HPLT Analytics report

HPLTAnalytics

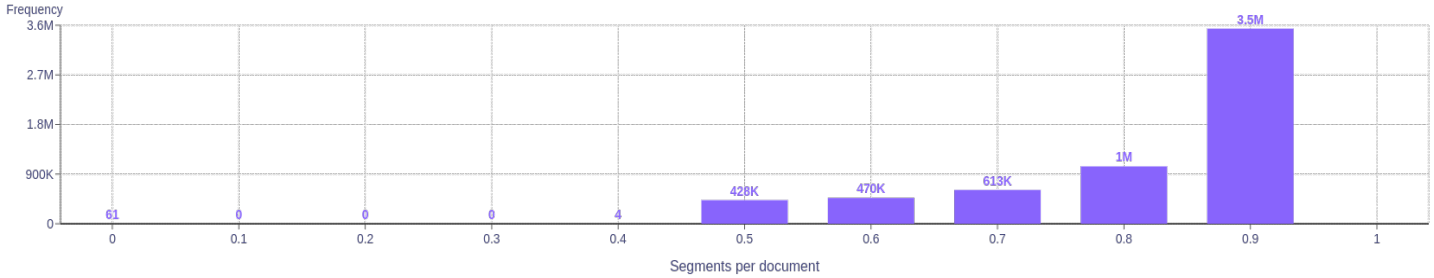
General overview

Corpus	Analytics date	Source language	Target language
HPLT.en-et	10/25/2023	English (en)	Estonian (et)

Volumes

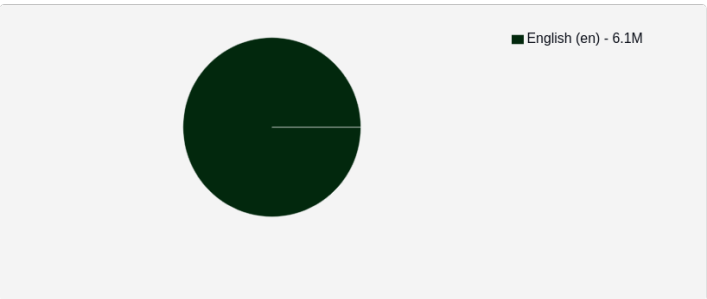
Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size	Src characters	Trg characters
6,089,852	6,089,792 (100.00 %)	111M	90M	564.07 MB	579.75 MB		

Translation likelihood

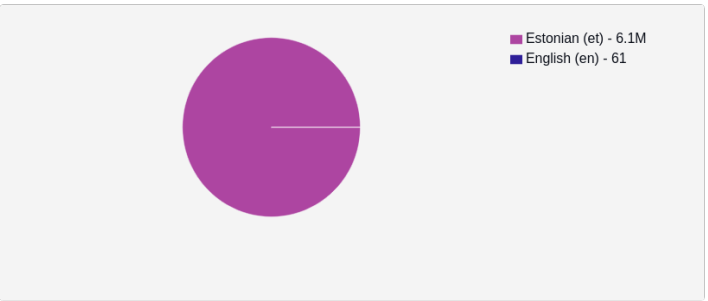


Language Distribution

Source



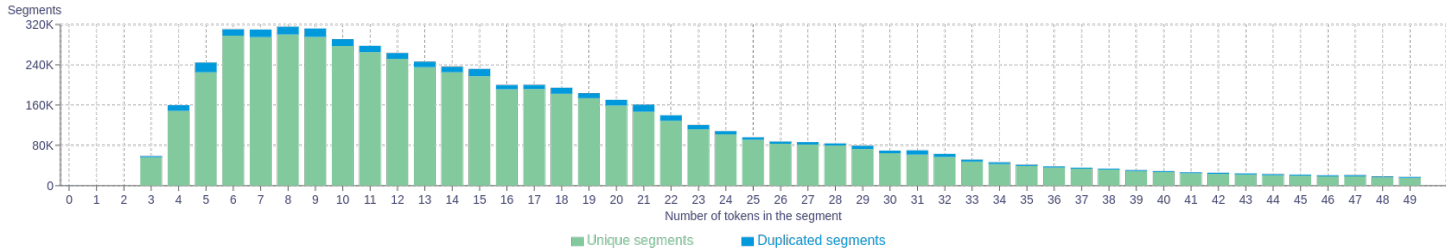
Target



Source segment length distribution by token

<= 49 tokens = **5.5M** segments | **339K** duplicates

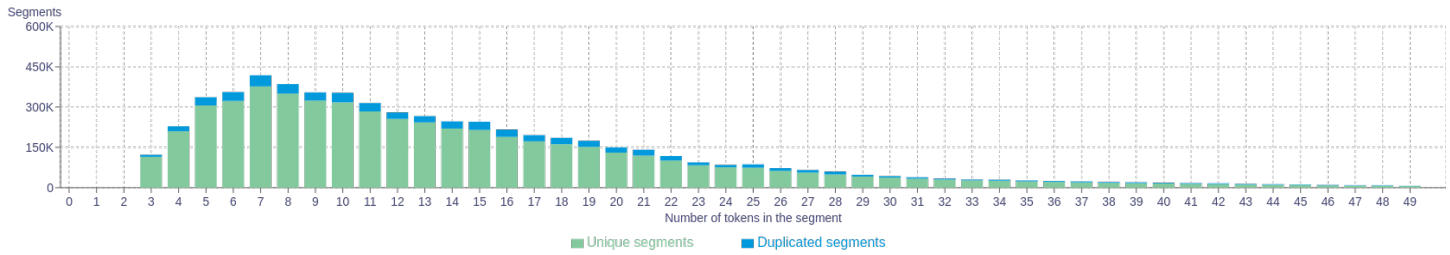
> 50 tokens = **214K** segments | **13K** duplicates



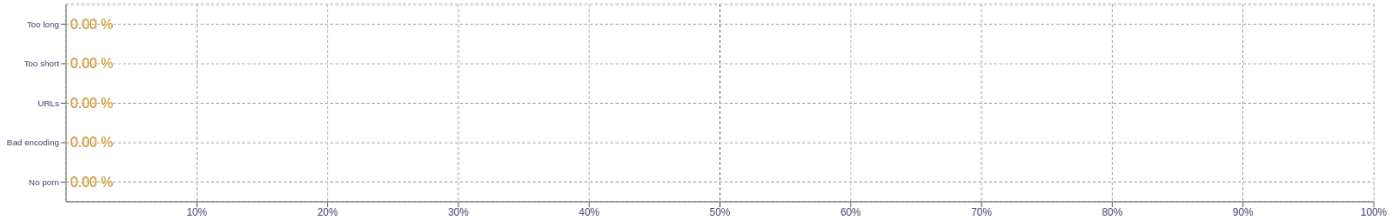
Target segment length distribution by token

<= 49 tokens = **5.4M** segments | **649K** duplicates

> 50 tokens = **74K** segments | **6.8K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	best | 319870 · airport | 299809 · books | 285365 · hotel | 276390 · car | 275518
2	car hire | 118255 · best prices | 103370 · best price | 88195 · second hand | 87976 · rare books | 87927
3	year of manufacture | 142172 · second hand books | 87908 · books and second | 87908 · available rare books | 87908 · amend your booking | 63983
4	used books and second | 87908 · books of the title | 87908 · books and second hand | 87908 · find you the best | 58984 · get the best price | 58956
5	used books and second hand | 87908 · hand books of the title | 87908 · books and second hand books | 87908 · amend your booking for free | 63978 · rentalcars.com and you can amend | 63977

Target n-grams

Size	n-grams
1	või | 503390 · ning | 343835 · kasutatud | 234275 · asukohas | 217472 · tasuta | 212514
2	kasutatud raamatute | 176436 · haruldased raamatud | 88219 · saadaval haruldased | 88218 · raamatute pealkiri | 88218 · täielikult loetletud | 85807
3	saadaval haruldased raamatud | 88218 · raamatute ja kasutatud | 88218 · kasutatud raamatute pealkiri | 88218 · saate oma broneeringut | 63981 · broneeringut tasuta muuta | 63979
4	raamatute ja kasutatud raamatute | 88218 · kasutatud raamatute ja kasutatud | 88218 · saate oma broneeringut tasuta | 63980 · kaudu ja te saate | 63880 · leida teile parimad hinnapakkumised | 58944
5	raamatute ja kasutatud raamatute pealkiri | 88218 · kasutatud raamatute ja kasutatud raamatute | 88218 · saate oma broneeringut tasuta muuta | 63979 · veebilehe kaudu ja te saate | 63880 · meie juures ja me garanteerime | 58844

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