

General overview

Corpus	Date	Language
hplt-v3-kaz_Cyrl	9/18/2025	Kazakh

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
5,119,572	100,617,148	71,967,855 (71.53 %)	2.7B	17,117,387,686	29.07 GB

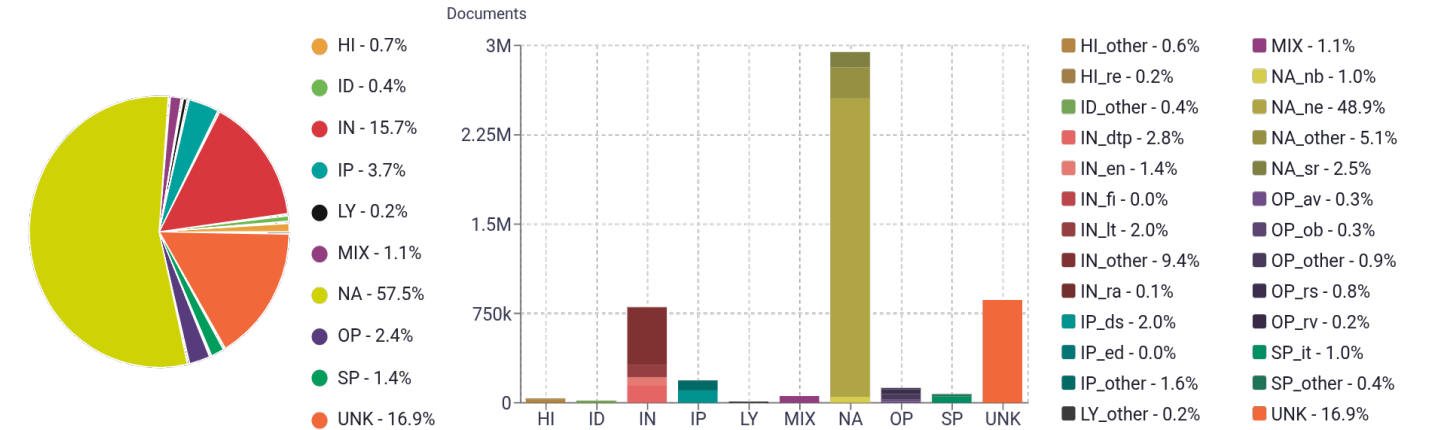
Top 10 domains

Domain	Docs	% of total
inform.kz	171K	3.34%
nur.kz	118K	2.31%
azattyq.org	110K	2.15%
egemen.kz	89K	1.75%
baq.kz	83K	1.62%
tengrinews.kz	74K	1.44%
stan.kz	72K	1.41%
wikipedia.org	68K	1.32%
24.kz	63K	1.23%
zakon.kz	56K	1.09%

Top 10 TLDs

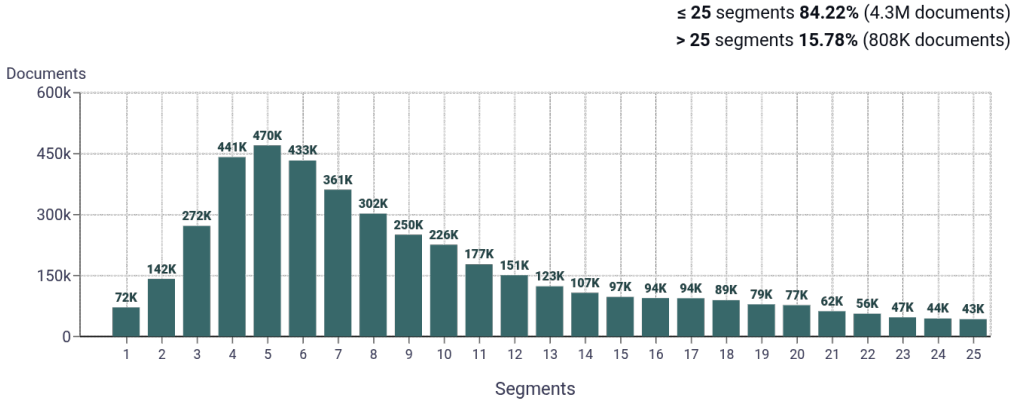
Domain	Docs	% of total
kz	3.7M	71.57%
com	484K	9.46%
org	285K	5.57%
ru	140K	2.73%
gov.kz	107K	2.09%
edu.kz	78K	1.52%
info	59K	1.16%
net	49K	0.96%
tv	31K	0.60%
uz	24K	0.48%

Register labels

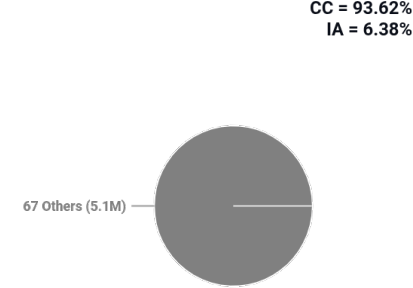


MT:11.8% | 604K Documents

Documents size (in segments) ⓘ

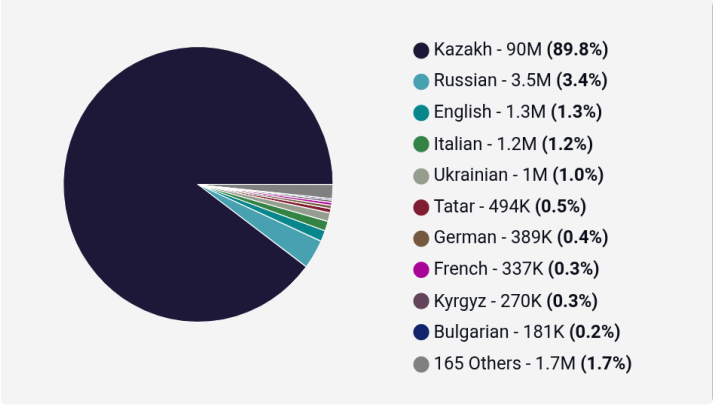


Document collections

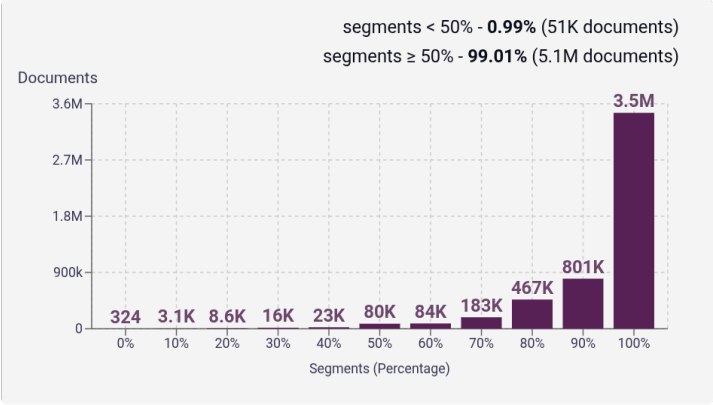


Language Distribution

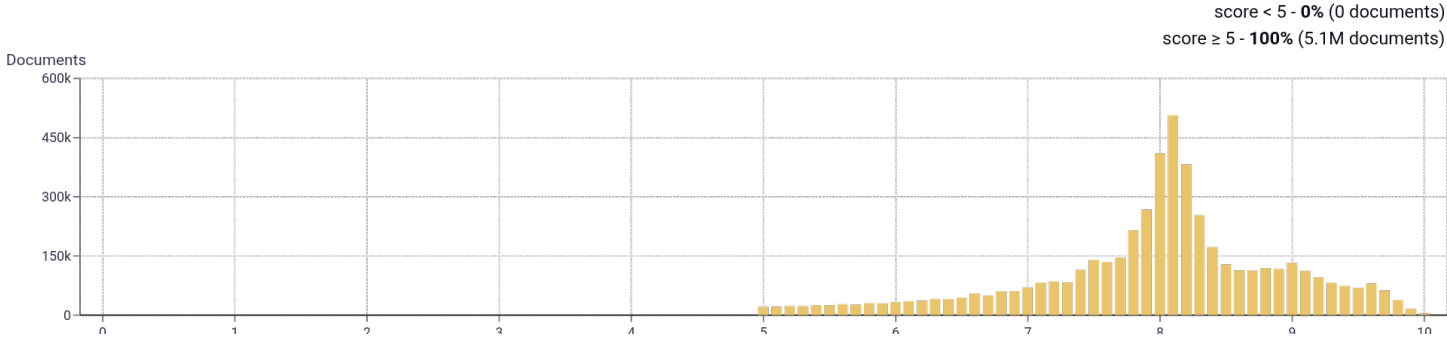
Number of segments in the Kazakh corpus



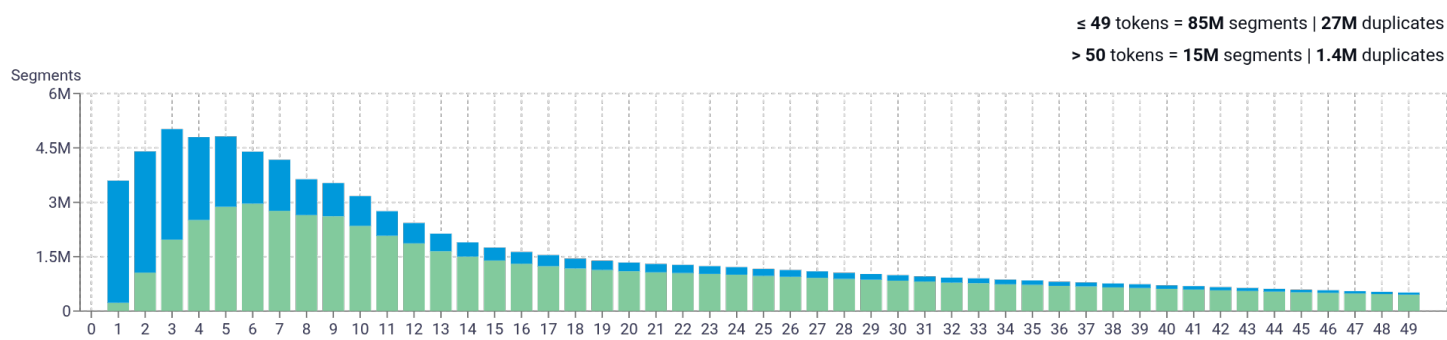
Percentage of segments in Kazakh inside documents



Distribution of documents by document score

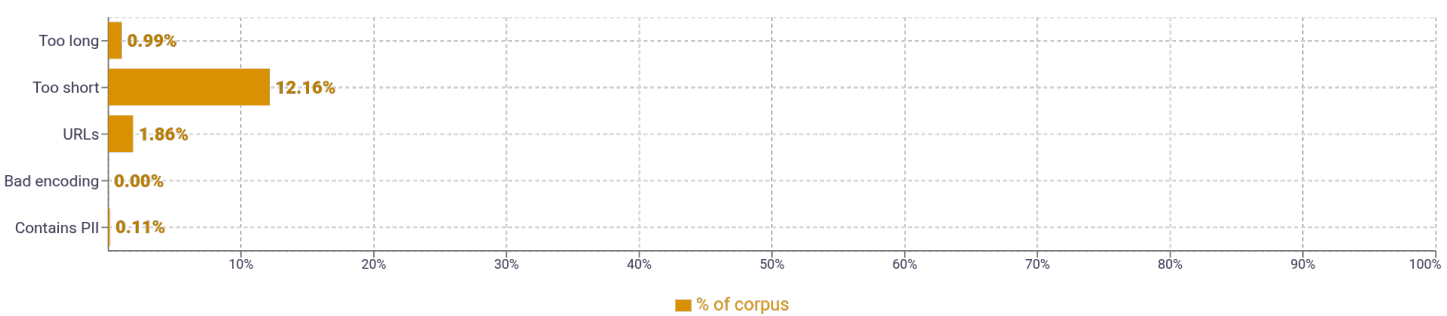


Segment length distribution by token



≤ 49 tokens = 85M segments | 27M duplicates
> 50 tokens = 15M segments | 1.4M duplicates

Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	және 28,495,646 деп 7,701,443 де 5,907,940 қазақстан 5,553,042 болады 5,275,924	
2	болып табылады 1,318,729 қазақстан республикасының 1,246,895 деп хабарлайды 1,054,858 білім беру 944,688 қазақстан республикасы 906,914	
3	сыбайлас жемқорлыққа қарсы 202,957 деп хабарлайды қазақпарат 184,347 білім және ғылым 158,280 хабарлайды қазақпарат тілшісі 105,836 деп атап өтті 105,833	
4	деп хабарлайды қазақпарат тілшісі 105,618 күн өткен соң қолданысқа 96,073 өткен соң қолданысқа енгізіледі 87,476 сыбайлас жемқорлыққа қарсы іс 85,519 алғашқы ресми жарияланған күнінен 79,408	
5	он күн өткен соң қолданысқа 85,615 күн өткен соң қолданысқа енгізіледі 83,912 ресми жарияланған күнінен кейін күнтізбелік 58,893 күнінен кейін күнтізбелік он күн 53,660 жарияланған күнінен кейін күнтізбелік он 53,452	

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				