

General overview

Corpus	Analytics date	Language
HPLT-docsite.bg.tsv	6/10/2024	Bulgarian (bg)

Volumes

Docs	Segments	Unique segments	Tokens	Size
3,583,094	538,385,172	67,352 (0.01 %)	5.8B	50.14 GB

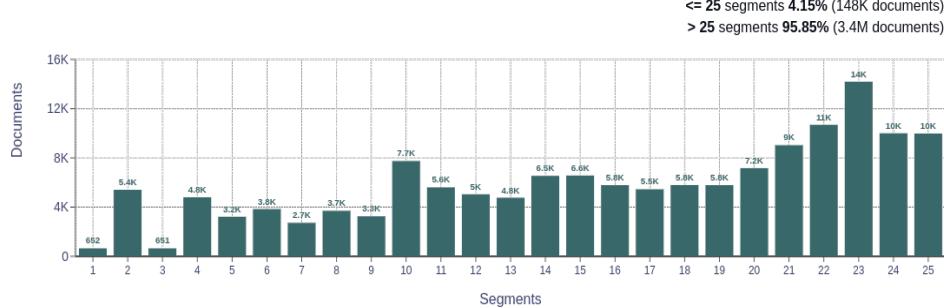
Top 10 domains

Domain	Docs	% of total
monitor.bg	39K	1.09
blogspot.com	38K	1.07
dtv-bg.com	32K	0.89
pomagalo.com	27K	0.76
chitanka.info	26K	0.74
utre.bg	26K	0.73
wikipedia.org	23K	0.65
petardanov.com	22K	0.61
europages.eu	20K	0.56
dnevnik.bg	20K	0.55

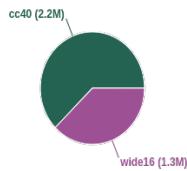
Top 10 TLDs

Domain	Docs	% of total
bg	1.5M	41.35
com	1.3M	37.05
net	220K	6.14
eu	146K	4.09
org	141K	3.92
info	119K	3.31
news	20K	0.57
pt	13K	0.37
top	8.7K	0.24
tv	7K	0.20

Documents size (in segments)

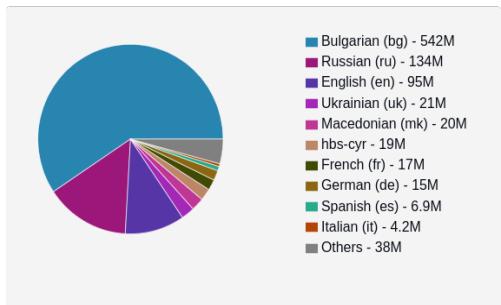


Documents by collection

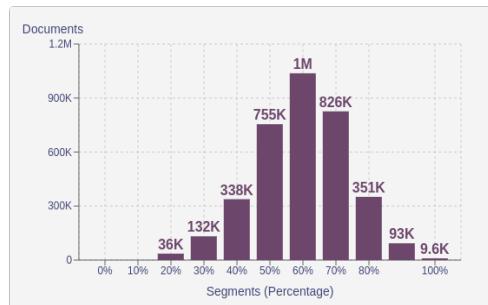


Language Distribution

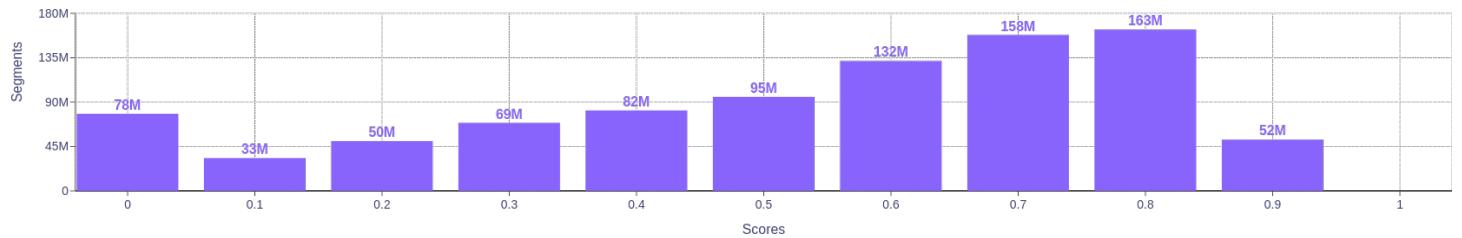
Number of segments



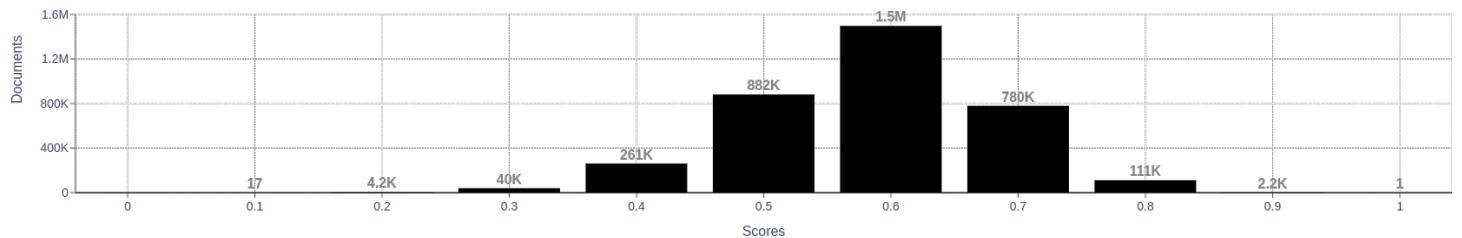
Percentage of segments in Bulgarian (bg) inside documents



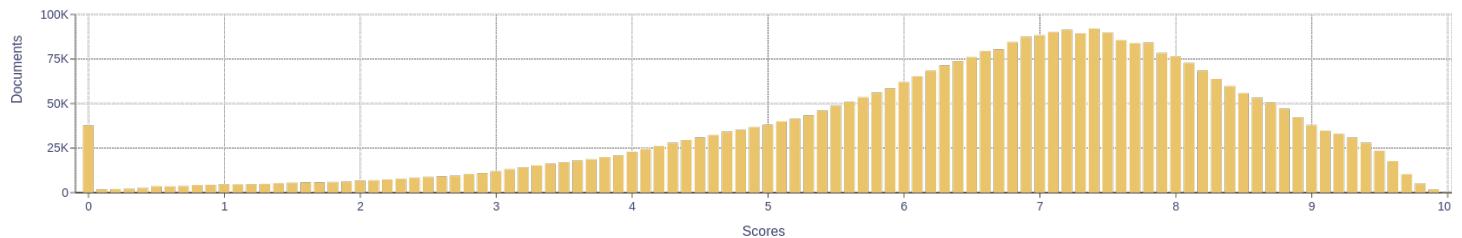
Distribution of segments by fluency score



Distribution of documents by average fluency score



Distribution of documents by document score



Segment length distribution by token

<= 49 tokens = 93M segments | 421M duplicates

> 50 tokens = 24M segments | 6.2M 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/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>