

General overview

Corpus	Analytics date	Language
HPLT-docslite.et.tsv	6/8/2024	Estonian (et)

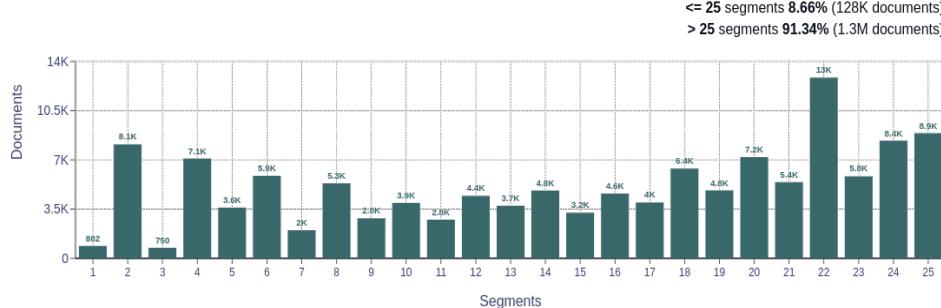
Volumes

Docs	Segments	Unique segments	Tokens	Size
1,475,951	195,463,992	57,255 (0.03 %)	2.1B	12.24 GB

Type-Token Ratio

Estonian (et)

Documents size (in segments)



Top 10 domains

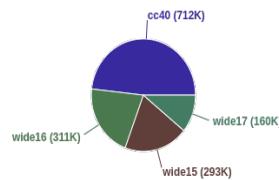
Top 10 domains

Domain	Docs	% of total
diebuchsuche.com	92K	6.22
delfi.ee	52K	3.54
blogspot.com	37K	2.54
postimees.ee	31K	2.07
blogspot.fi	28K	1.91
blogspot.com.ee	28K	1.88
ohuleht.ee	22K	1.52
wikipedia.org	19K	1.26
europages.org	13K	0.88
ituudised.ee	10K	0.68

Top 10 TLDs

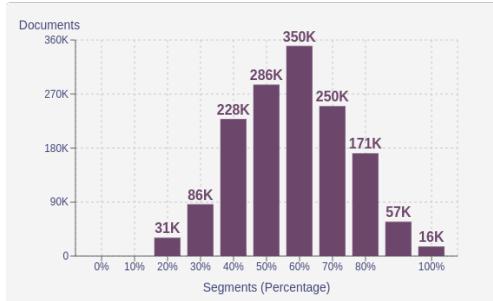
Domain	Docs	% of total
ee	856K	57.98
com	331K	22.41
org	52K	3.51
eu	49K	3.32
fi	32K	2.18
net	31K	2.12
com.ee	28K	1.90
edu.ee	6.5K	0.44
com.au	6.5K	0.44
info	5.6K	0.38

Documents by collection

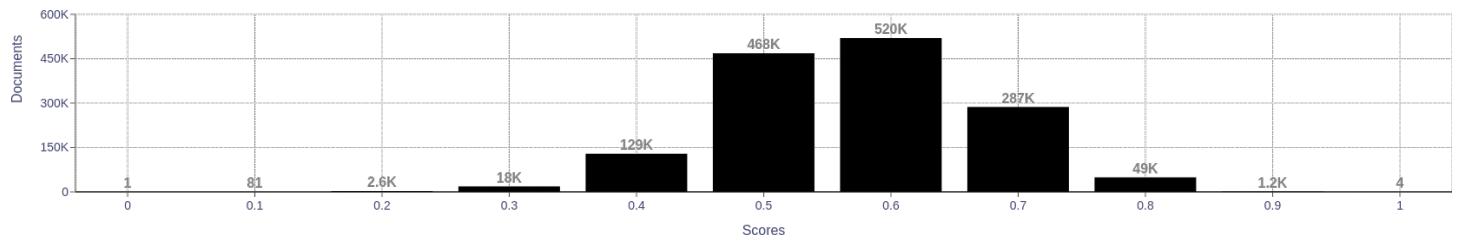


Language Distribution

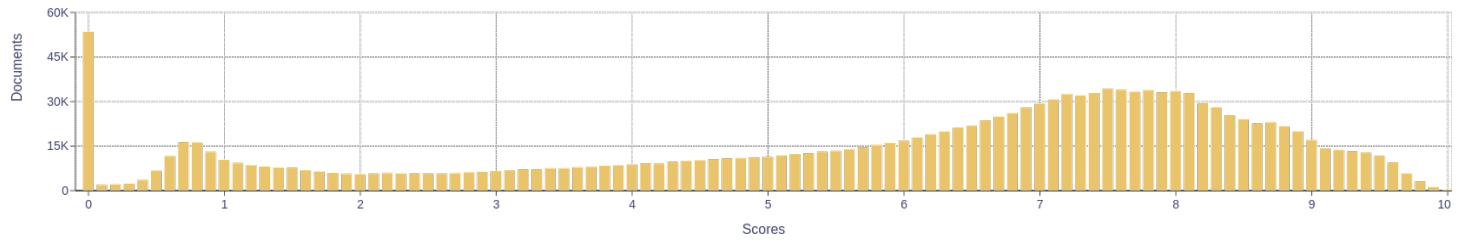
Percentage of segments in Estonian (et) inside documents



Distribution of documents by average fluency score

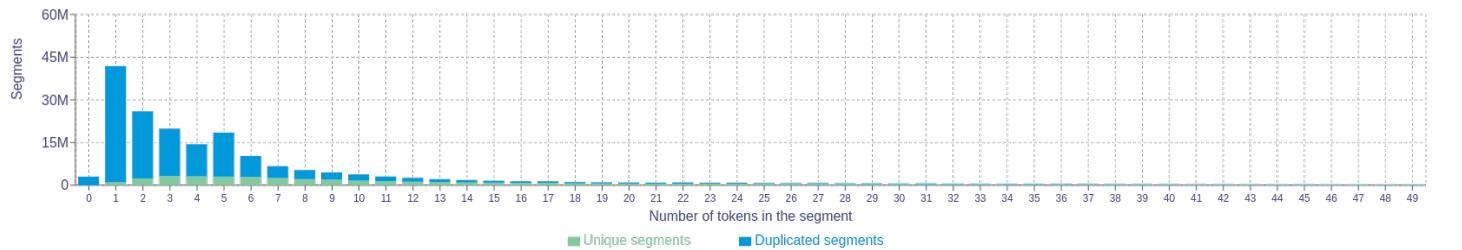


Distribution of documents by document score



Segment length distribution by token

<= 49 tokens = 43M segments | 143M duplicates
> 50 tokens = 8.6M segments | 2.3M duplicates



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>).

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>