

General overview

Corpus	Analytics date	Source language	Target language
HPLT.en-hi	10/31/2023	English (en)	Hindi (hi)

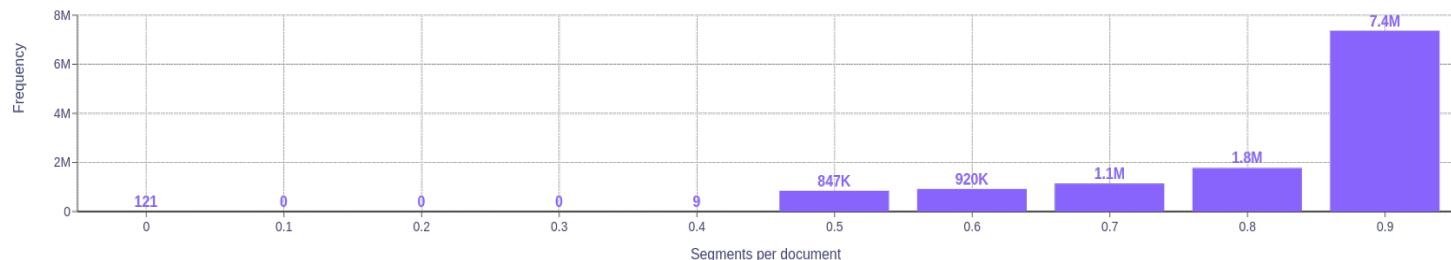
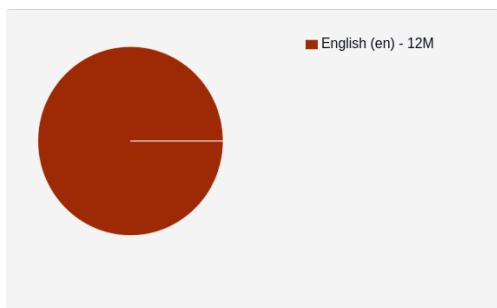
Volumes

Segments	Unique segments	Src tokens	Trg tokens	Src size	Trg size
12,043,190	2,839 (0.02 %)	195M	219M	969.09 MB	2.24 GB

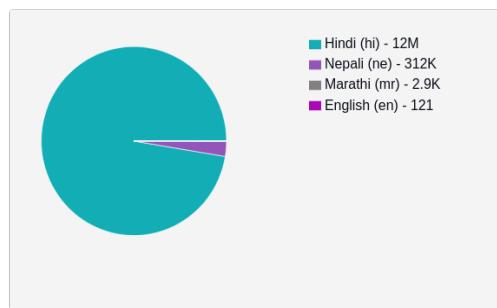
Type-Token Ratio

Source	Target
0.01	0.01

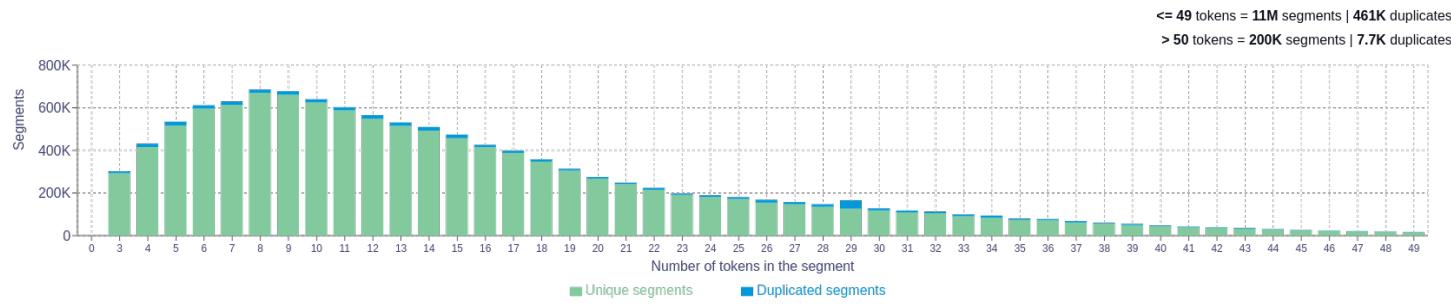
Translation likelihood

Language Distribution
Source

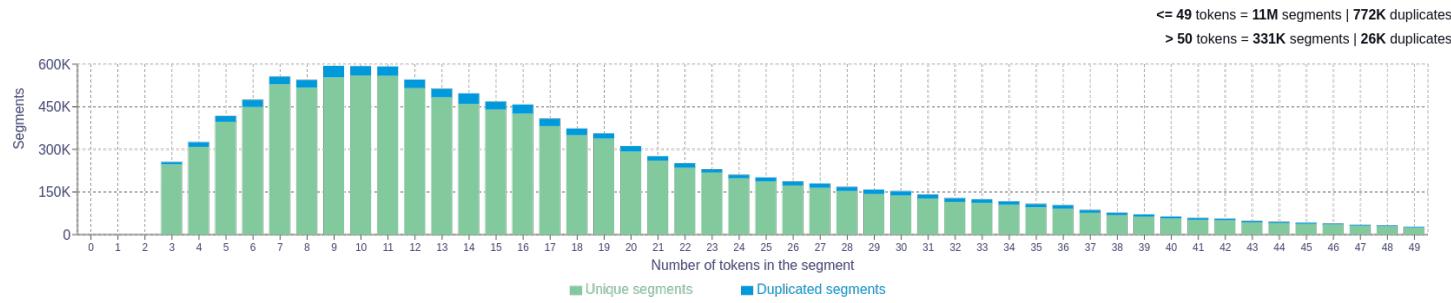
Target



Source segment length distribution by token



Target segment length distribution by token



Segment pair noise distribution



Source n-grams

Size	n-grams
1	(books 1652435) (used 714631) (available 641045) (hand 563667) (second 549358)
2	(second hand 520466) (rare books 519015) (used books 519003) (hand books 518986) (available rare 518982)
3	(second hand books 518985) (books and second 518982) (available rare books 518982) (view larger image 81168) (play the game 51602)
4	(used books and second 518982) (books of the title 518982) (books and second hand 518982) (posted over a year 71389) (click here to play 52076)
5	(used books and second hand 518982) (hand books of the title 518982) (books and second hand books 518982) (posted over a year ago 71243) (resources come from the dht 33453)

Target n-grams

Size	n-grams
1	(उपलब्ध 676099) (किताबें 589431) (पूरी 586918) (भी 586105) (हाथ 579621)
2	(दूसरा हाथ 525947) (दुर्लभ पुस्तकों 525858) (प्रयुक्त किताबें 525854) (उपलब्ध दुर्लभ 525854) (हाथ पुस्तकों 525852)
3	(पुस्तकों के शीर्षक 525857) (दूसरा हाथ पुस्तकों 525852) (किताबें और दूसरा 525852) (उपलब्ध दुर्लभ पुस्तकों 525852) (साल से अधिक 190381)
4	(हाथ पुस्तकों के शीर्षक 525852) (प्रयुक्त किताबें और दूसरा 525852) (किताबें और दूसरा हाथ 525852) (पूरी तरह से सूचीबद्ध 505892) (साल से अधिक पुस्तका 187304)
5	(प्रयुक्त किताबें और दूसरा हाथ 525852) (दूसरा हाथ पुस्तकों के शीर्षक 525852) (किताबें और दूसरा हाथ पुस्तकों 525852) (खेलने के लिए यहाँ फ़िक्र 49715) (गेम खेलने के लिए यहाँ 49710)

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 or types (uniques)/number of tokens*, after removing punctuation (<https://www.slinfo.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>