HPLT Analytics report

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
25,125,019	544M	2,832,218,431	2.65 GB

TL tokens	TL characters	TL size
496M	2,826,633,746	4.73 GB

Dataset top 10 domains

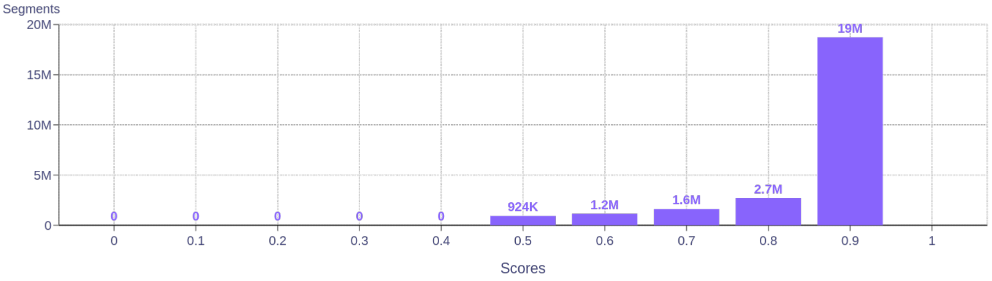
SL domain	Segments	TL domain	Segments
hotels.com	15.6%	hotels.com	6.0%
google.com	6.2%	wikipedia.org	4.4%
wikipedia.org	5.0%	google.com	2.7%
booking.com	2.2%	itsmygame.com.ua	1.8%
agoda.com	2.0%	khpg.org	1.5%
itsmygame.org	1.9%	agoda.com	1.4%
khpg.org	1.7%	booking.com	1.3%
microsoft.com	1.4%	studybible.info	1.2%
studybible.info	1.2%	biblegateway.com	1.1%
biblegateway.com	1.2%	shram.kiev.ua	1.0%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	90.5%	com	57.6%
org	21.1%	org	15.0%
ua	5.6%	com.ua	10.4%
com.ua	5.3%	ua	10.3%
net	5.0%	org.ua	4.9%
org.ua	3.0%	net	3.9%
info	2.7%	info	2.5%
ru	2.0%	ru	2.1%
kiev.ua	1.7%	in.ua	1.9%
co.uk	1.6%	kiev.ua	1.8%

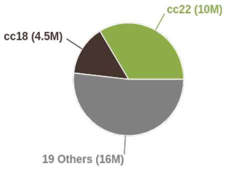
Translation likelihood

≥ 5 = 25M segments | **100.0%**
≥ 8 = 21M segments | **85.4%**
< 5 = 0 segments | **0.0%**



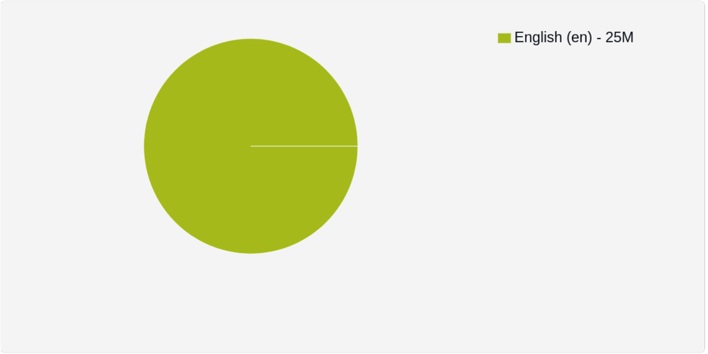
Collections

CC = 67.63%
IA = 32.37%

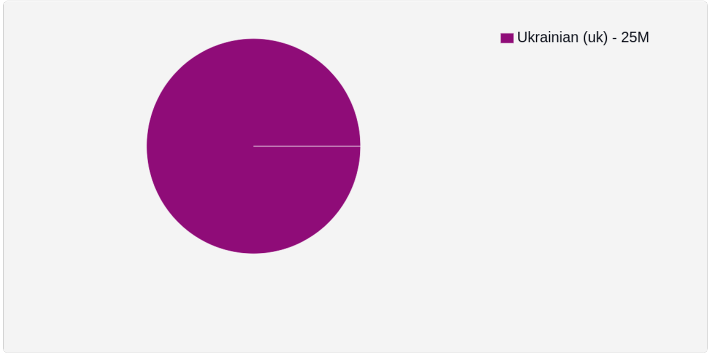


Language Distribution

Source

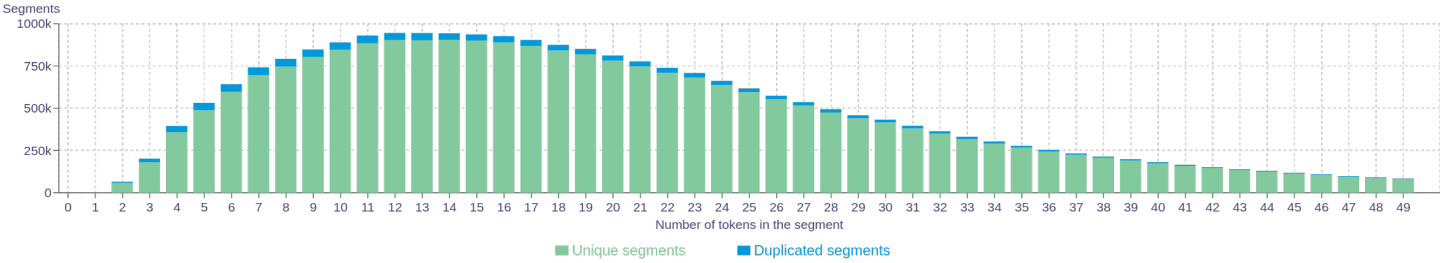


Target



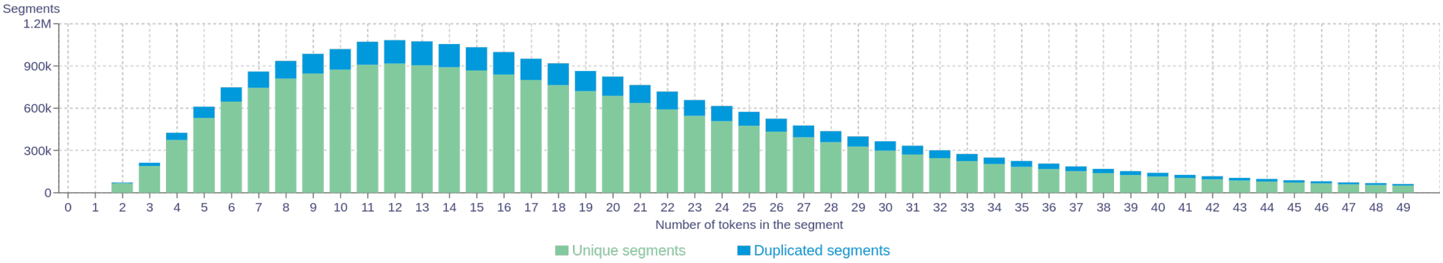
Source segment length distribution by token

<= 49 tokens = **23M** segments | **1.1M** duplicates
> 50 tokens = **1.1M** segments | **35K** duplicates

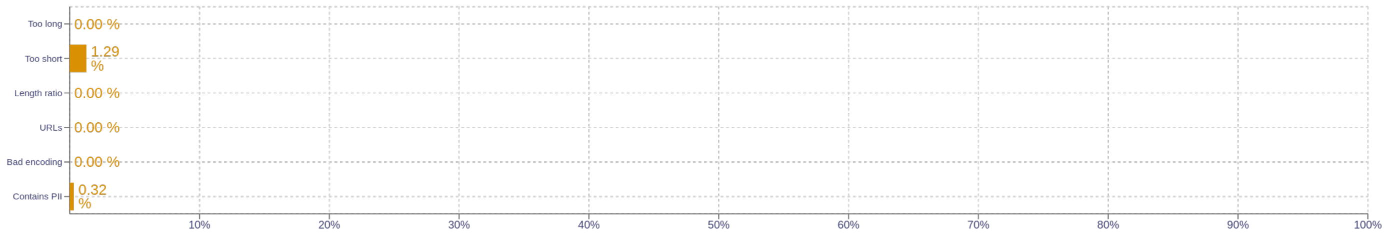


Target segment length distribution by token

<= 49 tokens = **20M** segments | **3.9M** duplicates
> 50 tokens = **789K** segments | **150K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	also | 973032 ukraine | 967696 one | 913549 use | 816295 time | 752347
2	personal data | 235884 personal information | 87224 united states | 86963 privacy policy | 80436 email address | 71236
3	like the game | 51969 see on map | 50517 protected from spambots | 49205 terms and conditions | 30946 play the game | 29376
4	address is being protected | 49283 code of your site | 29121 paste in the html | 29113 link to a friend | 28918 game with the world | 28902
5	email address is being protected | 46385 copy the code and paste | 29124 paste in the html code | 29113 html code of your site | 29113 copy and send the link | 28905

Target n-grams

Size	n-grams
1	від | 1897651 або | 1836785 щоб | 1304392 які | 1252201 також | 1176039
2	під час | 297437 може бути | 167748 персональних даних | 153272 км від | 135392 крім того | 120982
3	показати на мапі | 100496 якщо у вас | 50544 захищена від спам | 50072 адреса захищена від | 50070 електронна адреса захищена | 50052
4	адреса захищена від спам | 50069 електронна адреса захищена від | 50047 ця електронна адреса захищена | 49528 вам потрібно увімкнути javascript | 30658 вставте в html код | 29151
5	електронна адреса захищена від спам | 50047 ця електронна адреса захищена від | 49528 скопіюйте цей код і вставте | 29151 код і вставте в html | 29151 вставте в html код свого | 29151

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