HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-eo.tsv	2/11/2025	English (en)	Esperanto (eo)

Volumes

Segments	SL tokens	SL characters	SL size
1,521,821	39M	189,257,678	181.47 MB

TL tokens	TL characters	TL size
37M	187,362,538	181.04 MB

Dataset top 10 domains

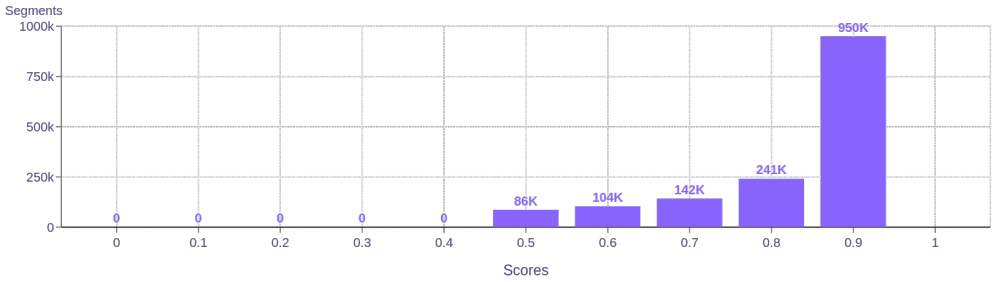
SL domain	Segments	TL domain	Segments
wikipedia.org	44.4%	wikipedia.org	36.1%
sacred-texts.com	15.1%	sacred-texts.com	12.8%
studybible.info	6.6%	studybible.info	6.3%
bibliaonline.com.br	4.8%	bibliaonline.com.br	5.7%
affiliatemarketingconsulting.net	4.2%	vessoft.com	1.6%
exactspy.com	3.9%	biblehub.com	1.6%
vessoft.com	2.0%	ebible.com	1.4%
educationbro.com	2.0%	exactspy.com	1.0%
printindustryhub.com	2.0%	educationbro.com	1.0%
biblehub.com	1.9%	doctrinepublishing.com	1.0%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	79.2%	com	53.2%
org	62.4%	org	45.4%
net	11.5%	info	6.6%
info	6.9%	com.br	6.3%
com.br	4.9%	net	5.5%
trade	1.0%	de	1.6%
de	0.9%	trade	0.9%
ru	0.9%	ru	1.4%
de	0.9%	trade	0.9%
plus	0.7%	eu	0.8%
org.uk	0.6%	be	0.5%

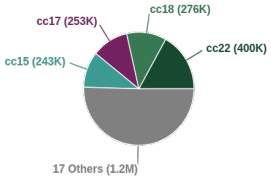
Translation likelihood

≥ 5 = 1.5M segments | **100.0%**
≥ 8 = 1.2M segments | **78.2%**
< 5 = 0 segments | **0.0%**



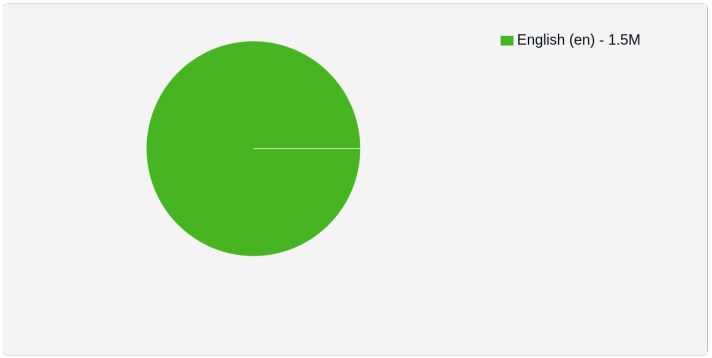
Collections

CC = 65.01%
IA = 34.99%

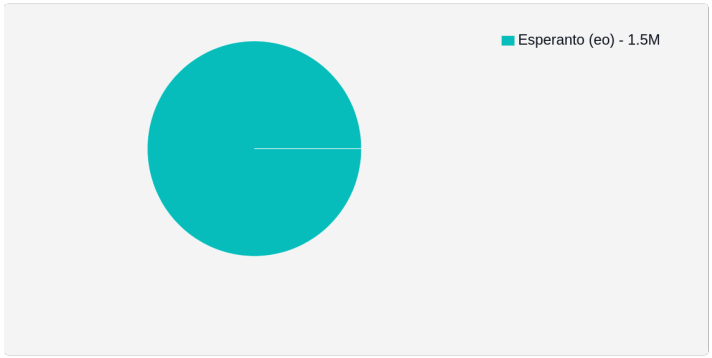


Language Distribution

Source

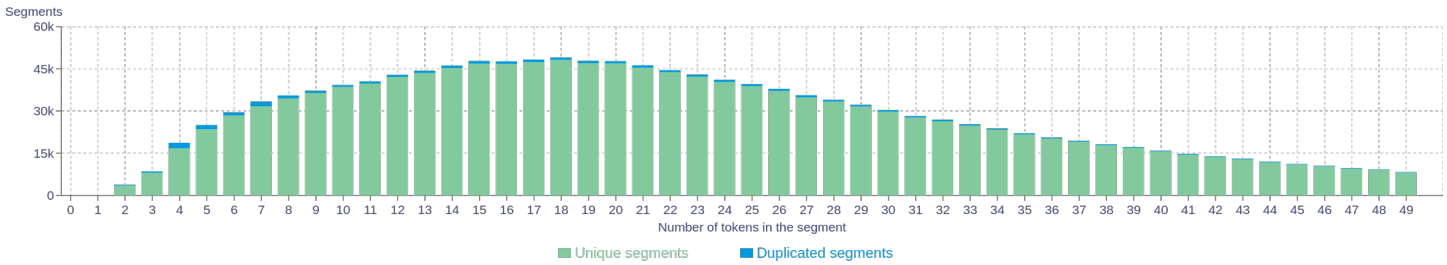


Target



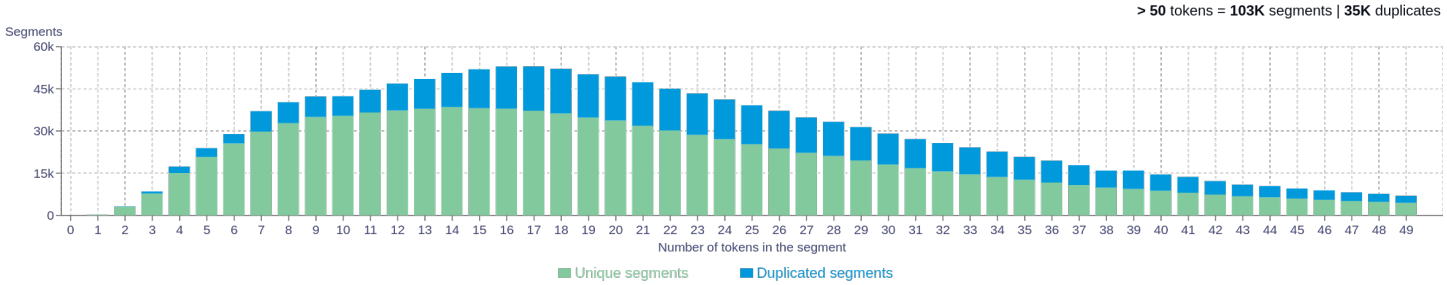
Source segment length distribution by token

≤ 49 tokens = **1.4M** segments | **32K** duplicates
> 50 tokens = **124K** segments | **2K** duplicates

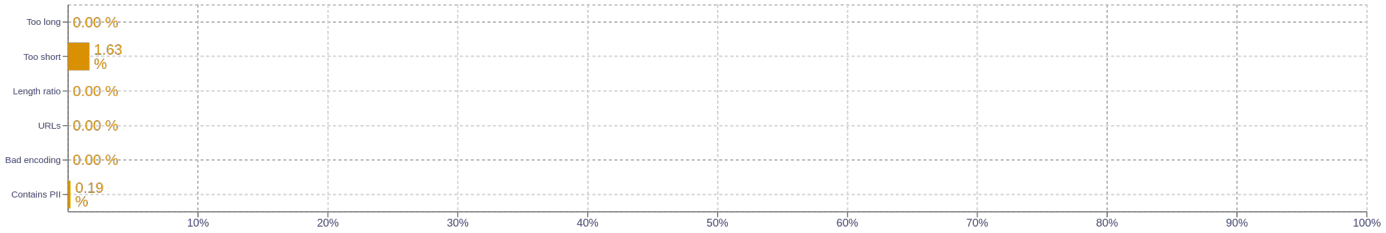


Target segment length distribution by token

≤ 49 tokens = **1M** segments | **419K** duplicates
> 50 tokens = **103K** segments | **35K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	one | 86974 · god | 77904 · also | 71565 · lord | 70901 · said | 69785
2	united states | 7547 · jesus christ | 5021 · cell phone | 3770 · new york | 3241 · lord god | 3021
3	children of israel | 5075 · son of man | 3289 · saith the lord | 2900 · land of egypt | 2681 · god of israel | 2533
4	word of the lord | 2085 · house of the lord | 1778 · name of the lord | 1073 · one of the best | 1005 · angel of the lord | 820
5	word of the lord came | 686 · one of the most important | 474 · hack and cheat for android | 432 · cheat for android and ios | 426 · years old when he began | 402

Target n-grams

Size	n-grams
1	kiel | 185841 · kun | 183837 · el | 149083 · povas | 118387 · pri | 107214
2	povas esti | 25118 · redakti fonton | 20757 · kune kun | 7610 · devas esti | 6737 · tiele diras | 5322
3	diras la eternulo | 7425 · antaux la eternulo | 3834 · filo de homo | 3154 · dio de izrael | 2459 · el la lando | 2125
4	tiele diras la eternulo | 3127 · vorto de la eternulo | 2442 · domo de la eternulo | 2427 · diras la eternulo cebaot | 1623 · el la lando egipta | 1246
5	tiele diras la eternulo cebaot | 897 · jarojn li regxis en jerusalem | 716 · donis al li la nomon | 660 · anstataux li ekregxis lia filo | 622 · cxar eterna estas lia boneco | 459

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