HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
3,448,296	93M	479,015,429	458.91 MB

TL tokens	TL characters	TL size
108M	481,767,006	802.31 MB

Dataset top 10 domains

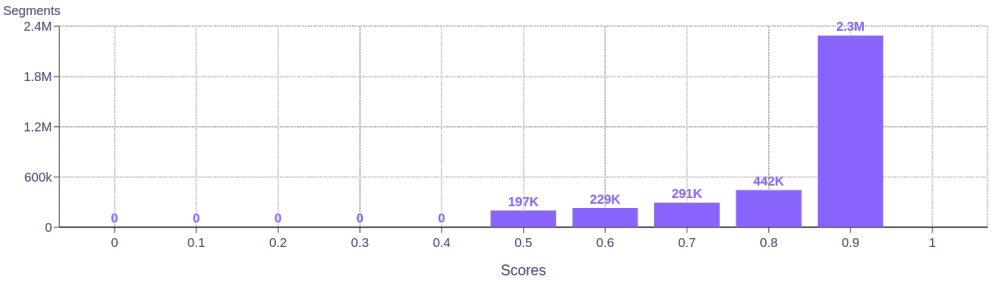
SL domain	Segments	TL domain	Segments
wikipedia.org	5.0%	wikipedia.org	4.3%
voanews.com	2.5%	voanews.com	2.4%
educationbro.com	1.6%	blogfa.com	2.1%
internationaldriversassociation.com	1.2%	internationaldriversassociation.com	1.1%
minghui.org	1.0%	destinia.ir	1.0%
wordpress.com	0.9%	minghui.org	1.0%
blogspot.com	0.9%	stepbible.org	0.9%
blogfa.com	0.9%	euronews.com	0.9%
itsmygame.org	0.9%	al-shorfa.com	0.8%
softoware.net	0.8%	itsmygame.org	0.8%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	84.2%	com	60.8%
org	18.8%	ir	21.7%
net	7.2%	org	14.7%
ir	4.2%	net	5.4%
co.uk	1.4%	info	1.0%
info	1.2%	ru	1.0%
ru	1.2%	eu	0.7%
eu	1.2%	ac.ir	0.6%
de	0.9%	de	0.6%
ca	0.8%	se	0.5%

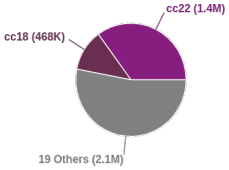
Translation likelihood

≥ 5 = 3.4M segments | **100.0%**
≥ 8 = 2.7M segments | **79.2%**
< 5 = 0 segments | **0.0%**



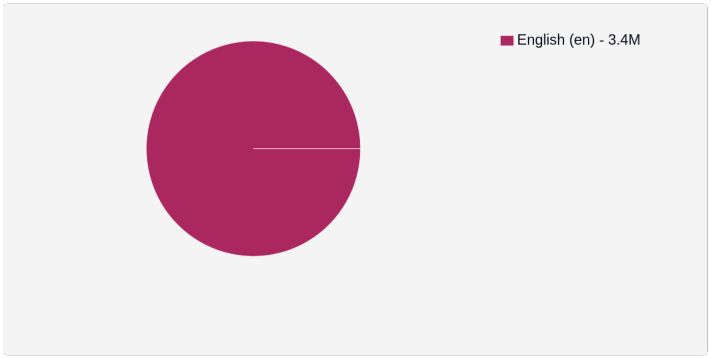
Collections

CC = 64.91%
IA = 35.09%

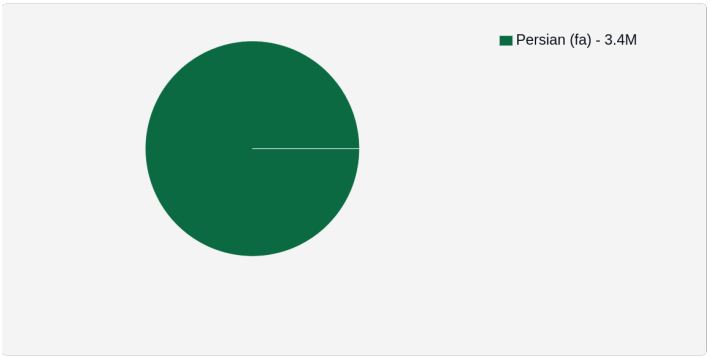


Language Distribution

Source

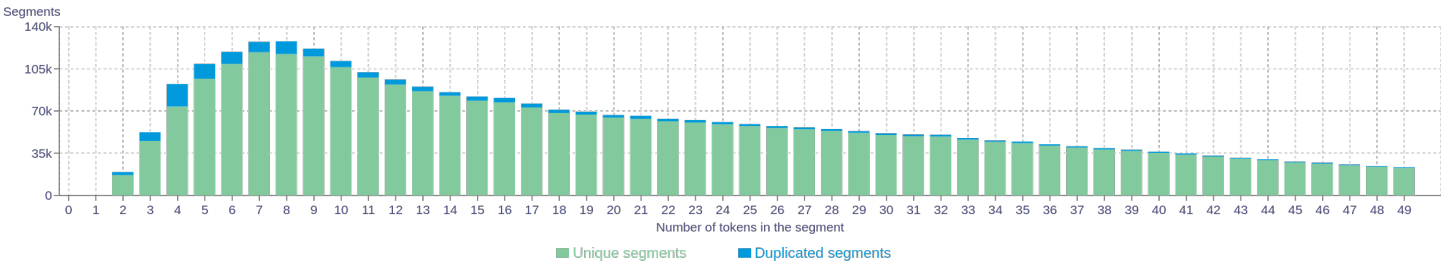


Target



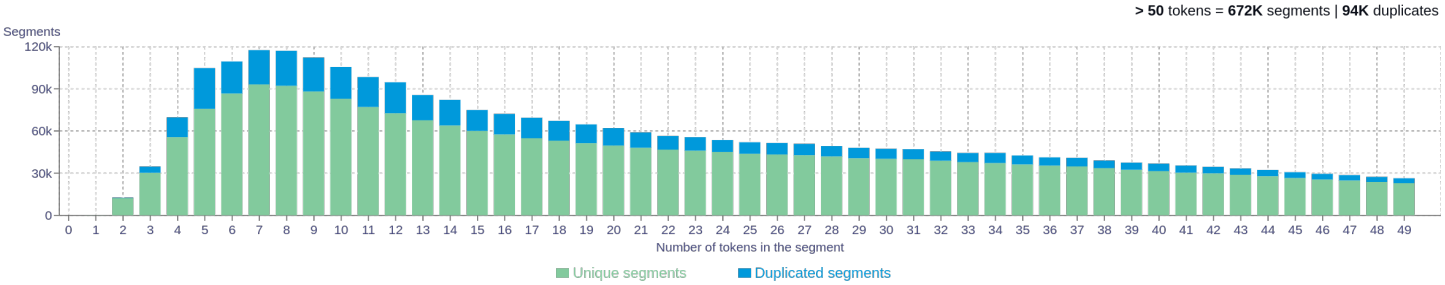
Source segment length distribution by token

<= 49 tokens = **2.8M** segments | **151K** duplicates
> 50 tokens = **469K** segments | **14K** duplicates

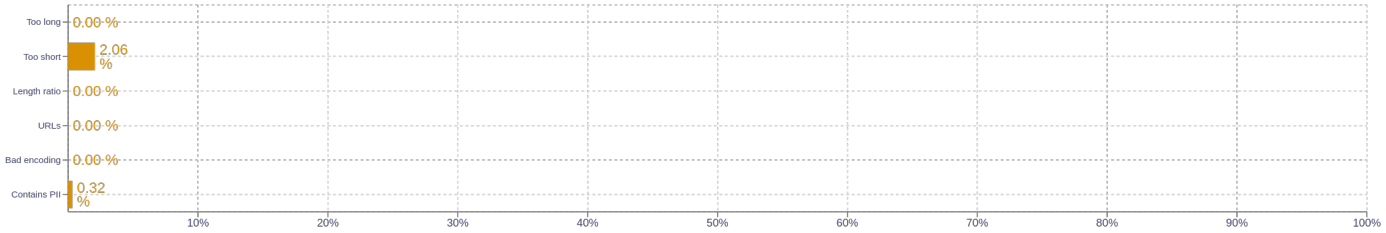


Target segment length distribution by token

<= 49 tokens = **2.3M** segments | **515K** duplicates
> 50 tokens = **672K** segments | **94K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	one | 178743 · said | 174584 · also | 156685 · new | 141242 · people | 131029
2	united states | 36331 · human rights | 29269 · high quality | 18623 · mobile phone | 15811 · quality mobile | 13333
3	call of duty | 19482 · high quality mobile | 13333 · thousands of free | 8398 · free mobile content | 8368 · quality mobile phone | 7719
4	site you can download | 8369 · download thousands of free | 8369 · thousands of free mobile | 8368 · high quality mobile phone | 7719 · islamic republic of iran | 6013
5	thousands of free mobile content | 8368 · site you can download thousands | 8368 · download thousands of free mobile | 8368 · symbian and java supported mobile | 5635 · quality mobile games and download | 5609

Target n-grams

Size	n-grams
1	می | 1223433 · این | 1103592 · برای | 950089 · های | 824351 · کنید | 449925
2	می کند | 123044 · می توانید | 112702 · نرم افزار | 66195 · می کنند | 57549 · بین المللی | 45007
3	این نرم افزار | 18992 · پروازها به مقصد | 18910 · call of duty | 18765 · کسب و کار | 15594 · حمل و نقل | 14902
4	کیفیت بالا سیار تلفن | 14072 · بازی call of duty | 9207 · این سایت می توانید | 8513 · موبایل خود دانلود کنید | 8499 · هزاران چیز رایگان برای | 8494
5	هزاران چیز رایگان برای موبایل | 8494 · می توانید هزاران چیز رایگان | 8494 · رایگان برای موبایل خود دانلود | 8494 · توانید هزاران چیز رایگان برای | 8494 · برای موبایل خود دانلود کنید | 8494

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