

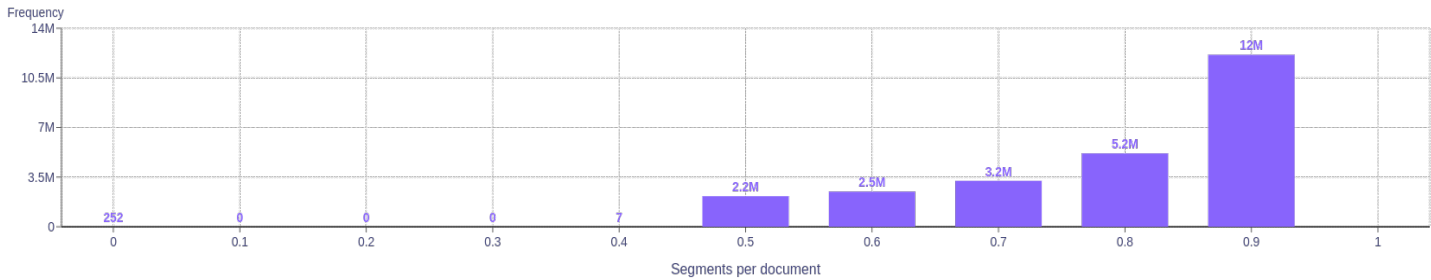
General overview

Corpus	Analytics date	Source language	Target language
HPLT.en-fi	10/30/2023	English (en)	Finnish (fi)

Volumes

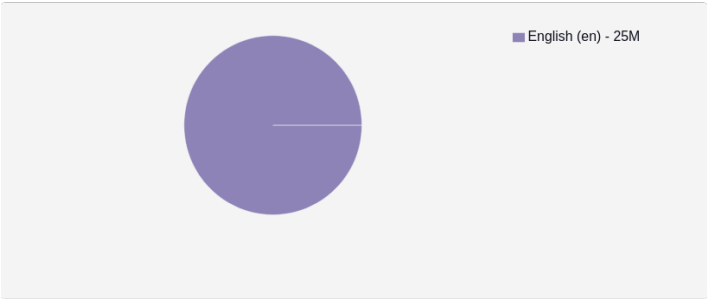
Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size	Src characters	Trg characters
25,176,714	25,176,463 (100.00 %)	397M	323M	1.95 GB	2.12 GB		

Translation likelihood



Language Distribution

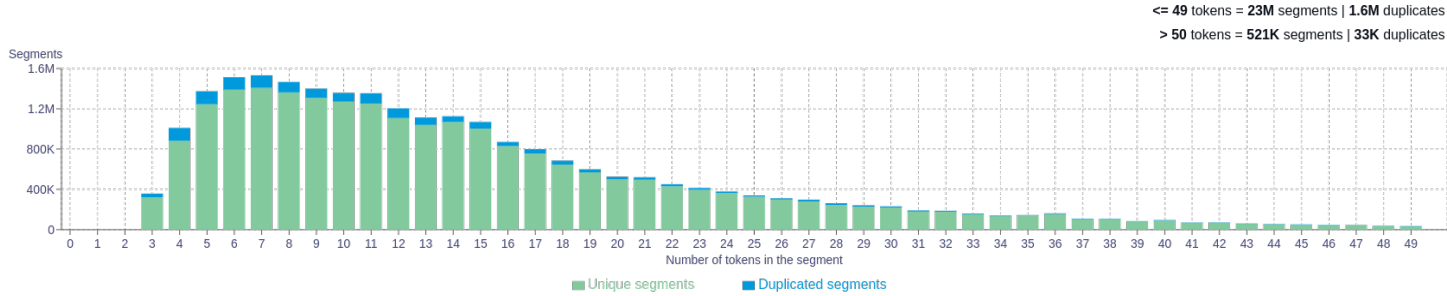
Source



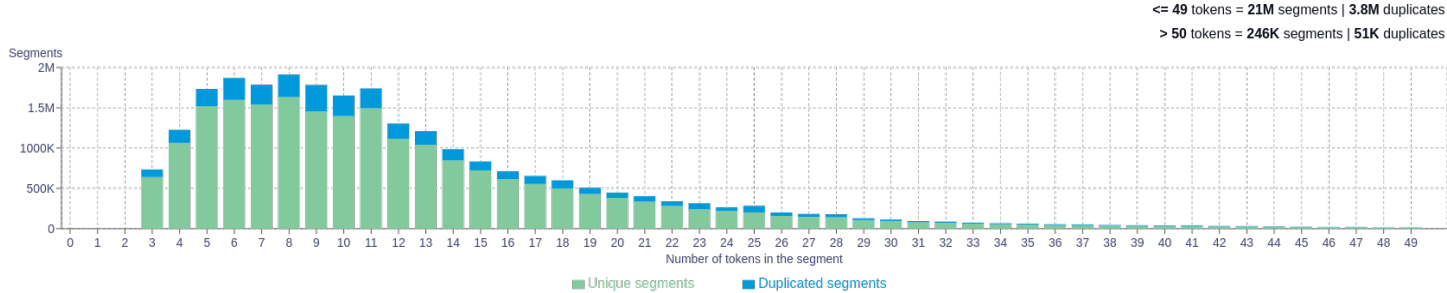
Target



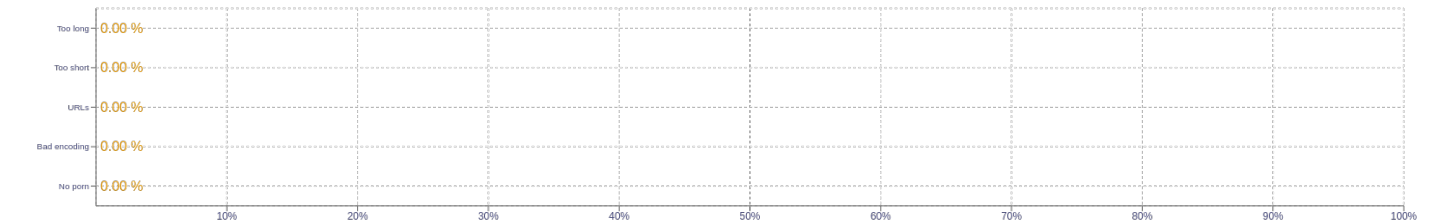
Source segment length distribution by token



Target segment length distribution by token



Segment pair noise distribution



Source n-grams

Size	n-grams
1	hotels 1434341 hotel 1171628 best 947170 near 911394 used 907998
2	hotels near 492361 weather forecast 319926 car hire 298363 love balls 242893 cheap flights 235100
3	year of manufacture 151215 condition not indicated 99083 weather forecast would 91964 quickly and easily 89491 find the best 88323
4	weather will be like 85133 get the best price 70732 things to do near 70415 prices from either machinery 69472 machinery dealers or private 69472
5	rentalcars.com and you can amend 71441 amend your booking for free 71440 prices from either machinery dealers 69472 machinery dealers or private sellers 69472 either machinery dealers or private 69472

Target n-grams

Size	n-grams
1	kohteessa 1345032 lähellä 994557 hotellit 966599 hotel 693268 käytetty 623117
2	lähellä paikkaa 456190 hotellit lähellä 429057 lähellä kohdetta 247757 halvat lennot 209498 tulee olemaan 181407
3	hotellit lähellä paikkaa 292322 ravintolat lähellä paikkaa 155035 vertaa lentojen hintoja 152479 halvinta hintaa lennoillesi 108073 haetko halvinta hintaa 108073
4	haetko halvinta hintaa lennoillesi 108073 sää malli tulee olemaan 87392 verkossa nopeasti ja helposti 86785 tehdä varauksen verkossa nopeasti 86785 lukea asiakasarvosteluita sekä tehdä 86785
5	voit lukea asiakasarvosteluita sekä tehdä 86785 varauksen verkossa nopeasti ja helposti 86785 lukea asiakasarvosteluita sekä tehdä varauksen 86785 asiakasarvosteluita sekä tehdä varauksen verkossa 86785 voit muuttaa varaustasi ilman muutoksuluja 71258

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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>