

General overview

Corpus	Analytics date	Source language	Target language
HPLT.en-zh_hant	10/28/2023	English (en)	Chinese (zh)

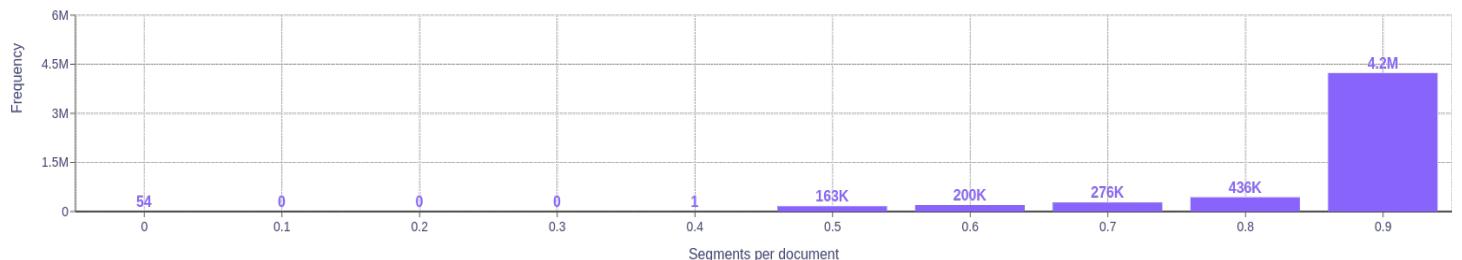
Volumes

Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size
5,306,624	2,987 (0.06 %)	98M	107M	508.07 MB	480.78 MB

Type-Token Ratio

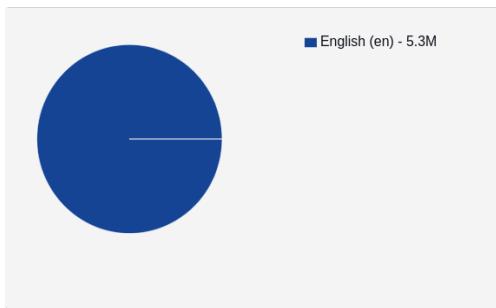
Source	Target
0.01	0.02

Translation likelihood

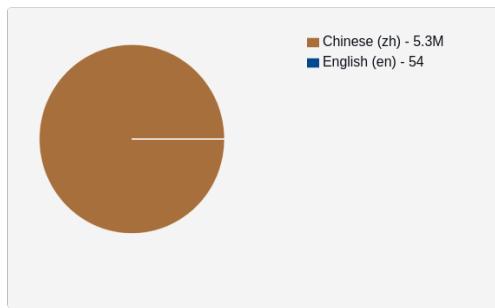


Language Distribution

Source



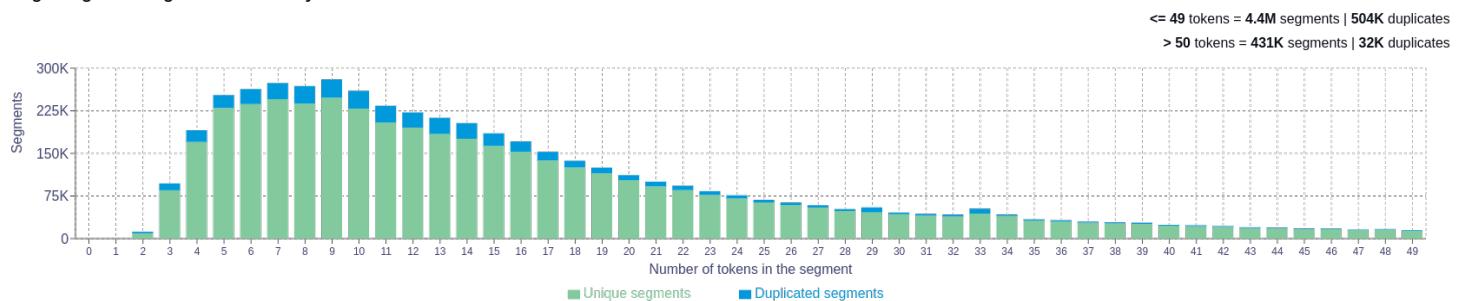
Target



Source segment length distribution by token



Target segment length distribution by token



Segment pair noise distribution



Source n-grams

Size	n-grams
1	(books 382909) (hotels 330601) (near 268940) (hotel 243874) (used 226419)
2	(hotels near 174165) (completely listed 119599) (second hand 115336) (rare books 115234) (used books 115232)
3	(second hand books 115220) (books and second 115219) (available rare books 115219) (ean or isbn 44036) (proud to partner 40723)
4	(used books and second 115219) (books of the title 115219) (books and second hand 115219) (find the perfect hotel 29277) (discounts and special offers 28754)
5	(used books and second hand 115219) (hand books of the title 115219) (books and second hand books 115219) (tripadvisor is proud to partner 40708) (month to find the perfect 28751)

Target n-grams

Size	n-grams
1	(們 447562) (中 434401) (酒店 360384) (個 352444) (提供 322665)
2	(英語 中 125681) (應用 程式 78280) (of t 74041) (micros of 72168) (國際 標準 68766)
3	(用過 的 書 115275) (書 和 二手書 115275) (可用 的 簡本 115275) (二手書 的 標題 115275) (micros of t 71090)
4	(ean 或 國際 標準 67834) (個 月 都 會 40753) (們 每 個 月 40745) (讓 您 可以 保證 40734) (隨時 提供 最 優惠 40730)
5	(用過 的 書 和 二手書 115275) (書 和 二手書 的 標題 115275) (ean 或 國際 標準 書號 67834) (tripadvisor 非常 荣幸 能 與 40744) (們 每 個 月 都 40737)

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>).

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>