HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-uzn_Latn	9/24/2025	Northern Uzbek

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
1,882,814	32,931,412	24,573,678 (74.62 %)	923M	6,480,285,390	6.2 GB

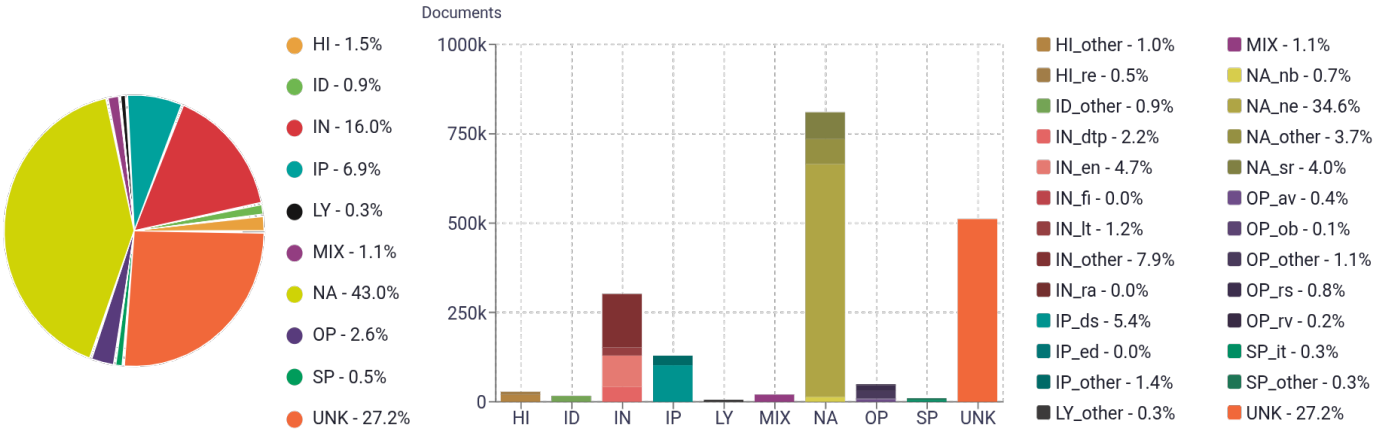
Top 10 domains

Domain	Docs	% of total
daryo.uz	72K	3.80%
wikipedia.org	69K	3.64%
kun.uz	56K	3.00%
gazeta.uz	38K	2.00%
xs.uz	36K	1.92%
sputniknews-uz.com	28K	1.49%
amerikaovozi.com	28K	1.49%
xit.uz	24K	1.26%
rfa.org	23K	1.21%
ozodlik.org	21K	1.10%

Top 10 TLDs

Domain	Docs	% of total
uz	971K	51.58%
com	415K	22.03%
org	180K	9.57%
net	80K	4.24%
ru	36K	1.93%
tv	20K	1.04%
info	14K	0.76%
net.tr	12K	0.65%
biz	9.1K	0.48%
pw	9.1K	0.48%

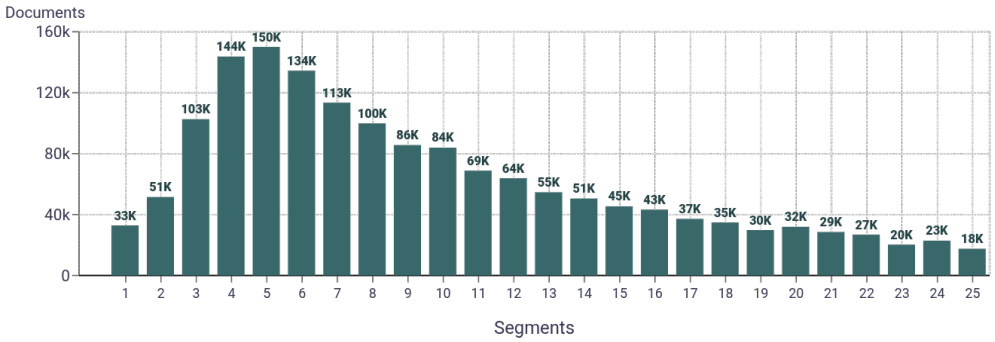
Register labels



MT:23.3% | 439K Documents

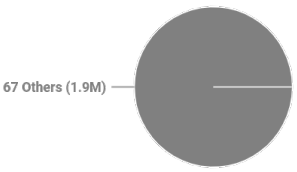
Documents size (in segments)

≤ 25 segments **83.55%** (1.6M documents)
> 25 segments **16.45%** (310K documents)



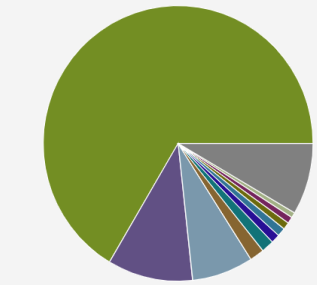
Document collections

CC = 95.83%
IA = 4.17%



Language Distribution

Number of segments in the Northern Uzbek corpus

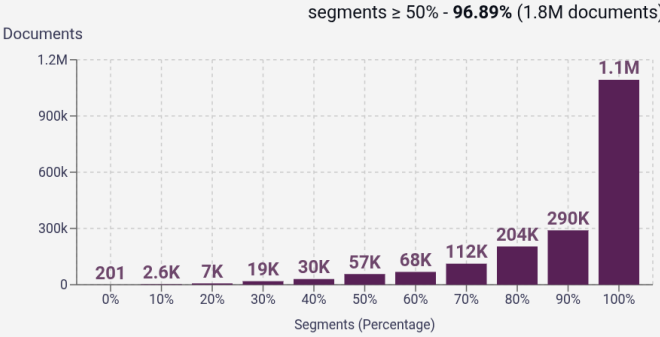


*Northern Uzbek identification might be inaccurate because it is not supported by FastSpell.

- Uzbek - 22M (**66.5%**)
- English - 3.4M (**10.2%**)
- Turkish - 2.4M (**7.4%**)
- Italian - 571K (**1.7%**)
- German - 499K (**1.5%**)
- Russian - 343K (**1.0%**)
- Azerbaijani - 328K (**1.0%**)
- Indonesian - 275K (**0.8%**)
- Portuguese - 255K (**0.8%**)
- Spanish - 204K (**0.6%**)
- 165 Others - 2.8M (**8.4%**)

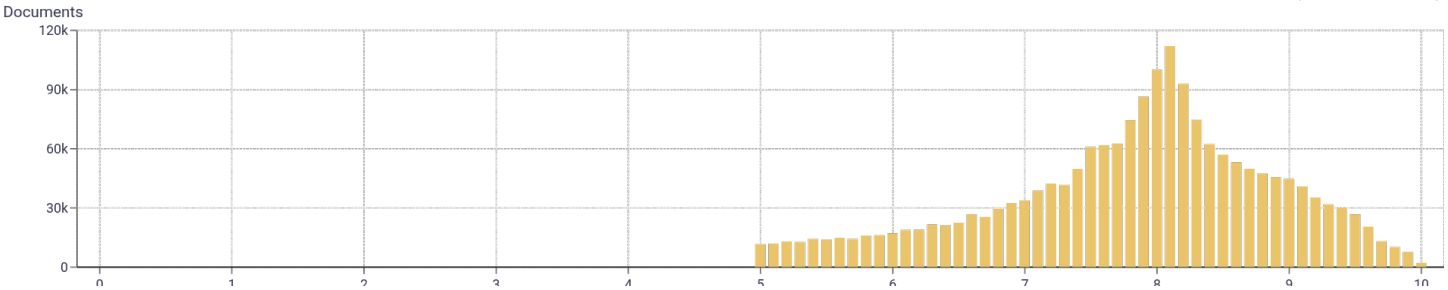
Percentage of segments in Northern Uzbek inside documents

segments < 50% - **3.11%** (59K documents)
segments ≥ 50% - **96.89%** (1.8M documents)



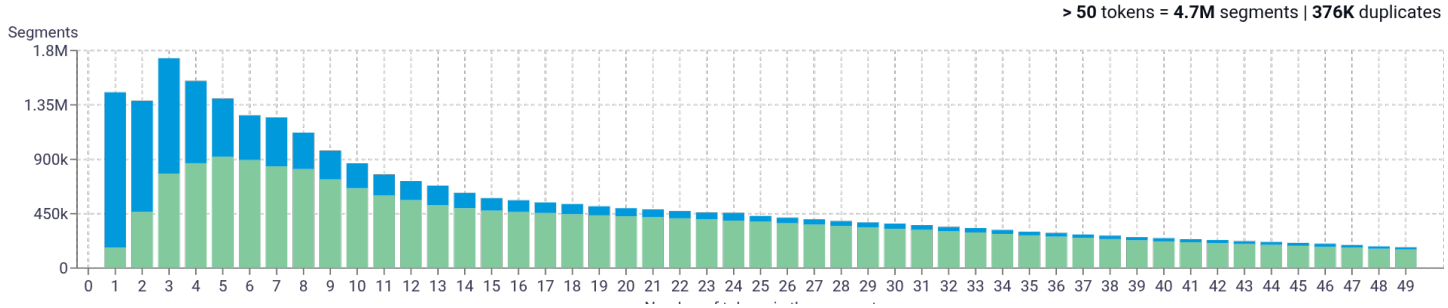
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (1.9M documents)

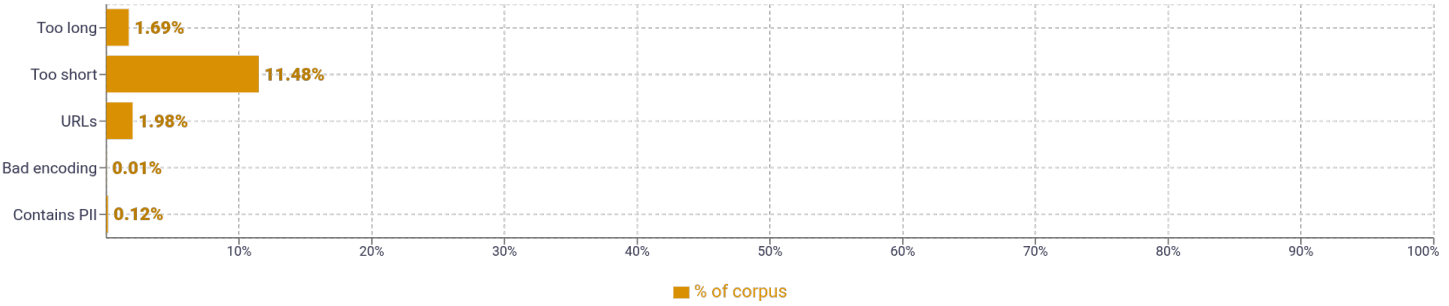


Segment length distribution by token

≤ 49 tokens = **28M** segments | **8M** duplicates
> 50 tokens = **4.7M** segments | **376K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	mumkin. | 1,660,303 · onlayn | 1,090,541 · yil | 1,069,236 · pul | 1,021,433 · katta | 1,016,859
2	uzbek tilida | 426,868 · batafsil ma | 323,558 · tarjima kino | 285,249 · o'zbekiston respublikasi | 259,087 · tilida o'zbekcha | 233,302
3	uzbek tilida o'zbekcha | 232,825 · o'zbekcha tarjima kino | 199,025 · tilida o'zbekcha tarjima | 187,843 · hd tas-ix skachat | 161,577 · full hd tas-ix | 87,124
4	uzbek tilida o'zbekcha tarjima | 187,376 · tilida o'zbekcha tarjima kino | 181,562 · full hd tas-ix skachat | 86,014 · tarjima kino full hd | 64,994 · asl nusxadan arxivlandi. qaraldi | 40,709
5	uzbek tilida o'zbekcha tarjima kino | 181,106 · tarjima kino full hd skachat | 31,123 · tarjima kino full hd tas-ix | 30,688 · kino full hd tas-ix skachat | 29,981 · premyera uzbek tilida o'zbekcha tarjima | 26,714

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				