

## General overview

Corpus	Analytics date	Language
tl_1.jsonl.tsv	3/24/2024	Filipino (tl)

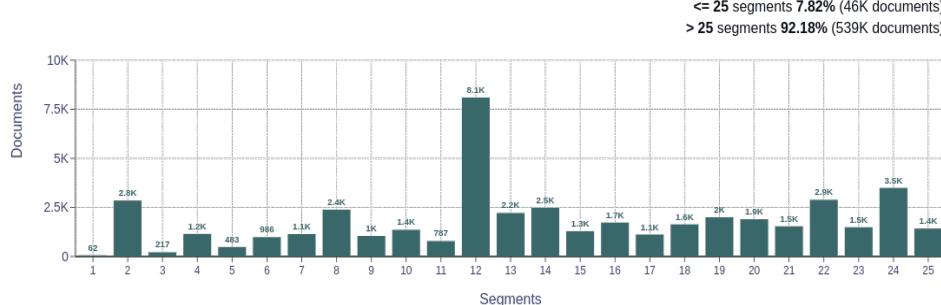
## Volumes

Docs	Segments	Unique segments	Tokens	Size
585,237	104,222,137	60,546 (0.06 %)	1.1B	5.11 GB

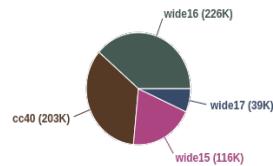
## Type-Token Ratio

Filipino (tl)
0.01

## Documents size (in segments)

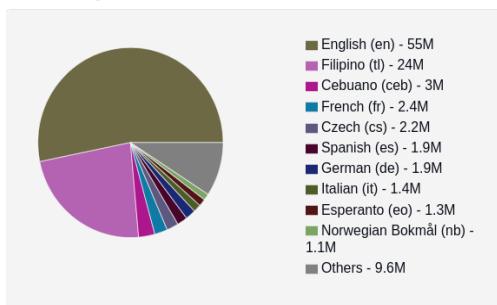


## Documents by collection

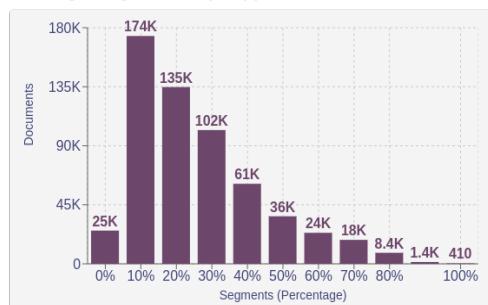


## Language Distribution

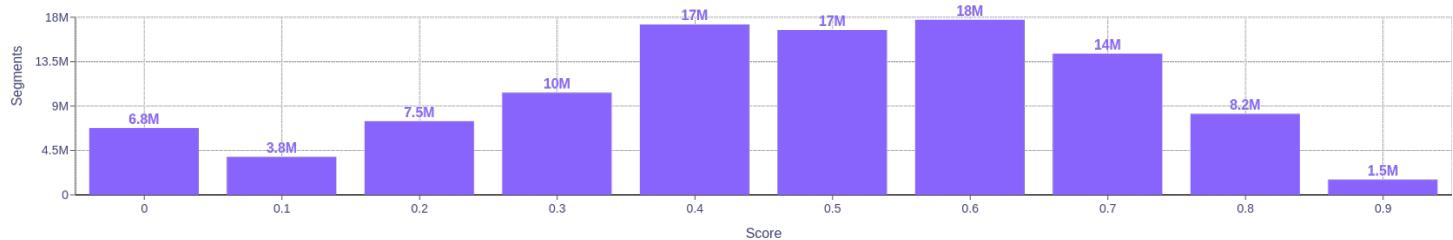
## Number of segments



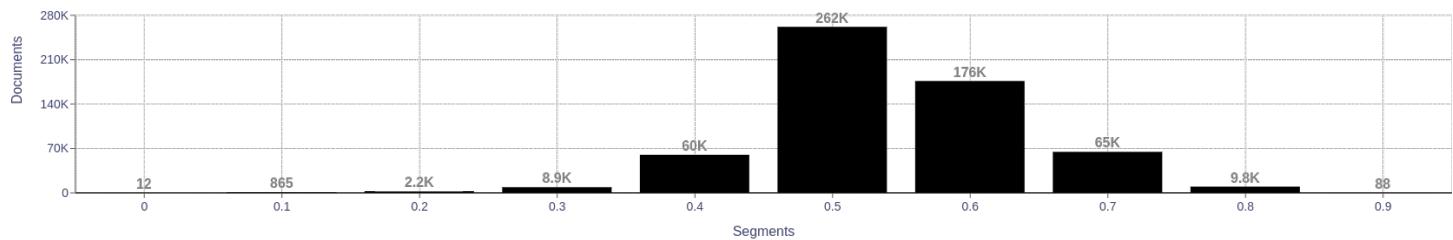
## Percentage of segments in Filipino (tl) inside documents



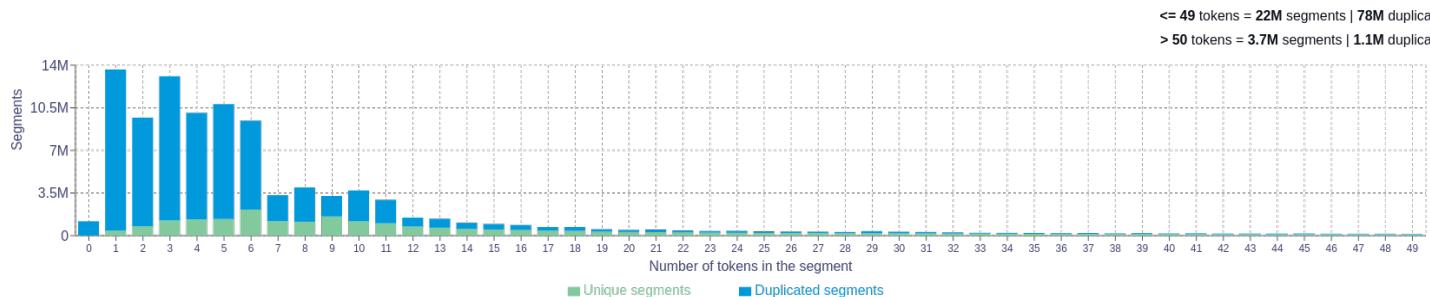
## Distribution of segments by fluency score



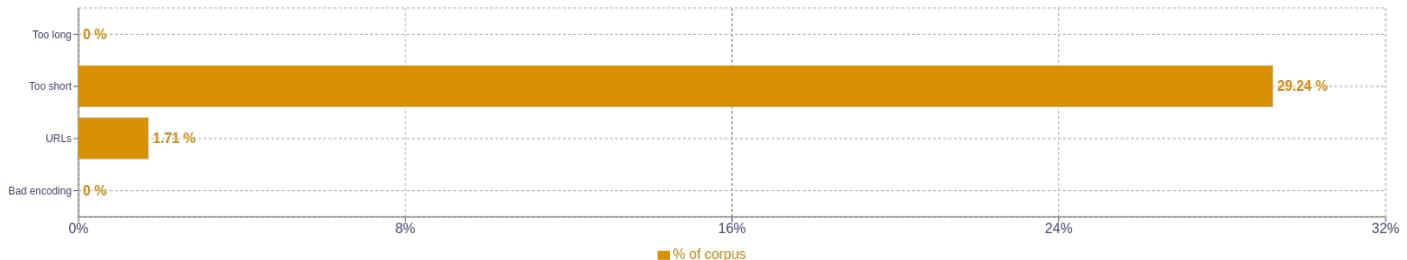
## Distribution of documents by average fluency score



## Segment length distribution by token



## Segment noise distribution



## Frequent n-grams

Size	n-grams
1	(the   8231358) (to   6859794) (and   5257716) (of   4599234) (code   4322041)
2	(zip code   1724206) (postal code   1516374) (digit zip   1028203) (-digit postal   1027687) (last line   981307)
3	(-digit postal code   1027686) (address pangunahing numero   980749) (zip code idagdag   980627) (buliding firm pangalan   980627) (address ikalawang number   980627)
4	(id ng carrier ruta   980626) (numero ng congressional district   980616) (preferred last line key   980612) (loob ng isang taon   785608) (taon na ang nakalipas   785435)
5	(accommodation photo ng accommodation photo   139366) (share to twittershare to facebookshare   126212) (twittershare to facebookshare to pinterest   125672) (to twittershare to facebookshare to   125672) (are you sure you want   121837)

## 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>