HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	SL	TL
hplt-v2-en-af.tsv	1/23/2025	English (en)	Afrikaans (af)

Volumes

Segments	SL tokens	SL characters	SL size
3,987,340	90M	450,451,134	431.75 MB

TL tokens	TL characters	TL size
94M	473,345,985	454.09 MB

Dataset top 10 domains

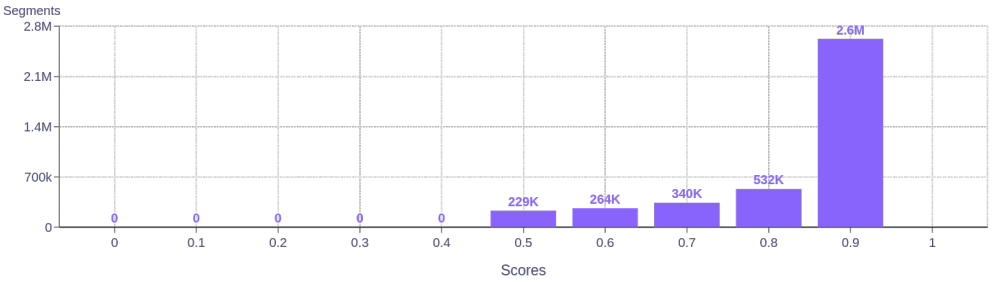
SL domain	Segments	TL domain	Segments
google.com	21.2%	wikipedia.org	9.9%
wikipedia.org	11.8%	google.com	8.4%
softoware.net	5.5%	sacred-texts.com	5.9%
sacred-texts.com	5.4%	softoware.net	4.6%
androware.net	2.9%	androware.net	2.3%
bibliaonline.com.br	2.3%	bibliaonline.com.br	1.9%
w3eacademy.com	1.8%	w3eacademy.com	1.8%
itsmygame.org	1.7%	jw.org	1.6%
jw.org	1.5%	itsmygame.org	1.3%
amightywind.com	1.4%	dualjuridik.org	1.2%

Dataset top 10 TLDs

SL domain	Segments	TL domain	Segments
com	95.6%	com	64.8%
org	29.7%	org	24.5%
net	15.3%	net	12.0%
co.za	5.6%	co.za	8.6%
com.br	2.5%	com.br	2.2%
ac.za	2.1%	ac.za	1.9%
org.za	1.7%	org.za	1.8%
name	1.4%	name	1.5%
co.uk	0.8%	nl	0.7%
nl	0.7%	today	0.6%

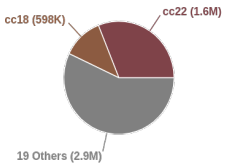
Translation likelihood

≥ 5 = 4M segments | **100.0%**
≥ 8 = 3.2M segments | **79.1%**
< 5 = 0 segments | **0.0%**



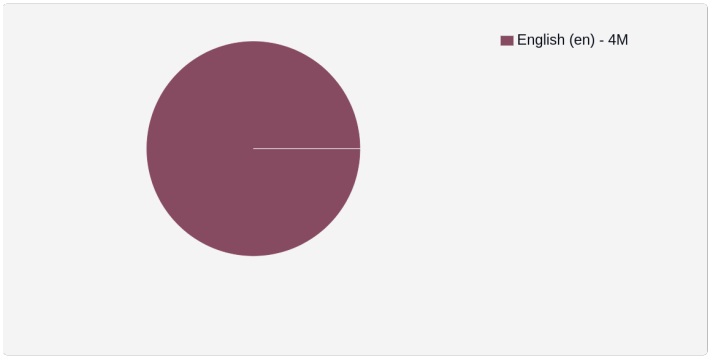
Collections

CC = 64.29%
IA = 35.71%

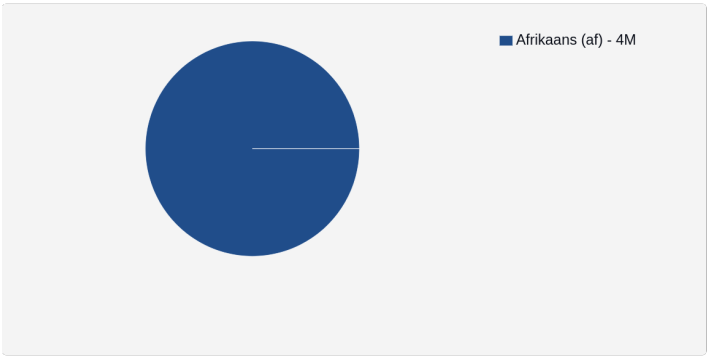


Language Distribution

Source

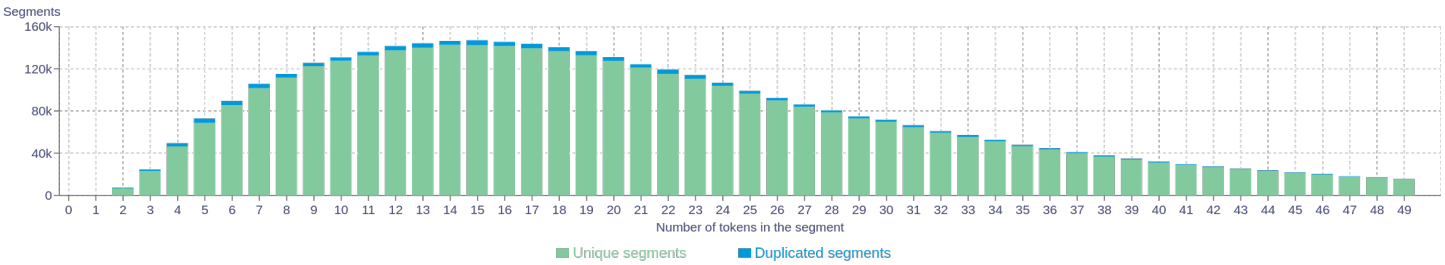


Target



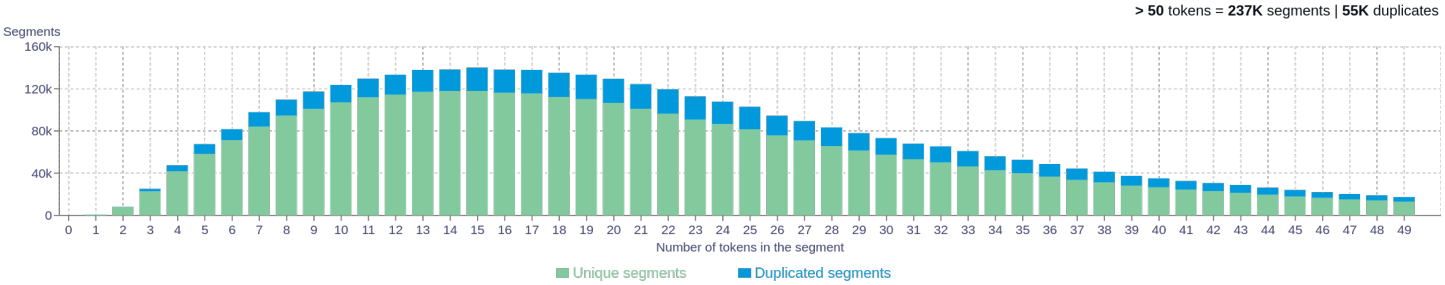
Source segment length distribution by token

<= 49 tokens = **3.7M** segments | **110K** duplicates
> 50 tokens = **210K** segments | **4.8K** duplicates

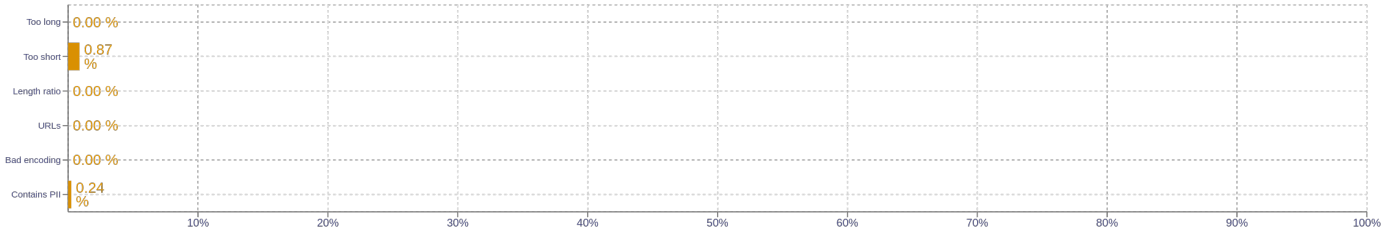


Target segment length distribution by token

<= 49 tokens = **3.1M** segments | **678K** duplicates
> 50 tokens = **237K** segments | **55K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	one | 190313 · also | 176070 · game | 150549 · god | 149943 · time | 131479
2	south africa | 17006 · jesus christ | 12647 · holy spirit | 12192 · united states | 10639 · online game | 9570
3	like the game | 9844 · share the game | 5983 · send the link | 5982 · copy and send | 5973 · forget to rate | 5803
4	link to a friend | 5976 · game with the world | 5972 · game with your best | 5734 · paste in the html | 4891 · code of your site | 4891
5	friend or all your friends | 5972 · copy and send the link | 5972 · game with your best friends | 5734 · forget to rate this game | 5733 · copy the code and paste | 4892

Target n-grams

Size	n-grams
1	word | 549041 · of | 427004 · hierdie | 400350 · deur | 363957 · tot | 238849
2	word deur | 26253 · gebruik word | 25087 · wysig bron | 15304 · toegang tot | 15183 · heilige gees | 14284
3	kinders van israel | 7342 · spreek die here | 6189 · stuur die skakel | 5977 · kopie en stuur | 5973 · wil die spel | 5962
4	deel van die wedstryd | 5977 · wedstryd met die wêreld | 5973 · of al jou vriende | 5973 · moenie vergeet om hierdie | 5810 · html-kode van jou site | 4897
5	vriend of al jou vriende | 5973 · kopie en stuur die skakel | 5973 · speletjie saam met jou beste | 5757 · kopieer die kode en plak | 4898 · plak dit in die html-kode | 4897

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