

General overview

Corpus	Date	Language
hplt-v3-als_Latn	9/16/2025	Albanian (sq)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
11,184,735	162,346,846	107,225,822 (66.05 %)	5.3B	27,558,921,146	27.6 GB

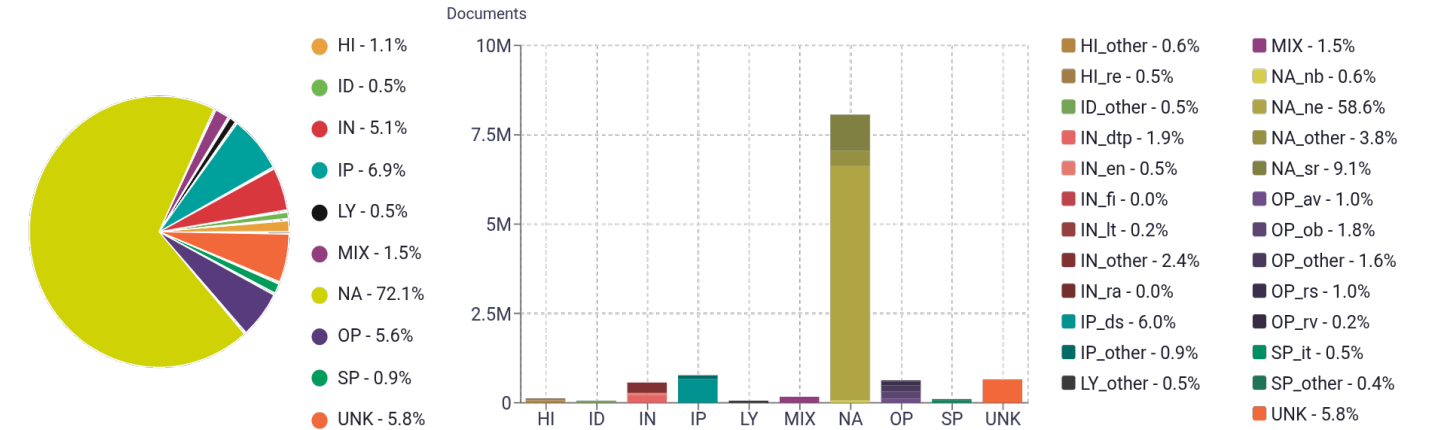
Top 10 domains

Domain	Docs	% of total
botasot.info	162K	1.45%
telegrafi.com	133K	1.19%
airbnb.com	133K	1.19%
top-channel.tv	133K	1.19%
evropaelire.org	130K	1.16%
albeu.com	126K	1.12%
cna.al	101K	0.90%
shqiptarja.com	100K	0.90%
koha.net	83K	0.74%
portalb.mk	83K	0.74%

Top 10 TLDs

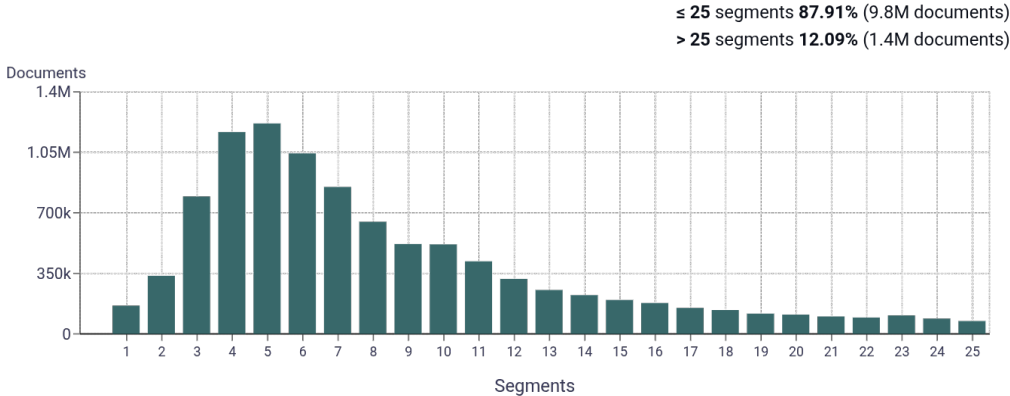
Domain	Docs	% of total
com	4.4M	39.67%
al	2.8M	24.63%
net	864K	7.73%
info	682K	6.10%
org	495K	4.42%
tv	494K	4.41%
mk	440K	3.93%
com.al	182K	1.62%
ch	110K	0.98%
gov.al	99K	0.89%

Register labels

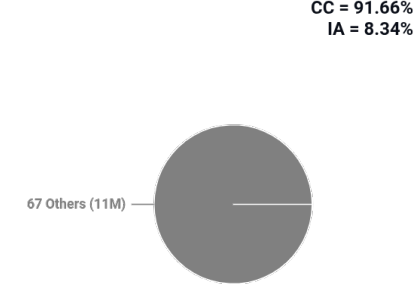


MT:3.6% | 403K Documents

Documents size (in segments) ⓘ

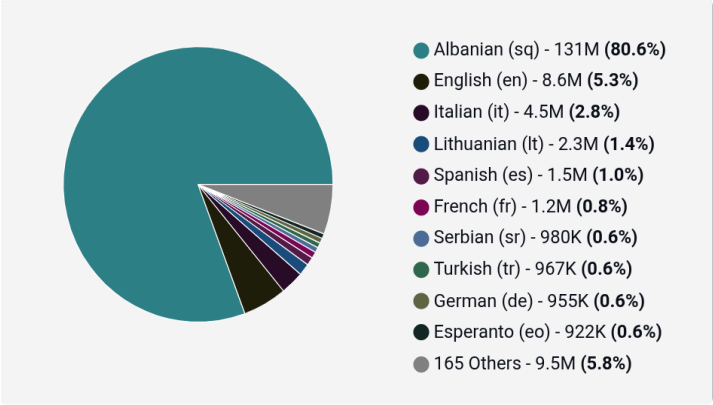


Document collections

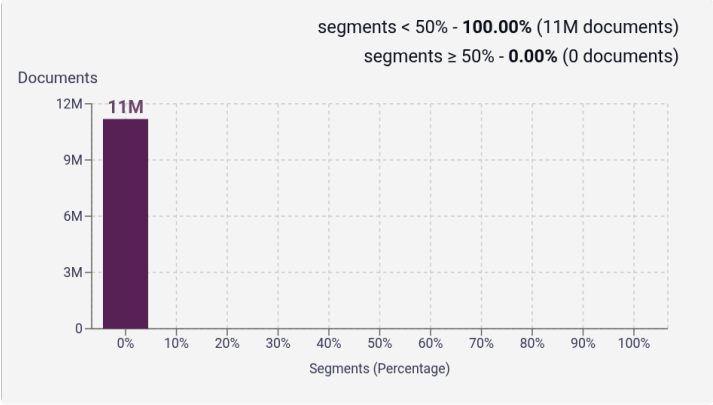


Language Distribution

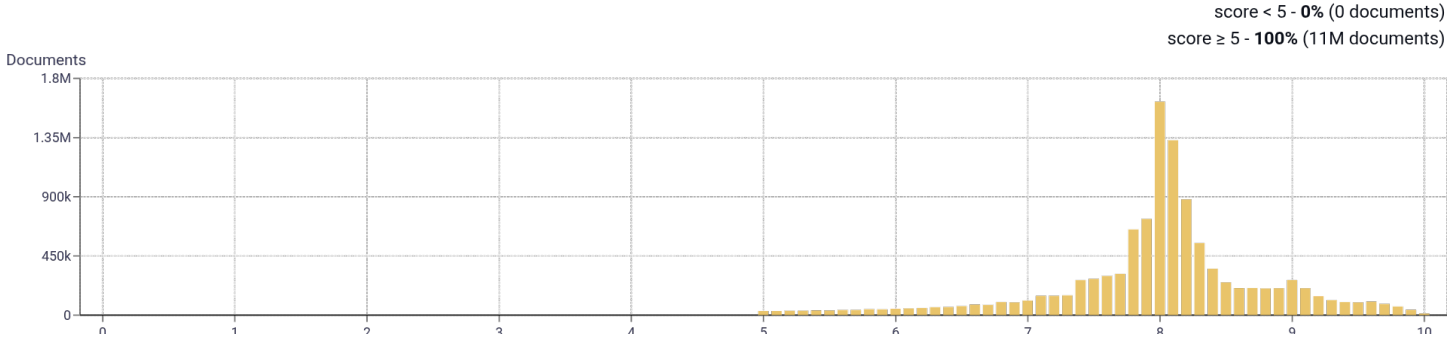
Number of segments in the Albanian (sq) corpus



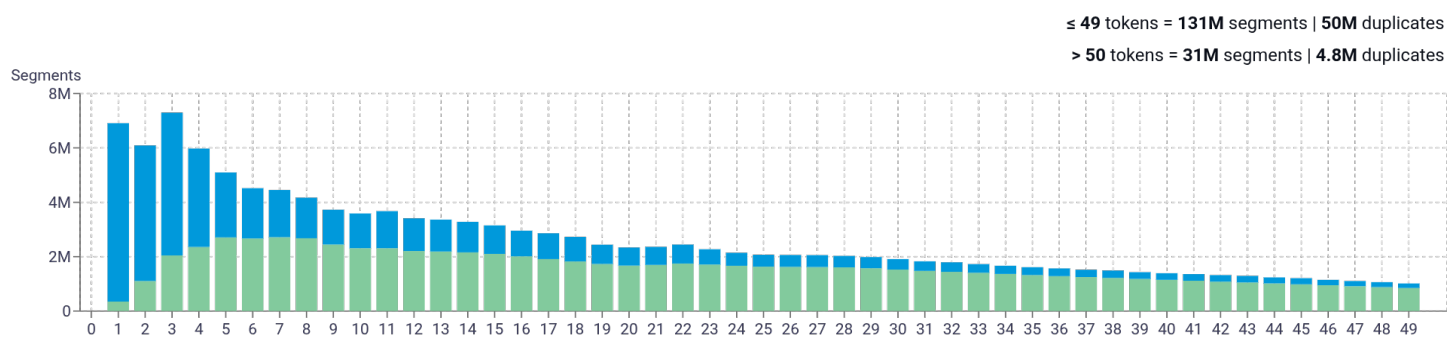
Percentage of segments in Albanian (sq) inside documents



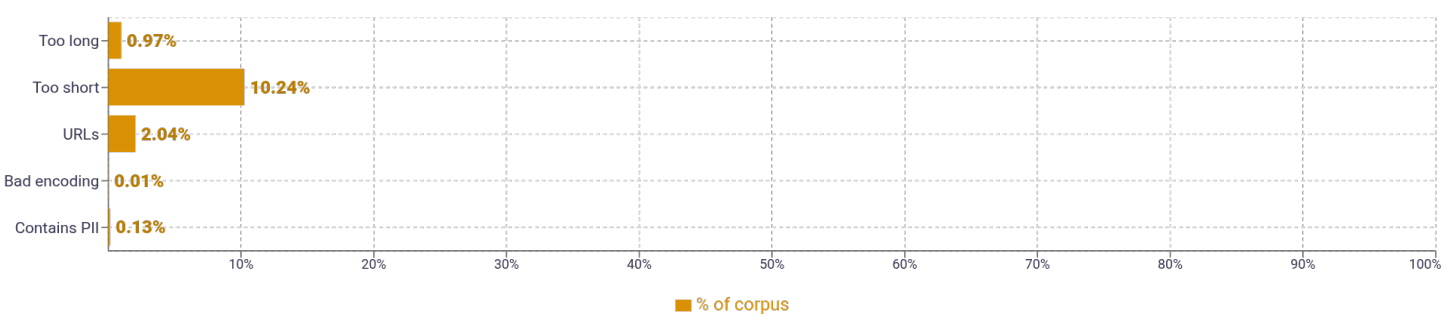
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	kanë 12,091,041parë 7,301,854vetëm 6,901,338gjatë 6,164,769tha 5,551,862	
2	vlerësimi mesatar 1,153,675edi rama 781,913shiko detajet 713,114kanë qenë 657,241milionë euro 482,915	
3	herë të parë 522,478shtetet e bashkuara 435,308ditë më parë 355,913republikës së kosovës 351,041vitet e fundit 249,806	
4	shtëpi pushimesh me qira 142,997luftës së dytë botërore 80,714kanë vlerësime të larta 78,704hapësirë sa të jetë 78,527qira për çdo stil 78,525	
5	shtetet e bashkuara të amerikës 109,423shteteve të bashkuara të amerikës 86,145vlerësime të larta për vendndodhjen 78,669pushimesh me qira për çdo 78,525o i o i o 55,720	

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				