HPLT Analytics report

HPLTAnalytics

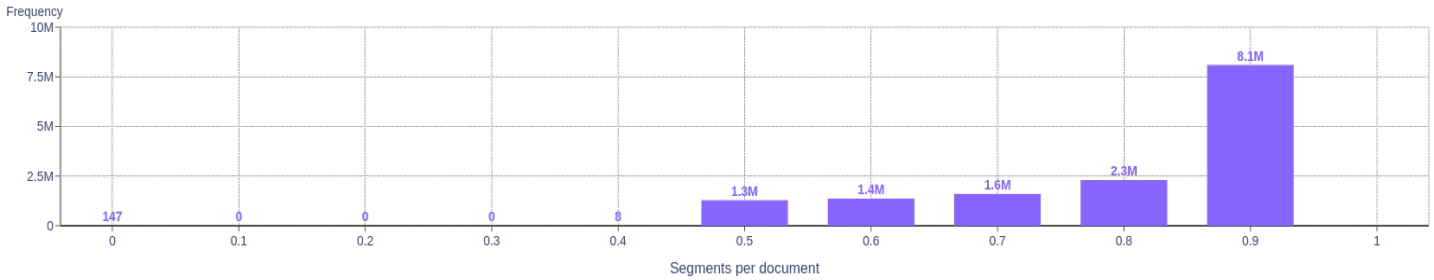
General overview

Corpus	Analytics date	Source language	Target language
HPLT.en-ar	10/28/2023	English (en)	Arabic (ar)

Volumes

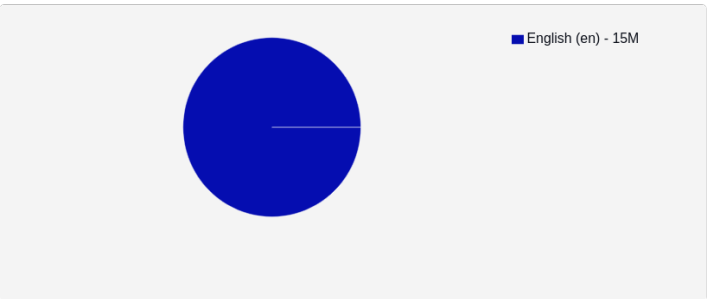
Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size	Src characters	Trg characters
14,645,275	14,645,129 (100.00 %)	282M	283M	1.4 GB	2.16 GB		

Translation likelihood



Language Distribution

Source

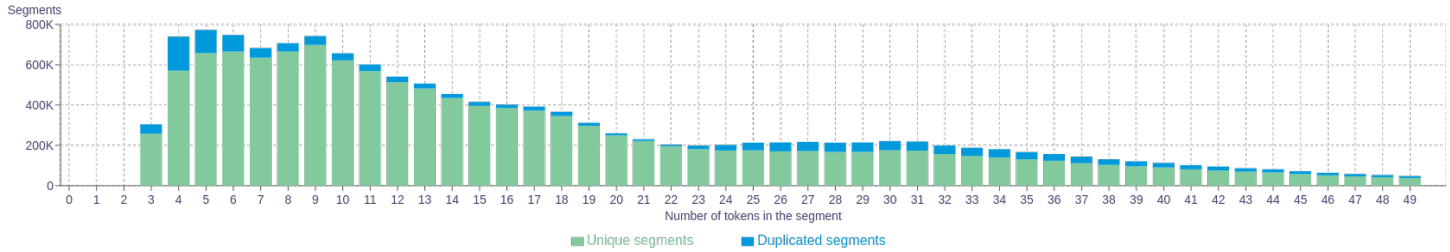


Target

Source segment length distribution by token

<= 49 tokens = **12M** segments | **1.6M** duplicates

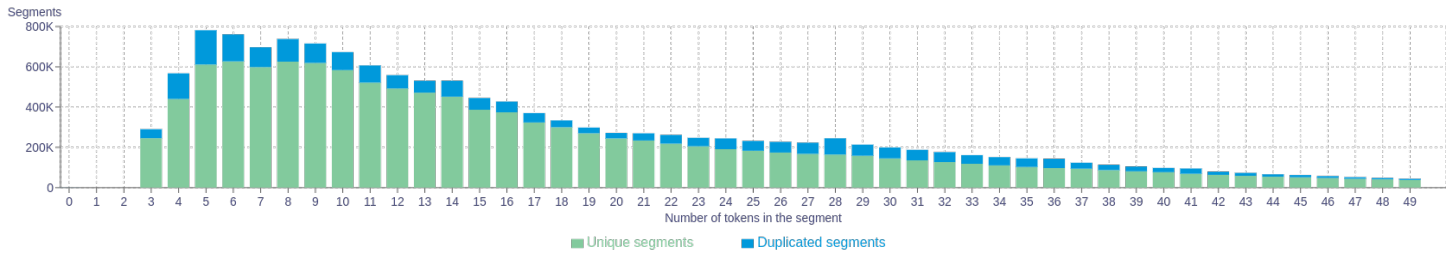
> 50 tokens = **629K** segments | **97K** duplicates



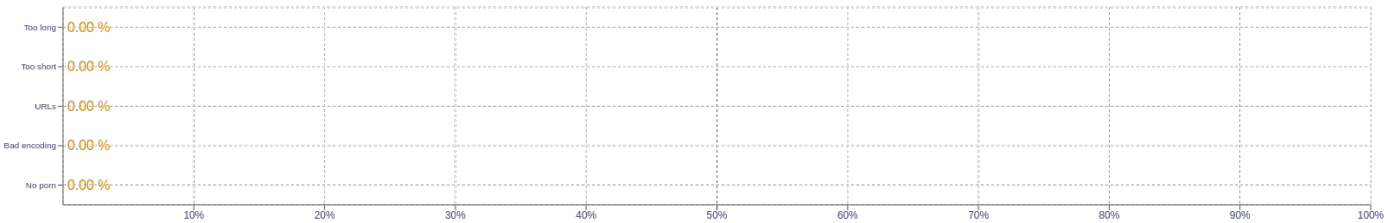
Target segment length distribution by token

<= 49 tokens = **12M** segments | **2.4M** duplicates

> 50 tokens = **695K** segments | **83K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	books | 6081325 · used | 2113903 · available | 2022226 · second | 1929174 · hand | 1904473
2	second hand | 1884215 · rare books | 1884038 · used books | 1883979 · hand books | 1883949 · available rare | 1883945
3	second hand books | 1883949 · books and second | 1883945 · available rare books | 1883945 · compare every offer | 74397 · offer archive entry | 74145
4	used books and second | 1883945 · books of the title | 1883945 · books and second hand | 1883945 · things to do near | 80302 · compare every offer archive | 74145
5	used books and second hand | 1883945 · hand books of the title | 1883945 · books and second hand books | 1883945 · compare every offer archive entry | 74145 · tripadvisor is proud to partner | 45679

Target n-grams

Size	n-grams
1	3804053 | والكتب · 2269116 | الكتب · 2089717 | يتم · 1927928 | المتاحة · 1917519 | المستخدمة
2	1901642 | جهة ثانية · 1901513 | الكتب النادرة · 1901471 | النادرة المتاحة · 1901470 | والكتب المستخدمة · 1901470 | ثانية للعنوان
3	1901470 | والكتب من جهة · 1901470 | والكتب المستخدمة والكتب · 1901470 | جهة ثانية للعنوان · 1901470 | النادرة المتاحة والكتب · 1901470 | المتاحة والكتب المستخدمة
4	1901470 | والكتب من جهة ثانية · 1901470 | النادرة المتاحة والكتب المستخدمة · 1901470 | المستخدمة والكتب من جهة · 1901470 | المتاحة والكتب المستخدمة والكتب · 1901470 | الكتب النادرة المتاحة والكتب
5	1901470 | والكتب من جهة ثانية للعنوان · 1901470 | والكتب المستخدمة والكتب من جهة · 1901470 | النادرة المتاحة والكتب المستخدمة والكتب · 1901470 | المستخدمة والكتب من جهة ثانية · 1901470 | الكتب النادرة المتاحة والكتب المستخدمة

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