HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-bg.tsv	1/30/2025	English (en)	Bulgarian (bg)

Volumes

Segments	SL tokens	SL characters	SL size
22,725,326	503M	2,607,600,799	2.44 GB

TL tokens	TL characters	TL size
490M	2,706,292,613	4.5 GB

Dataset top 10 domains

SL domain	Segments	TL domain	Segments
europa.eu	7.6%	europa.eu	5.8%
google.com	6.5%	wikipedia.org	4.6%
wikipedia.org	5.4%	agoda.com	3.2%
agoda.com	4.3%	google.com	2.9%
booking.com	4.1%	booking.com	2.0%
bulgarianproperties.com	2.2%	biblegateway.com	1.9%
biblegateway.com	2.0%	bulgarianproperties.bg	1.8%
microsoft.com	1.6%	office.com	1.3%
office.com	1.4%	microsoft.com	1.0%
softoware.net	1.1%	softoware.net	0.9%

Dataset top 10 TLDs

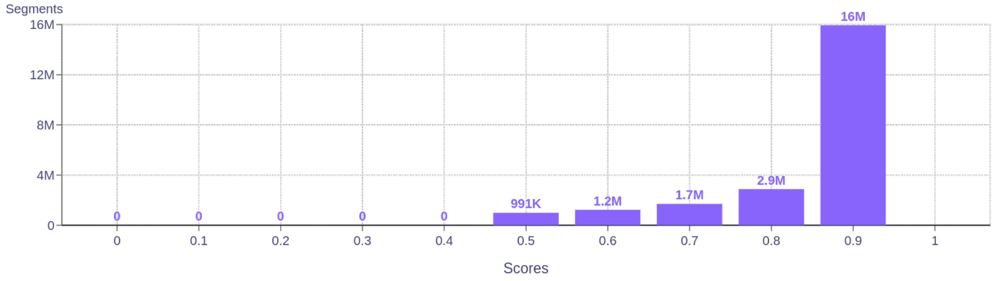
SL domain	Segments	TL domain	Segments
com	95.3%	com	63.7%
org	14.6%	bg	26.3%
eu	12.1%	org	11.3%
bg	7.7%	eu	10.4%
net	5.9%	net	5.0%
co.uk	3.3%	info	2.6%
info	2.8%	de	0.7%
de	1.5%	ru	0.5%
ie	1.0%	biz	0.5%
biz	0.7%	ro	0.4%

Translation likelihood

≥ 5 = 23M segments | **100.0%**

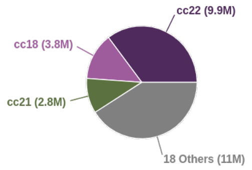
≥ 8 = 19M segments | **82.8%**

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



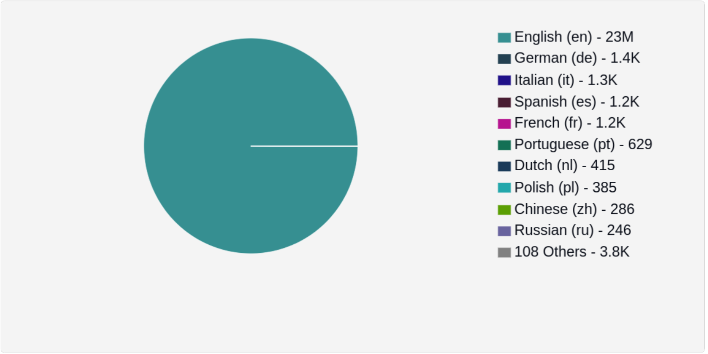
Collections

CC = 67.84%
IA = 32.16%

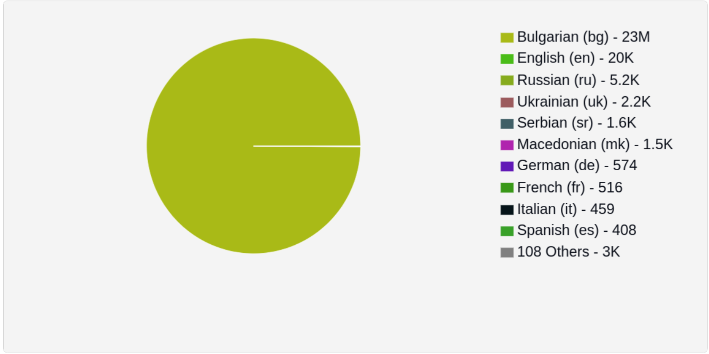


Language Distribution

Source



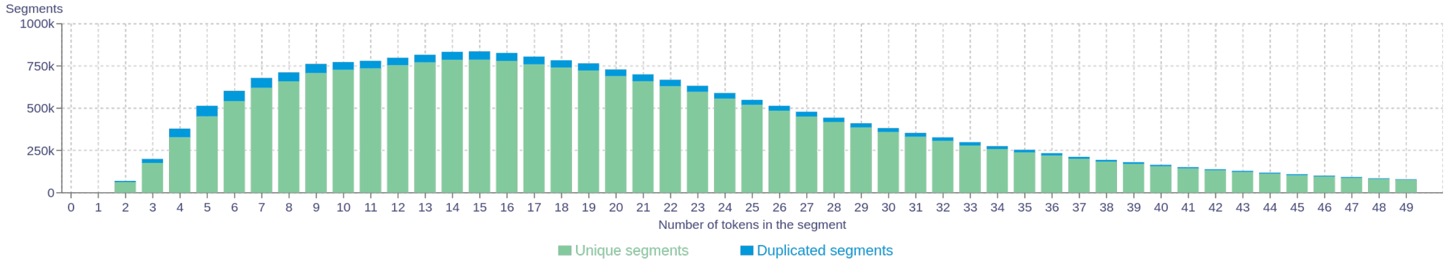
Target



Source segment length distribution by token

<= 49 tokens = **20M** segments | **1.4M** duplicates

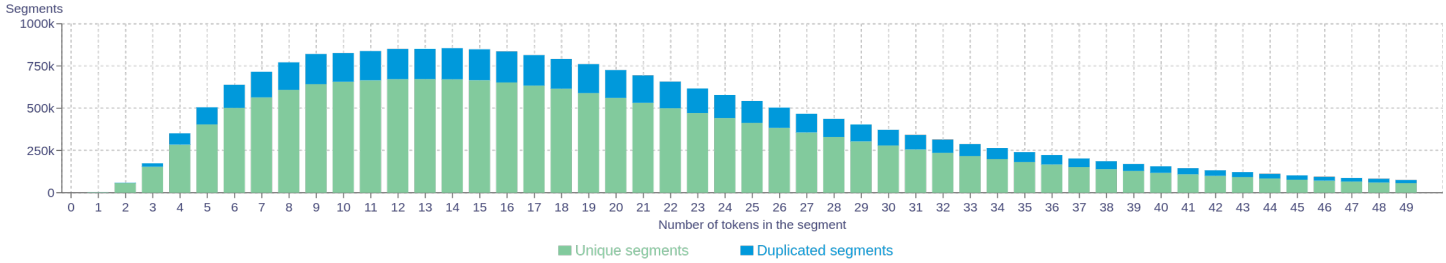
> 50 tokens = **1.2M** segments | **53K** duplicates



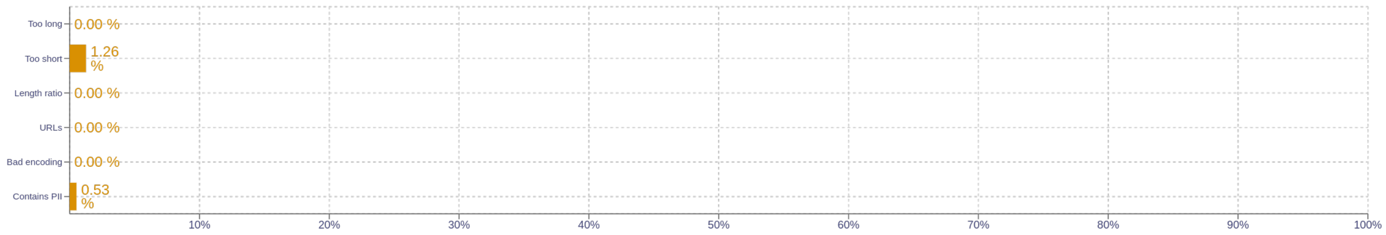
Target segment length distribution by token

<= 49 tokens = **17M** segments | **4.9M** duplicates

> 50 tokens = **1.1M** segments | **276K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	data | 1504994 use | 996449 personal | 943246 information | 894708 also | 893223
2	personal data | 707122 personal information | 145341 privacy policy | 118703 data protection | 117785 data subject | 85038
3	processing of personal | 67407 terms and conditions | 65112 wi-fi in public | 43464 protected from spambots | 42351 right to object | 38693
4	processing of personal data | 66560 processing of your personal | 61731 wi-fi in public areas | 43428 address is being protected | 42374 wi-fi in all rooms | 38879
5	processing of your personal data | 56006 email address is being protected | 39959 free wi-fi in all rooms | 38862 parliament and of the council | 28649 need javascript enabled to view | 12907

Target n-grams

Size	n-grams
1	данни | 1394111 можете | 758128 информация | 713062 г. | 687379 лични | 508433
2	лични данни | 487125 личните данни | 245548 вашите лични | 165883 имате право | 114457 трети страни | 80329
3	вашите лични данни | 164331 личните ви данни | 117505 защита на личните | 85280 защита на данните | 52527 редактиране на кода | 46778
4	защита на личните данни | 82442 защитен от спам ботове | 41118 имейл адрес е защитен | 40015 връзка в общите части | 39354 wi-fi връзка в общите | 39346
5	адрес е защитен от спам | 40071 wi-fi връзка в общите части | 39346 wifi достъп във всички стаи | 38769 европейския парламент и на съвета | 31860 обработването на личните ви данни | 16587

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