

General overview

Corpus	Date	Language
hplt-v3-kat_Geor	9/18/2025	Georgian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
6,129,289	105,852,687	65,997,447 (62.35 %)	2.6B	16,909,859,110	41.51 GB

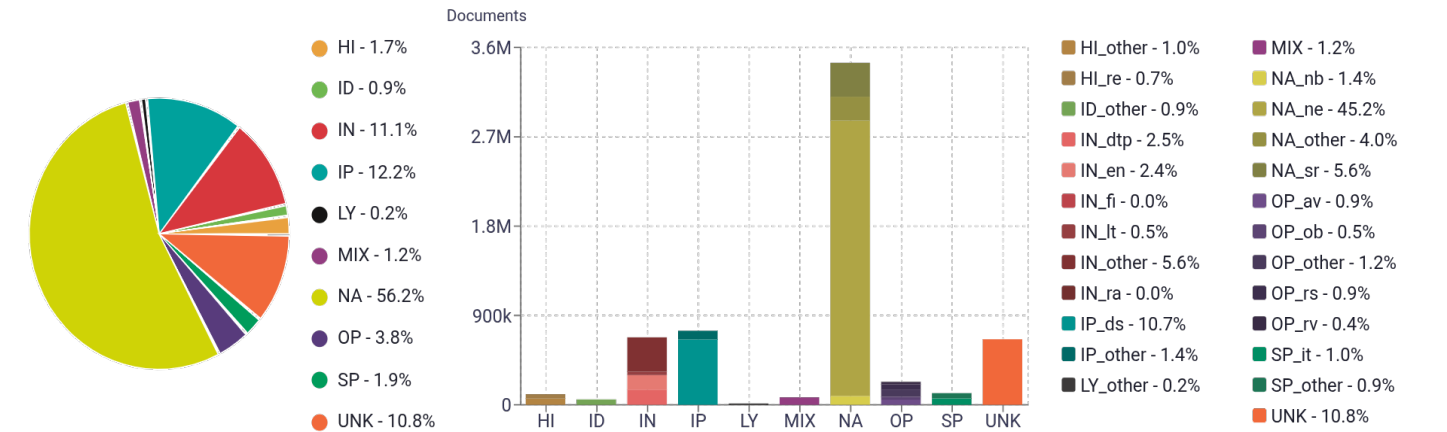
Top 10 domains

Domain	Docs	% of total
radiotavisupleb...	157K	2.56%
interpressnews.ge	103K	1.68%
airbnb.com	99K	1.62%
gancxadebebi.ge	99K	1.61%
netgazeti.ge	97K	1.59%
wikipedia.org	95K	1.55%
sputnik-georgia...	90K	1.47%
wordpress.com	74K	1.21%
bm.ge	73K	1.20%
on.ge	66K	1.08%

Top 10 TLDs

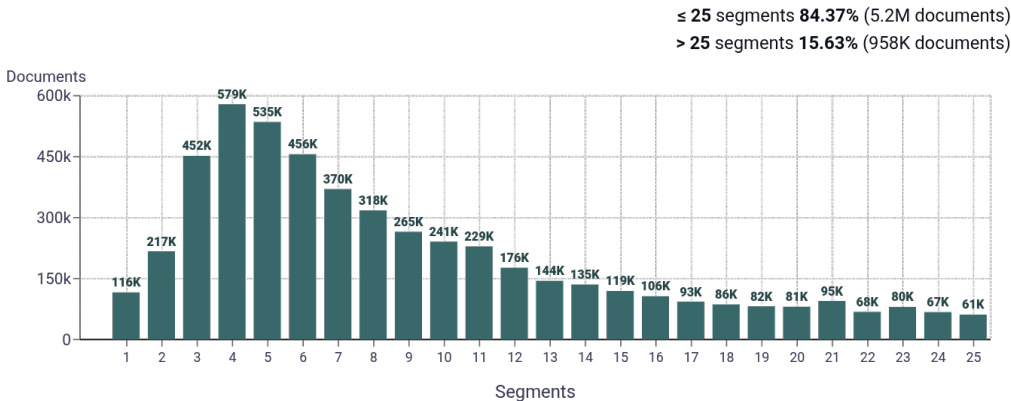
Domain	Docs	% of total
ge	4.2M	68.31%
com	1M	16.97%
org	211K	3.44%
gov.ge	127K	2.08%
net	117K	1.91%
edu.ge	88K	1.43%
com.ge	48K	0.78%
org.ge	28K	0.46%
info	24K	0.39%
eu	20K	0.33%

Register labels

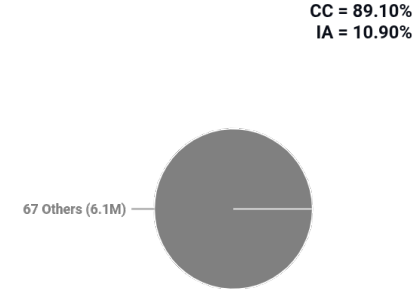


MT:6.5% | 398K Documents

Documents size (in segments) ⓘ

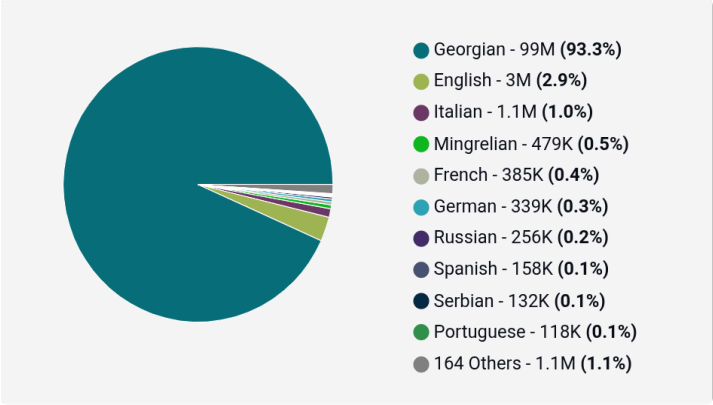


Document collections

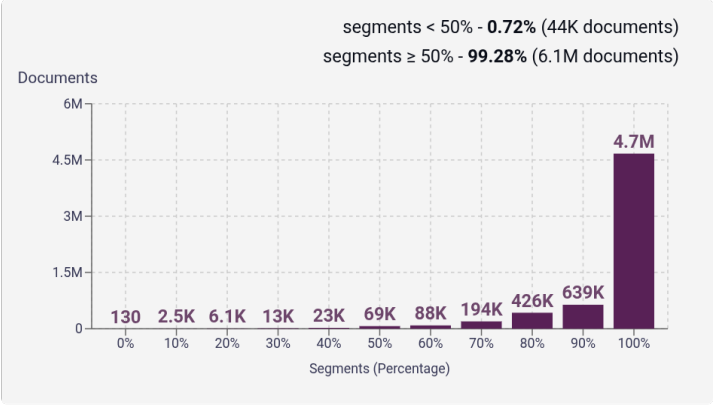


Language Distribution

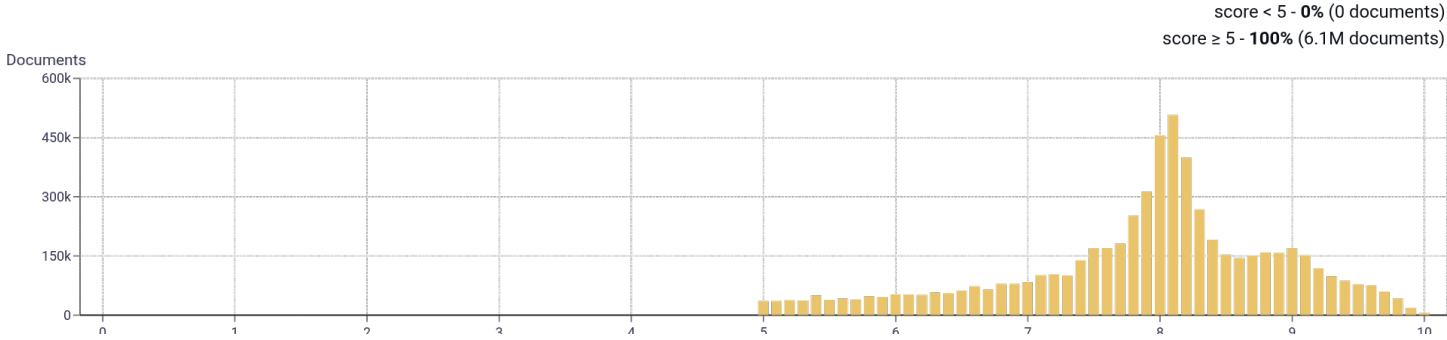
Number of segments in the Georgian corpus



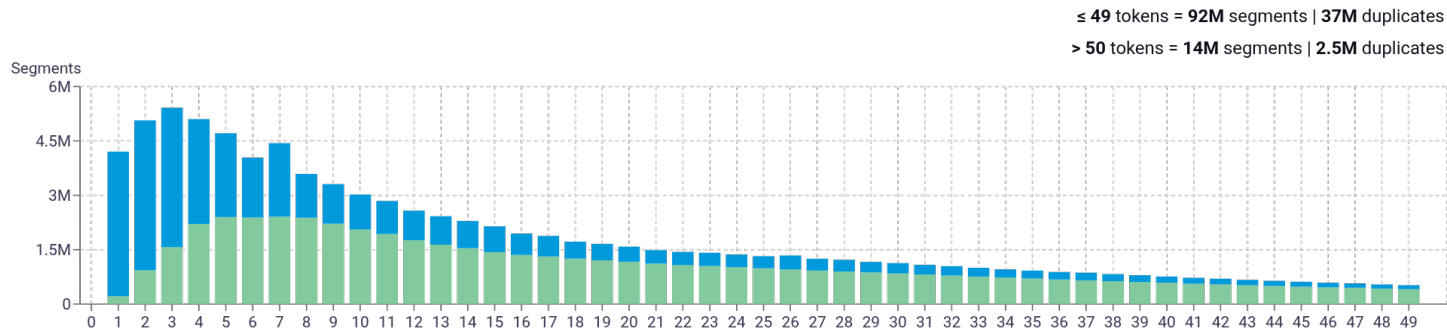
Percentage of segments in Georgian inside documents



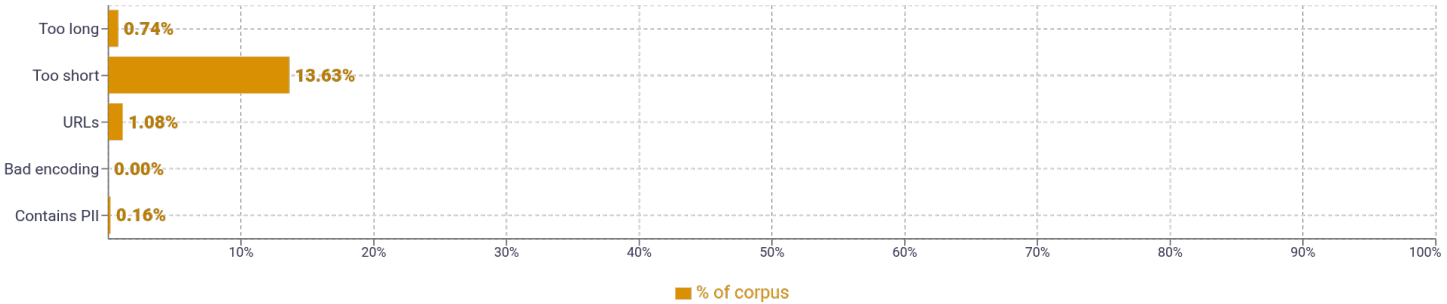
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	ამ 9,613,507 საქართველოს 6,097,485 წლის 5,566,834 მისი 4,756,804 ერთი 4,477,069	
2	სამუალო შეფასება 935,695 რა თქმა 583,414 მიუხედავად იმისა 477,972 იხილეთ მეტი 461,822 წლის განმავლობაში 387,980	
3	შინაგან საქმეთა სამინისტროს 125,970 ცოტა ხნის წინ 119,049 სამედიცინო ენციკლოპედიური განმარტებითი 95,989 ენციკლოპედიური განმარტებითი ლექსიკონი 95,771 ამა თუ იმ 80,573	
4	სამედიცინო ენციკლოპედიური განმარტებითი ლექსიკონი 95,755 ეკონომიკისა და მდგრადი განვითარების 71,858 სისუფთავისა და სხვა მახასიათებლების 60,381 მიღებული აქვს მაღალი შეფასებები 60,381 აქვს მაღალი შეფასებები მდებარეობის 60,381	
5	სისუფთავისა და სხვა მახასიათებლების მხრივ 60,381 მიღებული აქვს მაღალი შეფასებები მდებარეობის 60,381 იოვეთ თქვენთვის შესაფერისი ფართობის საცხოვრებელი 60,332 ჯინეტის განვითარებისთვის საქართველოში რამდენიმე არაკომერციული 56,584 ძირითადად ფრანგულ ბაზარზე იყო ორიენტირებული 56,584	

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>

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	How-to or instructions	HI	Description of a thing or person	dtp
Lyrical	LY	Recipe	re	FAQ	fi
Spoken	SP	Informational persuasion	IP	Legal terms & conditions	lt
Interview	it	Description with intent to sell	ds	Opinion	OP
Interactive discussion	ID	News & opinion blog or editorial	ed	Review	rv
Narrative	NA	Informational description	IN	Opinion blog	ob
News report	ne	Enciclopedia article	en	Denominational religious blog or sermon	rs
Sports report	sr	Research article	ra	Advice	av
Narrative blog	nb				