HPLT Analytics report

⊚ HPLT Analytics

General overview

Corpus	Date	Language
hplt-v3-bul_Cyrl	9/18/2025	Bulgarian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
42,967,942	978,252,958	539,310,730 (55.13 %)	27B	144,827,849,190	240.67 GB

Top 10 domains

Domain	Docs	% of total
blogspot.com	391K	0.91%
offnews.bg	315K	0.73%
capital.bg	307K	0.71%
blog.bg	297K	0.69%
dnevnik.bg	286K	0.67%
dir.bg	277K	0.65%
blitz.bg	260K	0.61%
mediapool.bg	249K	0.58%
fakti.bg	228K	0.53%
sportal.bg	224K	0.52%

Top 10 TLDs

Domain	Docs	% of total
bg	20M	46.18%
com	15M	33.77%
net	2.1M	4.91%
org	2.1M	4.78%
eu	1.3M	2.94%
info	1.1M	2.55%
co.uk	186K	0.43%
news	185K	0.43%
ru	125K	0.29%
pt	115K	0.27%

Register labels

Documents

- HI - 3.3%
- ID - 2.4%
- IN - 8.4%
- IP - 23.9%
- LY - 0.1%
- MIX - 3.4%
- NA - 43.5%
- OP - 5.4%
- SP - 0.8%
- UNK - 8.8%

- HI_other - 1.7%
- HI_re - 1.6%
- ID_other - 2.4%
- IN_dtp - 3.3%
- IN_en - 0.7%
- IN_fi - 0.0%
- IN_lt - 0.6%
- IN_other - 3.7%
- IN_ra - 0.0%
- IP_ds - 21.8%
- IP_ed - 0.0%
- IP_other - 2.1%
- LY_other - 0.1%

- MIX - 3.4%
- NA_nb - 2.3%
- NA_ne - 32.8%
- NA_other - 3.3%
- NA_sr - 5.1%
- OP_av - 1.4%
- OP_ob - 1.2%
- OP_other - 1.4%
- OP_rs - 0.6%
- OP_rv - 0.8%
- SP_it - 0.6%
- SP_other - 0.2%
- UNK - 8.8%
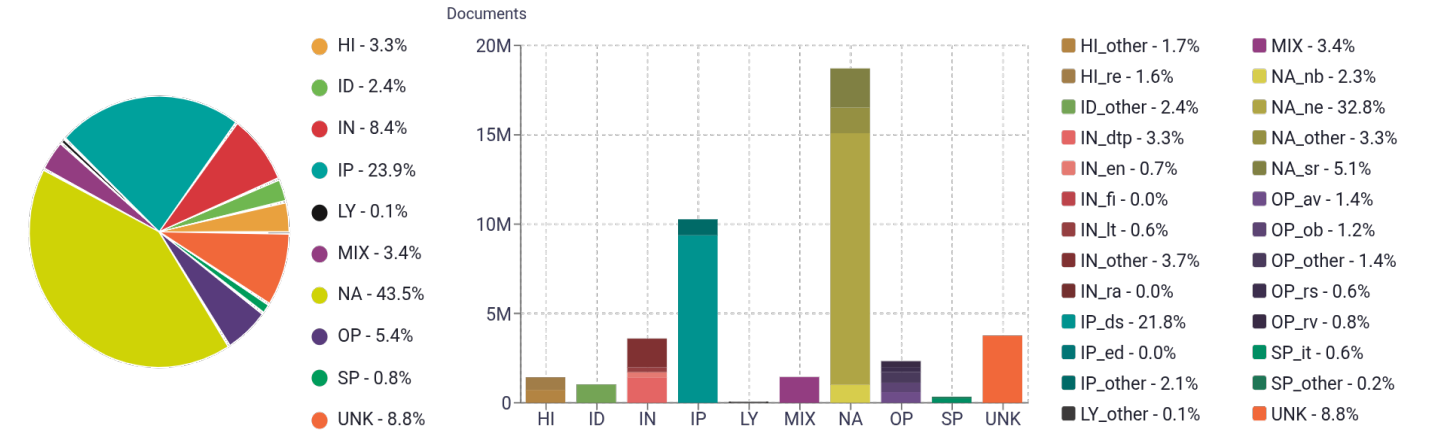
🤖 **MT**:5.7% | 2.4M Documents

Documents size (in segments) ⓘ

≤ 25 segments **79.55%** (34M documents)
> 25 segments **20.45%** (8.8M documents)

Documents

Segments

Document collections

CC = 87.01%
IA = 12.99%

67 Others (43M)
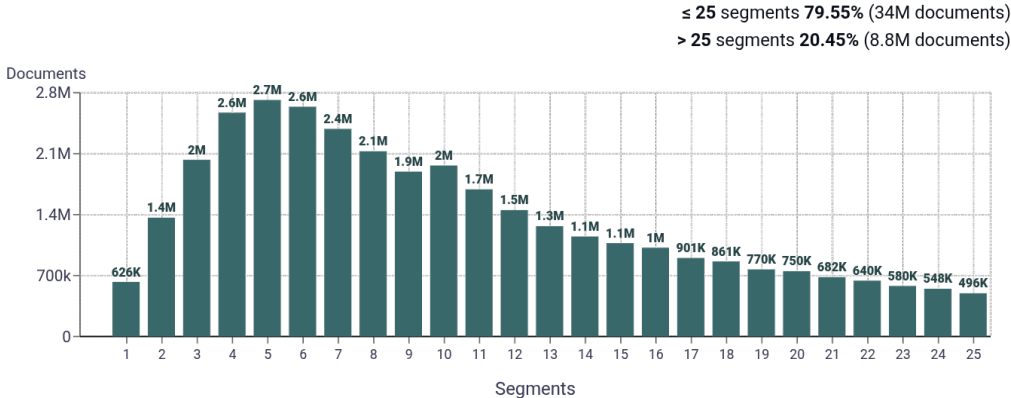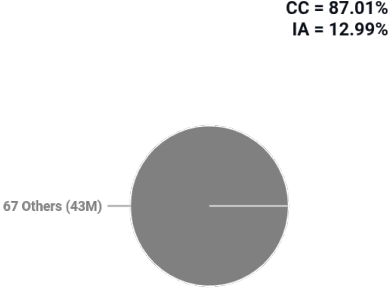
Language Distribution

Number of segments in the Bulgarian corpus

- Bulgarian - 861M **(88.0%)**
- English - 27M **(2.7%)**
- Russian - 24M **(2.5%)**
- Italian - 15M **(1.5%)**
- Ukrainian - 8.1M **(0.8%)**
- Serbian - 8M **(0.8%)**
- Macedonian - 5.8M **(0.6%)**
- German - 5.2M **(0.5%)**
- French - 4.9M **(0.5%)**
- Spanish - 1.8M **(0.2%)**
- 165 Others - 19M **(1.9%)**
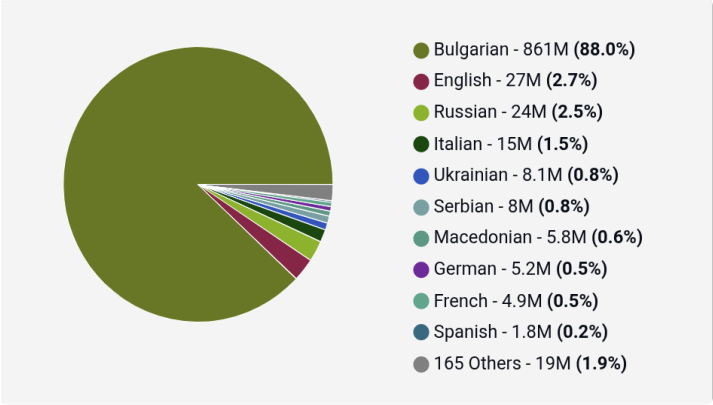
Percentage of segments in Bulgarian inside documents
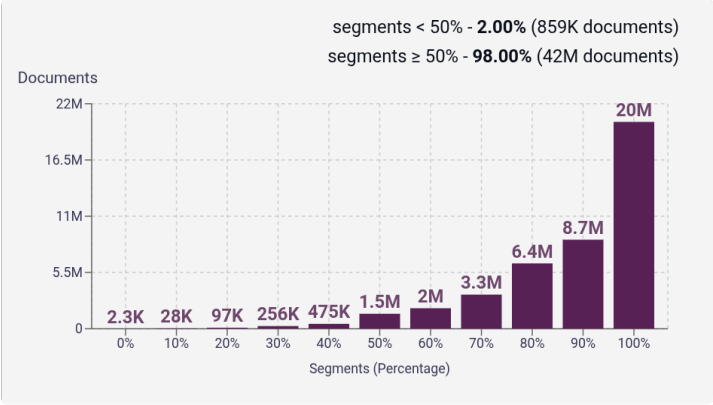
segments < 50% - **2.00%** (859K documents)
segments ≥ 50% - **98.00%** (42M documents)

Documents

Segments (Percentage)	Value
0%	2.3K
10%	28K
20%	97K
30%	256K
40%	475K
50%	1.5M
60%	2M
70%	3.3M
80%	6.4M
90%	8.7M
100%	20M

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (43M documents)

Documents

Segment length distribution by token

≤ 49 tokens = **820M** segments | **405M** duplicates

> 50 tokens = **158M** segments | **35M** duplicates

Segments

Segment noise distribution

- Too long — 0.65%
- Too short — 13.83%
- URLs — 1.59%
- Bad encoding — 0.00%
- Contains PII — 0.50%

■ % of corpus
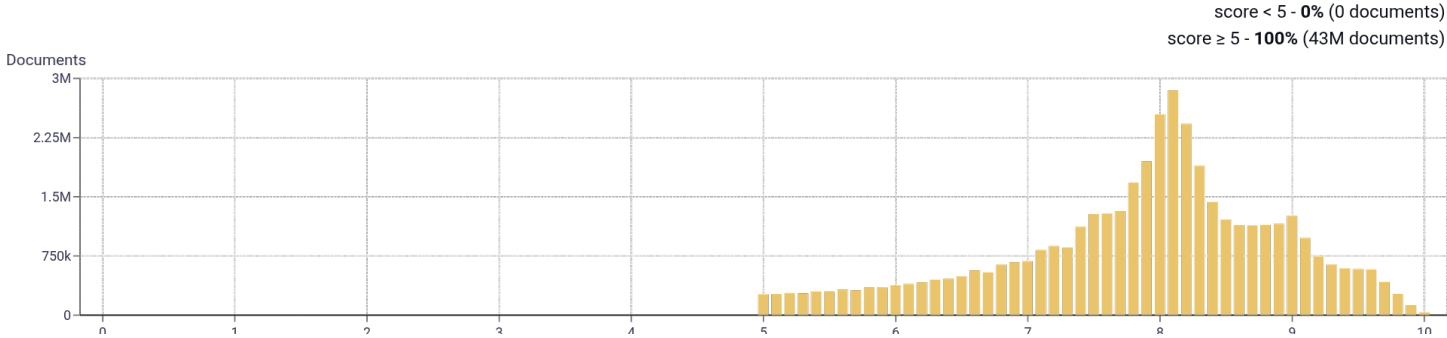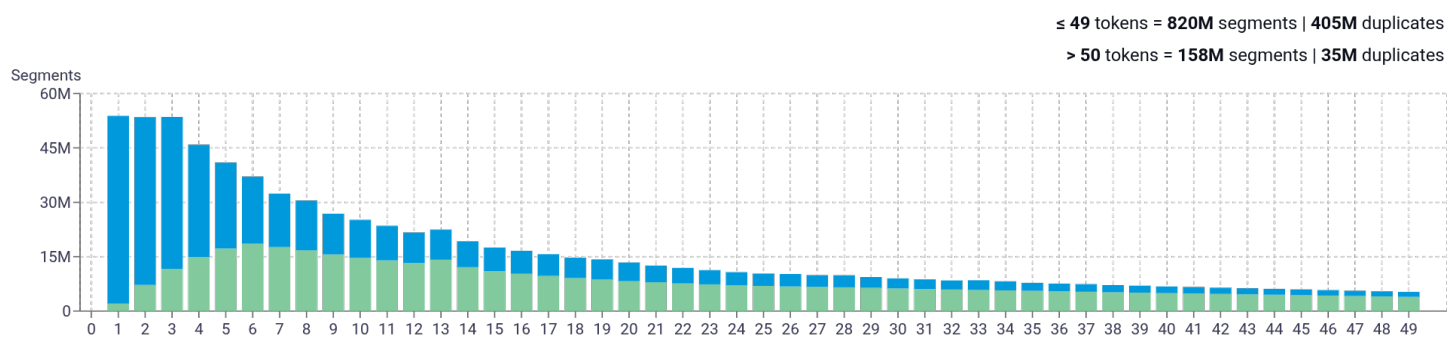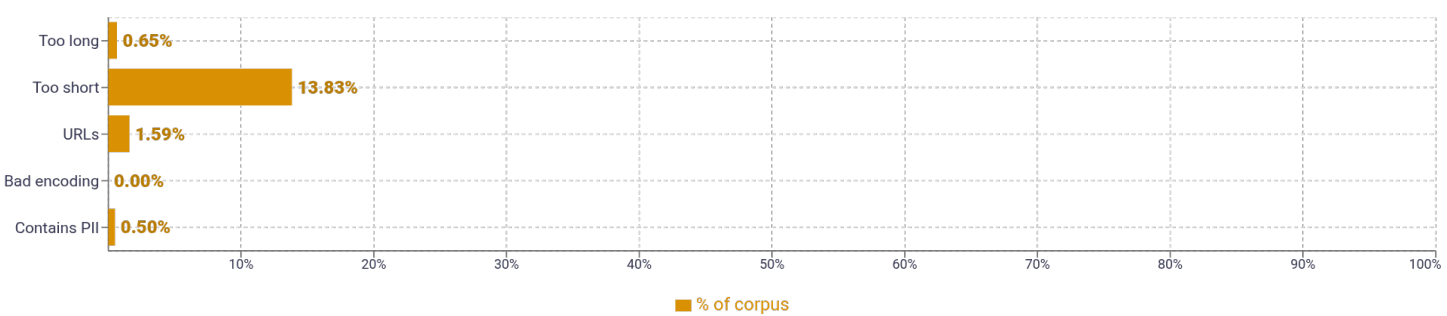
Frequent n-grams

SIZE	N-GRAMS	
1	г. | 38,753,143 българия | 22,182,363 част | 19,052,047 хора | 16,800,290 лв | 15,984,244	⧉
2	т. н. | 2,727,646 т. е. | 2,048,043 крайна сметка | 1,741,286 стара загора | 1,724,884 неутралнодо коментар | 1,704,238	⧉
3	що се отнася | 778,655 течение на времето | 487,531 загуба на тегло | 466,811 коментарът беше изтрит | 417,839 изтрит от модераторите | 406,146	⧉
4	етническа или верска основа | 389,987 верска основа или призиви | 276,760 адрес на конкретни лица | 267,420 регистрирайте се от тук | 261,405 обидни или нецензурни квалификации | 258,654	⧉
5	коментарът беше изтрит от модераторите | 405,817 призиви към насилие по адрес | 276,792 основа или призиви към насилие | 276,758 насилие по адрес на конкретни | 267,963 съдържаше обидни или нецензурни квалификации | 258,582	⧉

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				