HPLT Analytics report

HPLTAnalytics

General overview

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

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
67,808	8,788,276	3,352,630 (38.15 %)	131M	635.3 MB	

Top 10 domains

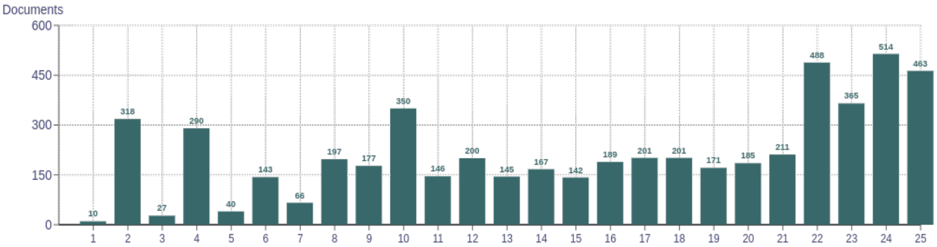
Domain	Docs	% of total
wikipedia.org	21K	30.38
uea.org	1.6K	2.40
pola-retradio.org	1.2K	1.82
mondediplo.com	1.1K	1.63
wordpress.org	1K	1.50
ipernity.com	968	1.43
wikitrans.net	953	1.41
wikiquote.org	917	1.35
wordpress.com	901	1.33
blogspot.ro	846	1.25

Top 10 TLDs

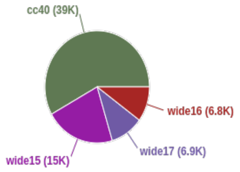
Domain	Docs	% of total
org	33K	48.60
com	13K	18.93
net	4.1K	5.99
ru	2.1K	3.05
de	1.4K	2.02
be	886	1.31
eu	881	1.30
cn	875	1.29
info	871	1.28
ro	855	1.26

Documents size (in segments)

<= 25 segments **7.98%** (5.4K documents)
> 25 segments **92.02%** (62K documents)

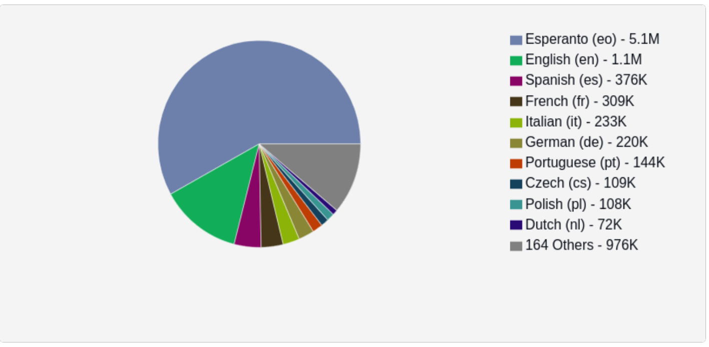


Documents by collection

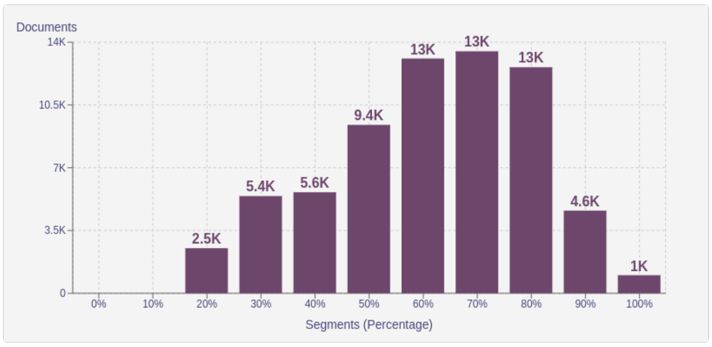


Language Distribution

Number of segments

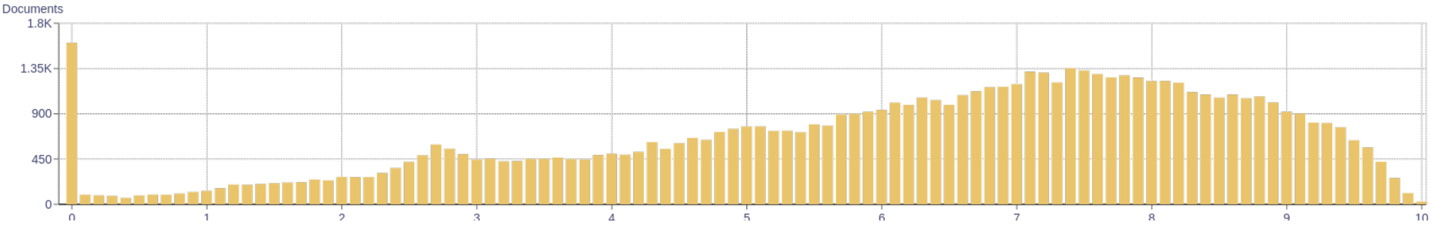


Percentage of segments in Esperanto (eo) inside documents



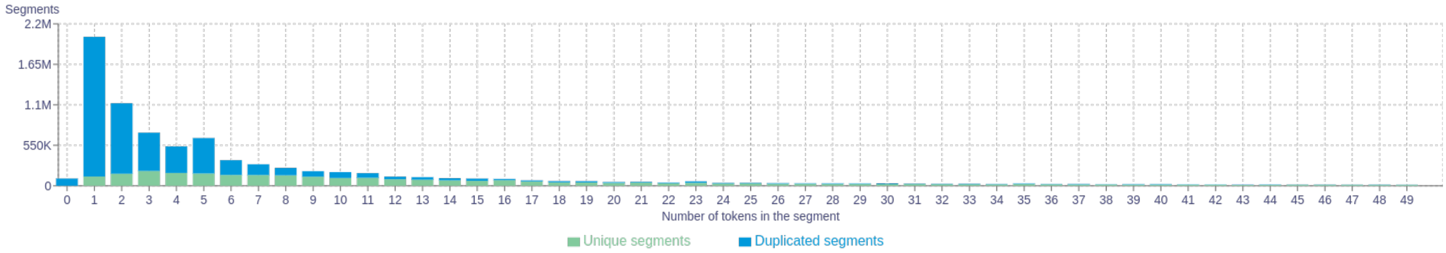
Distribution of documents by document score

score < 5 - **28.1%** (19K documents)
score >= 5 - **71.9%** (49K documents)

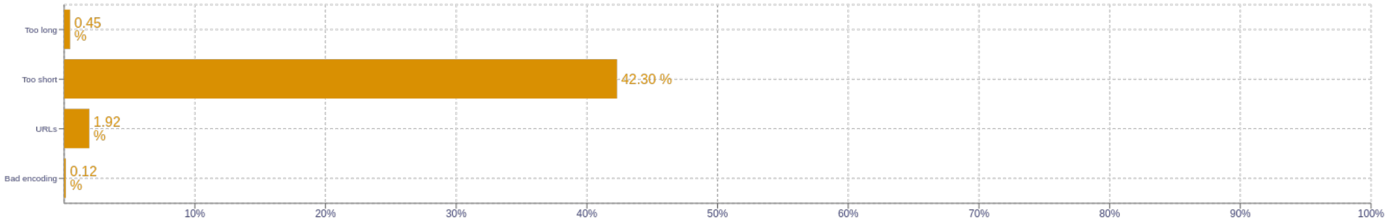


Segment length distribution by token

<= 49 tokens = **2.8M** segments | **5.3M** duplicates
> 50 tokens = **607K** segments | **89K** duplicates



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 respekto | 22872 deklaro pri kuketoj | 22102 paĝo estis lastafoje | 18223 laŭ la permesilo | 13996 ligiloj ĉi tien | 13265
4	respekto de la privateco | 22877 paĝo estis lastafoje redaktita | 18223 informoj pri la paĝo | 13081 laŭ la permesilo krea | 10565 permesilo krea komunaĵo atribuite-samkondiĉe | 9931
5	pri respekto de la privateco | 22872 laŭ la permesilo krea komunaĵo | 10565 vidu la uzkondiĉojn por detaloj | 9811 disponeblas laŭ la permesilo krea | 9786 teksto disponeblas laŭ la permesilo | 9771

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 or 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).

Distribution of documents by document score

Obtained with Web Docs Scorer (https://github.com/pablop16n/web-docs-scorer/).

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 (https://github.com/bitextor/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