HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-als_Latn	9/23/2025	Albanian (als)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
11,184,735	162,346,846	107,225,822 (66.05 %)	5.4B	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

- HI - 1.1%
- ID - 0.5%
- IN - 5.1%
- IP - 6.9%
- LY - 0.5%
- MIX - 1.5%
- NA - 72.1%
- OP - 5.6%
- SP - 0.9%
- UNK - 5.8%

- HI_other - 0.6%
- HI_re - 0.5%
- ID_other - 0.5%
- IN_dtp - 1.9%
- IN_en - 0.5%
- IN_fi - 0.0%
- IN_lt - 0.2%
- IN_other - 2.4%
- IN_ra - 0.0%
- IP_ds - 6.0%
- IP_other - 0.9%
- LY_other - 0.5%
- MIX - 1.5%
- NA_nb - 0.6%
- NA_ne - 58.6%
- NA_other - 3.8%
- NA_sr - 9.1%
- OP_av - 1.0%
- OP_ob - 1.8%
- OP_other - 1.6%
- OP_rs - 1.0%
- OP_rv - 0.2%
- SP_it - 0.5%
- SP_other - 0.4%
- UNK - 5.8%
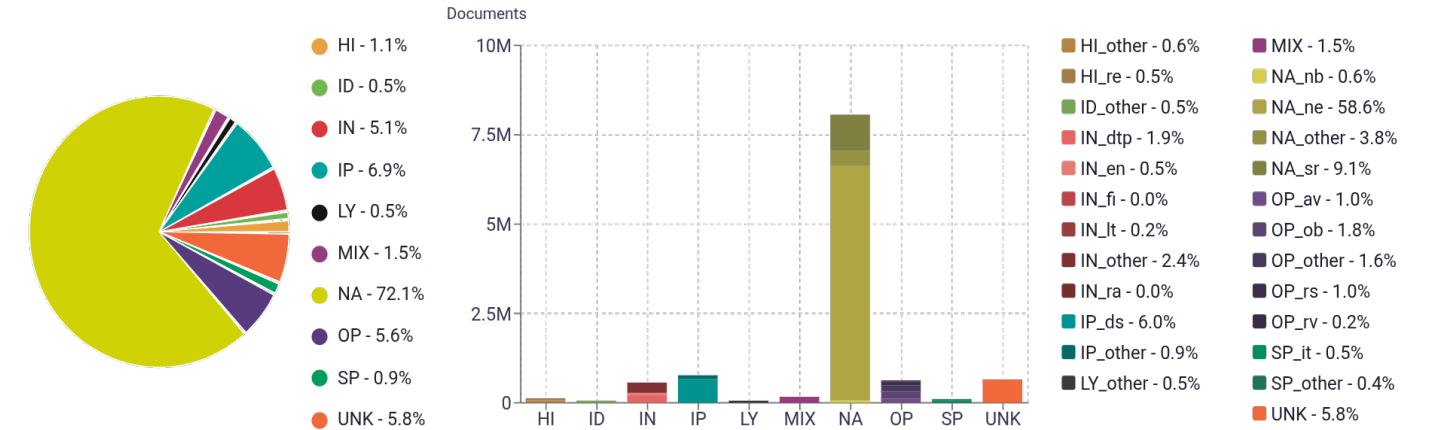
🤖 **MT**:3.6% | 403K Documents

Documents size (in segments) ⓘ

≤ 25 segments **87.91%** (9.8M documents)
> 25 segments **12.09%** (1.4M documents)

Document collections

CC = 91.66%
IA = 8.34%

67 Others (11M)
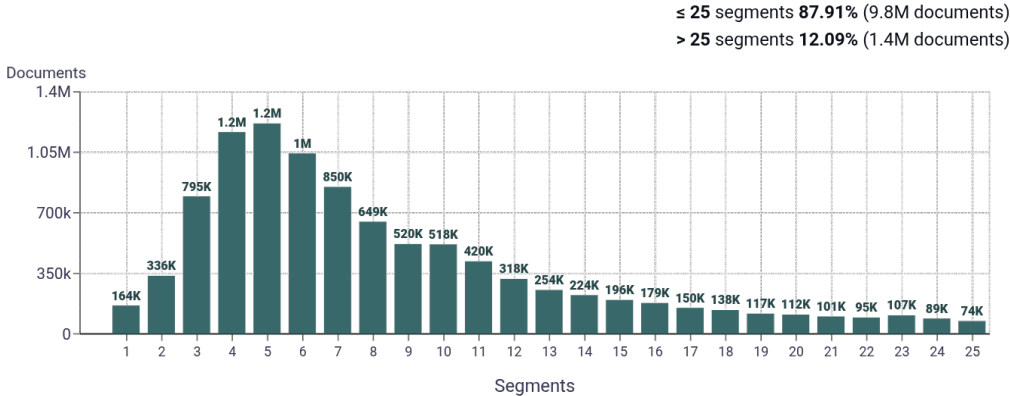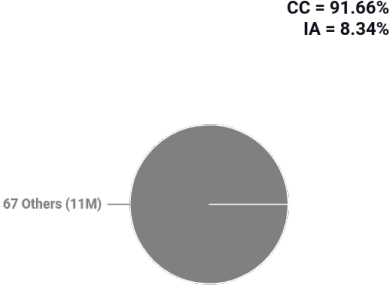
Language Distribution

Number of segments in the Albanian (als) corpus

- Albanian (sq) - 131M **(80.6%)**
- English (en) - 8.6M **(5.3%)**
- Italian (it) - 4.5M **(2.8%)**
- Lithuanian (lt) - 2.3M **(1.4%)**
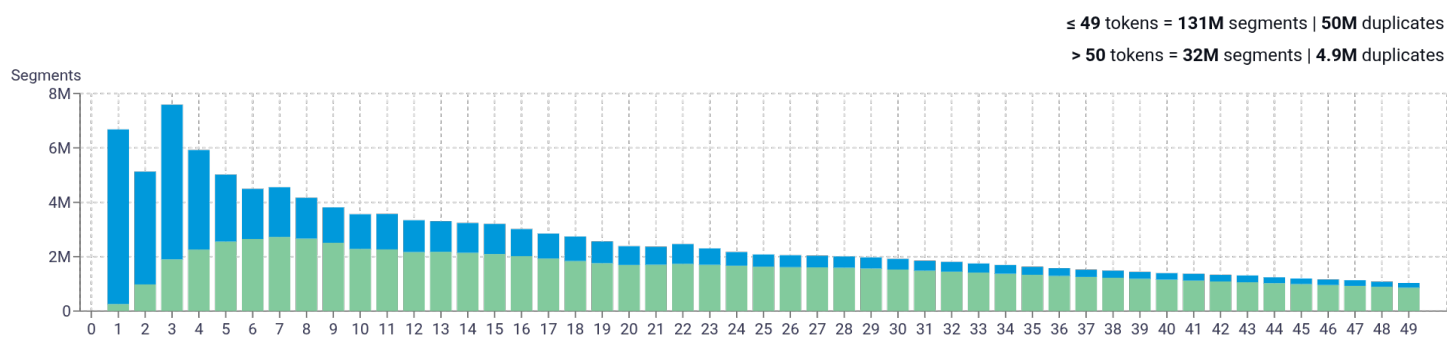- Spanish (es) - 1.5M **(1.0%)**
- French (fr) - 1.2M **(0.8%)**
- Serbian (sr) - 980K **(0.6%)**
- Turkish (tr) - 967K **(0.6%)**
- German (de) - 955K **(0.6%)**
- Esperanto (eo) - 922K **(0.6%)**
- 165 Others - 9.5M **(5.8%)**
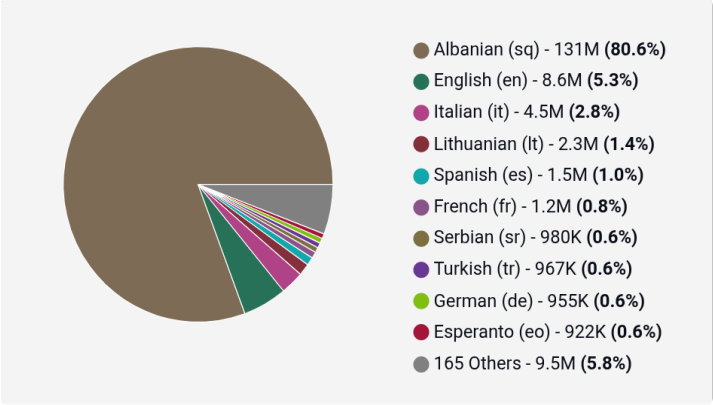
Percentage of segments in Albanian (als) inside documents
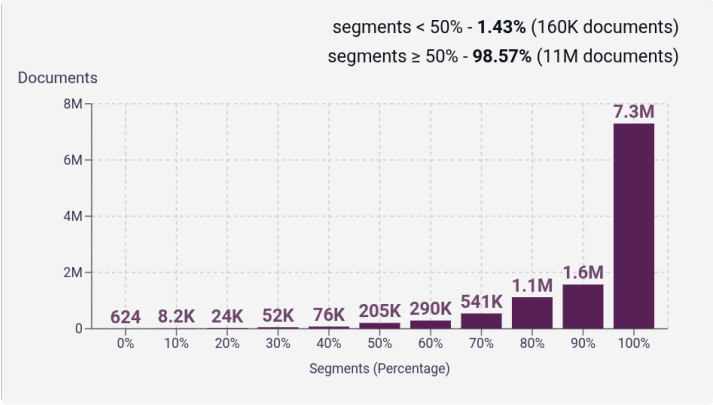
segments < 50% - **1.43%** (160K documents)
segments ≥ 50% - **98.57%** (11M documents)

Documents

0%	624	
10%	8.2K	
20%	24K	
30%	52K	
40%	76K	
50%	205K	
60%	290K	
70%	541K	
80%	1.1M	
90%	1.6M	
100%	7.3M	

Segments (Percentage)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (11M documents)

Documents

Segment length distribution by token

≤ 49 tokens = **131M** segments | **50M** duplicates
> 50 tokens = **32M** segments | **4.9M** duplicates

Segments

Segment noise distribution

Too long	0.97%
Too short	10.24%
URLs	2.04%
Bad encoding	0.01%
Contains PII	0.08%

■ % of corpus
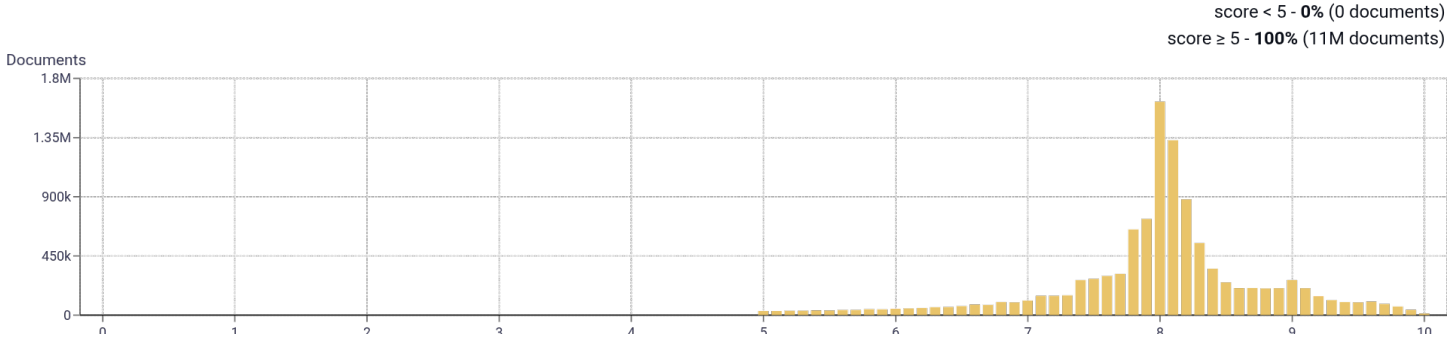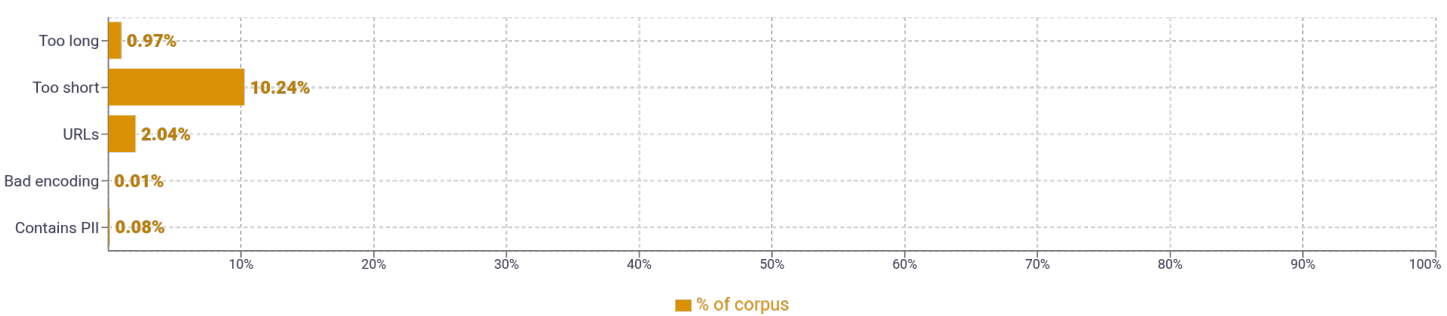
Frequent n-grams

SIZE	N-GRAMS	
1	kanë | 12,099,790 parë | 7,311,344 vetëm | 6,909,099 gjatë | 6,183,055 tha | 5,602,443	⧉
2	vlerësimi mesatar | 1,153,675 edi rama | 785,797 shiko detajet | 713,123 kanë qenë | 657,705 milionë euro | 483,876	⧉
3	herë të parë | 522,911 shtetet e bashkuara | 436,330 ditë më parë | 356,460 republikës së kosovës | 351,956 vitet e fundit | 250,993	⧉
4	shtëpi pushimesh me qira | 142,998 luftës së dytë botërore | 81,247 kanë vlerësime të larta | 78,704 hapësirë sa të jetë | 78,527 qira për çdo stil | 78,525	⧉
5	shtetet e bashkuara të amerikës | 109,899 shteteve të bashkuara të amerikës | 86,305 vlerësime të larta për vendndodhjen | 78,669 pushimesh me qira për çdo | 78,525 o i o i o | 55,720	⧉

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.	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				