

General overview

| Corpus | Analytics date | Language |
|----------------|----------------|----------------|
| eo_1.jsonl.tsv | 3/16/2024 | Esperanto (eo) |

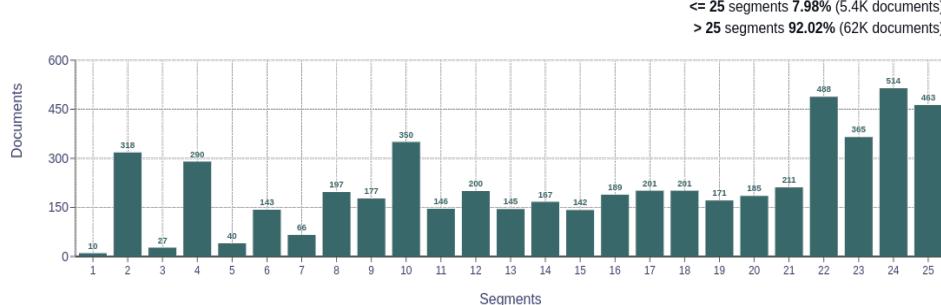
Volumes

| Docs | Segments | Unique segments | Tokens | Size |
|--------|-----------|-----------------|--------|----------|
| 67,808 | 8,788,276 | 26,273 (0.30 %) | 131M | 635.3 MB |

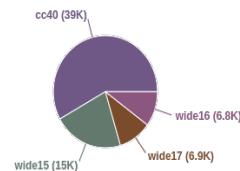
Type-Token Ratio

| |
|----------------|
| Esperanto (eo) |
| 0.03 |

Documents size (in segments)

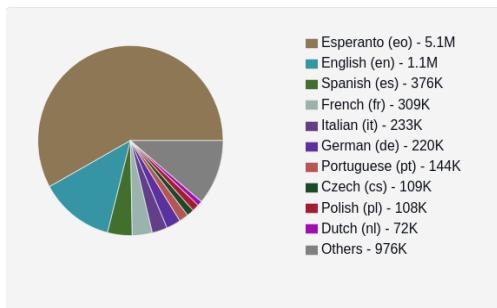


Documents by collection

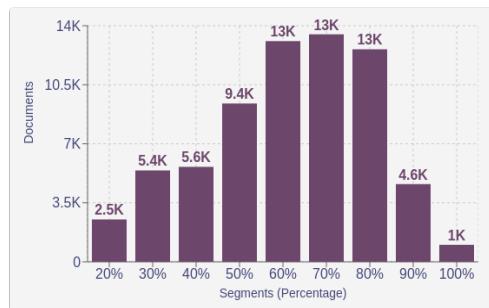


Language Distribution

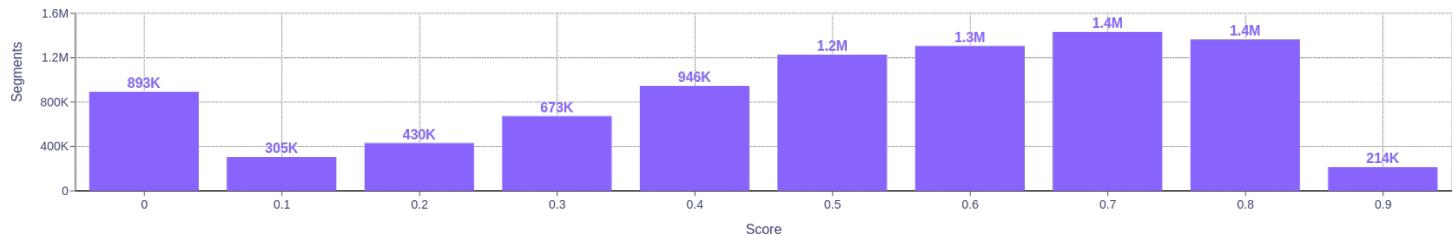
Number of segments



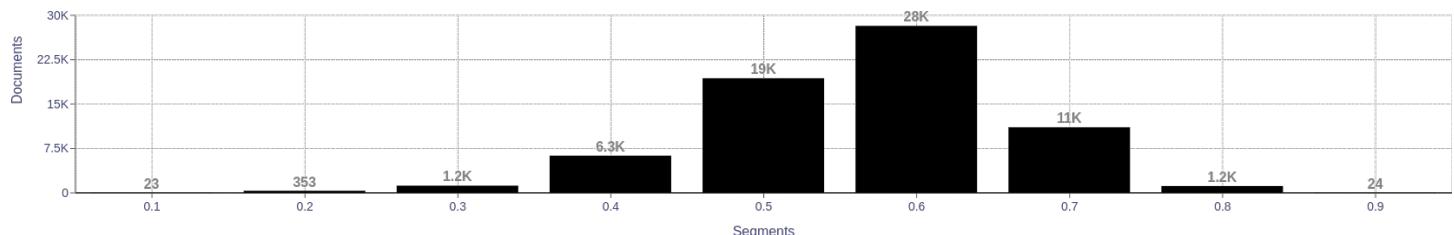
Percentage of segments in Esperanto (eo) inside documents



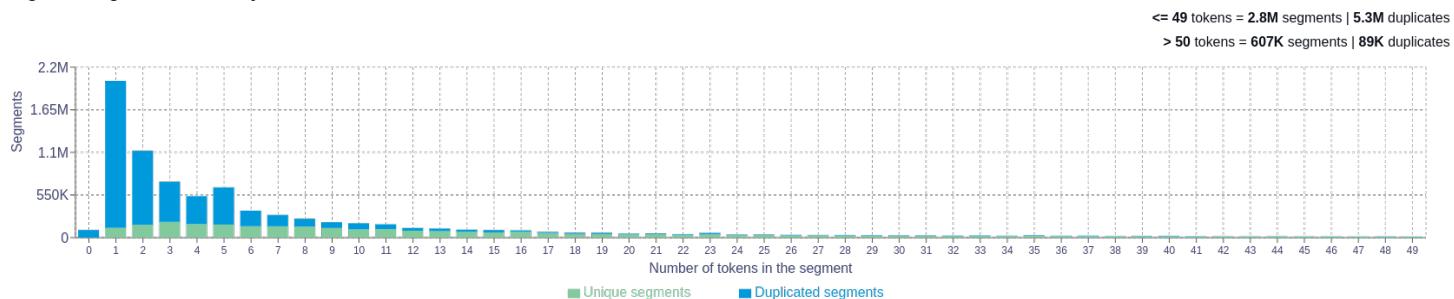
Distribution of segments by fluency score



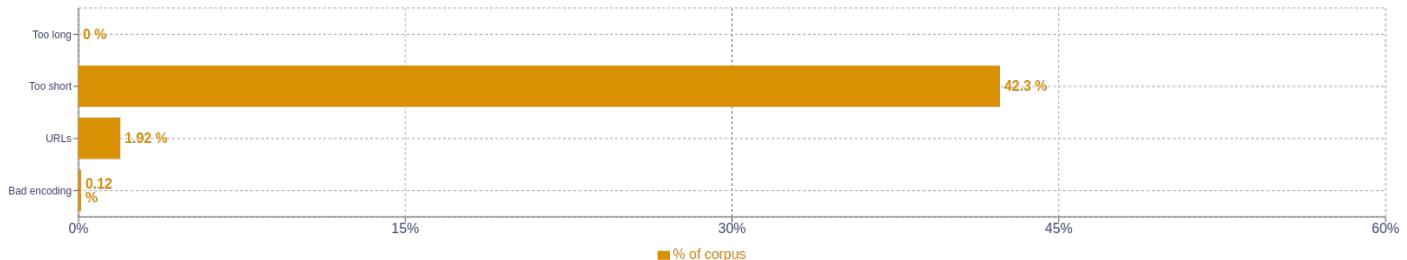
Distribution of documents by average fluency score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

| Size | n-grams |
|------|--|
| 1 | (pri 634446) (el 424536) (kun 413006) (kiel 409979) (per 261033) |
| 2 | (redakti fonton 82508) (povas esti 31378) (of the 29855) (pri vikipedio 29413) (regularo pri 23222) |
| 3 | (regularo pri respeko 22872) (deklaro pri kuketoj 22102) (pa o estis lastafoje 18223) (la  la permesilo 13996) (ligiloj  i tien 13265) |
| 4 | (respeko de la privateco 22877) (pa o estis lastafoje redaktita 18223) (informoj pri la pa o 13081) (la  la permesilo krea 10565) (permesilo krea komunajo atribuive-samkondice 9931) |
| 5 | (pri respeko de la privateco 22872) (la  la permesilo krea komunaajo 10565) (vidu la uzkondi ojn por detaloj 9811) (disponebolas la  la permesilo krea 9786) (teksto disponeblas la  la permesilo 9771) |

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/mabanon/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>