HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
bel_Cyrl.tsv	9/16/2024	Belarusian (be)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
2,319,619	48,844,196	23,307,246 (47.72 %)	1.5B	8,493,165,091	14.23 GB

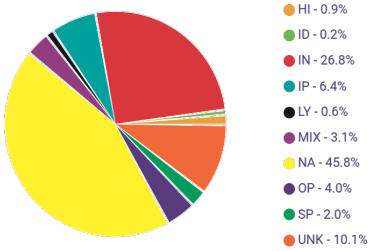
Top 10 domains

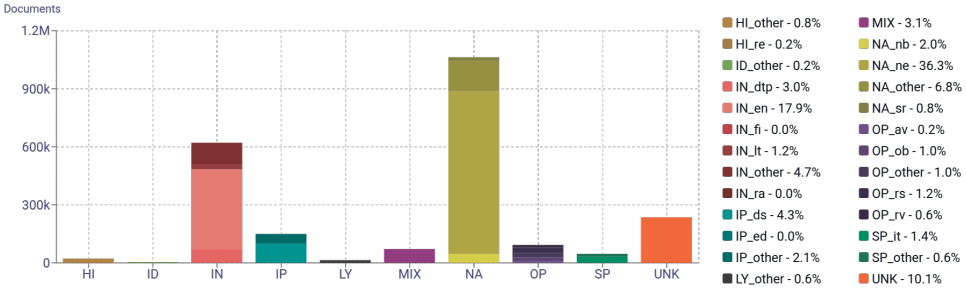
Domain	Docs	% of total
wikipedia.org	407K	17.54%
svaboda.org	219K	9.43%
spring96.org	87K	3.73%
racyja.com	62K	2.66%
nashaniva.com	62K	2.66%
belsat.eu	40K	1.73%
novychas.by	34K	1.46%
zviazda.by	27K	1.15%
skarnik.by	24K	1.02%
nn.by	23K	1.01%

Top 10 TLDs

Domain	Docs	% of total
org	858K	36.98%
by	632K	27.24%
com	337K	14.55%
ru	126K	5.44%
net	72K	3.09%
eu	63K	2.74%
info	55K	2.39%
gov.by	32K	1.38%
fm	25K	1.06%
online	16K	0.71%

Register labels

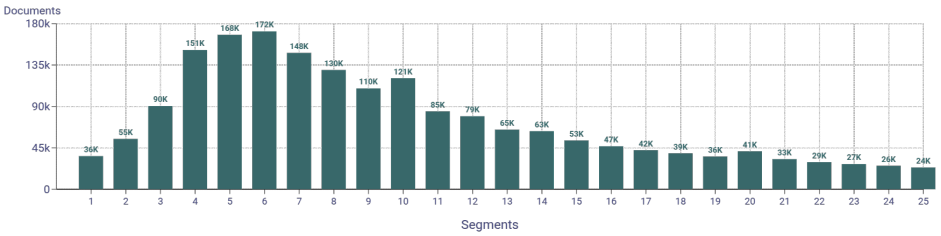




MT:5.7% | 133K Documents

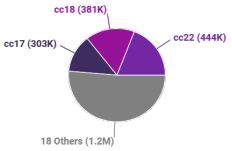
Documents size (in segments)

<= 25 segments **80.56%** (1.9M documents)
> 25 segments **19.44%** (451K documents)



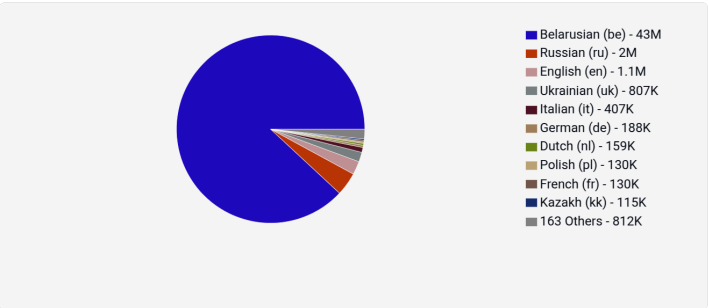
Documents by collection

CC = 63.83%
IA = 36.17%

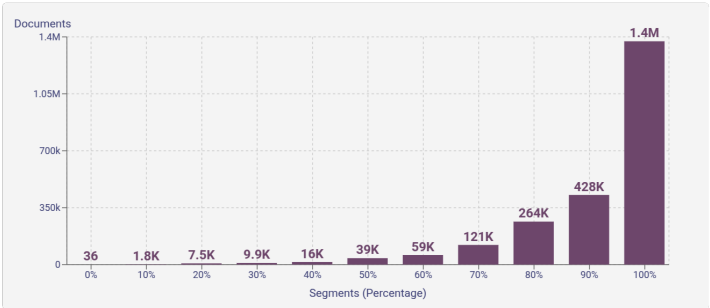


Language Distribution

Number of segments in the Belarusian (be) corpus

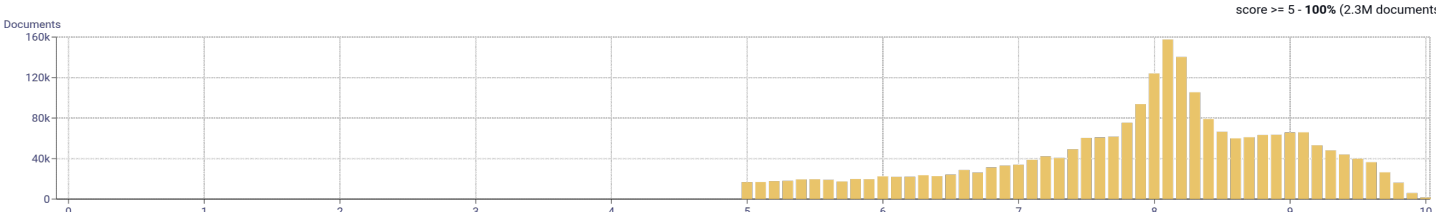


Percentage of segments in Belarusian (be) inside documents



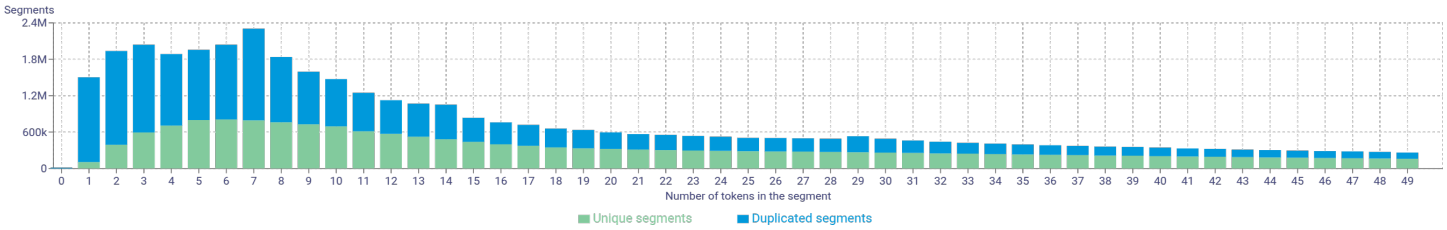
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (2.3M documents)

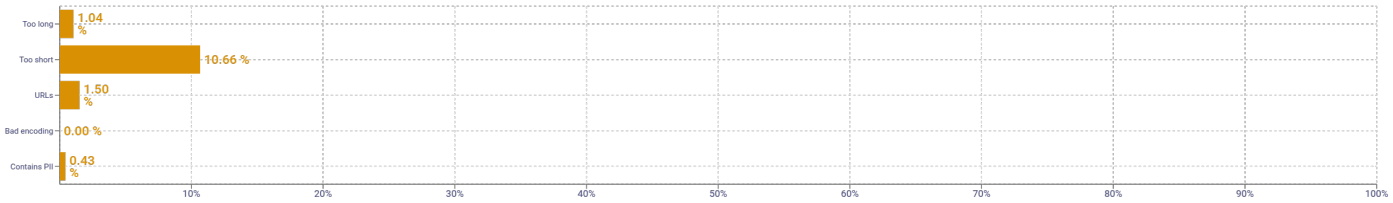


Segment length distribution by token

≤ 49 tokens = **18M** segments | **22M** duplicates
> 50 tokens = **9.8M** segments | **4M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	ў | 22444103 · да | 7444290 · як | 5589025 · гэта | 4896701 · ад | 4025399
2	правіць зыходнік | 710669 · рэдагаваць крыніцу | 396147 · ў беларусі | 388317 · рэспублікі беларусь | 379246 · тым ліку | 324698
3	лістапад кастрычнік верасень | 126015 · сьнежань лістапад кастрычнік | 125959 · сакавік люты студзень | 123100 · красавік сакавік люты | 123093 · травень красавік сакавік | 123060
4	сьнежань лістапад кастрычнік верасень | 125949 · красавік сакавік люты студзень | 123079 · травень красавік сакавік люты | 123047 · чэрвень травень красавік сакавік | 122839 · ліпень чэрвень травень красавік | 122628
5	травень красавік сакавік люты студзень | 123033 · чэрвень травень красавік сакавік люты | 122826 · ліпень чэрвень травень красавік сакавік | 122601 · жнівень ліпень чэрвень травень красавік | 119774 · верасень жнівень ліпень чэрвень травень | 119745

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av