HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hye_Armn.jsonl.tsv	9/6/2024	Armenian (hy)

Volumes

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

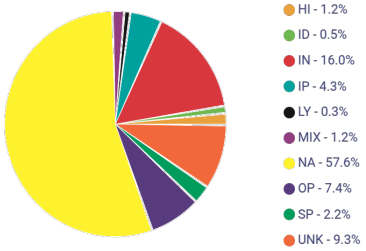
Top 10 domains

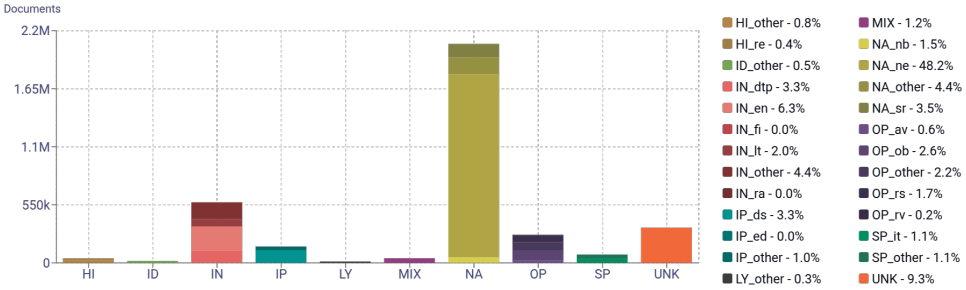
Domain	Docs	% of total
wikipedia.org	213K	5.92%
azatutyun.am	112K	3.10%
armeniasputnik.am	104K	2.89%
aravot.am	99K	2.76%
epress.am	67K	1.85%
armtimes.com	57K	1.59%
armlur.am	53K	1.47%
168.am	52K	1.44%
top-news.am	50K	1.39%
newsarmenia.am	47K	1.31%

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%

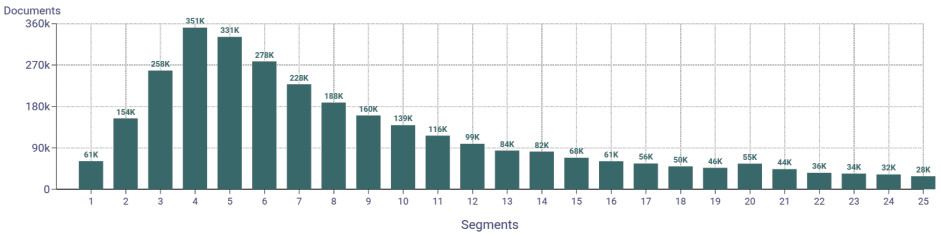
Register labels





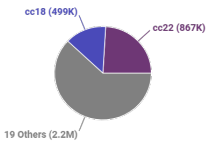
MT:3.2% | 116K Documents

Documents size (in segments)



<= 25 segments **84.45%** (3M documents)
> 25 segments **15.55%** (560K documents)

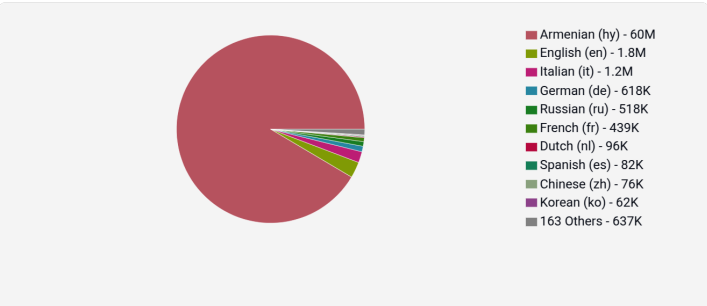
Documents by collection



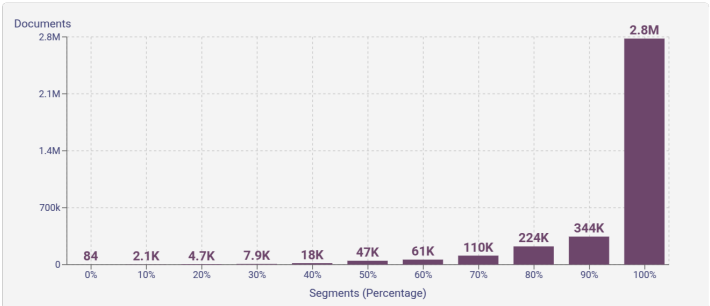
CC = 57.89%
IA = 42.11%

Language Distribution

Number of segments in the Armenian (hy) corpus

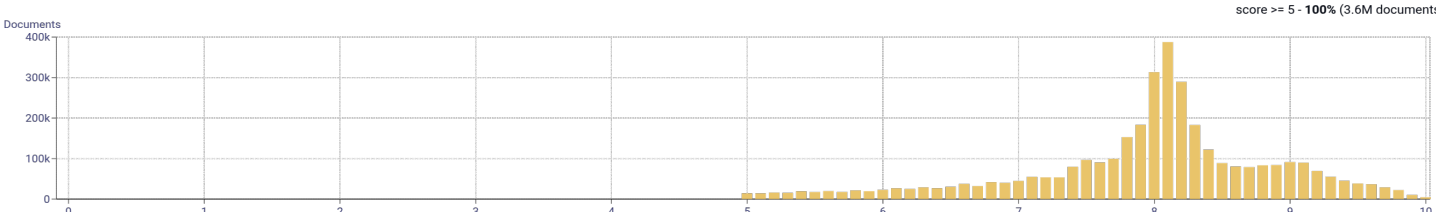


Percentage of segments in Armenian (hy) inside documents



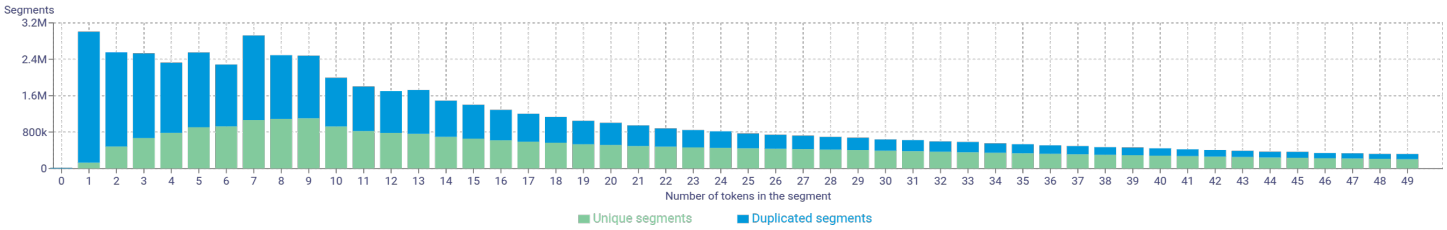
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (3.6M documents)

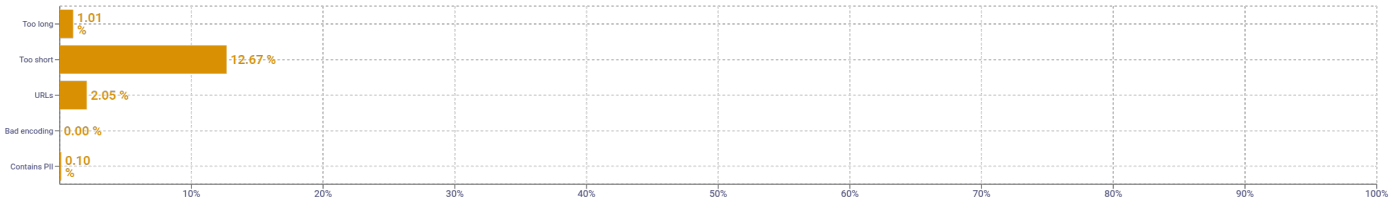


Segment length distribution by token

≤ 49 tokens = **24M** segments | **31M** duplicates
> 50 tokens = **10M** segments | **3.4M** duplicates



Segment noise distribution



Frequent n-grams

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

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 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>, <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

Register labels

Name	Abbr.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av