HPLT Analytics report

HPLTAnalytics

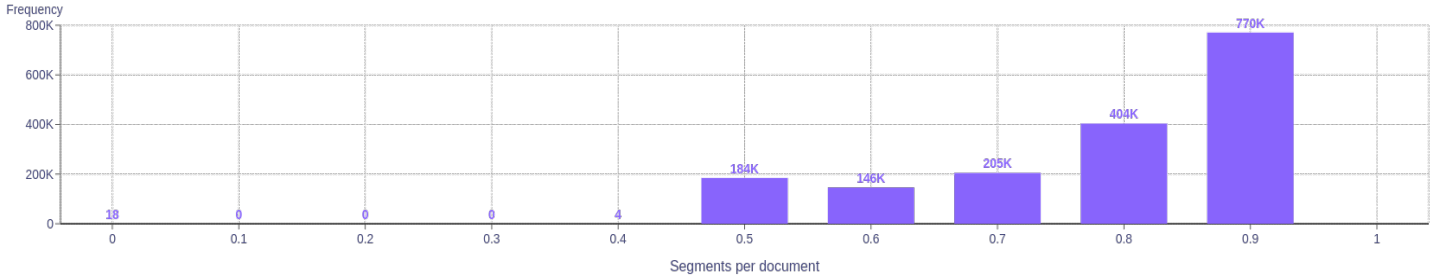
General overview

Corpus	Analytics date	Source language	Target language
HPLT.en-sw	10/23/2023	English (en)	Swahili (sw)

Volumes

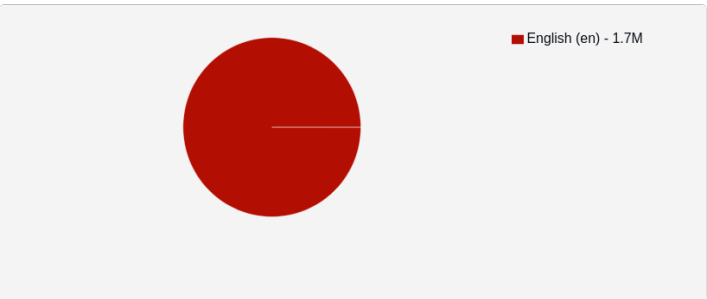
Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size	Src characters	Trg characters
1,710,223	1,710,206 (100.00 %)	25M	26M	112.98 MB	130.75 MB		

Translation likelihood

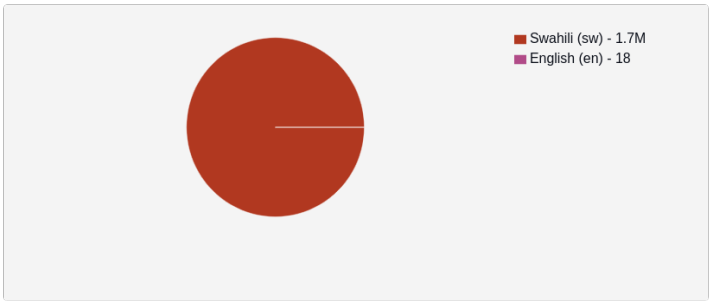


Language Distribution

Source



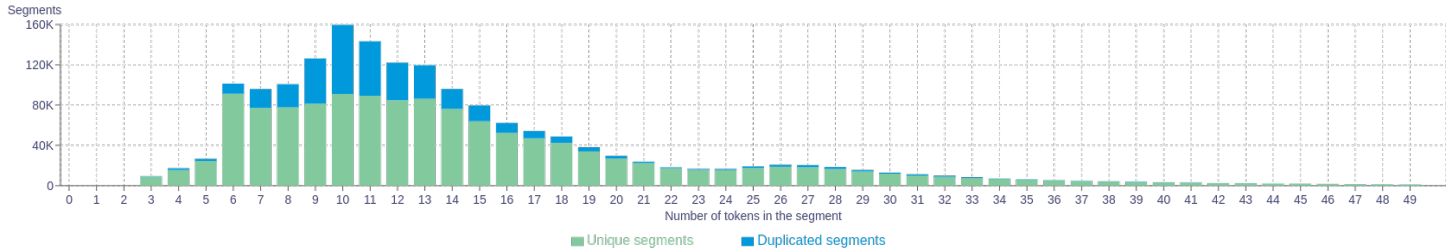
Target



Source segment length distribution by token

<= 49 tokens = **1.3M** segments | **380K** duplicates

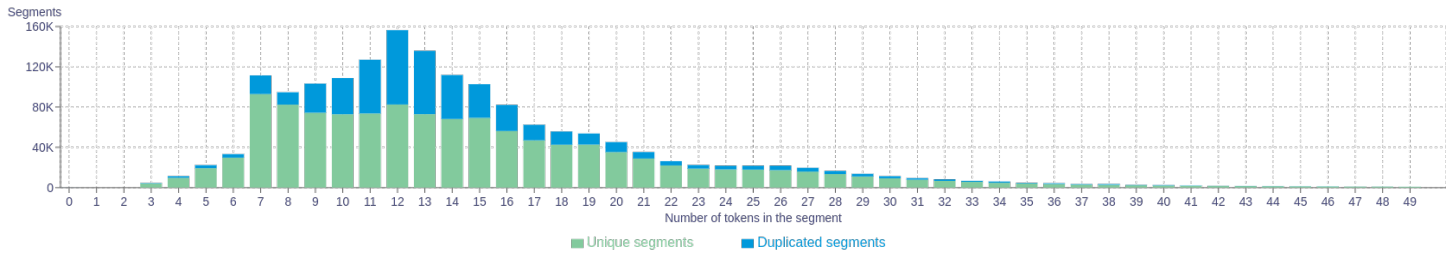
> 50 tokens = **15K** segments | **201** duplicates



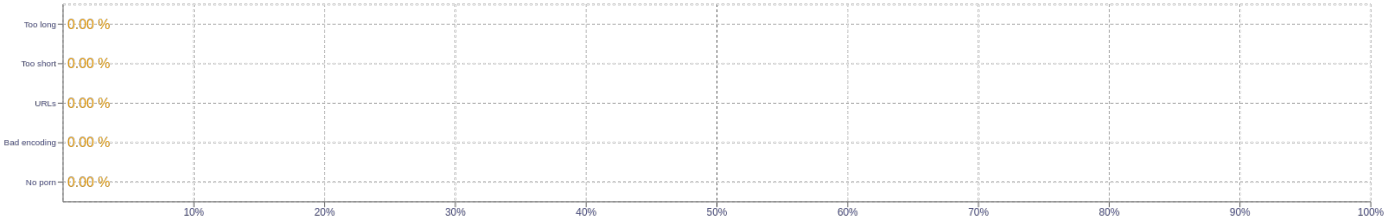
Target segment length distribution by token

<= 49 tokens = **1.2M** segments | **495K** duplicates

> 50 tokens = **10K** segments | **504** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	english | 545088 · used | 408366 · books | 318411 · year | 295517 · ago | 293229
2	year ago | 290486 · rare books | 84158 · used books | 84146 · second hand | 84140 · hand books | 84121
3	second hand books | 84121 · books and second | 84121 · available rare books | 84121 · apps on google | 12118 · visit website email | 10151
4	posted over a year | 96028 · used books and second | 84121 · books of the title | 84121 · books and second hand | 84121 · posts have been made | 14153
5	posted over a year ago | 95793 · used books and second hand | 84121 · hand books of the title | 84121 · books and second hand books | 84121 · android apps on google play | 12116

Target n-grams

Size	n-grams
1	lugha | 594814 · kiingereza | 568720 · kutumika | 401071 · mwaka | 313639 · uliopita | 304802
2	zimeorodheshwa kabisa | 85837 · vitabu vya | 84577 · vya kichwa | 84144 · vitabu kutumika | 84126 · pili vitabu | 84126
3	lugha ya kiingereza | 566170 · mwaka mmoja uliopita | 304319 · mkono wa pili | 84135 · kutumika na mkono | 84127 · vitabu vya kichwa | 84126
4	ilitumwa zaidi ya mwaka | 86552 · vitabu kutumika na mkono | 84126 · pili vitabu vya kichwa | 84126 · mkono wa pili vitabu | 84126 · mwaka mmoja uliopita by | 53485
5	kutumika na mkono wa pili | 84127 · mkono wa pili vitabu vya | 84126 · hakuna chapisho zilizowekwa kwa ukuta | 14368 · programu za android kwenye google | 12125 · tembelea tovuti tuma barua pepe | 10242

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