

## General overview

Corpus	Analytics date	Language
ne_1.jsonl.tsv	3/23/2024	Nepali (ne)

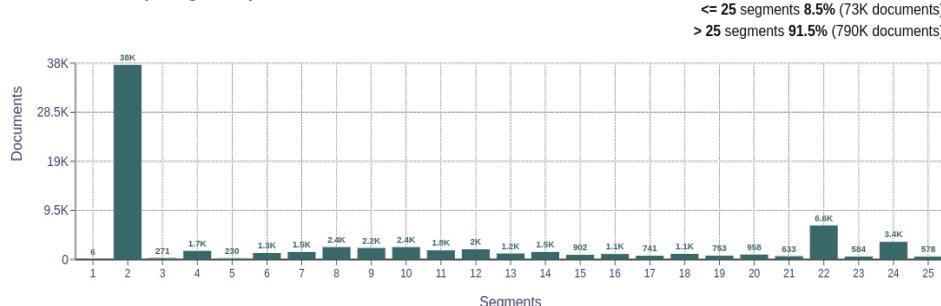
## Volumes

Docs	Segments	Unique segments	Tokens	Size
863,349	78,392,421	43,583 (0.06 %)	795M	10.41 GB

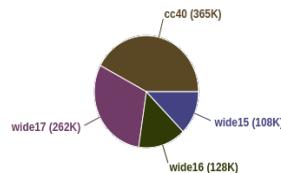
## Type-Token Ratio

Nepali (ne)
0.01

## Documents size (in segments)

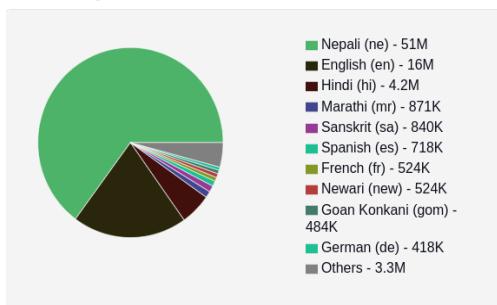


## Documents by collection

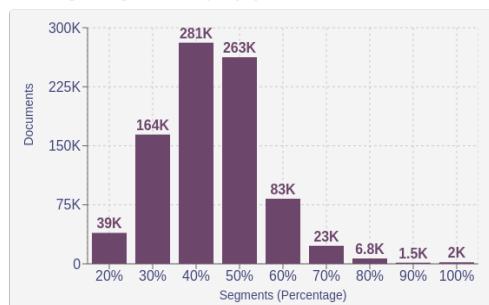


## Language Distribution

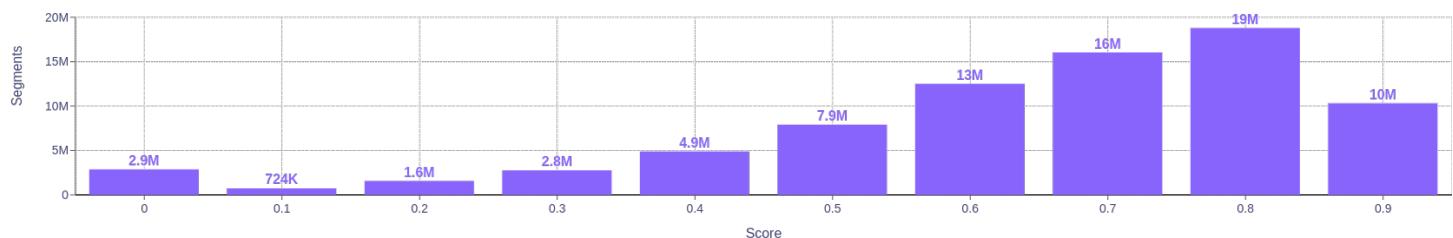
## Number of segments



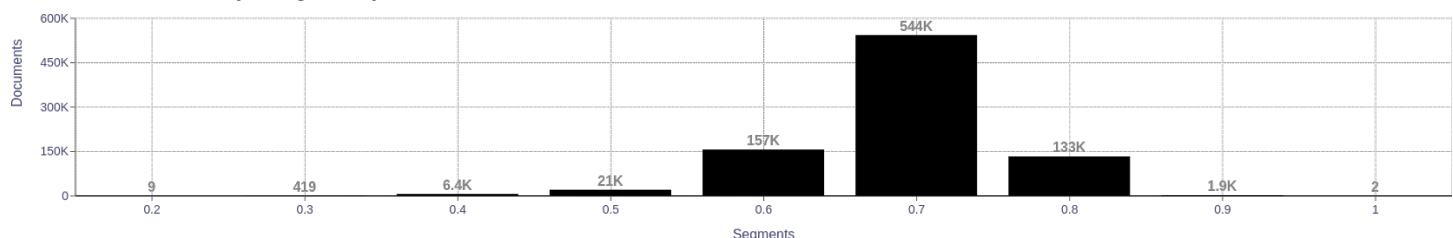
## Percentage of segments in Nepali (ne) inside documents



## Distribution of segments by fluency score



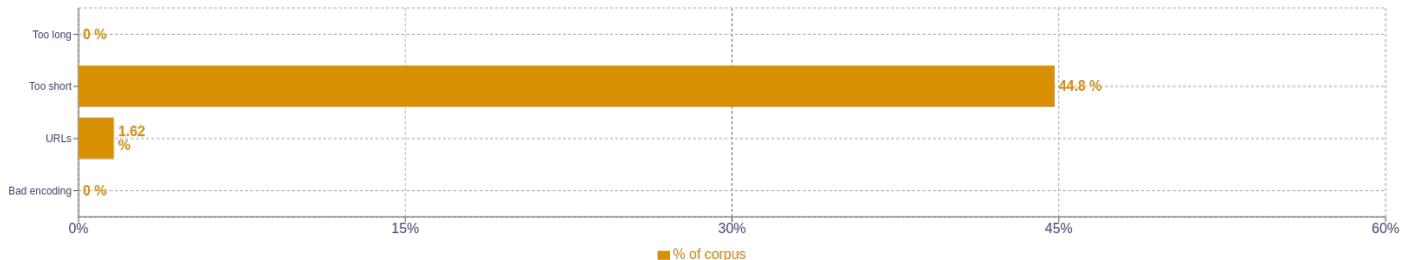
## Distribution of documents by average fluency score



## Segment length distribution by token



## Segment noise distribution



## Frequent n-grams

Size	n-grams
1	समाचार   2112597 (नेपाल   1771404) (com   1651501) (नेपाली   1468742) (प्रदेश   1274807)
2	(email protected   325587) (rights reserved   303236) (all rights   303005) (read more   292234) (हातो बारेमा   248969)
3	(all rights reserved   300070) (सुलना विभाग दर्ता   115639) (विभाग दर्ता ने   104792) (leave a reply   97600) (your email address   94972)
4	(leave a reply cancel   86228) (a reply cancel reply   85758) (will not be published   84327) (address will not be   81861) (your email address will   81827)
5	(leave a reply cancel reply   85758) (address will not be published   81854) (email address will not be   81826) (your email address will not   81824) (twittershare to facebookshare to pinterest   59874)

## 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 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>, <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>).

### 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>