HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-hr.tsv	1/28/2025	English (en)	Croatian (hr)

Volumes

Segments	SL tokens	SL characters	SL size
14,263,908	324M	1,686,313,719	1.58 GB

TL tokens	TL characters	TL size
288M	1,660,624,244	1.59 GB

Dataset top 10 domains

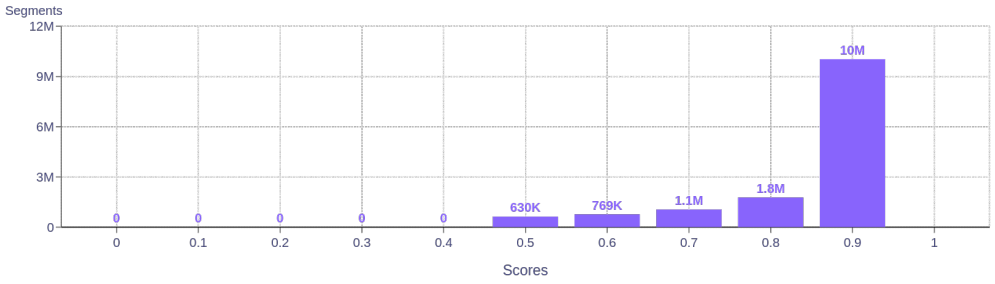
SL domain	Segments	TL domain	Segments
hotels.com	14.9%	europa.eu	6.9%
europa.eu	9.1%	hotels.com	5.5%
google.com	7.4%	google.com	3.3%
agoda.com	4.6%	wikipedia.org	2.9%
booking.com	3.9%	agoda.com	2.8%
wikipedia.org	3.3%	booking.com	1.8%
microsoft.com	1.7%	bibliacatolica.com.br	1.8%
office.com	1.7%	office.com	1.6%
bibliacatolica.com.br	1.6%	microsoft.com	1.1%
sacred-texts.com	0.8%	sacred-texts.com	0.9%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	106.8%	com	65.2%
eu	13.0%	hr	23.9%
org	12.1%	eu	10.0%
hr	10.2%	org	9.1%
net	5.1%	net	4.1%
co.uk	3.2%	com.hr	2.1%
de	2.1%	ba	2.0%
com.br	1.8%	com.br	2.0%
ie	1.4%	info	1.4%
info	1.3%	de	0.9%

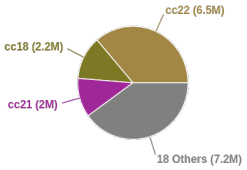
Translation likelihood

≥ 5 = 14M segments | **100.0%**
≥ 8 = 12M segments | **82.8%**
< 5 = 0 segments | **0.0%**



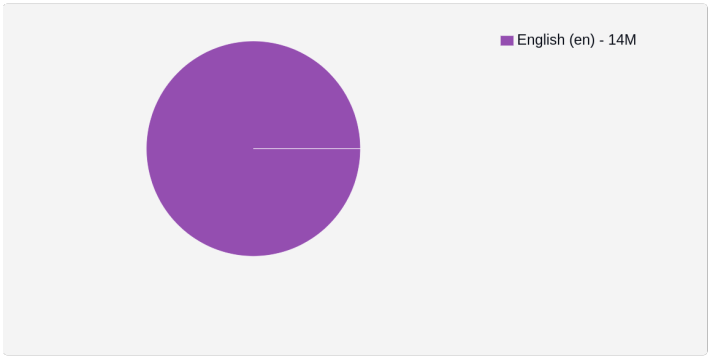
Collections

CC = 70.53%
IA = 29.47%

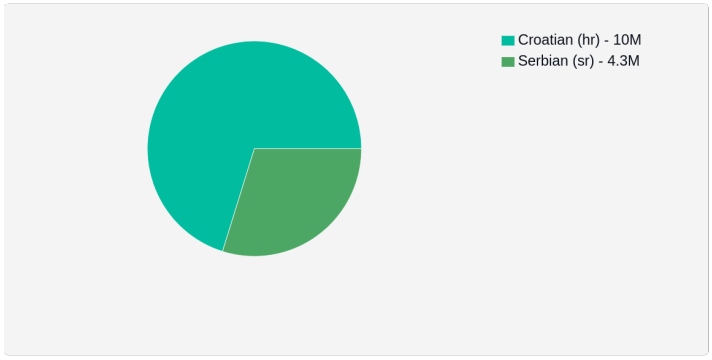


Language Distribution

Source

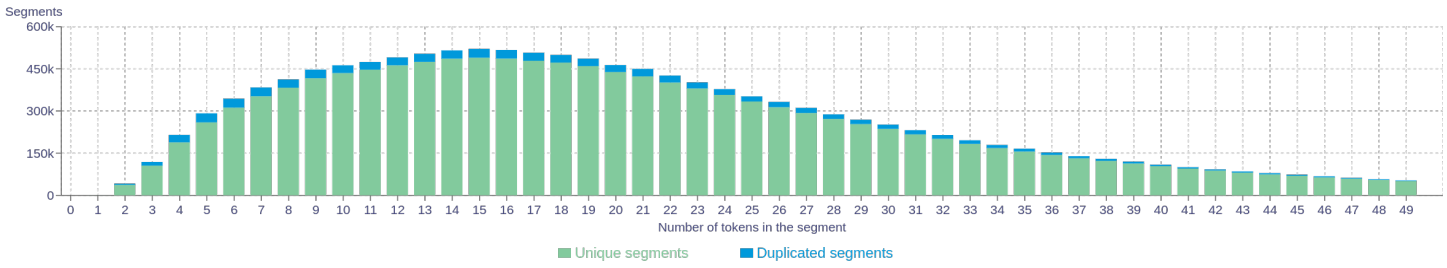


Target



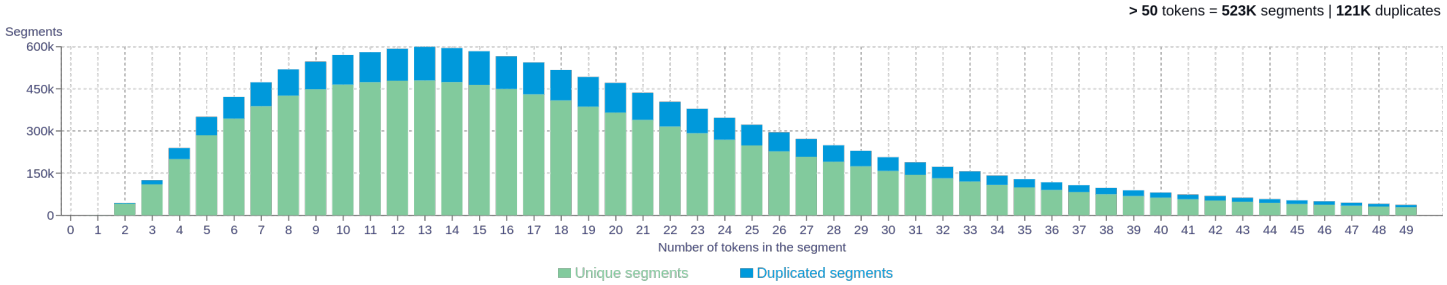
Source segment length distribution by token

<= 49 tokens = **13M** segments | **838K** duplicates
> 50 tokens = **797K** segments | **32K** duplicates

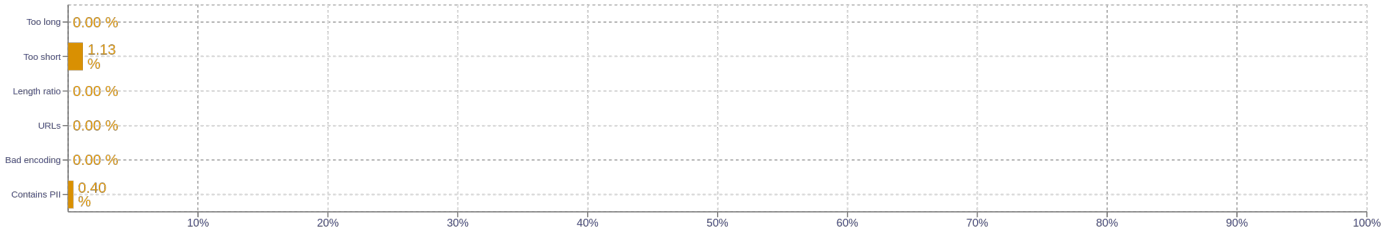


Target segment length distribution by token

<= 49 tokens = **11M** segments | **2.8M** duplicates
> 50 tokens = **523K** segments | **121K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	data | 1007703 · use | 678943 · also | 605970 · information | 579453 · personal | 567952
2	personal data | 433644 · data protection | 82790 · personal information | 76703 · privacy policy | 69859 · member states | 63170
3	processing of personal | 49557 · wi-fi in public | 29844 · terms and conditions | 29745 · right to object | 26959 · personal data concerning | 19940
4	processing of personal data | 49046 · processing of your personal | 37605 · wi-fi in public areas | 29836 · wi-fi in all rooms | 29047 · use of the website | 26701
5	processing of your personal data | 34611 · free wi-fi in all rooms | 29033 · parliament and of the council | 17614 · price ratio with competitive prices | 10736 · part of our preferred property | 10736

Target n-grams

Size	n-grams
1	podataka | 651920 · može | 604700 · možete | 589950 · više | 486026 · mogu | 443780
2	osobnih podataka | 290381 · osobne podatke | 150449 · web stranice | 128282 · vaših osobnih | 71387 · imate pravo | 68659
3	vaših osobnih podataka | 69701 · bežični pristup internetu | 45886 · wi-fi u svim | 31542 · obrade osobnih podataka | 26452 · obradu osobnih podataka | 24240
4	besplatan wi-fi u svim | 31536 · wi-fi u svim sobama | 31534 · europskog parlamenta i vijeća | 19396 · obradu vaših osobnih podataka | 15724 · nalazi se u blizini | 15555
5	besplatan wi-fi u svim sobama | 31532 · nalazi se u blizini znamenitosti | 11466 · odličnom omjeru cijene i kvalitete | 10801 · usluzi i odličnom omjeru cijene | 10796 · sudjelovati u našem programu prioritetnih | 10796

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