

General overview

Corpus	Date	Language
hplt-v3-uzn_Latn	9/18/2025	Uzbek (uz)

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

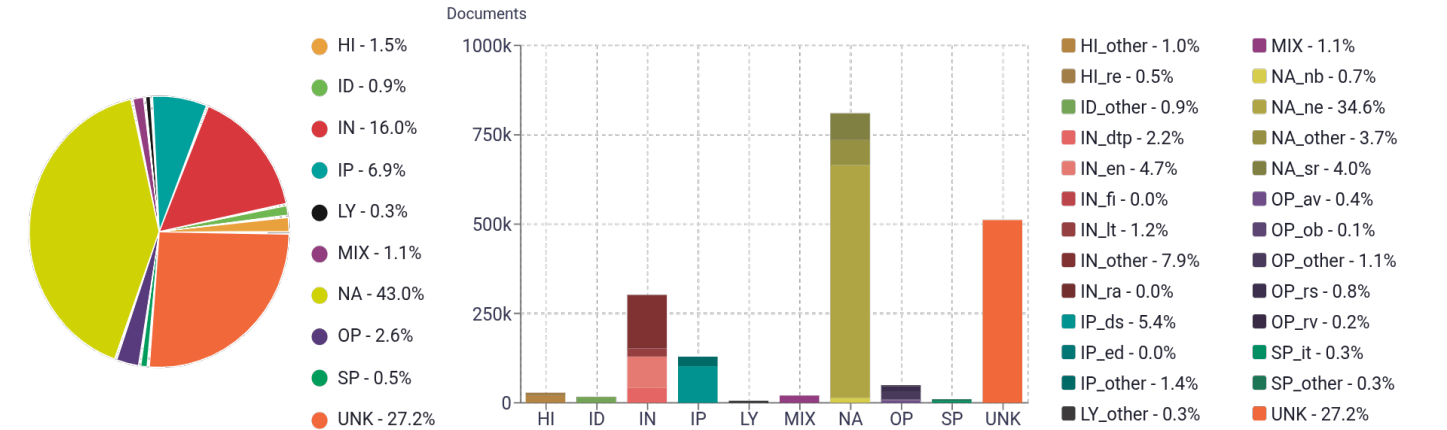
Top 10 domains

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

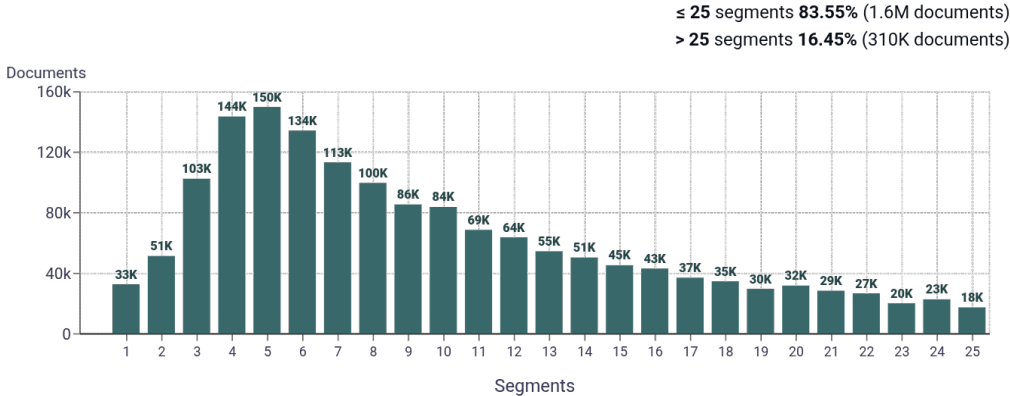
Domain	Docs	% of total
uz	971K	51.58%
com	415K	22.03%
org	180K	9.57%
net	80K	4.24%
ru	36K	1.93%
tv	20K	1.04%
info	14K	0.76%
net.tr	12K	0.65%
biz	9.1K	0.48%
pw	9.1K	0.48%

Register labels

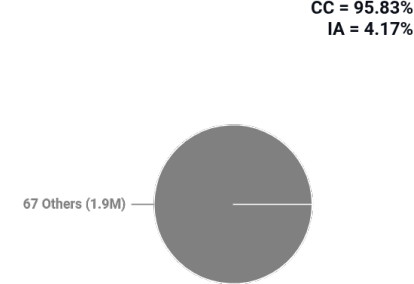


MT:23.3% | 439K Documents

Documents size (in segments) ⓘ

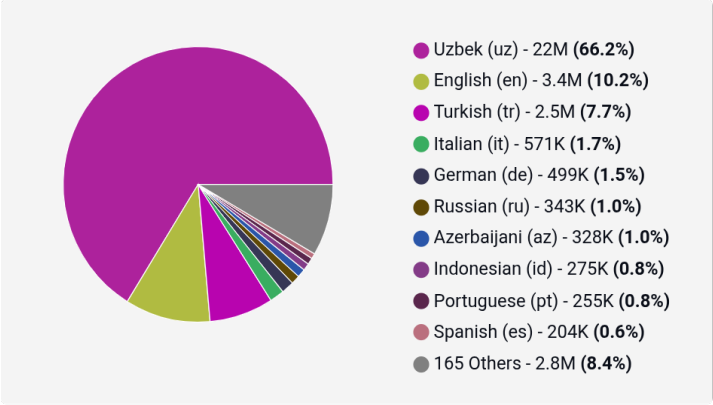


Document collections

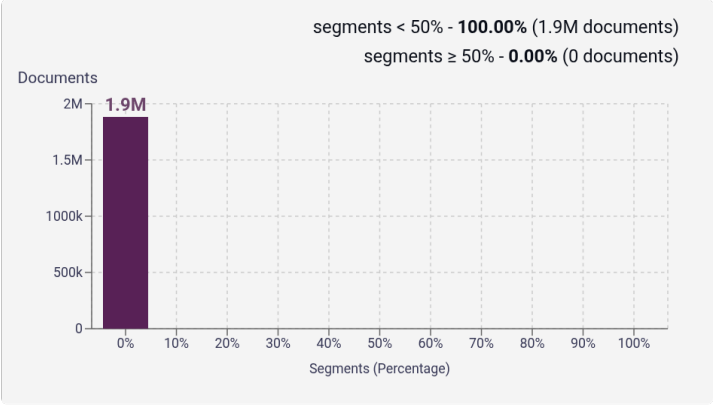


Language Distribution

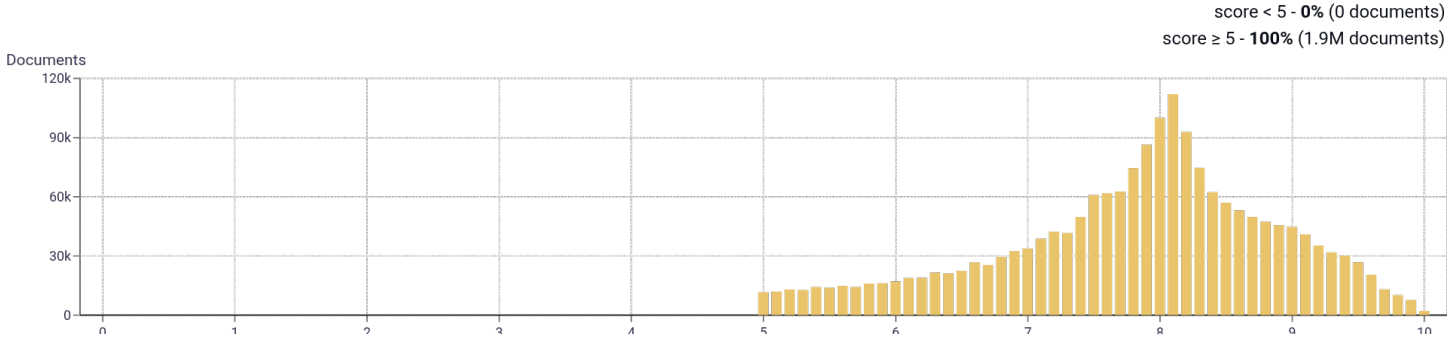
Number of segments in the Uzbek (uz) corpus



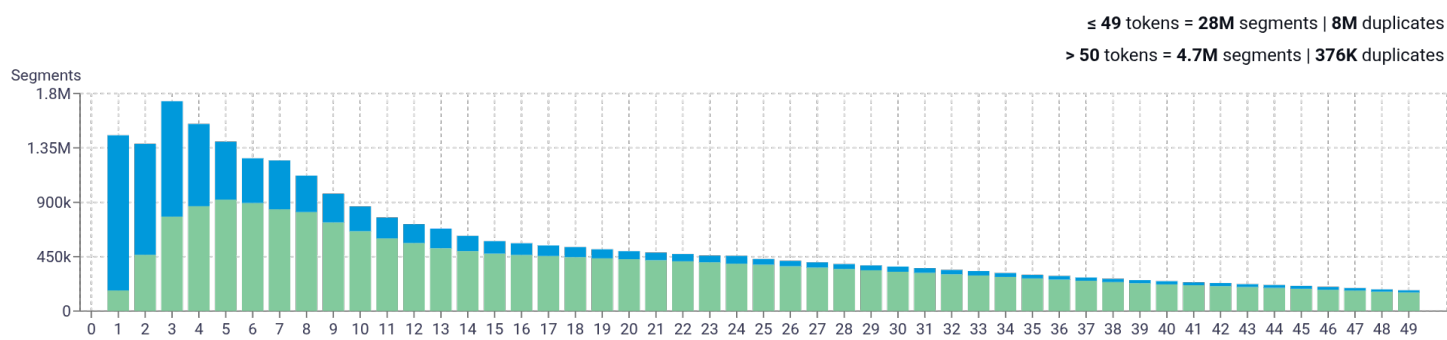
Percentage of segments in Uzbek (uz) inside documents



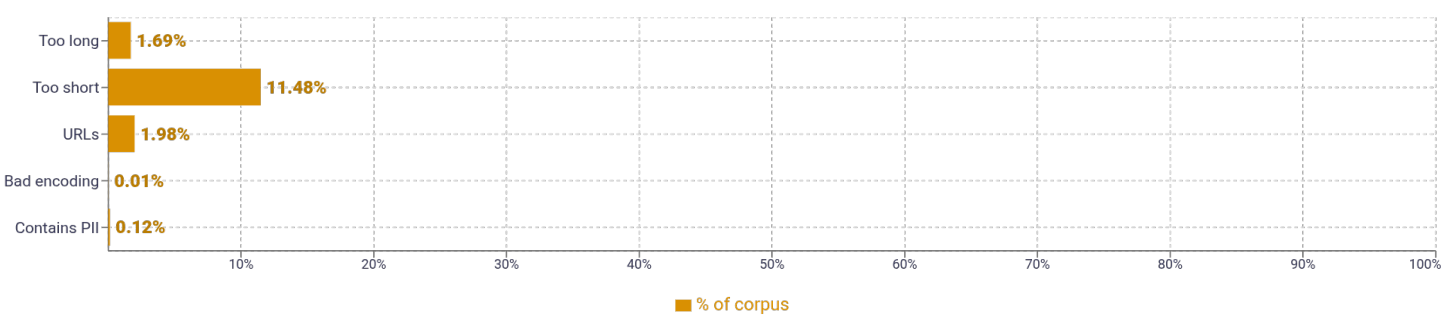
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

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				