HPLT Analytics report

HPLTAnalytics

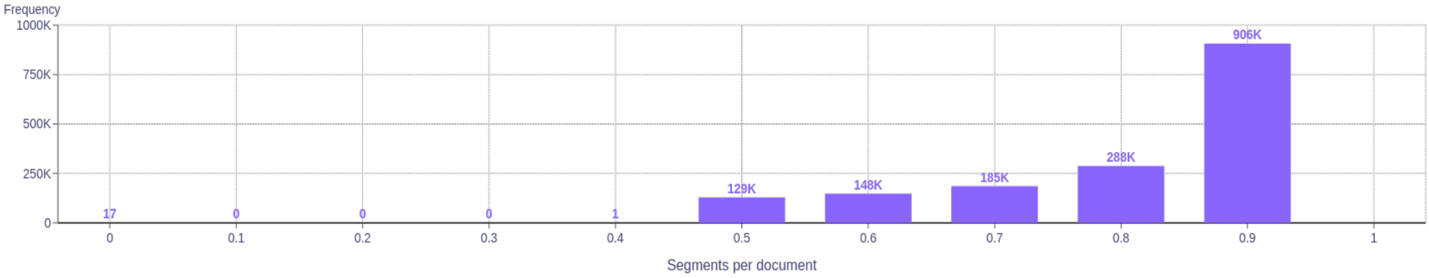
General overview

Corpus	Analytics date	Source language	Target language
HPLT.en-sq	10/26/2023	English (en)	Albanian (sq)

Volumes

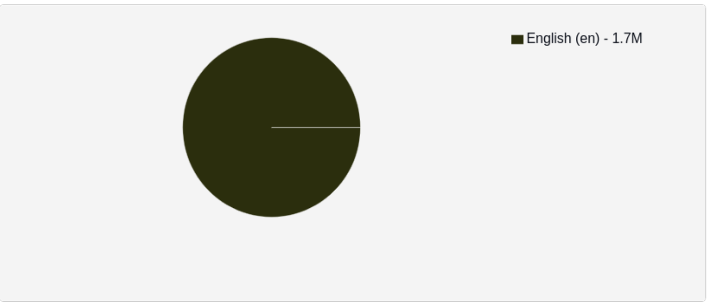
Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size	Src characters	Trg characters
1,655,975	1,655,959 (100.00 %)	29M	31M	147.06 MB	168.64 MB		

Translation likelihood

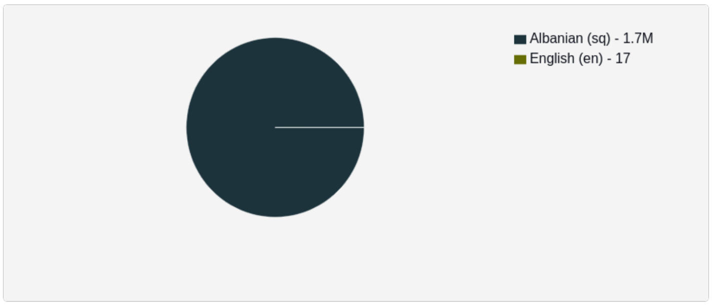


Language Distribution

Source



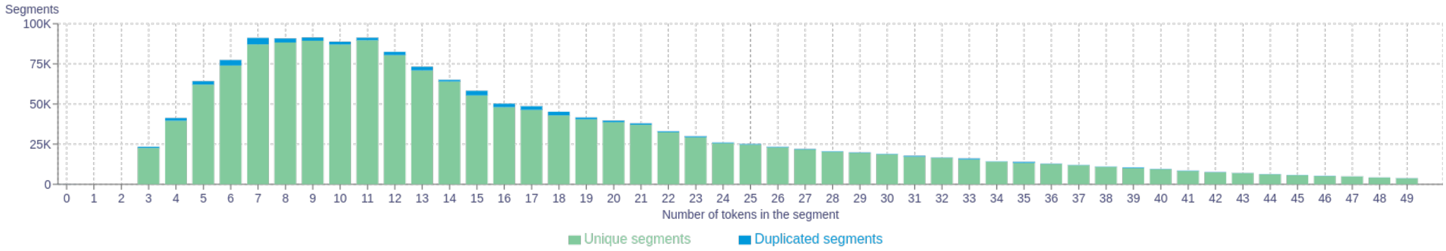
Target



Source segment length distribution by token

<= 49 tokens = **1.6M** segments | **44K** duplicates

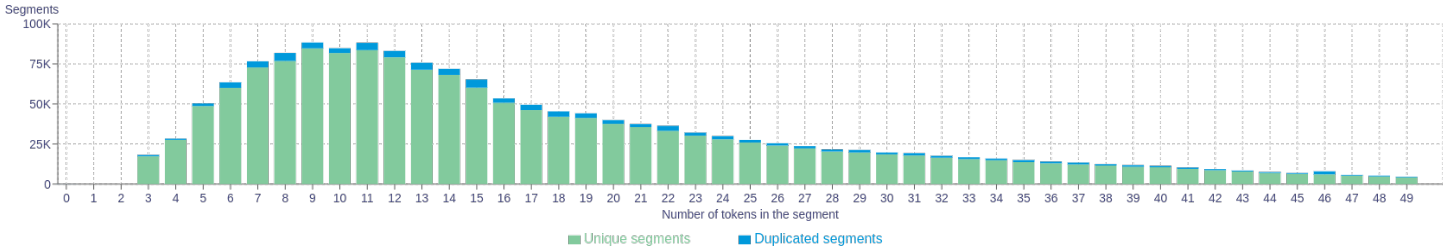
> 50 tokens = **47K** segments | **586** duplicates



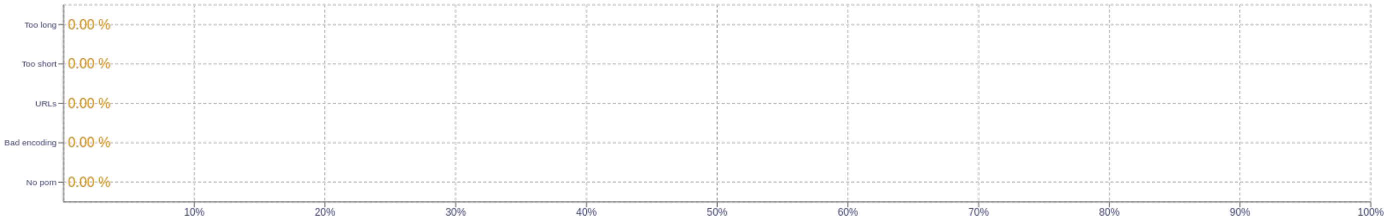
Target segment length distribution by token

<= 49 tokens = **1.5M** segments | **94K** duplicates

> 50 tokens = **56K** segments | **4.4K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	porn | 135689 free | 106398 quality | 91707 video | 84257 hd | 82912
2	good quality | 24513 excellent quality | 24050 hd excellent | 22821 quality hd | 18660 porn video | 16502
3	hd excellent quality | 22798 video in good | 20806 good quality hd | 11911 hd great quality | 10918 site porno365.mobi close | 10755
4	video in good quality | 20795 name of this movie | 16291 site porno365.mobi close category | 10670 eyes of the operator | 7921 hd quality for free | 7866
5	video in good quality hd | 11615 video is in the categories | 9310 original name of this movie | 9310 watch and download in hd | 7833 look free and without registration | 6480

Target n-grams

Size	n-grams
1	të | 2022829 në | 949203 për | 538651 një | 412997 që | 255314
2	për të | 161195 më të | 72291 cilësi të | 64223 të mirë | 45392 në një | 45124
3	hd me cilësi | 34564 cilësi të shkëlqyer | 23296 cilësi të mirë | 21788 për të lira | 21087 video me cilësi | 14770
4	hd me cilësi të | 34562 video me cilësi të | 14759 në hd me cilësi | 11669 faqja porno365.mobi afër kategoria | 10670 falas dhe pa regjistrim | 10045
5	hd me cilësi të shkëlqyer | 22818 hd me cilësi të mirë | 11732 në hd me cilësi të | 11669 video në hd me cilësi | 11392 emri origjinal i këtij filmi | 9310

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