

General overview

Corpus	Analytics date	Language
hy_1.jsonl.tsv	3/21/2024	Armenian (hy)

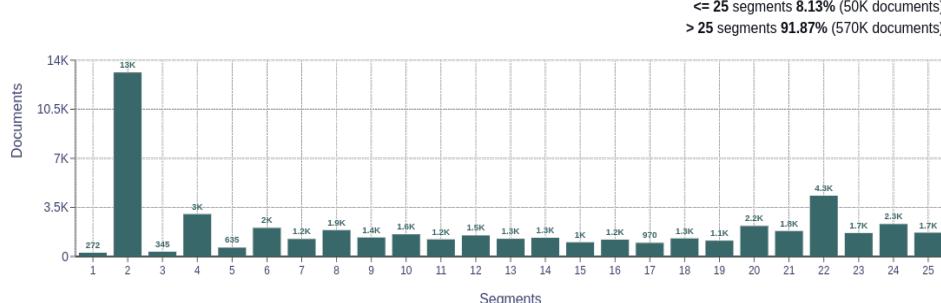
Volumes

Docs	Segments	Unique segments	Tokens	Size
621,465	67,013,428	43,201 (0.06 %)	794M	7.21 GB

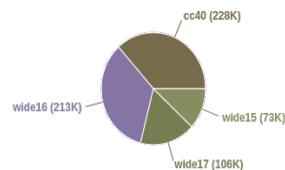
Type-Token Ratio

Armenian (hy)
0.01

Documents size (in segments)

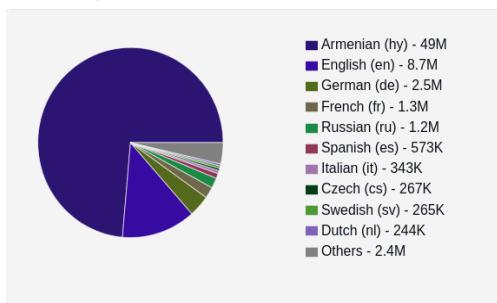


Documents by collection

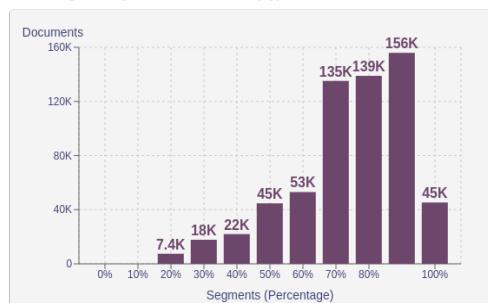


Language Distribution

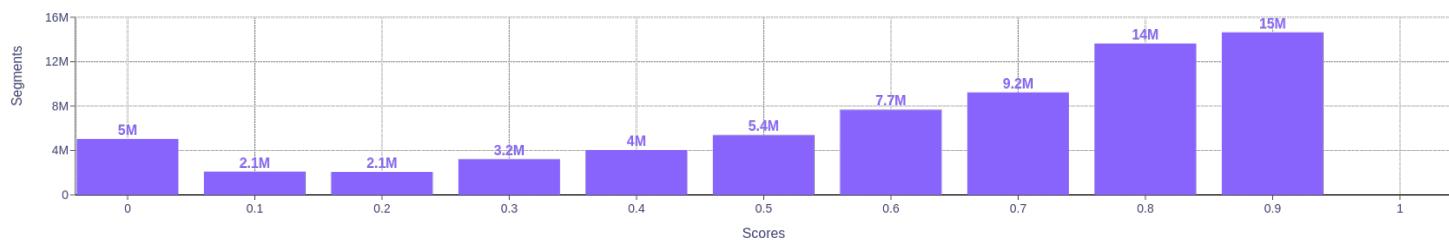
Number of segments



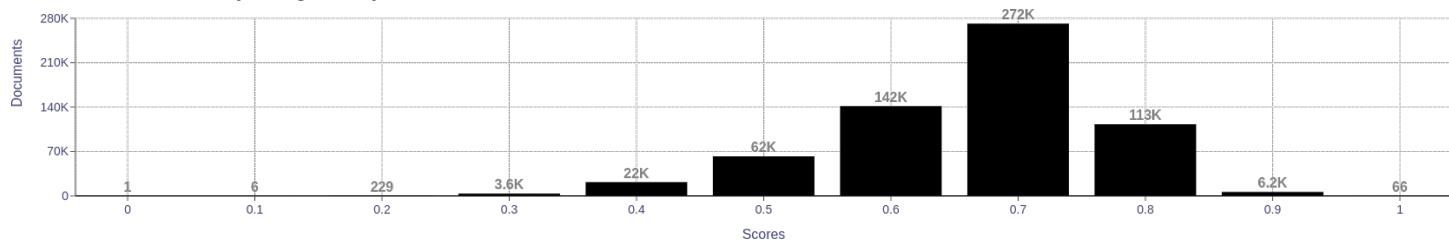
Percentage of segments in Armenian (hy) inside documents



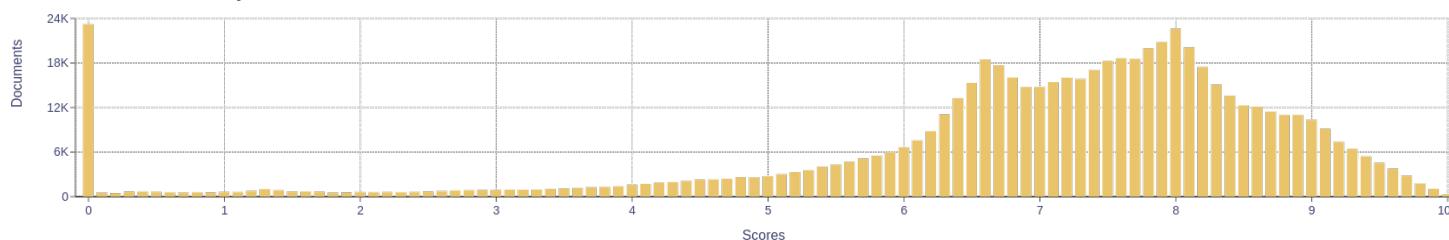
Distribution of segments by fluency score



Distribution of documents by average fluency score

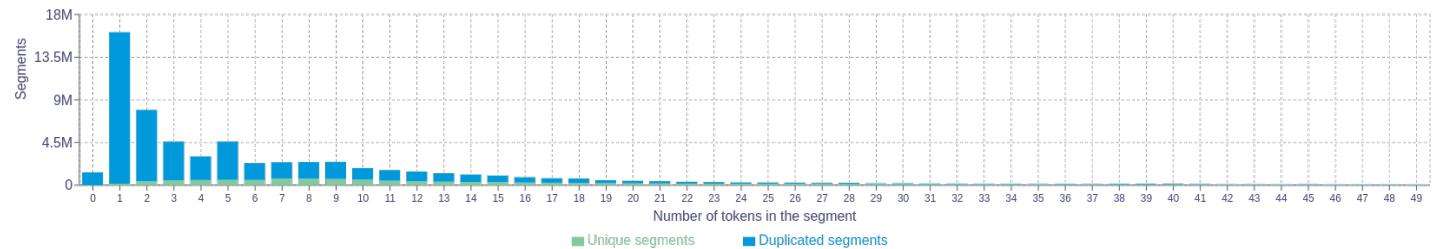


Distribution of documents by document score

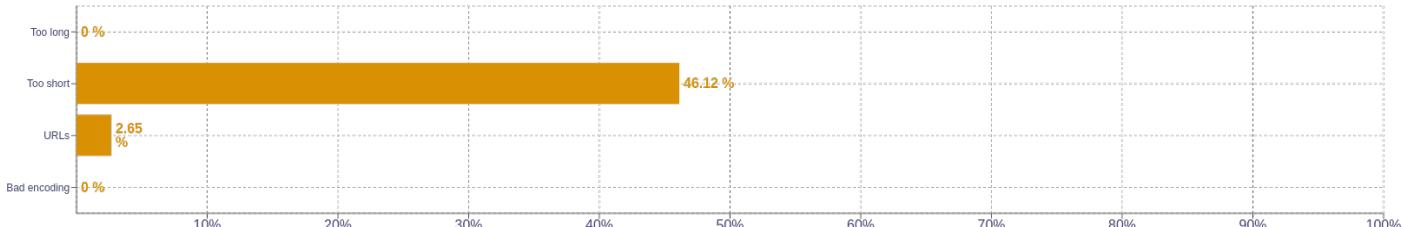


Segment length distribution by token

<= 49 tokens = 12M segments | 53M duplicates
 > 50 tokens = 2.9M segments | 750K duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	մասին 2076616 եւ 2055212 մի 1673183 ամ 1560627 չի 1435923
2	մի քանի 246348 մեր մասին 224034 հայաստանի համբաւության 220727 իրավունքները պաշտպանված 203535 բոլոր իրավունքները 200671
3	բոլոր իրավունքները պաշտպանված 194283 պատասխանակիրայուն չի կրում 113511 skip to content 85556 նորաց համար չեկավ 83803 զոհվել ե մ 83790
4	նորաց համար չեկավ զարում 83803 հովհաննիսյանը զոհվել է մ 83789 առ բոլոր իրավունքները պաշտպանված 56777 կայքը պատասխանավորյուն չի կրում 54338
5	զարա հովհաննիսյանը զոհվել է մ 83789 կայքի սյուբերի ամբողջական կամ մասնակի օգագործման 49039 սյուբերի ամբողջական կամ մասնակի օգագործման 48870 պատասխանակիրայուն չի կրում կայքում արտահայտված 48778

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>