2%	com	61.5%
org	30.3%	org	25.3%
net	6.8%	rs	14.2%
rs	6.4%	net	5.8%
co.uk	2.1%	gov.rs	2.0%
gov.rs	2.0%	info	1.2%
eu	1.4%	ru	1.1%
ru	1.2%	eu	1.0%
info	1.2%	nu	0.9%
de	1.0%	de	0.8%

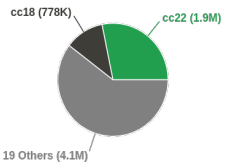
Translation likelihood

≥ 5 = 5.3M segments | **100.0%**
≥ 8 = 4.3M segments | **80.7%**
< 5 = 0 segments | **0.0%**



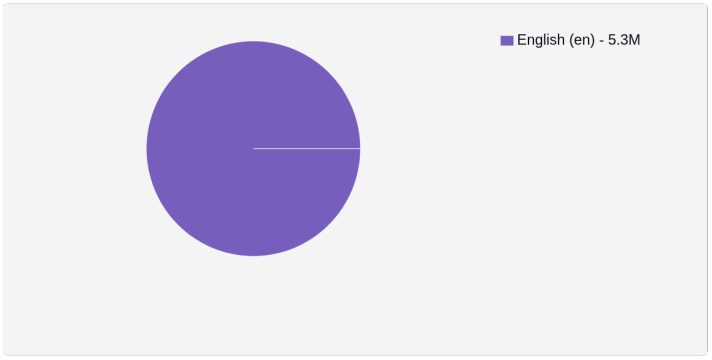
Collections

CC = 58.22%
IA = 41.78%

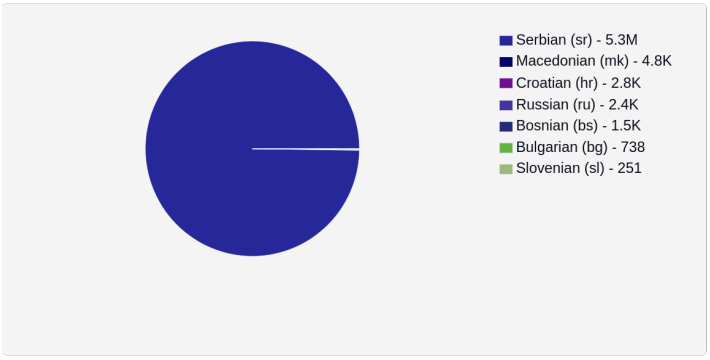


Language Distribution

Source

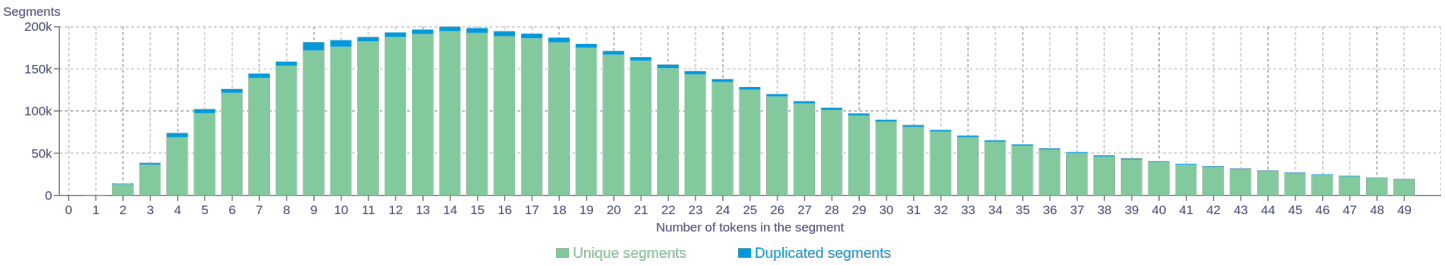


Target



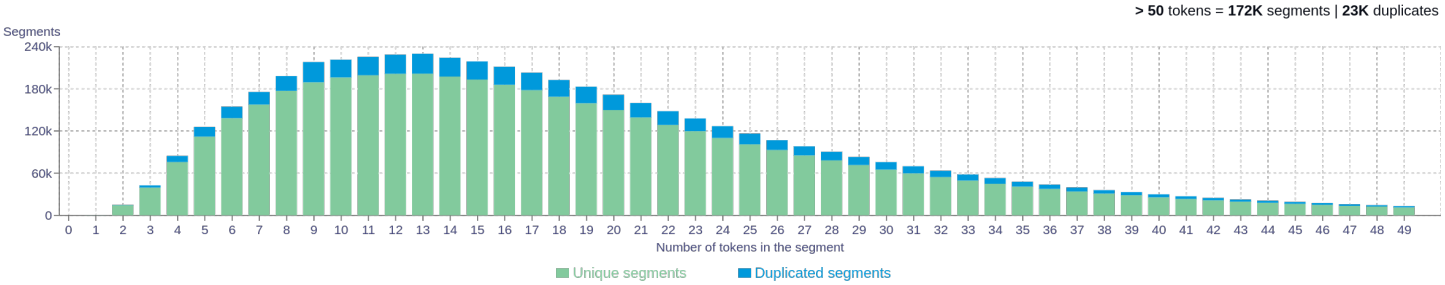
Source segment length distribution by token

<= 49 tokens = **4.9M** segments | **149K** duplicates
> 50 tokens = **272K** segments | **6.1K** duplicates

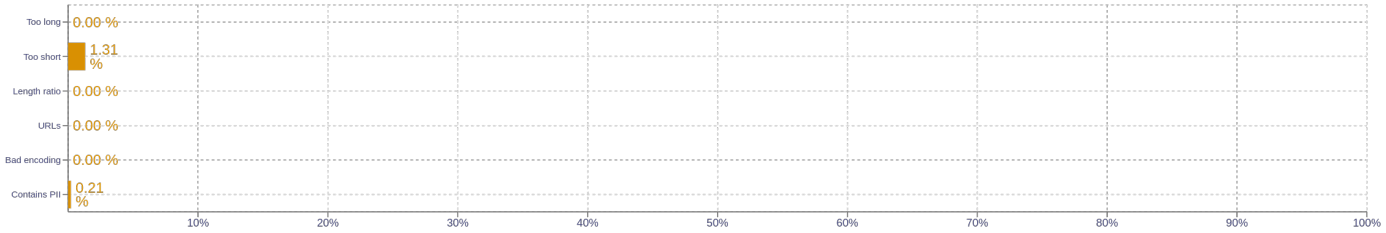


Target segment length distribution by token

<= 49 tokens = **4.5M** segments | **634K** duplicates
> 50 tokens = **172K** segments | **23K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	one | 237634 · also | 228866 · time | 166318 · new | 150877 · use | 147535
2	personal data | 22441 · united states | 21667 · national assembly | 13893 · online game | 10540 · make sure | 9909
3	republic of serbia | 22280 · bosnia and herzegovina | 10978 · republic of srpska | 10445 · around the world | 6076 · copy the code | 4519
4	president of the republic | 5168 · one of the best | 4638 · assembly of the republic | 4620 · code of your site | 4496 · paste in the html | 4494
5	national assembly of the republic | 4548 · html code of your site | 4495 · copy the code and paste | 4495 · paste in the html code | 4494 · email address is being protected | 4018

Target n-grams

Size	n-grams
1	који | 517003 · као | 449279 · које | 251468 · може | 207248 · године | 203753
2	може бити | 33566 · пре него | 19927 · републике србије | 17519 · изговорио корисник | 12639 · због тога | 11708
3	босне и херцеговине | 5742 · имајте на уму | 5485 · још много тога | 5081 · сједињене америчке државе | 5071 · који је био | 4989
4	копирајте код и налепите | 4499 · хтмл код вашег сајта | 4497 · налепите у хтмл код | 4497 · висок квалитет mobitel телефон | 3773 · народне скупштине републике србије | 2932
5	налепите у хтмл код вашег | 4497 · код и налепите у хтмл | 4497 · поште је заштићена од спамботова | 3720 · заборавите да оцени ову игру | 2012 · ову игру са својим најбољим | 1819

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