HPLT Analytics report

HPLTAnalytics

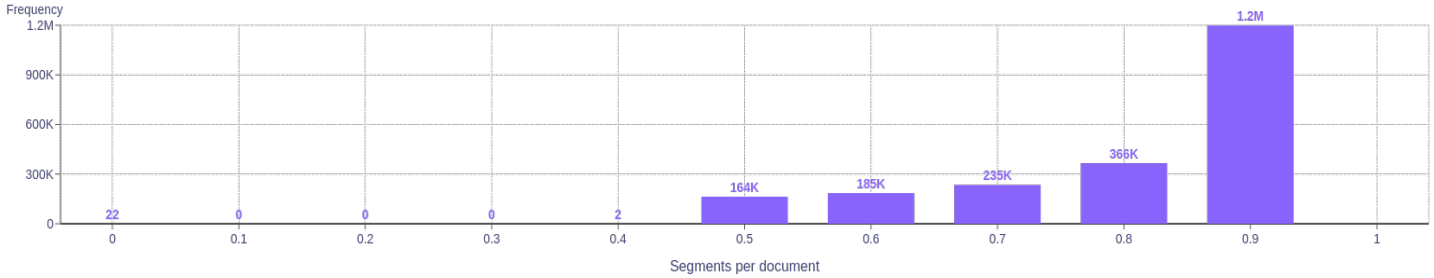
General overview

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

Volumes

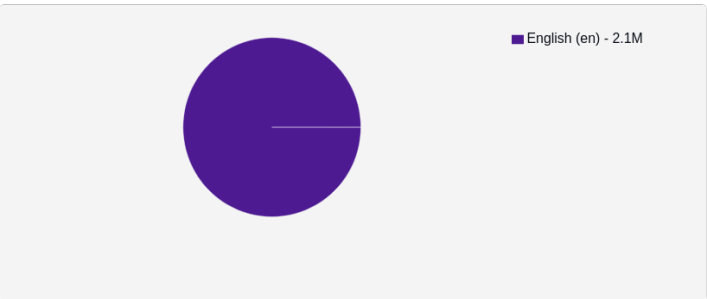
Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size	Src characters	Trg characters
2,148,876	2,148,855 (100.00 %)	33M	33M	167.14 MB	195.03 MB		

Translation likelihood

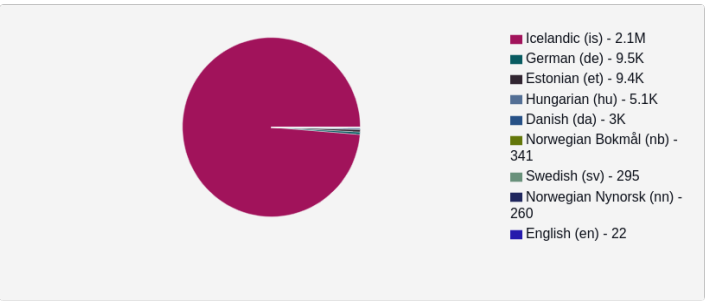


Language Distribution

Source



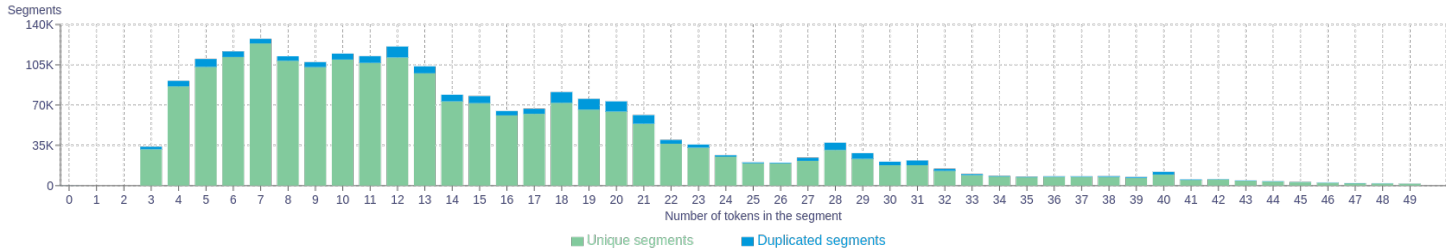
Target



Source segment length distribution by token

<= 49 tokens = **2M** segments | **152K** duplicates

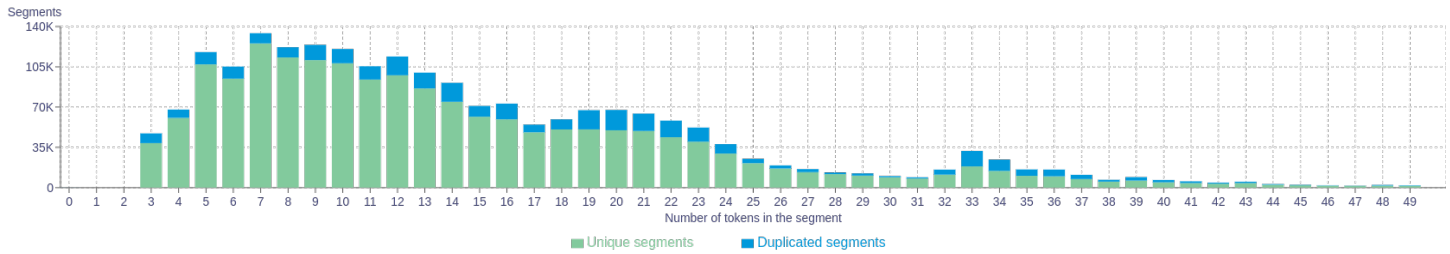
> 50 tokens = **28K** segments | **769** duplicates



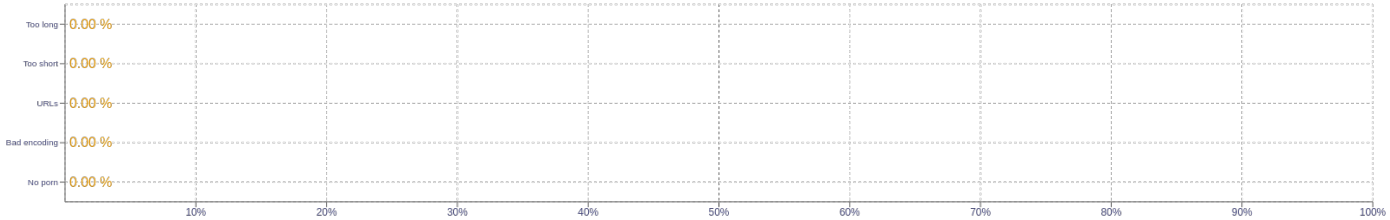
Target segment length distribution by token

<= 49 tokens = **1.8M** segments | **332K** duplicates

> 50 tokens = **23K** segments | **3.2K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	car | 263826 · airport | 260383 · book | 241892 · best | 227192 · prices | 198555
2	car hire | 148479 · best prices | 81098 · best price | 78228 · find great | 77495 · great prices | 77359
3	quickly and easily | 77351 · find great prices | 77285 · see customer ratings | 77284 · booking for free | 76364 · amend your booking | 76364
4	find you the best | 75885 · get the best price | 75857 · work hard to find | 75854 · opens in new window | 40326 · welcoming booking.com guests since | 22128
5	amend your booking for free | 76362 · rentalcars.com and you can amend | 76361 · find you the best prices | 75854 · book with us and get | 75854 · rent a car car hire | 19334

Target n-grams

Size	n-grams
1	bókaðu | 235024 · bílaleigubíl | 177800 · airport | 141736 · hotel | 128558 · verð | 119887
2	besta verðið | 78427 · finndu frábær | 77517 · frábær verð | 77376 · getur breytt | 77002 · breytt bókun | 76871
3	bókaðu á netinu | 77347 · finndu frábær verð | 77338 · verðið á bílaleigubíl | 76935 · getur breytt bókun | 76870 · fáðu besta verðið | 75875
4	besta verðið á bílaleigubíl | 76935 · rentalcars.com og þú getur | 76870 · bókun þinni án endurgjalds | 76870 · airport í gegnum rentalcars.com | 62591 · opnast í nýjum glugga | 40047
5	rentalcars.com og þú getur breytt | 76870 · breytt bókun þinni án endurgjalds | 76870 · fáðu besta verðið á bílaleigubíl | 75873 · tekið á móti gestum booking.com | 22128 · verðs ef litið er tilhlutfallsins | 18199

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