HPLT Analytics report

◉ HPLT Analytics

General overview

Corpus	Date	Language
pol_Latn.jsonl.tsv	8/28/2025	Polish (pl)

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
175,410,669	4,459,632,710	1,456,680,416 (32.66 %)	67.34%	105B	627,308,372,672	615.5 GB

Top 10 domains

Domain	Docs	% of total
blogspot.com	3.7M	2.12%
onet.pl	1.8M	1.05%
wikipedia.org	1.5M	0.84%
wp.pl	1M	0.59%
gazeta.pl	992K	0.57%
interia.pl	923K	0.53%
sfd.pl	761K	0.43%
wordpress.com	722K	0.41%
docplayer.pl	620K	0.35%
frazeo.pl	440K	0.25%

Top 10 TLDs

Domain	Docs	% of total
pl	124M	70.75%
com	17M	9.54%
com.pl	6.2M	3.53%
org	4.2M	2.37%
eu	4M	2.26%
net	2.7M	1.56%
info	2.2M	1.27%
org.pl	1.6M	0.91%
edu.pl	1.5M	0.85%
net.pl	1.2M	0.68%

Register labels

- HI - 3.1%
- ID - 5.5%
- IN - 13.1%
- IP - 35.4%
- LY - 0.1%
- MIX - 5.5%
- NA - 25.1%
- OP - 6.4%
- SP - 0.5%
- UNK - 5.2%

- HI_other - 1.5%
- HI_re - 1.6%
- ID_other - 5.5%
- IN_dtp - 4.3%
- IN_en - 1.3%
- IN_fi - 0.0%
- IN_lt - 1.3%
- IN_other - 6.2%
- IN_ra - 0.1%
- IP_ds - 32.0%
- IP_ed - 0.0%
- IP_other - 3.4%
- LY_other - 0.1%
- MIX - 5.5%
- NA_nb - 6.6%
- NA_ne - 12.9%
- NA_other - 3.3%
- NA_sr - 2.3%
- OP_av - 1.3%
- OP_ob - 1.0%
- OP_other - 1.1%
- OP_rs - 0.8%
- OP_rv - 2.2%
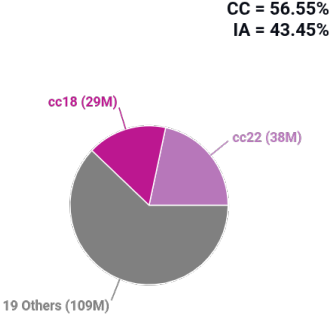- SP_it - 0.4%
- SP_other - 0.1%
- UNK - 5.2%
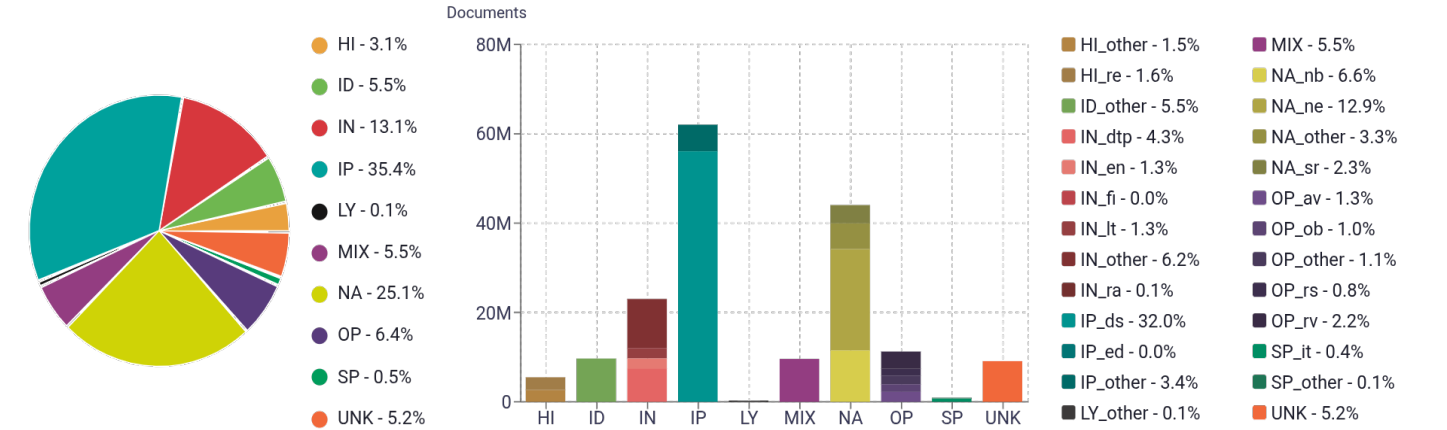
♟ **MT**:1.8% | 3.2M Documents

Documents size (in segments) ⓘ
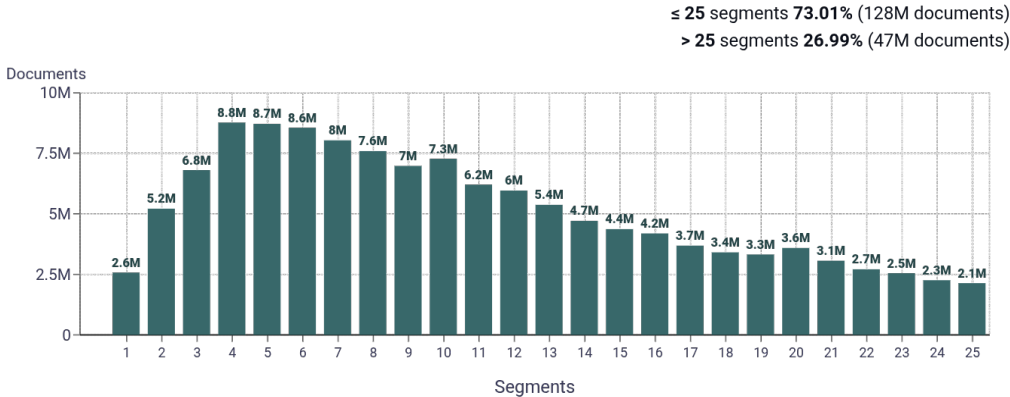
≤ 25 segments **73.01%** (128M documents)
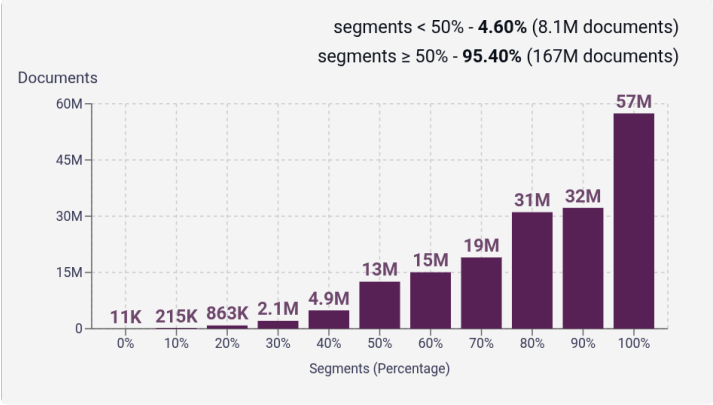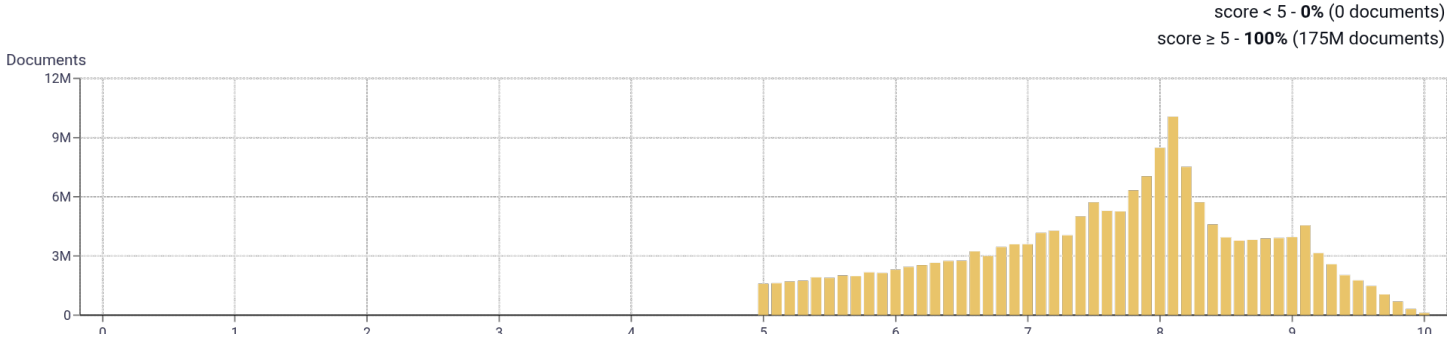**> 25** segments **26.99%** (47M documents)

Document collections

CC = 56.55%
IA = 43.45%

cc18 (29M)
cc22 (38M)
19 Others (109M)

Language Distribution

Number of segments in the Polish (pl) corpus

- Polish (pl) - 3.8B **(85.3%)**
- Italian (it) - 199M **(4.5%)**
- English (en) - 185M **(4.1%)**
- French (fr) - 37M **(0.8%)**
- German (de) - 37M **(0.8%)**
- Spanish (es) - 25M **(0.6%)**
- Dutch (nl) - 20M **(0.4%)**
- Czech (cs) - 13M **(0.3%)**
- Portuguese (pt) - 10M **(0.2%)**
- Swedish (sv) - 10M **(0.2%)**
- 165 Others - 119M **(2.7%)**
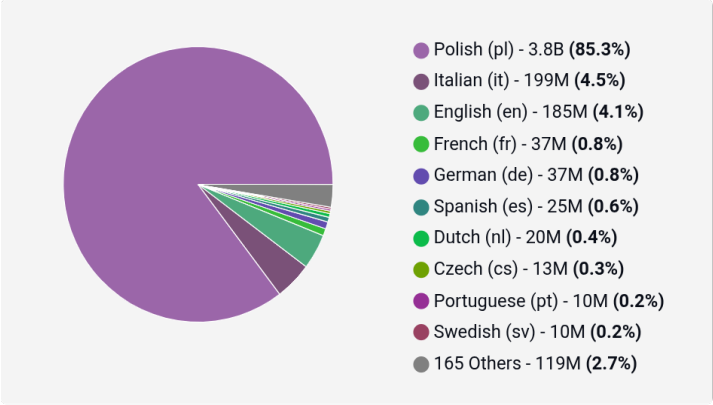
Percentage of segments in Polish (pl) inside documents

segments < 50% - **4.60%** (8.1M documents)
segments ≥ 50% - **95.40%** (167M documents)

Documents

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
11K	215K	863K	2.1M	4.9M	13M	15M	19M	31M	32M	57M

Segments (Percentage)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (175M documents)

Documents

Segment length distribution by token

≤ **49** tokens = **3.9B** segments | **2.7B** duplicates
> **50** tokens = **592M** segments | **300M** duplicates

Segments

Segment noise distribution

Too long	0.97%
Too short	18.30%
URLs	2.81%
Bad encoding	0.00%
Contains PII	0.56%

■ % of corpus
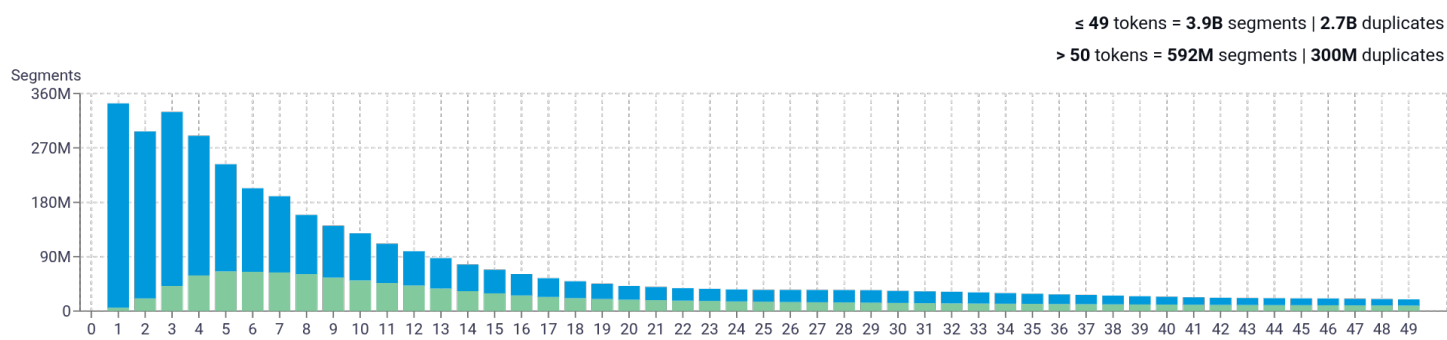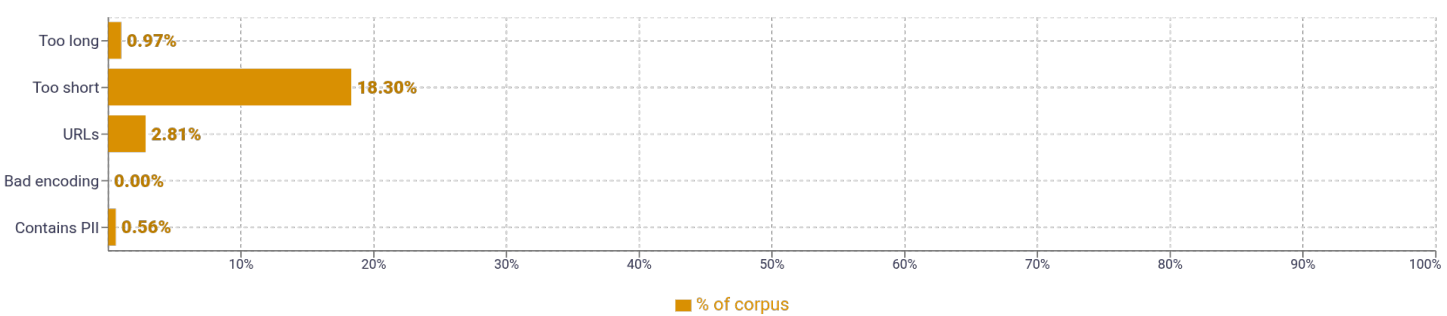
Frequent n-grams

SIZE	N-GRAMS
1	przez | 255,225,646 za | 252,645,284 ale | 214,932,192 tym | 213,833,068 które | 166,132,979
2	szczegóły wpisu | 48,895,369 przede wszystkim | 25,144,439 data dodania | 21,619,040 za pomocą | 15,105,365 tym samym | 11,920,405
3	link nie działa | 3,016,296 jana pawła ii | 2,721,197 jesteśmy w stanie | 2,597,997 dzieci i młodzieży | 2,425,678 informacje na temat | 2,368,517
4	bądź widoczny w katalogu | 2,895,230 kliknij i zobacz pełną | 1,902,764 bądź pierwszy i skomentuj | 1,290,744 których mowa w art. | 1,089,092 dochodowym od osób fizycznych | 928,561
5	kliknij i zobacz pełną galerię | 1,902,557 bądź pierwszy i skomentuj wpis | 1,287,134 gratis bez minimalnej wartości zamówienia | 613,566 konieczne jest włączenie w przeglądarce | 506,401 darmowa wysyłka płatność za pobraniem | 497,849

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				