HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Segments	SL tokens	SL characters	SL size
4,166,536	102M	516,075,912	494.85 MB

TL tokens	TL characters	TL size
107M	551,579,636	568.04 MB

Dataset top 10 domains

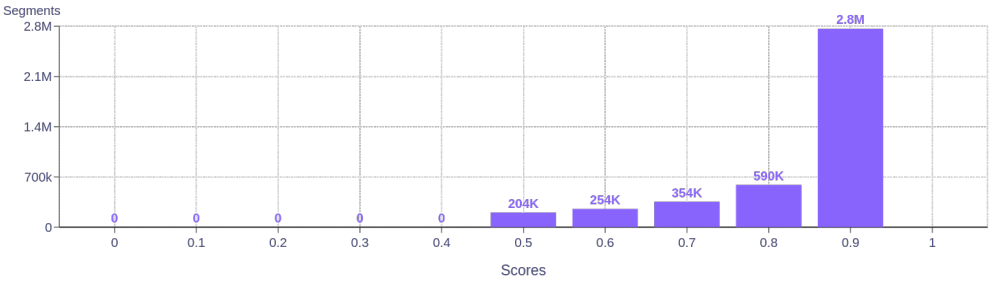
SL domain	Segments	TL domain	Segments
biblegateway.com	7.1%	biblegateway.com	6.4%
wikipedia.org	6.1%	sacred-texts.com	5.4%
sacred-texts.com	5.1%	wikipedia.org	5.1%
memorie.al	2.8%	memorie.al	2.8%
itsmygame.org	2.2%	kosovotwopointzero.com	1.9%
educationbro.com	2.1%	studybible.info	1.7%
kosovotwopointzero.com	1.9%	itsmygame.org	1.7%
studybible.info	1.8%	skolarbete.nu	1.2%
skolarbete.nu	1.3%	game-game.com	1.2%
game-game.com	1.2%	sot.com.al	1.0%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	75.0%	com	58.2%
org	27.2%	org	21.3%
al	7.5%	al	11.1%
net	6.1%	net	4.7%
info	3.0%	info	3.0%
eu	2.7%	mk	2.3%
mk	2.2%	eu	2.2%
nu	1.3%	com.al	1.8%
gov	1.2%	nu	1.2%
de	1.2%	gov	1.0%

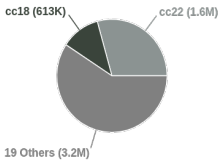
Translation likelihood

≥ 5 = 4.2M segments | **100.0%**
≥ 8 = 3.4M segments | **80.5%**
< 5 = 0 segments | **0.0%**



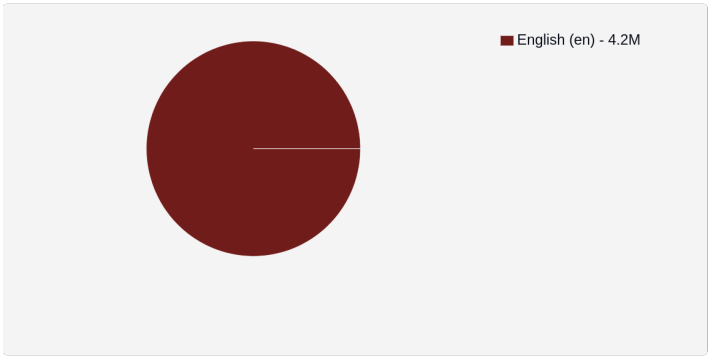
Collections

CC = 64.26%
IA = 35.74%

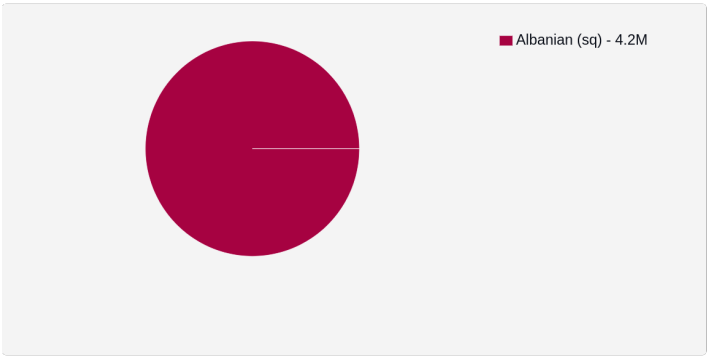


Language Distribution

Source

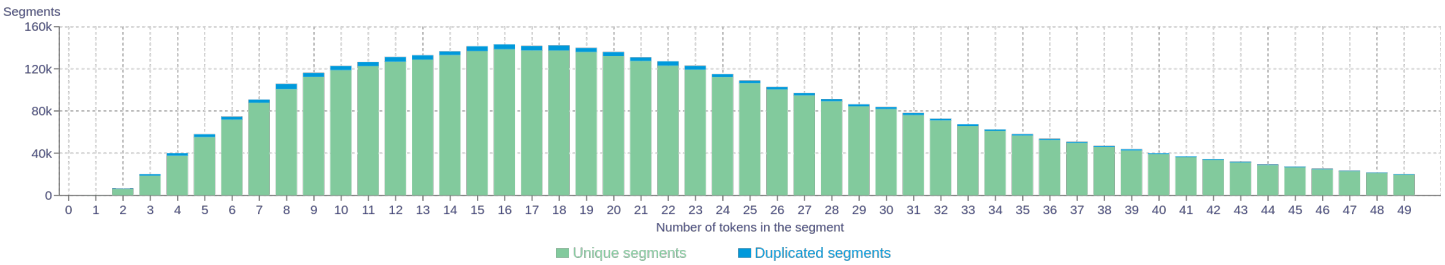


Target



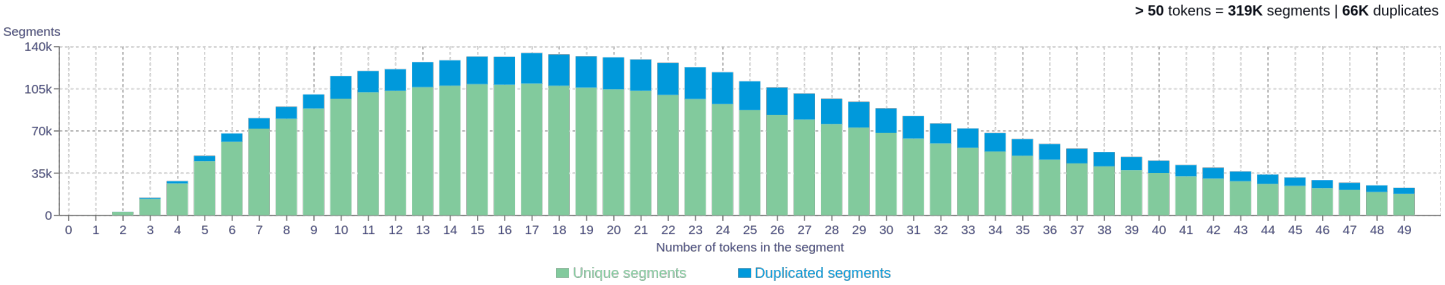
Source segment length distribution by token

<= 49 tokens = **3.8M** segments | **112K** duplicates
> 50 tokens = **270K** segments | **4.4K** duplicates

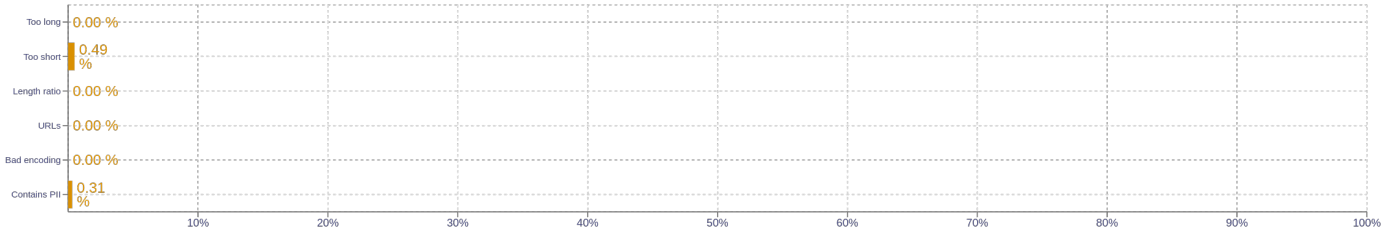


Target segment length distribution by token

<= 49 tokens = **3.1M** segments | **729K** duplicates
> 50 tokens = **319K** segments | **66K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	one | 198656 · also | 190213 · people | 155527 · game | 147417 · said | 144411
2	personal data | 27855 · united states | 16345 · prime minister | 13227 · online game | 12620 · personal information | 12448
3	online flash game | 9400 · play online flash | 9353 · forget to rate | 8981 · rate this game | 8958 · like the game | 8936
4	play online flash game | 9349 · game with your best | 8885 · link to a friend | 7497 · game with the world | 7496 · code of your site | 7389
5	forget to rate this game | 8954 · game with your best friends | 8885 · friend or all your friends | 7496 · copy and send the link | 7496 · paste in the html code | 7387

Target n-grams

Size	n-grams
1	është | 901877 · shumë | 349381 · duhet | 188864 · gjithë | 161478 · kanë | 159302
2	është shumë | 18685 · tuaja personale | 14094 · web faqen | 13686 · çdo gjë | 13574 · thotë zoti | 13480
3	cilësi të lartë | 13472 · duhet të jetë | 12436 · shtetet e bashkuara | 10477 · luaj online flash | 9298 · është e nevojshme | 9145
4	cilësi të lartë mobile | 8794 · luaj online flash lojë | 8064 · ndarë lojën me botën | 7496 · mik apo të gjithë | 7496 · lidhje për një mik | 7496
5	miqtë tuaj më të mirë | 9049 · ndajnë këtë lojë me miqtë | 9018 · cilësi të lartë mobile telefon | 8794 · harroni të vlerësoni këtë game | 7607 · qoftë se ju si lojë | 7505

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