HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-bel_Cyrl	9/17/2025	Belarusian (be)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
2,999,651	55,967,243	41,254,779 (73.71 %)	1.8B	10,127,434,031	17.02 GB

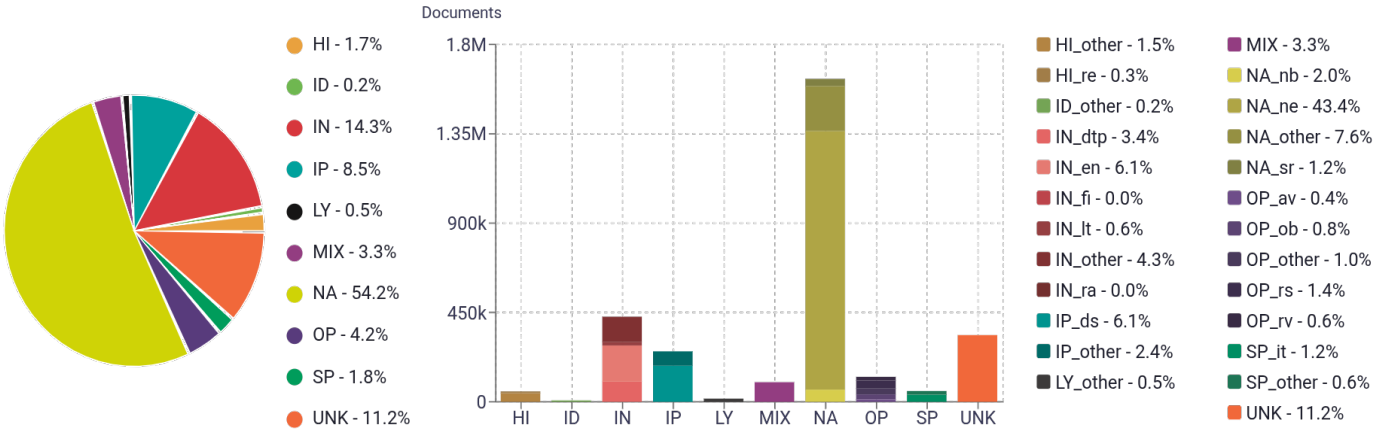
Top 10 domains

Domain	Docs	% of total
svaboda.org	239K	7.98%
wikipedia.org	167K	5.58%
cloudfront.net	111K	3.72%
zviazda.by	77K	2.58%
racyja.com	74K	2.45%
belsat.eu	66K	2.21%
sputnik.by	66K	2.21%
euroradio.fm	63K	2.11%
spring96.org	50K	1.67%
nn.by	43K	1.44%

Top 10 TLDs

Domain	Docs	% of total
by	1M	33.91%
org	654K	21.79%
com	496K	16.55%
net	173K	5.76%
ru	133K	4.44%
eu	98K	3.25%
info	86K	2.88%
fm	72K	2.41%
gov.by	50K	1.67%
life	20K	0.68%

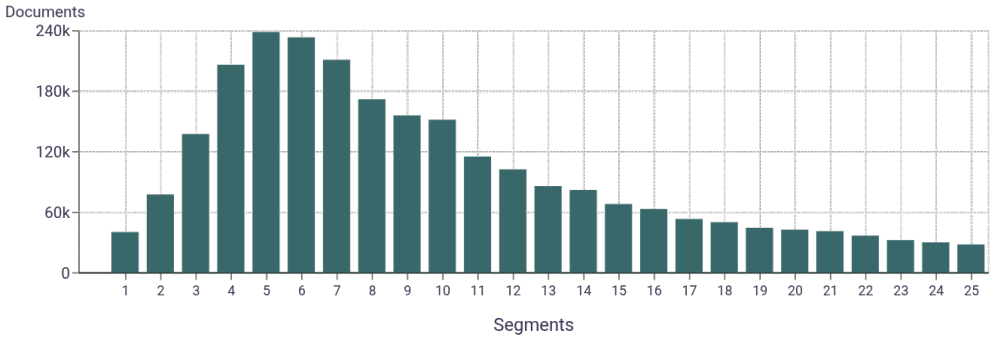
Register labels



MT:6.8% | 204K Documents

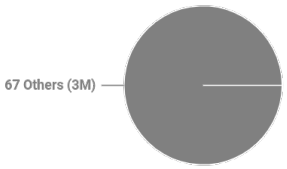
Documents size (in segments)

≤ 25 segments **83.47%** (2.5M documents)
> 25 segments **16.53%** (496K documents)



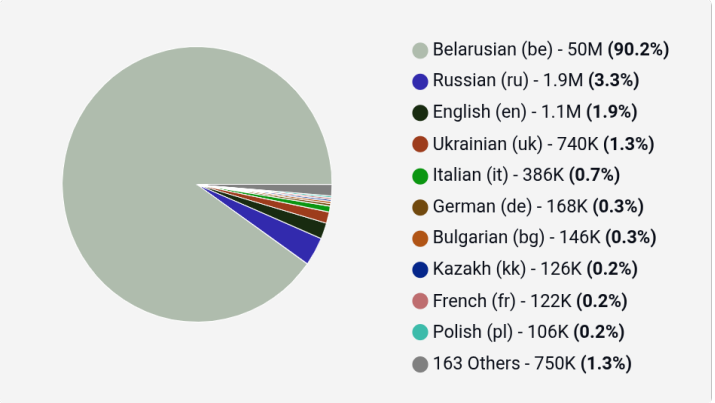
Document collections

CC = 89.67%
IA = 10.33%

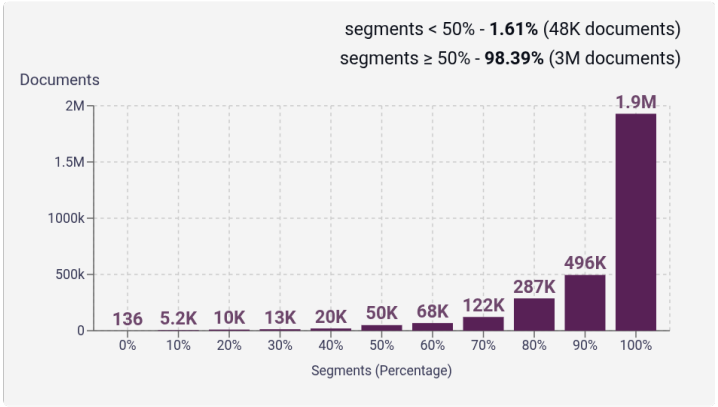


Language Distribution

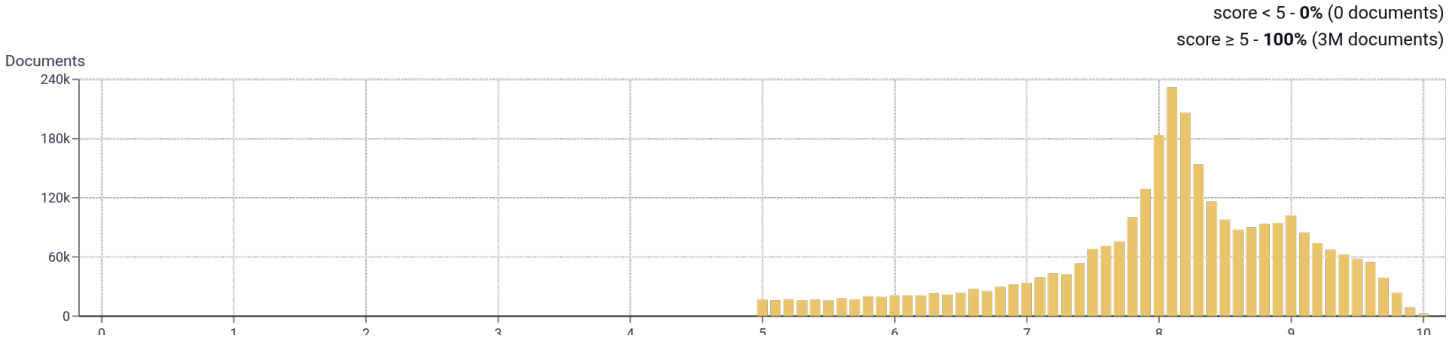
Number of segments in the Belarusian (be) corpus



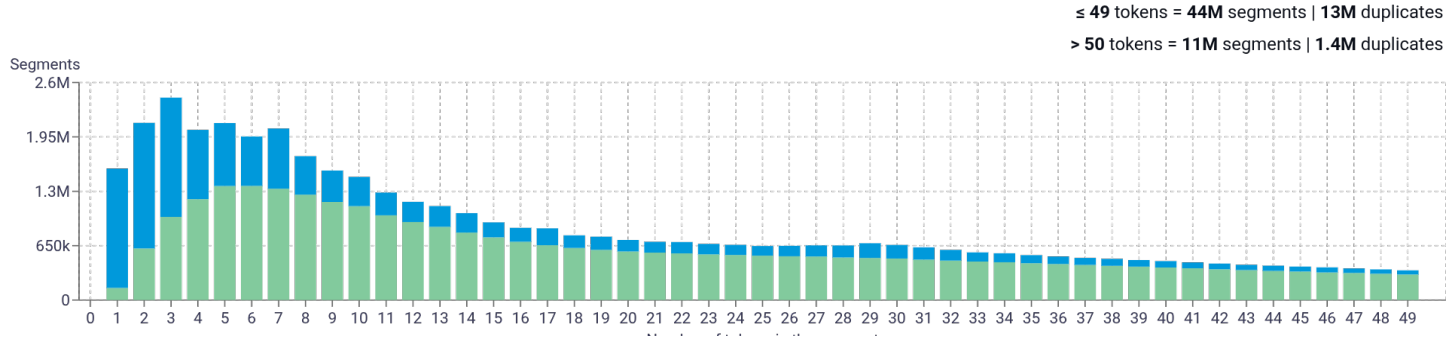
Percentage of segments in Belarusian (be) inside documents



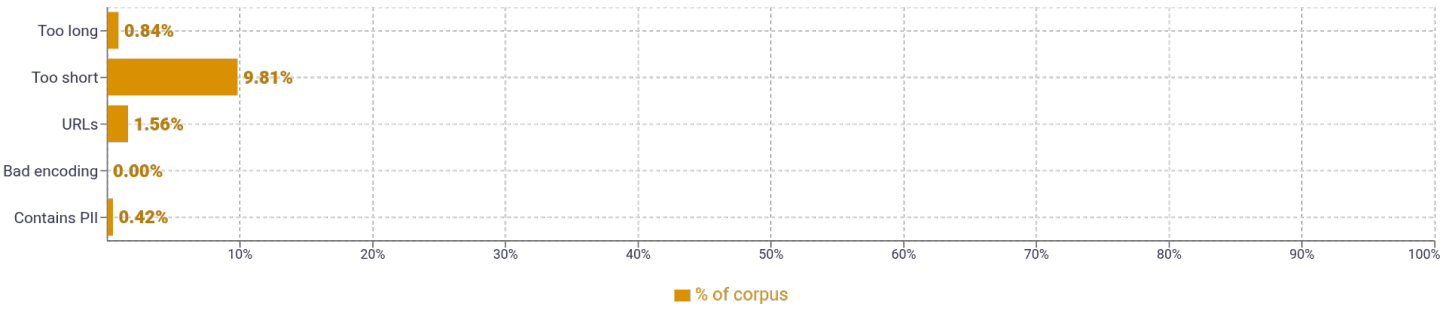
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS				
1	гэта | 6,916,281	але | 4,620,042	яго | 4,439,646	якія | 4,231,146	ён | 4,128,840
2	можа быць | 447,056	рэспублікі беларусь | 433,358	тым ліку | 430,841	такім чынам | 405,498	пра тое | 323,939
3	тым не менш | 103,699	перш за ўсё | 93,802	нягледзячы на тое | 79,290	вялікай айчыннай вайны | 72,957	хутчэй за ўсё | 70,559
4	б в г д | 42,523	г д е ё | 41,306	д е ё ж | 41,064	к л м н | 40,594	л м н о | 40,571
5	б в г д е | 41,768	г д е ё ж | 41,064	к л м н о | 40,570	л м н о п | 40,543	м н о п р | 40,531

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				