

General overview

Corpus	Analytics date	Language
hye_Arn.jsonl.tsv	9/6/2024	Armenian (hy)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
3,598,520	65,242,460	31,079,599 (47.64 %)	1.8B	17.99 GB	10,657,868,503

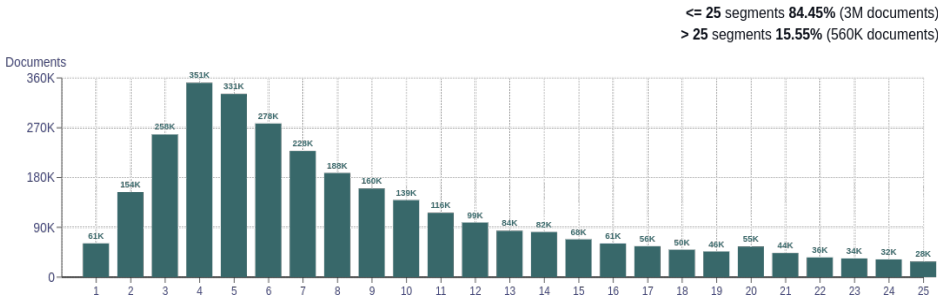
Top 10 domains

Domain	Docs	% of total
wikipedia.org	213K	5.92
azatutyun.am	112K	3.10
armeniasputnik.am	104K	2.89
aravot.am	99K	2.76
epress.am	67K	1.85
armtimes.com	57K	1.59
armhur.am	53K	1.47
168.am	52K	1.44
top-news.am	50K	1.39
newsarmenia.am	47K	1.31

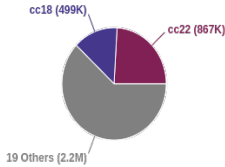
Top 10 TLDs

Domain	Docs	% of total
am	2.4M	65.89
com	540K	15.00
org	324K	9.01
ru	88K	2.45
net	67K	1.87
info	65K	1.81
news	17K	0.48
blog	13K	0.37
ir	12K	0.33
tv	10K	0.29

Documents size (in segments)

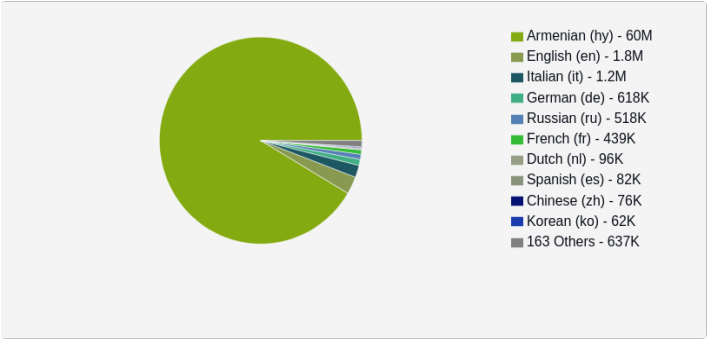


Documents by collection

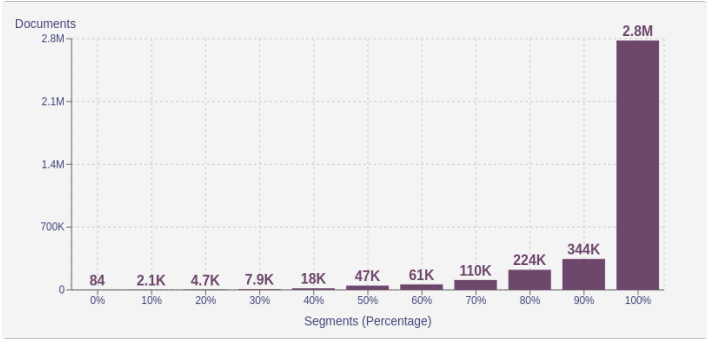


Language Distribution

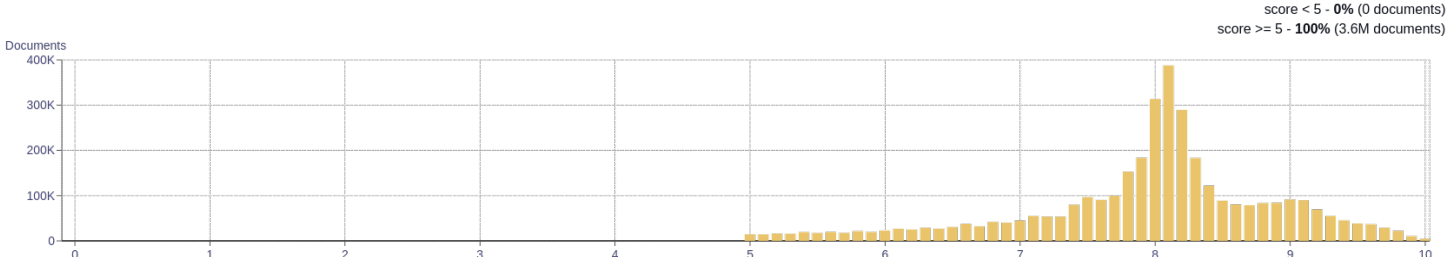
Number of segments



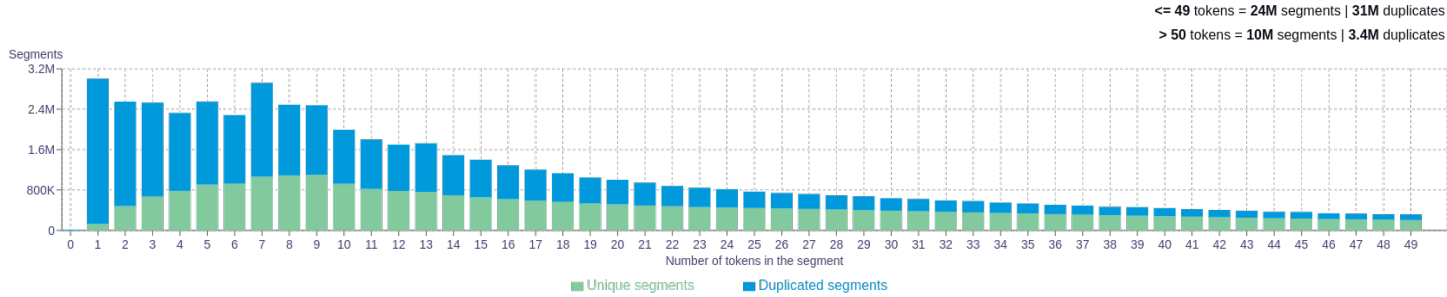
Percentage of segments in Armenian (hy) inside documents



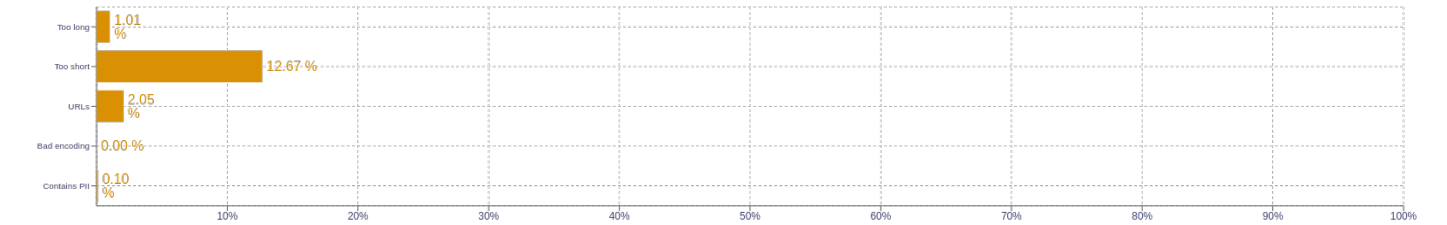
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>ել   7890678</div> <div>մի   5480619</div> <div>մասին   4770899</div> <div>թե   3994145</div> <div>էլ   3959184</div>
2	<div>հայաստանի հանրապետության   1002855</div> <div>մի քանի   853368</div> <div>խմբագրել կողմ   517252</div> <div>ոչ թե   492398</div> <div>ոչ միայն   480803</div>
3	<div>տեղի է ունեցել   182981</div> <div>հայաստանի հանրապետության կառավարության   110538</div> <div>պետք է լինի   102335</div> <div>թույլ է տալիս   95637</div> <div>կարող է լինել   90920</div>
4	<div>ուժի մեջ է մտնում   40514</div> <div>հարուցվել է քրեական գործ   36643</div> <div>արտակարգ և լիազոր դեսպան   26440</div> <div>at the wayback machine   24336</div> <div>աշխատանքի և սոցիալական հարցերի   23667</div>
5	<div>արևմտահայերեն բացառական բառարան հիմանիչներ ունսերեն   21121</div> <div>ուժի մեջ է մտնում պաշտոնական   19363</div> <div>տեղեկատվության և հասարակայնության հետ կապերի   18211</div> <div>սպասվում է առանց տեղումների եղանակ   17085</div> <div>ը ո չ ու մ   16892</div>

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, <ul>, <ol>, 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.stinfo.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>, <ul>, <ol>, 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>