HPLT Analytics report

HPLTAnalytics

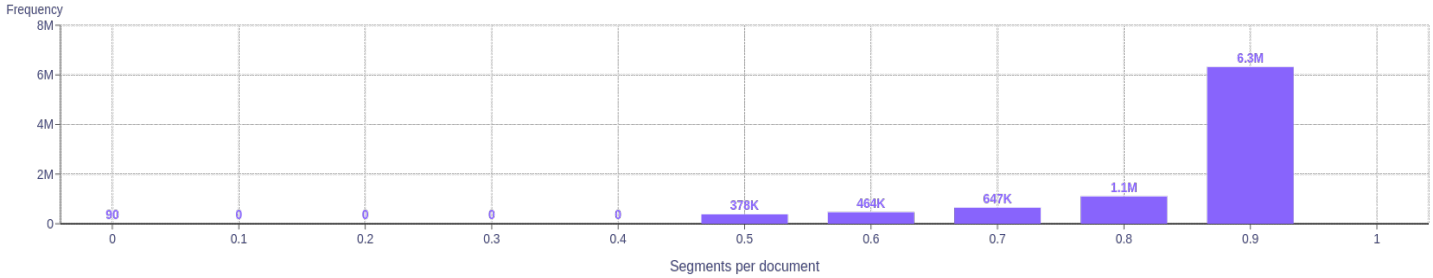
General overview

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

Volumes

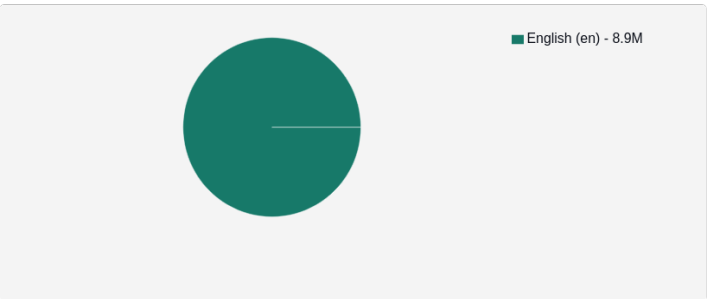
Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size	Src characters	Trg characters
8,905,979	8,905,890 (100.00 %)	165M	184M	834.16 MB	913.2 MB		

Translation likelihood

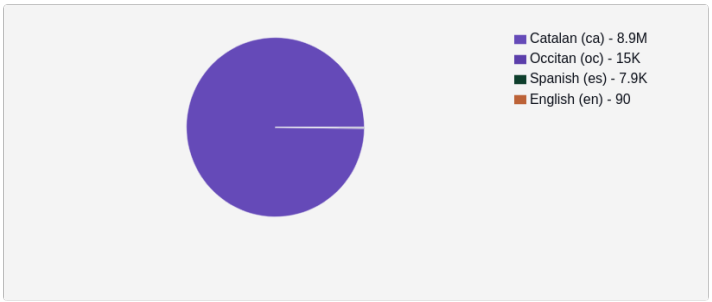


Language Distribution

Source



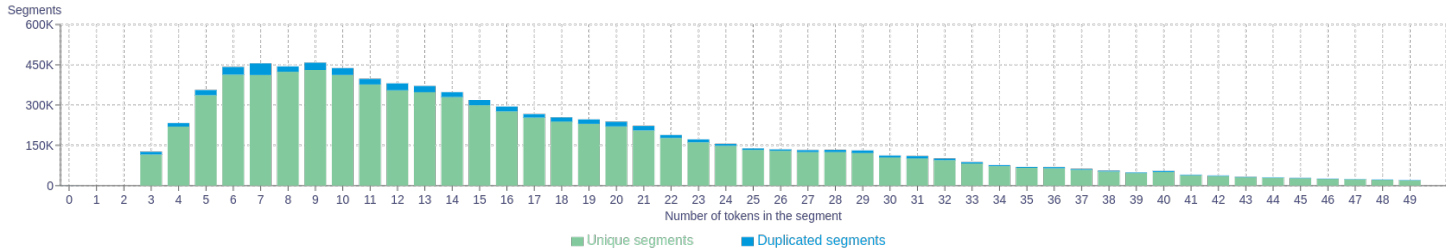
Target



Source segment length distribution by token

<= 49 tokens = **8.1M** segments | **522K** duplicates

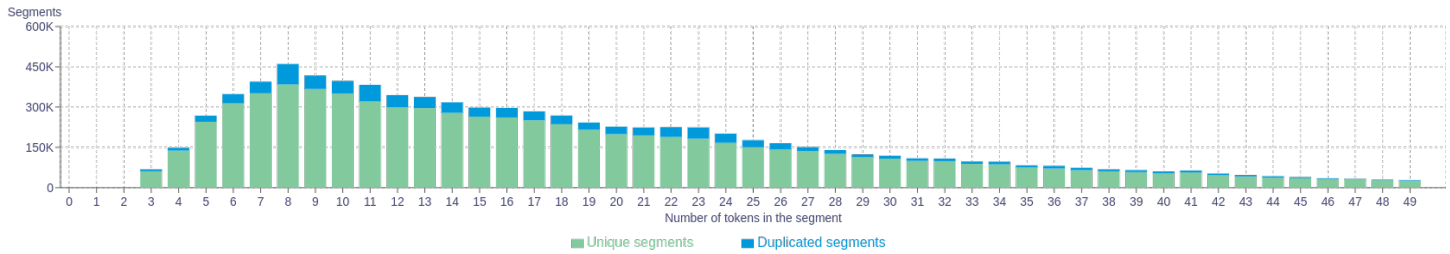
> 50 tokens = **315K** segments | **23K** duplicates



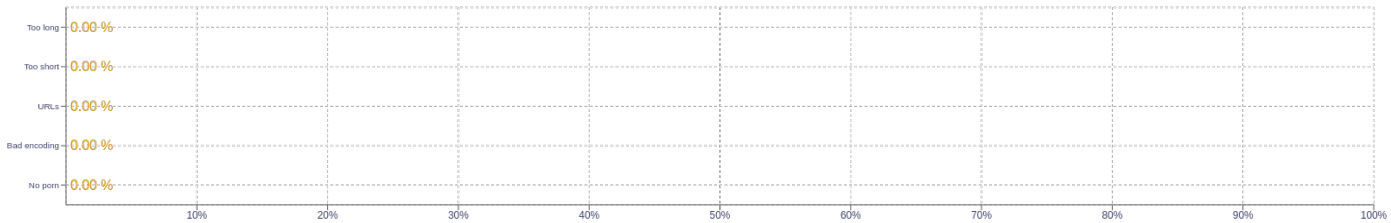
Target segment length distribution by token

<= 49 tokens = **7.4M** segments | **1M** duplicates

> 50 tokens = **440K** segments | **52K** duplicates



Segment pair noise distribution



Source n-grams

Size	n-grams
1	hotel | 576694 · car | 509062 · best | 474642 · airport | 384841 · see | 360587
2	car hire | 259758 · remote control | 183445 · best price | 166453 · vat included | 135162 · universal remote | 98705
3	universal remote control | 97614 · see available equivalences | 90547 · rent a car | 83577 · quickly and easily | 79182 · see customer ratings | 78522
4	models universal remote control | 71234 · get the best price | 61101 · find you the best | 60968 · work hard to find | 60954 · rent a car car | 34191
5	amend your booking for free | 65914 · rentalcars.com and you can amend | 65913 · find you the best prices | 60950 · book with us and get | 60950 · android apps on google play | 30939

Target n-grams

Size	n-grams
1	hotel | 550673 · lloguer | 487274 · hotels | 468082 · millors | 362903 · preus | 297181
2	millors preus | 146450 · millors descomptes | 103549 · distància universal | 96361 · lloc web | 96255 · veure equivalències | 90532
3	lloguer de cotxes | 198059 · comandament a distància | 156846 · descomptes en línia | 103482 · veure equivalències disponibles | 90532 · cotxe de lloguer | 76050
4	millors descomptes en línia | 103480 · hotels amb els millors | 103480 · comandament a distància universal | 96354 · models comandament a distància | 71243 · forma fàcil i ràpida | 71158
5	hotels amb els millors descomptes | 103480 · models comandament a distància universal | 71240 · reserveu online de forma fàcil | 71114 · qualificacions dels clients i reserveu | 71114 · consulteu les qualificacions dels clients | 71114

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