

General overview

Corpus	Date	Language
hplt-v3-rus_Cyrl-SAMPLED	9/24/2025	Russian (ru)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
878,000	27,190,260	23,535,178 (86.56 %)	717M	4,285,368,912	7.23 GB

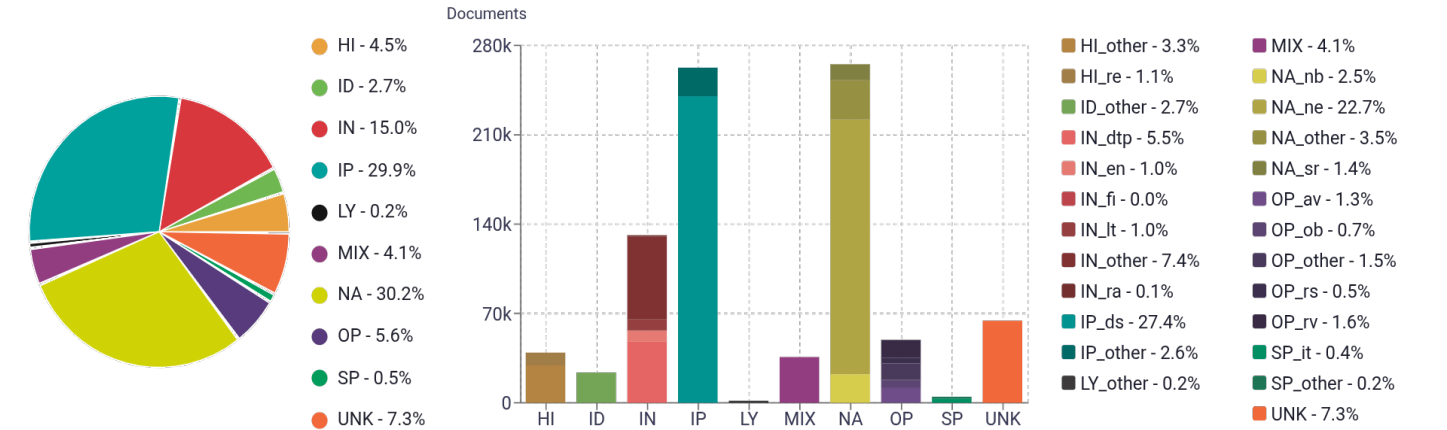
Top 10 domains

Domain	Docs	% of total
livejournal.com	7.3K	0.83%
aif.ru	4.9K	0.56%
blogspot.com	2.3K	0.26%
wikipedia.org	1.6K	0.18%
spb.ru	1.5K	0.17%
mail.ru	1.5K	0.17%
webcindario.com	1.3K	0.15%
academic.ru	1.2K	0.14%
rbc.ru	1.2K	0.13%
gov.ru	951	0.11%

Top 10 TLDs

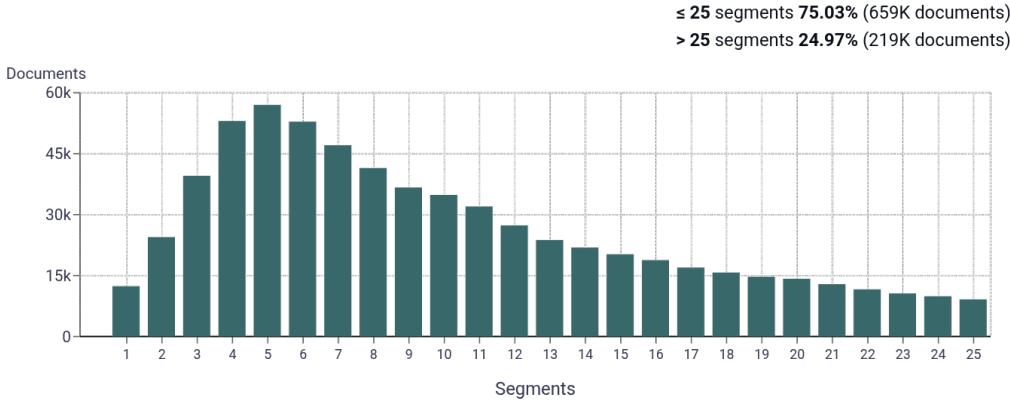
Domain	Docs	% of total
ru	559K	63.68%
com	96K	10.98%
net	24K	2.79%
com.ua	23K	2.64%
org	19K	2.12%
info	18K	2.03%
by	16K	1.84%
ua	16K	1.79%
pφ	12K	1.39%
kz	10K	1.19%

Register labels

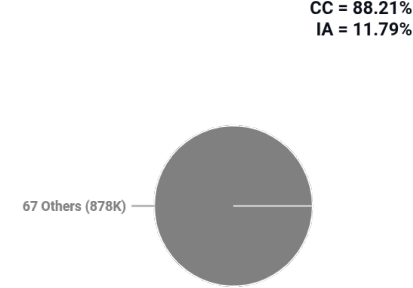


MT:3.2% | 28K Documents

Documents size (in segments) ⓘ

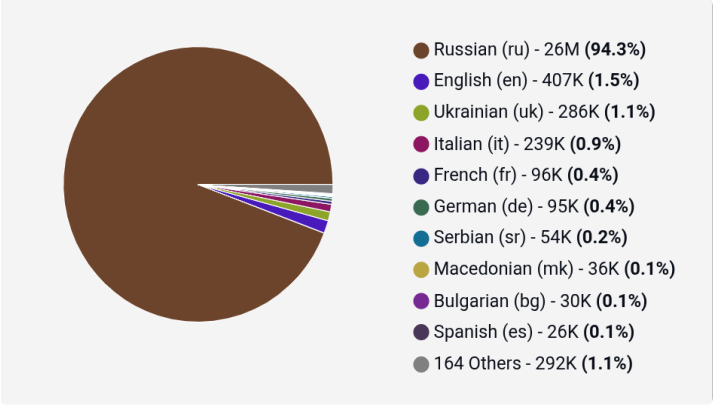


Document collections

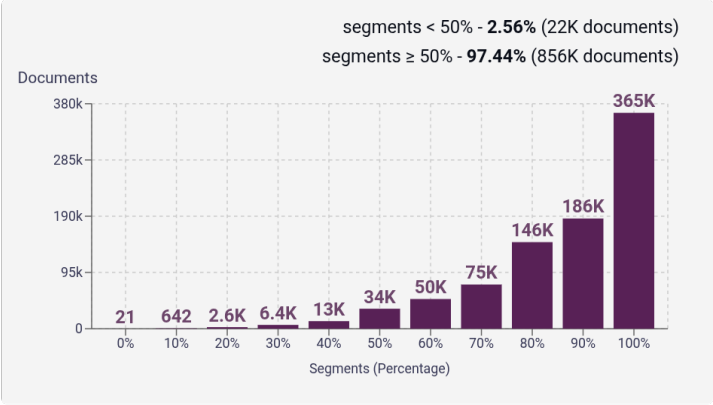


Language Distribution

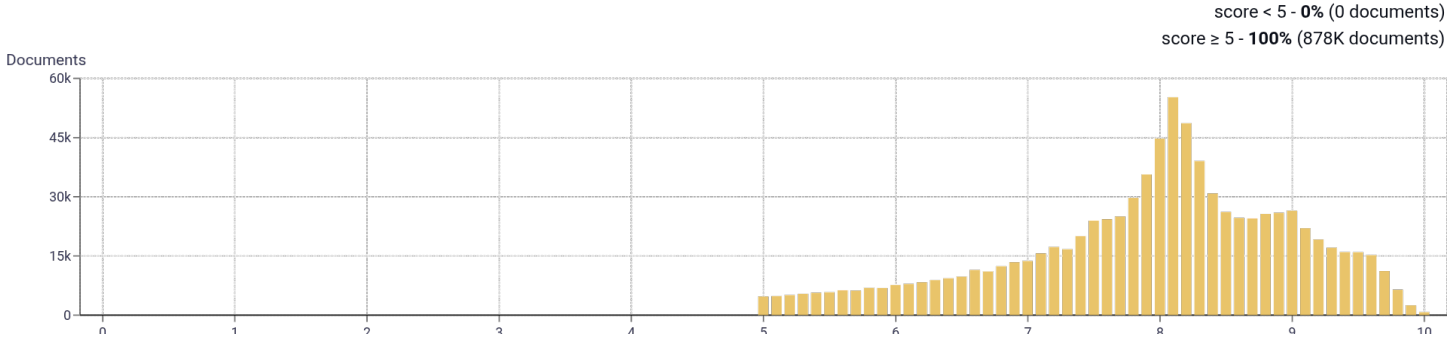
Number of segments in the Russian (ru) corpus



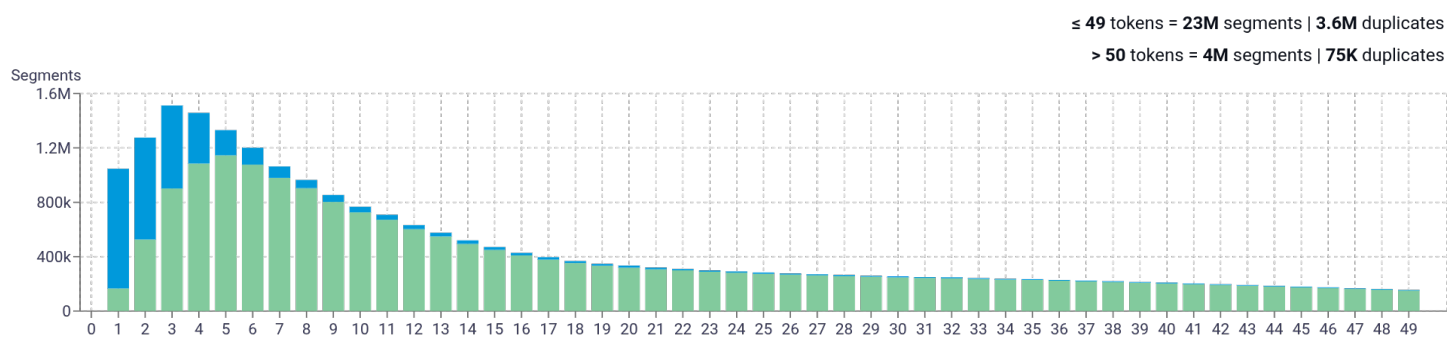
Percentage of segments in Russian (ru) inside documents



Distribution of documents by document score

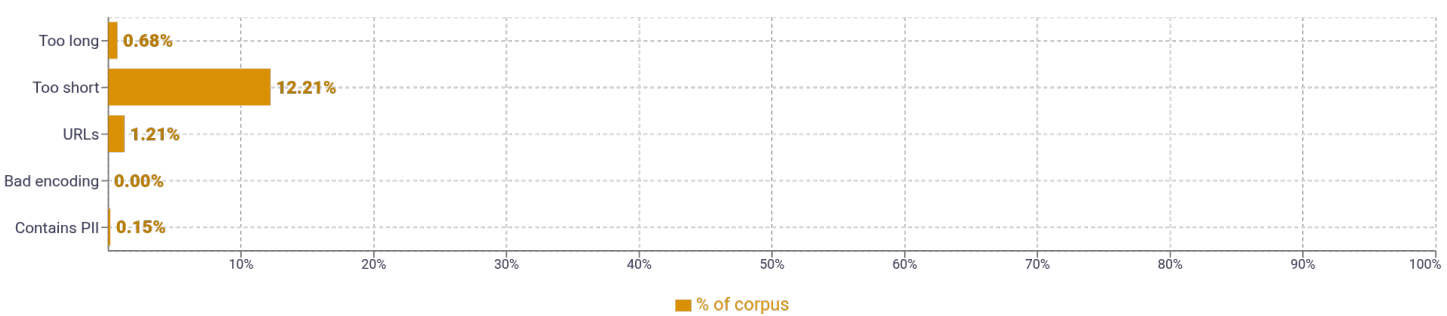


Segment length distribution by token



≤ 49 tokens = 23M segments | 3.6M duplicates  
> 50 tokens = 4M segments | 75K duplicates

Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	также   1,087,473    время   878,917    которые   876,279    очень   730,570    года   687,444	
2	русской федерации   116,777    таким образом   116,244    самом деле   60,537    первую очередь   53,613    настоящее время   53,239	
3	одним из самых   16,842    раза в день   14,298    связи с этим   14,225    друг с другом   13,271    друг от друга   12,344	
4	оставляет за собой право   5,814    президент россии владимир путин   5,411    бесплатно и без регистрации   4,623 является одним из самых   4,410    процессуального кодекса российской федерации   3,404	
5	арбитражного процессуального кодекса российской федерации   3,099    рассказывают о типовых способах решения   2,494 статьи рассказывают о типовых способах   2,490    наши статьи рассказывают о типовых   2,490 каждый случай носит уникальный характер   2,487	

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 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>, <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	ntp
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				