

General overview

Corpus	Analytics date	Language
mk_1.jsonl.tsv	3/22/2024	Macedonian (mk)

Volumes

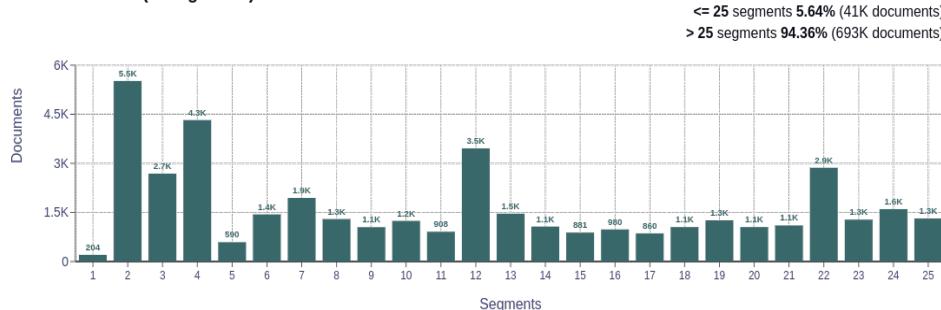
Docs	Segments	Unique segments	Tokens	Size
734,687	86,189,730	40,966 (0.05 %)	845M	7.59 GB

Type-Token Ratio

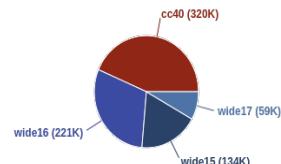
Macedonian (mk)

0.01

Documents size (in segments)

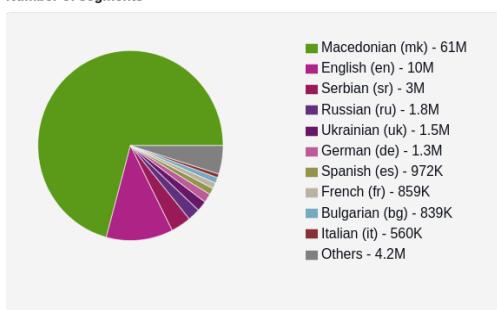


Documents by collection

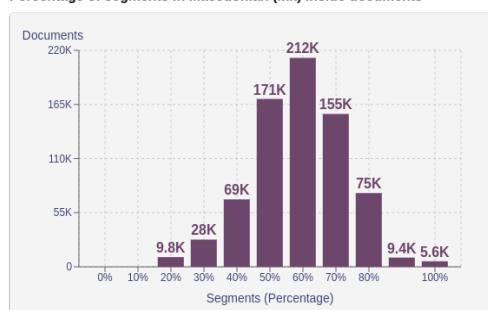


Language Distribution

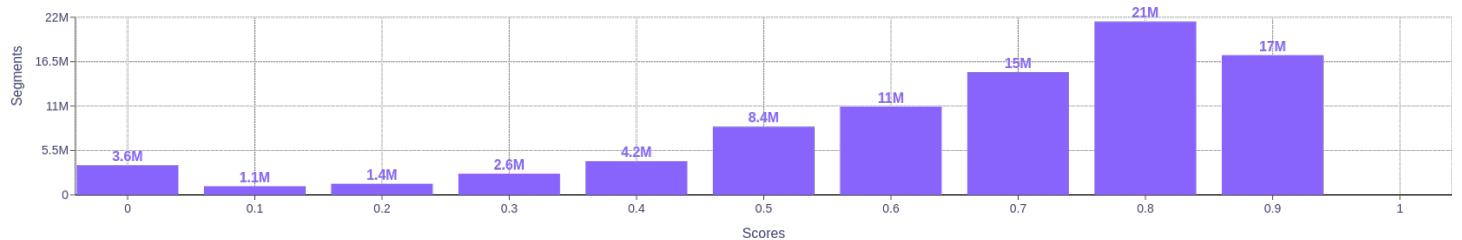
Number of segments



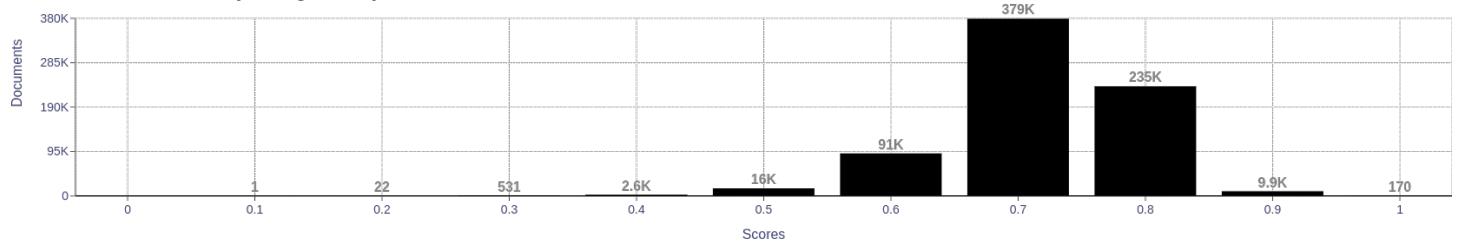
Percentage of segments in Macedonian (mk) inside documents



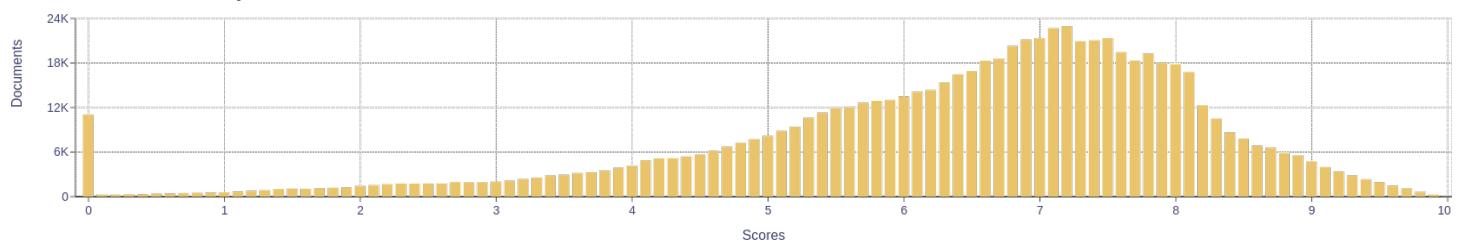
Distribution of segments by fluency score



Distribution of documents by average fluency score

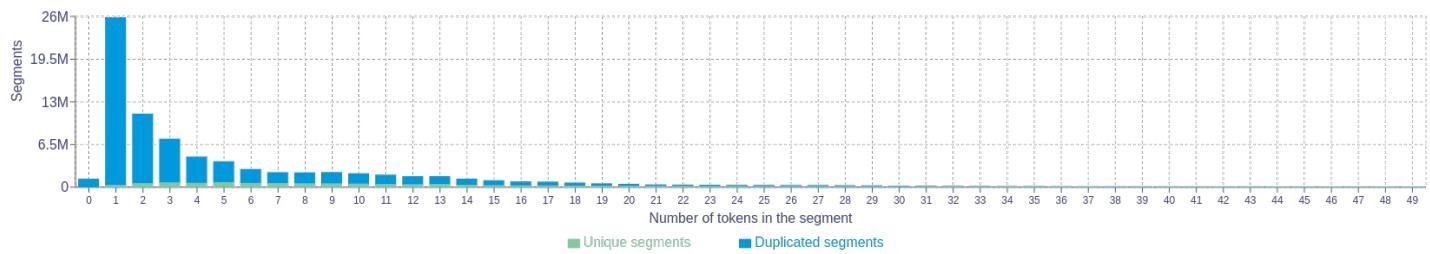


Distribution of documents by document score

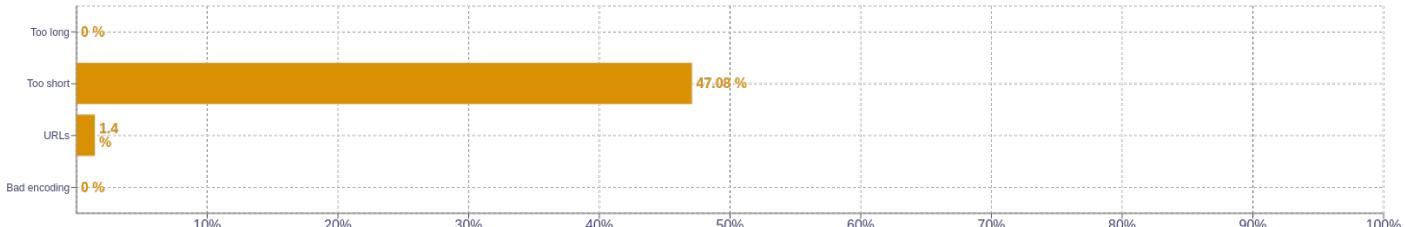


Segment length distribution by token

<= 49 tokens = 12M segments | 71M duplicates
 > 50 tokens = 3M segments | 810K duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	македонија 2546789 вести 1408817 година 1313749 скопје 1187774 видео 1138123
2	република македонија 237582 најнови вести 171533 of the 142448 милиони евра 124160 read more 122921
3	права се задржани 127603 услови за користење 120153 јубов и секс 96559 all rights reserved 83602 skip to content 68997
4	cookie is set by 47536 40187 the user consent for 40018 user consent for the 39935 consent for the cookies 39935
5	user consent for the cookies 39935 the user consent for the 39935 for the cookies in the 39933 consent for the cookies in 39933 the cookies in the category 38982

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.slinfo.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>