HPLT Analytics report

HPLTAnalytics

General overview

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

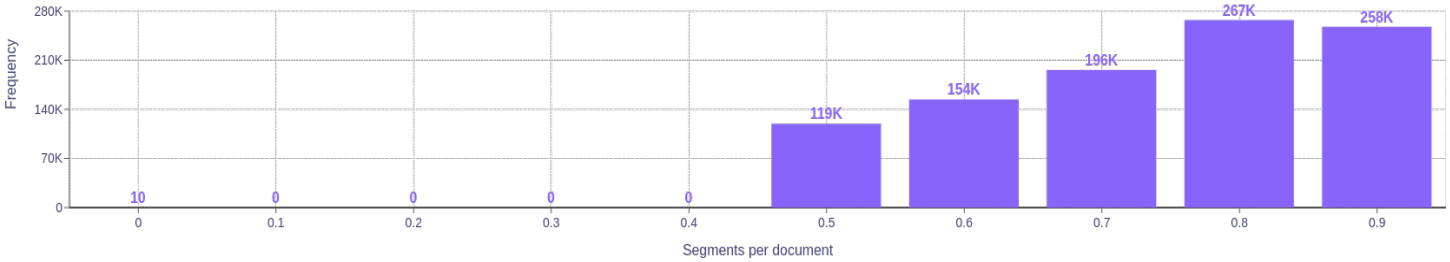
Volumes

Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size
994,756	2,209 (0.22 %)	18M	20M	94.79 MB	113.35 MB

Type-Token Ratio

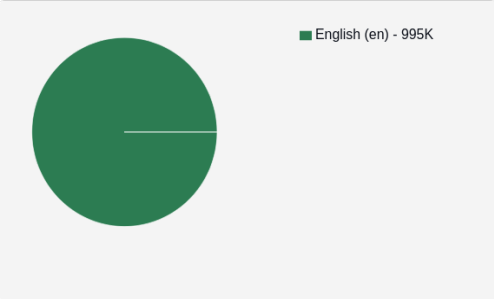
Source	Target
0.01	0.02

Translation likelihood

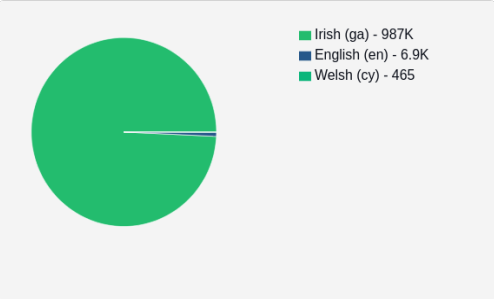


Language Distribution

Source



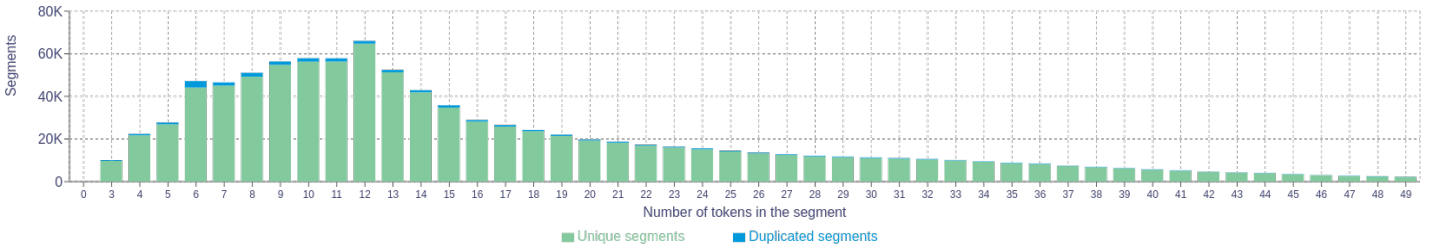
Target



Source segment length distribution by token

<= 49 tokens = **931K** segments | **27K** duplicates

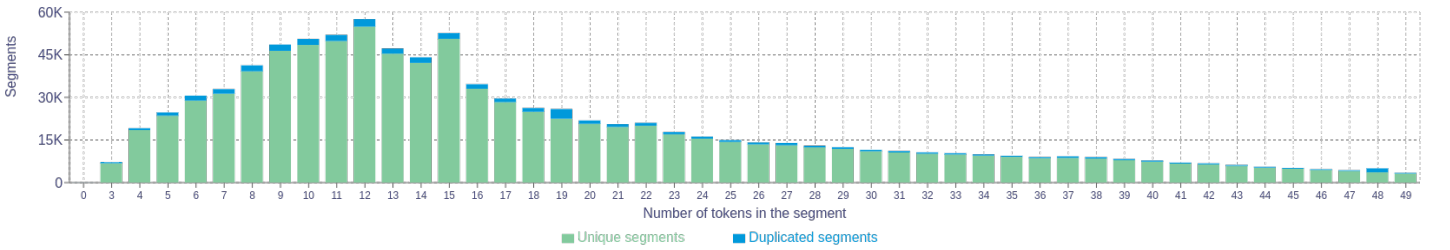
> 50 tokens = **37K** segments | **1.1K** duplicates



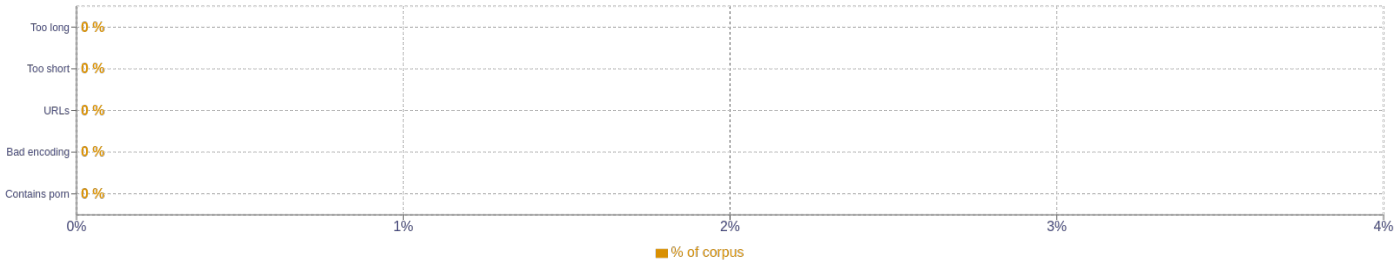
Target segment length distribution by token

<= 49 tokens = **899K** segments | **47K** duplicates

> 50 tokens = **49K** segments | **4.1K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	porn | 168845 · free | 137838 · quality | 119879 · hd | 108388 · site | 103134
2	excellent quality | 38113 · hd excellent | 36807 · good quality | 35092 · quality hd | 34752 · without registration | 29869
3	hd excellent quality | 36779 · video in good | 29731 · good quality hd | 22720 · free and without | 17869 · hd great quality | 12000
4	video in good quality | 29727 · name of this movie | 25821 · free and without registration | 17454 · look free and without | 16169 · eyes of the operator | 11834
5	video in good quality hd | 22412 · look free and without registration | 16169 · video in high quality hd | 11008 · video is in the categories | 8618 · original name of this movie | 8618

Target n-grams

Size	n-grams
1	porn | 177398 · saor | 137800 · aisce | 135445 · hd | 108810 · suíomh | 105501
2	féidir leat | 36056 · caighdeán maith | 34370 · hd cáilíochta | 26015 · maith hd | 22398 · porn saor | 22056
3	saor in aisce | 134561 · físeán i caighdeán | 30133 · caighdeán maith hd | 22396 · chaighdeán den scoth | 22242 · cáilíochta den scoth | 15817
4	físeán i caighdeán maith | 29987 · aisce agus gan chlárú | 22824 · porn saor in aisce | 21962 · hd chaighdeán den scoth | 21563 · t-ainm ar an scannán | 17310
5	físeán i caighdeán maith hd | 22342 · físeán i chaighdeán ard hd | 10683 · aisce agus gan chlárú catagóir | 10305 · féach ar saor in aisce | 9290 · féach ar agus a íoslódáil | 8555

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).

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.

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