

General overview

Corpus	Date	Language
hplt-v3-pol_Latn	9/18/2025	Polish (pl)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
255,893,058	5,641,340,884	2,625,053,779 (46.53 %)	147B	878,329,025,420	862.24 GB

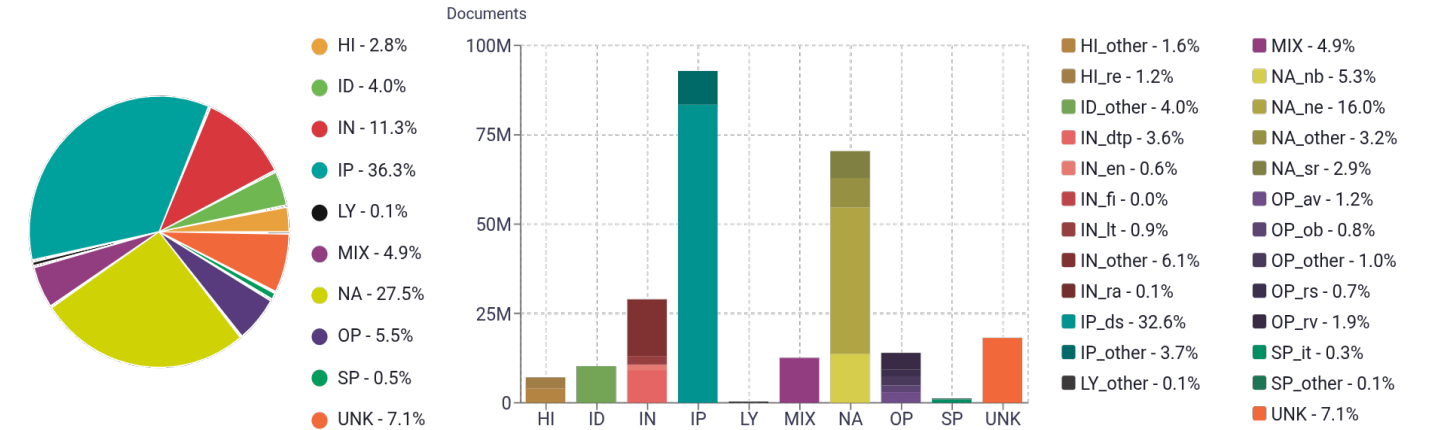
Top 10 domains

Domain	Docs	% of total
blogspot.com	4.5M	1.76%
onet.pl	2M	0.80%
wp.pl	1.8M	0.72%
interia.pl	1.5M	0.60%
gazeta.pl	1.1M	0.44%
naszemiasto.pl	1.1M	0.42%
wordpress.com	852K	0.33%
docplayer.pl	796K	0.31%
rp.pl	734K	0.29%
sfd.pl	686K	0.27%

Top 10 TLDs

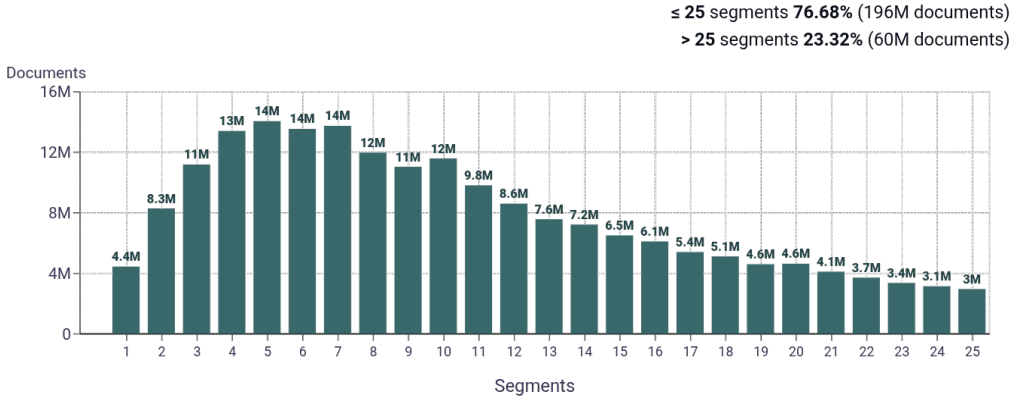
Domain	Docs	% of total
pl	181M	70.65%
com	25M	9.67%
com.pl	9.8M	3.82%
eu	6.1M	2.37%
org	4.1M	1.60%
net	3.7M	1.46%
info	3.1M	1.19%
edu.pl	2.3M	0.88%
org.pl	2.2M	0.85%
net.pl	1.6M	0.64%

Register labels

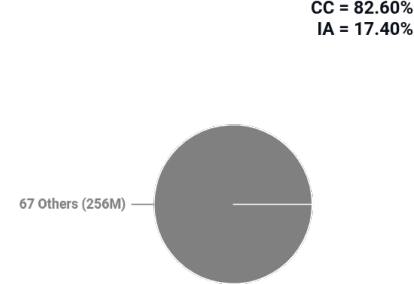


MT:3.4% | 8.8M Documents

Documents size (in segments) ⓘ

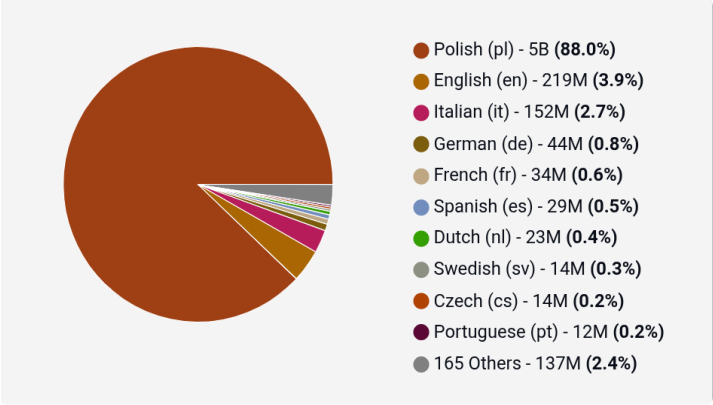


Document collections

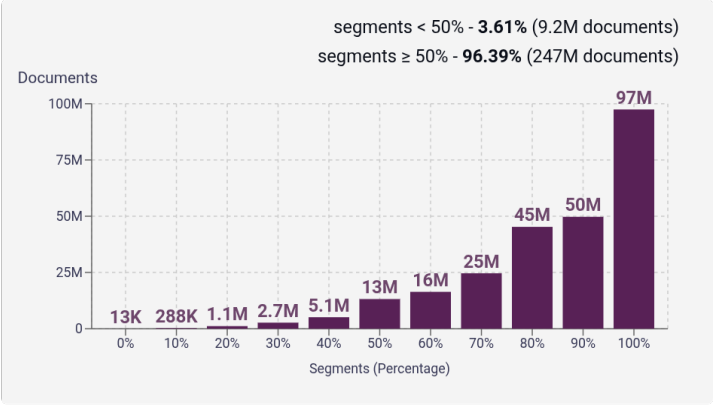


Language Distribution

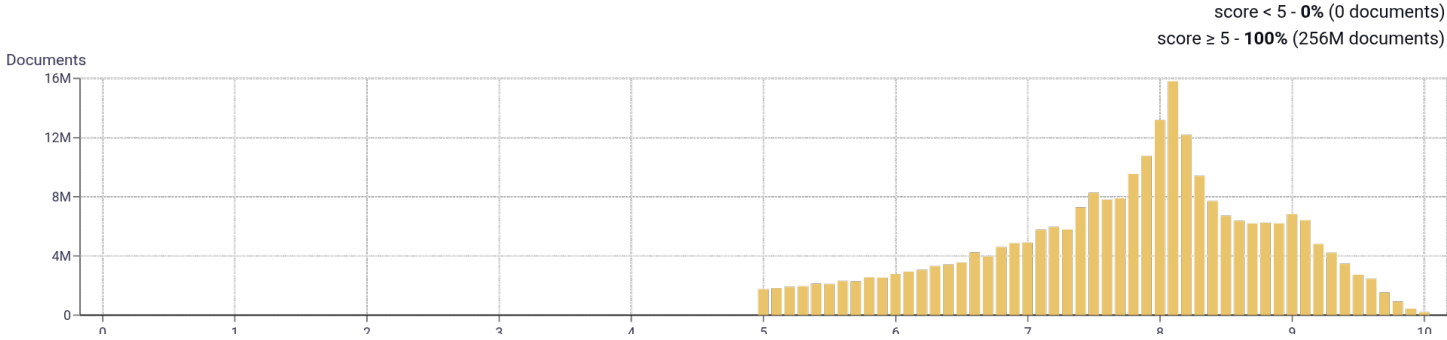
Number of segments in the Polish (pl) corpus



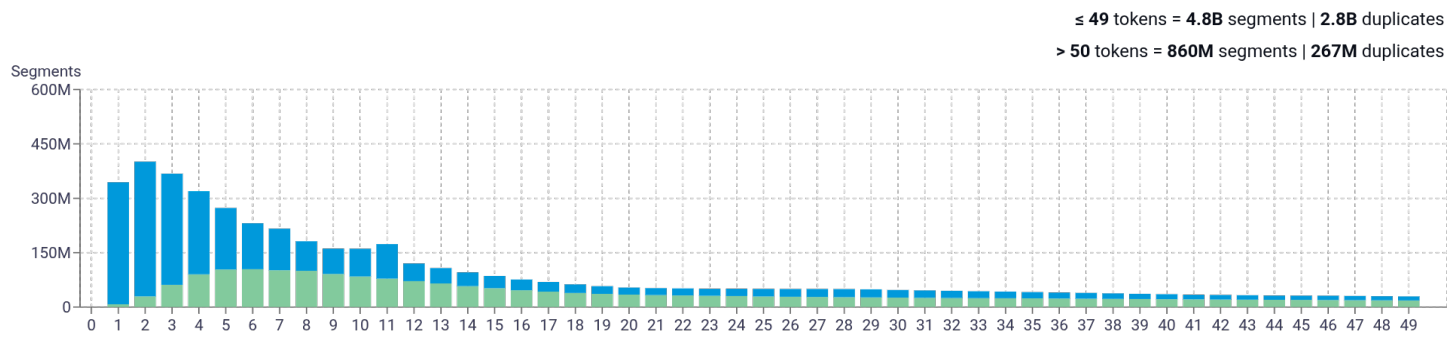
Percentage of segments in Polish (pl) inside documents



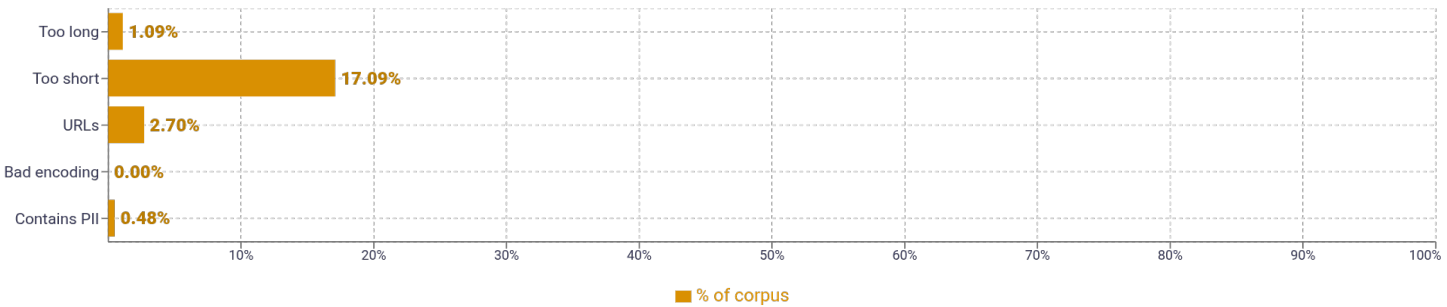
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	za 348,441,099 przez 340,226,929 ale 307,857,162 tym 307,371,396 które 278,434,341	
2	data dodania 65,794,660 szczegóły wpisu 61,856,530 bardziej szczegółowo 39,332,959 przede wszystkim 36,996,004 za pomocą 22,632,283	
3	sp. z o.o. 9,839,879 związku z tym 6,698,618 ustawy z dnia 5,578,705 r. w sprawie 5,296,507 zobacz podobne wpisy 5,011,846	
4	wpisy w tej kategorii 4,990,561 podobne wpisy w tej 4,975,761 bądź widoczny w katalogu 4,040,287 strony związane z hasłem 1,748,504 zapoznania się z naszą 1,698,421	
5	podobne wpisy w tej kategorii 4,975,269 zobacz podobne wpisy w tej 4,970,990 zapoznania się z naszą ofertą 1,164,984 podatku od towarów i usług 1,156,665 gra za darmo bez rejestracji 1,093,004	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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>, , , 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				