HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-xh.tsv	1/21/2025	English (en)	Xhosa (xh)

Volumes

Segments	SL tokens	SL characters	SL size
405,605	11M	49,345,473	47.32 MB

TL tokens	TL characters	TL size
8.1M	50,914,883	48.66 MB

Dataset top 10 domains

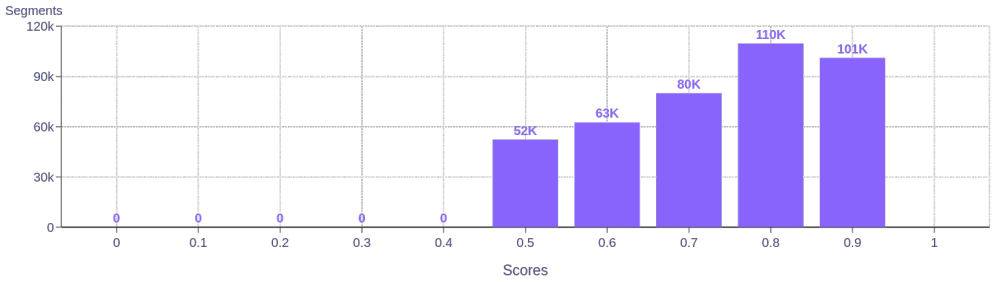
SL domain	Segments	TL domain	Segments
sacred-texts.com	41.9%	sacred-texts.com	45.8%
jw.org	12.8%	jw.org	14.2%
educationbro.com	3.5%	wordplanet.org	3.9%
learn2.trade	3.1%	learn2.trade	2.9%
wordpress.com	2.9%	educationbro.com	2.2%
basic-english.org	1.8%	airbnb.com	1.8%
airbnb.com	1.7%	wordproject.org	1.7%
oremus.org	1.7%	martech.zone	1.3%
martech.zone	1.4%	wikipedia.org	0.9%
blogspot.com	1.1%	plumamazing.com	0.8%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	93.4%	com	80.3%
org	23.4%	org	24.6%
trade	3.1%	trade	2.9%
net	3.0%	net	2.0%
top	1.8%	top	1.8%
zone	1.4%	zone	1.3%
info	1.1%	co.za	1.1%
org.au	0.9%	de	0.9%
co.za	0.8%	info	0.8%
de	0.7%	kz	0.4%

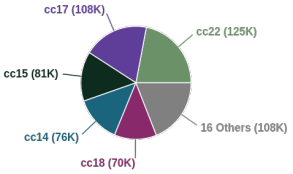
Translation likelihood

≥ 5 = 406K segments | **100.0%**
≥ 8 = 211K segments | **52.0%**
< 5 = 0 segments | **0.0%**



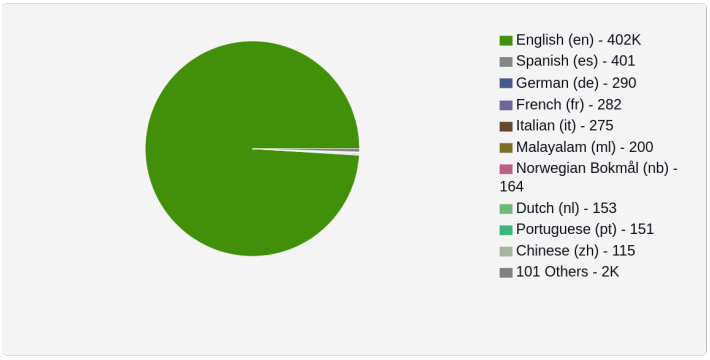
Collections

CC = 86.88%
IA = 13.12%

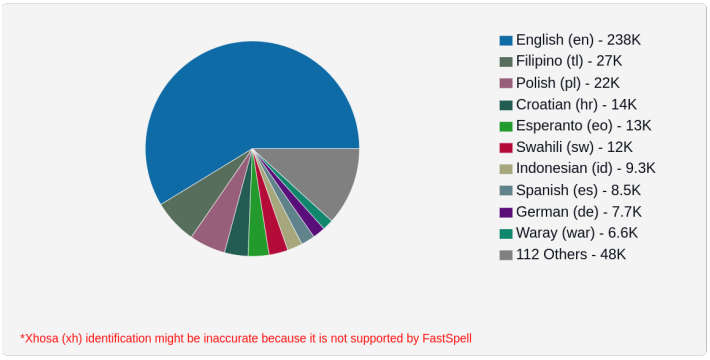


Language Distribution

Source



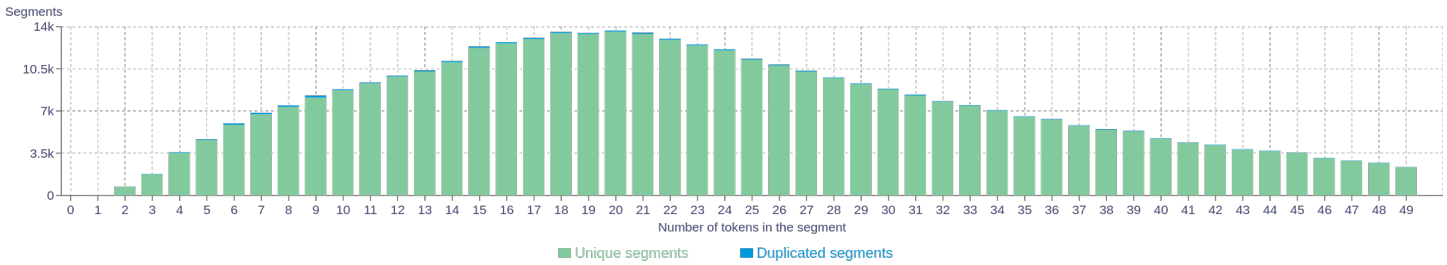
Target



*Xhosa (xh) identification might be inaccurate because it is not supported by FastSpell

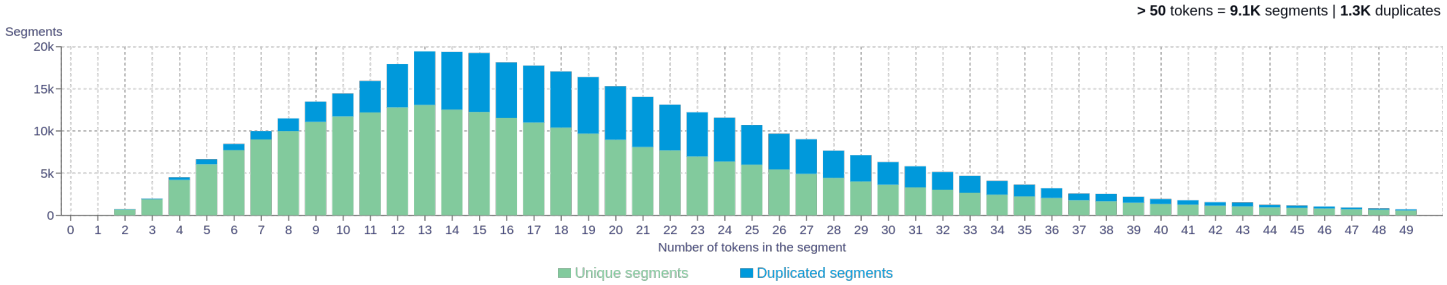
Source segment length distribution by token

<= 49 tokens = **372K** segments | **3.1K** duplicates
> 50 tokens = **30K** segments | **211** duplicates

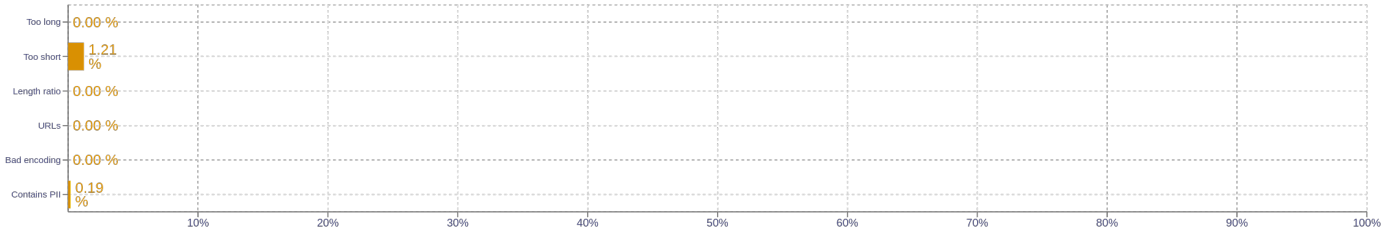


Target segment length distribution by token

<= 49 tokens = **265K** segments | **132K** duplicates
> 50 tokens = **9.1K** segments | **1.3K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	god | 36539 · lord | 27397 · said | 26522 · one | 22988 · people | 20611
2	jesus christ | 1914 · lord god | 1122 · holy spirit | 1084 · good news | 1035 · thy god | 1014
3	children of israel | 2579 · saith the lord | 1661 · god of israel | 1400 · son of man | 1391 · land of egypt | 1303
4	word of the lord | 813 · house of the lord | 662 · name of the lord | 381 · written in the book | 358 · saith the lord god | 304
5	saith the lord of hosts | 261 · word of the lord came | 256 · prices and availability information displayed | 150 · place of worship of yahweh | 150 · informationany prices and availability information | 150

Target n-grams

Size	n-grams
1	i | 47524 · uyehova | 24281 · unyana | 11380 · ukumkani | 10164 · uyesu | 9135
2	utsho uyehova | 4157 · oonyana bakasirayeli | 1845 · uyehova ukuthi | 1664 · itsho inkosi | 1542 · inkosi uyehova | 1455
3	itsho inkosi uyehova | 1348 · utsho uyehova wemikhosi | 963 · utsho uyehova ukuthi | 859 · inkosi uyehova ukuthi | 758 · nawuphi na umbuzo | 576
4	itsho inkosi uyehova ukuthi | 758 · waba ngukumkani esikhundleni sakhe | 342 · utsho uyehova wemikhosi ukuthi | 323 · encwadini yemicimbi yemihla yookumkani | 250 · kwafika ilizwi likayehova kum | 214
5	azibhalwanga na encwadini yemicimbi yemihla | 186 · naliphi na ulwazi malunga nalo | 150 · encwadini yemicimbi yemihla yookumkani bakwasirayeli | 130 · encwadini yemicimbi yemihla yookumkani bakwayuda | 113 · ngako oko itsho inkosi uyehova | 90

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