

General overview

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

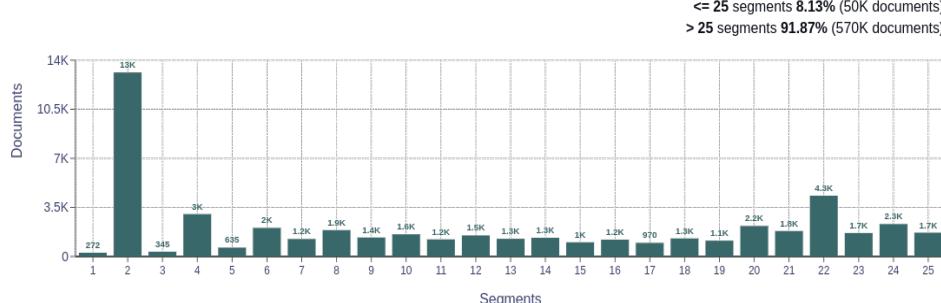
Volumes

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

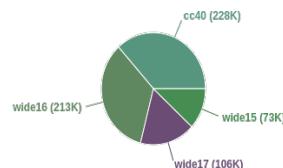
Type-Token Ratio

Armenian (hy)
0.01

Documents size (in segments)

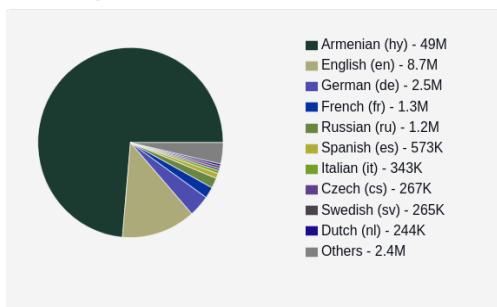


Documents by collection

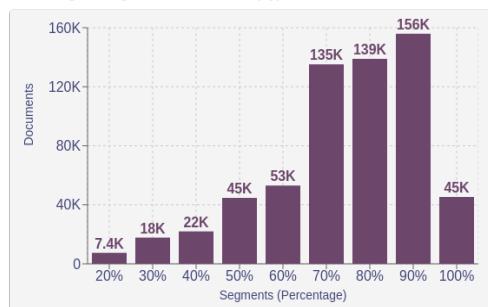


Language Distribution

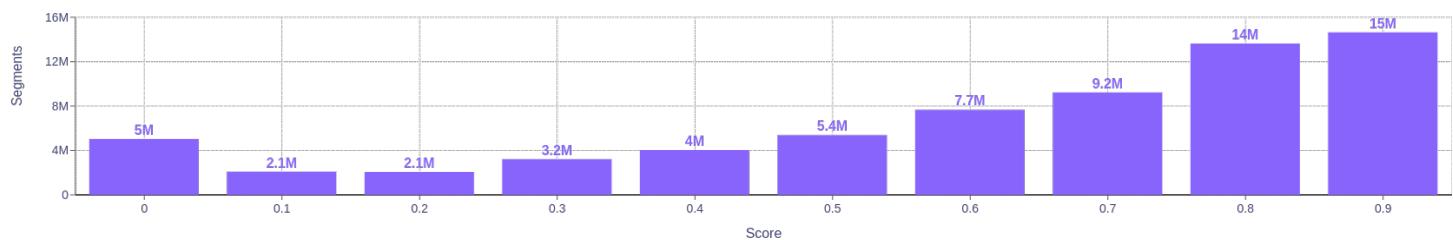
Number of segments



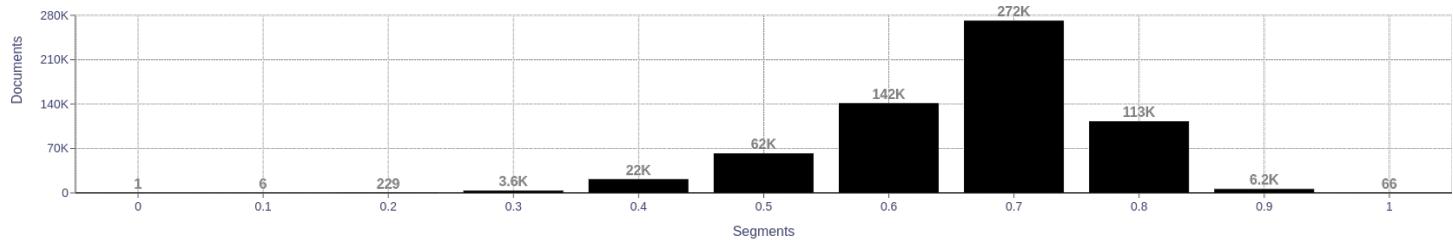
Percentage of segments in Armenian (hy) inside documents



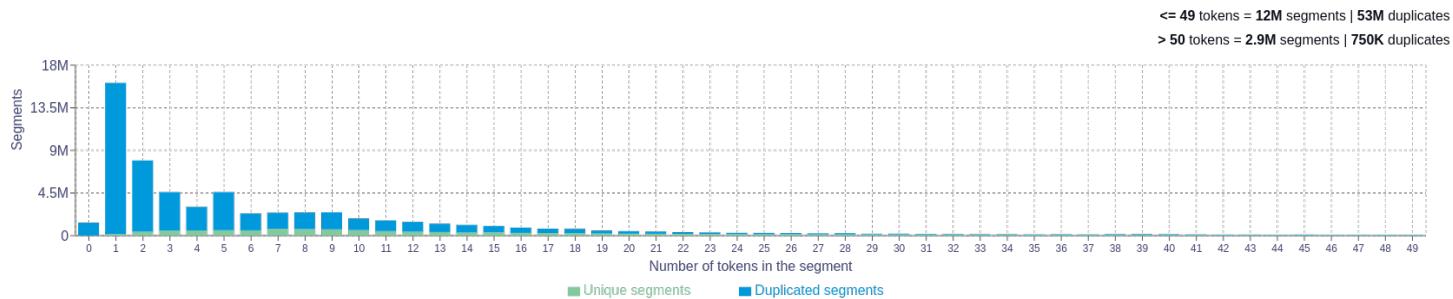
Distribution of segments by fluency score



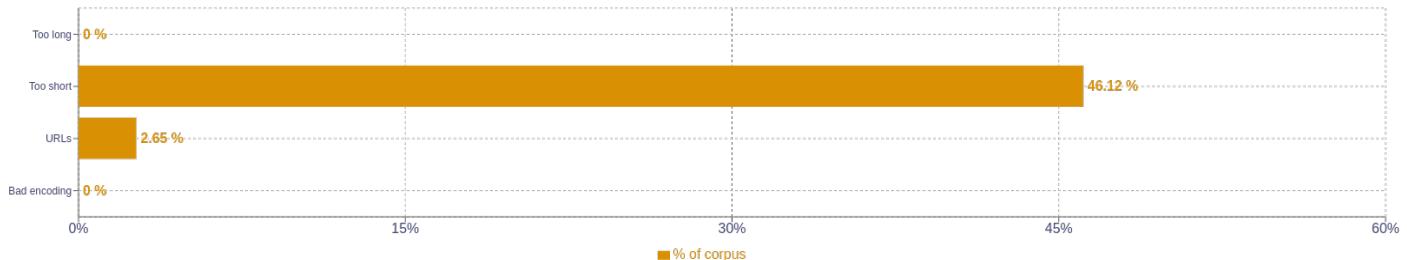
Distribution of documents by average fluency score



Segment length distribution by token



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) (հայաստակ համբառելուր յան 220727) (իրավունքները պաշտպանեած 203535) (բոլոր իրավունքները 200671)
3	(բոլոր իրավունքները պաշտպանեած 194283) (պատասխանավորություն չի կրում 113511) (skip to content 85556) (նրանց համար չճիշգ 83803) (զոհվել է ձ 83790)
4	(նրանց համար չեկալ զրուս 83803) (հովհաննիսյանը զոհվել է ձ 83789) (ամ բոլոր իրավունքները պաշտպանեած 56777) (կայքը պատասխանավորություն չի կրում 54338) (նյութերի ամբողջական կամ մասնակի 50988)
5	(զարդ հովհաննիսյանը զրիմել է ձ 83789) (կայքի նյութերի ամբողջական կամ մասնակի 49039) (նյութերի ամբողջական կամ մասնակի օգտագործման 48870) (ամբողջական կամ մասնակի օգտագործման դեպքում 48870) (պատասխանավորություն չի կրում կայքում պատահայտված 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 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>, , , 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>).

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.

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>