HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-sw.tsv	1/22/2025	English (en)	Swahili (sw)

Volumes

Segments	SL tokens	SL characters	SL size
1,985,899	49M	244,727,985	234.81 MB

TL tokens	TL characters	TL size
46M	254,326,507	243.19 MB

Dataset top 10 domains

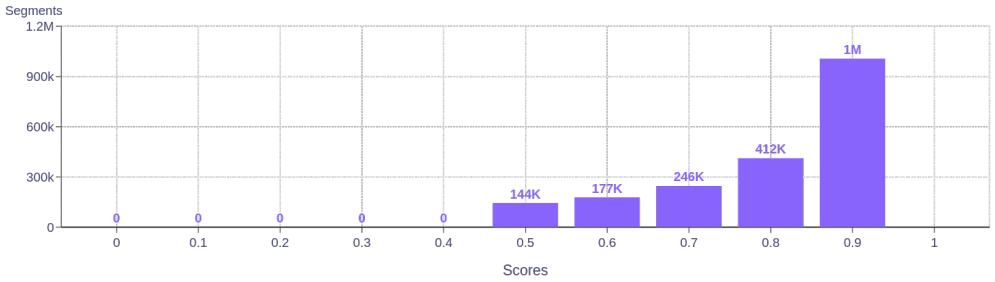
SL domain	Segments	TL domain	Segments
google.com	20.2%	google.com	8.7%
al-islam.org	5.9%	al-islam.org	6.5%
jw.org	5.0%	jw.org	4.9%
wikipedia.org	4.3%	wikipedia.org	3.5%
educationbro.com	3.1%	tuko.co.ke	2.9%
tuko.co.ke	2.9%	w3eacademy.com	2.8%
w3eacademy.com	2.9%	sacred-texts.com	2.4%
godfootsteps.org	2.7%	bjnewlife.org	2.0%
sabahionline.com	2.4%	sabahionline.com	2.0%
sacred-texts.com	2.2%	godfootsteps.org	1.8%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	95.8%	com	65.8%
org	42.2%	org	40.1%
net	5.6%	net	5.2%
co.ke	3.1%	co.ke	4.1%
info	1.5%	ws	1.0%
ws	1.4%	co.tz	0.9%
com.br	0.8%	info	0.9%
de	0.7%	com.br	0.8%
ai	0.7%	de	0.7%
top	0.7%	ai	0.6%

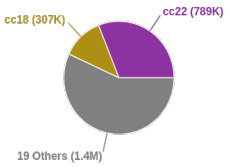
Translation likelihood

≥ 5 = 2M segments | **100.0%**
≥ 8 = 1.4M segments | **71.4%**
< 5 = 0 segments | **0.0%**



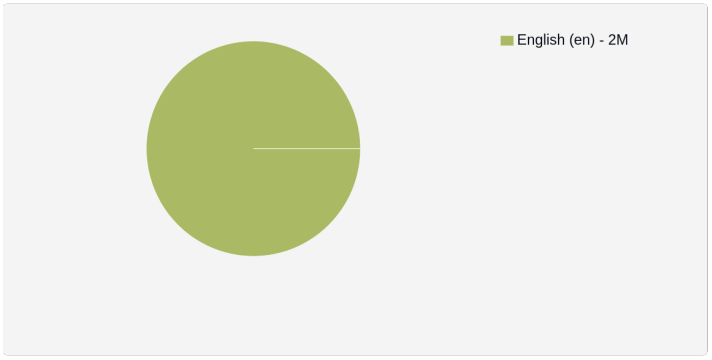
Collections

CC = 65.40%
IA = 34.60%

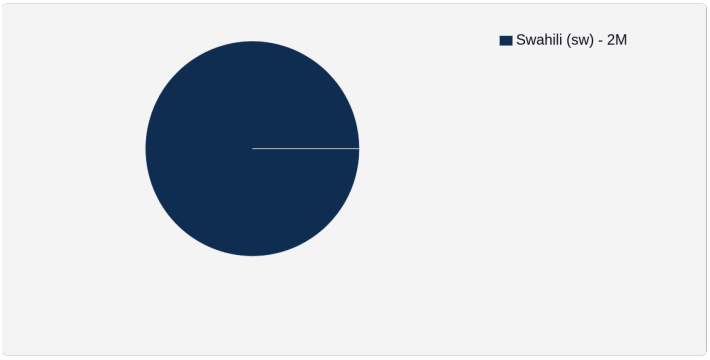


Language Distribution

Source

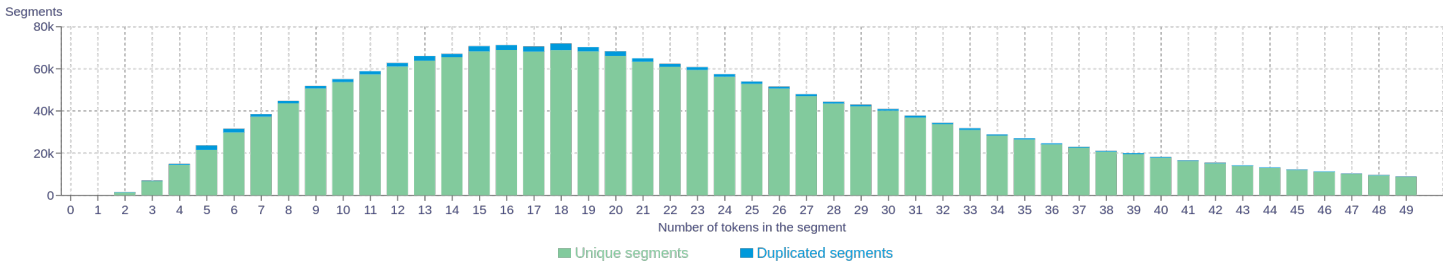


Target



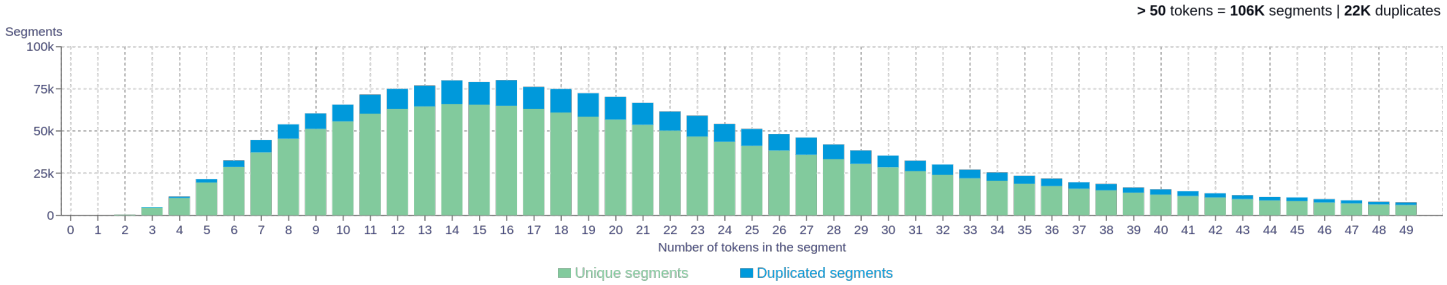
Source segment length distribution by token

<= 49 tokens = **1.8M** segments | **50K** duplicates
> 50 tokens = **132K** segments | **2.1K** duplicates

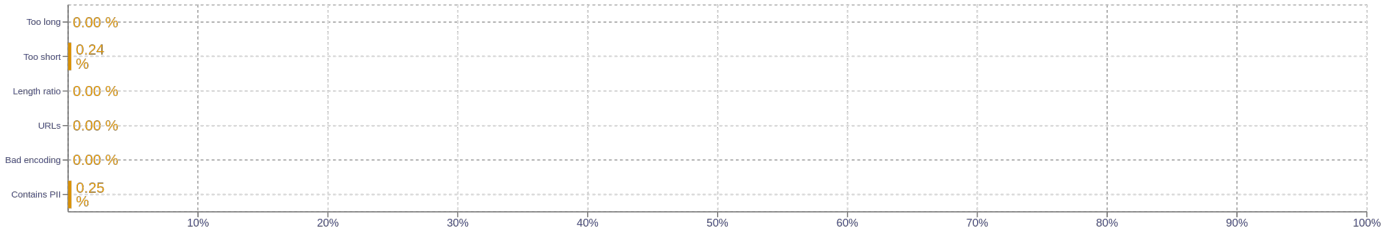


Target segment length distribution by token

<= 49 tokens = **1.5M** segments | **339K** duplicates
> 50 tokens = **106K** segments | **22K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	god | 141494 · one | 110844 · also | 91869 · people | 87092 · game | 77360
2	jesus christ | 10955 · holy spirit | 9421 · united states | 6326 · human rights | 5815 · personal information | 5610
3	copy the code | 5341 · code and paste | 5333 · insert the game | 5253 · like the game | 4960 · forget to rate | 4636
4	water and the spirit | 6197 · gospel of the water | 5475 · paste in the html | 5330 · code of your site | 5330 · game with your best | 4573
5	copy the code and paste | 5331 · paste in the html code | 5330 · html code of your site | 5330 · game with your best friends | 4573 · forget to rate this game | 4567

Target n-grams

Size	n-grams
1	au | 185847 · yako | 171838 · mungu | 166221 · hii | 142189 · ambayo | 115520
2	tovuti yako | 18611 · chuo kikuu | 17721 · mwenyezi mungu | 16555 · mchezo huu | 16014 · kufanya kazi | 14887
3	umoja wa mataifa | 12296 · simu ya mkono | 9064 · umri wa miaka | 7415 · vyombo vya habari | 6521 · hali ya hewa | 5872
4	html ya tovuti yako | 5333 · kuweka katika kanuni html | 5332 · kanuni html ya tovuti | 5332 · huu kwa rafiki yako | 4615 · mchezo huu kwa rafiki | 4580
5	nakala ya kanuni na kuweka | 5332 · kanuni na kuweka katika kanuni | 5332 · kanuni html ya tovuti yako | 5332 · mchezo huu kwa rafiki yako | 4579 · kushiriki mchezo huu kwa rafiki | 4578

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