

General overview

Corpus	Date	Language
hplt-v3-szl_Latn	9/18/2025	Silesian (szl)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
48,394	639,717	440,777 (68.90 %)	23M	126,002,821	128 MB

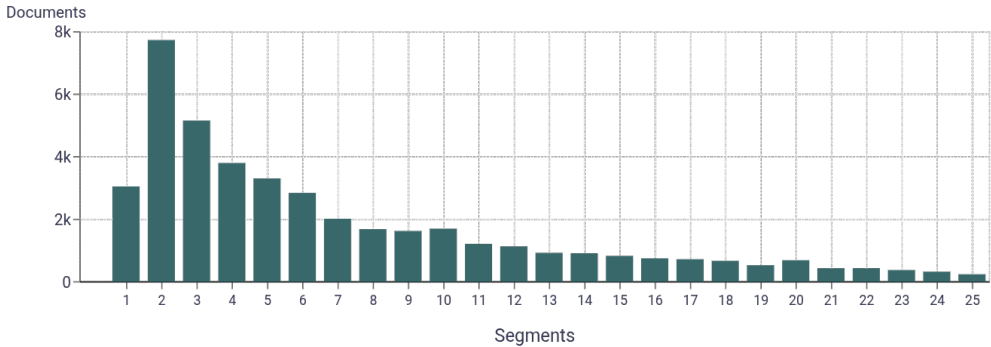
Top 10 domains

Domain	Docs	% of total
serbske-nowiny.de	16K	32.76%
wikipedia.org	4.5K	9.24%
slonskogodka.com	2.8K	5.81%
wachtyrz.eu	1.3K	2.69%
rozhlad.de	894	1.85%
nowycasnik.de	806	1.67%
chopwkuchni.pl	779	1.61%
mdr.de	693	1.43%
wordpress.com	530	1.10%
uj.edu.pl	434	0.90%

Top 10 TLDs

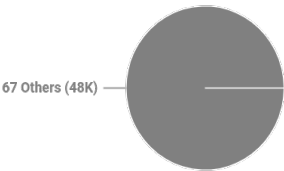
Domain	Docs	% of total
de	22K	45.21%
pl	9.2K	19.00%
org	5.3K	11.04%
com	5.2K	10.72%
eu	2.4K	4.91%
edu.pl	854	1.76%
com.pl	545	1.13%
cz	452	0.93%
net.pl	390	0.81%
info	365	0.75%

Documents size (in segments) ⓘ



≤ 25 segments **89.22%** (43K documents)
> 25 segments **10.78%** (5.2K documents)

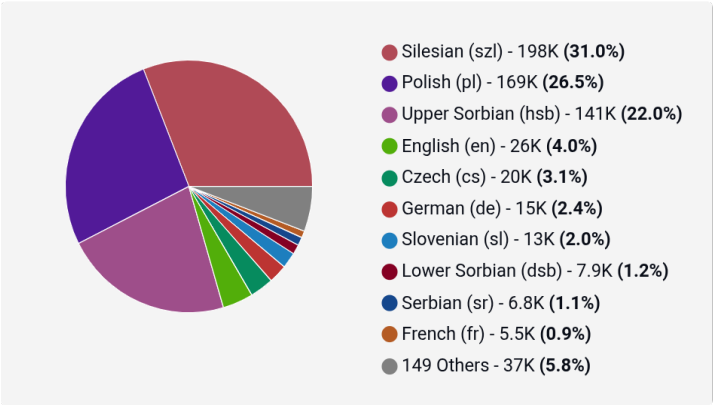
Document collections



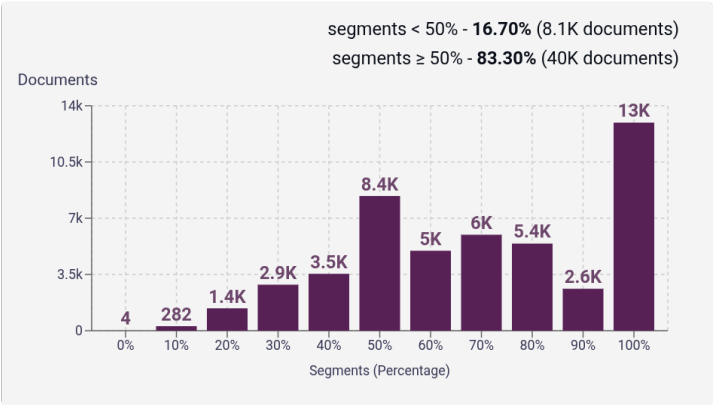
CC = 88.64%
IA = 11.36%

Language Distribution

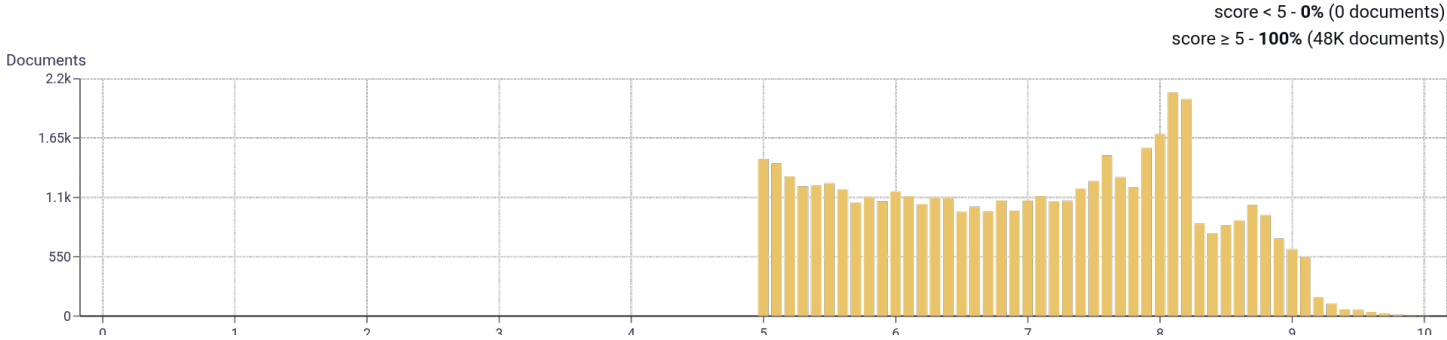
Number of segments in the Silesian (szl) corpus



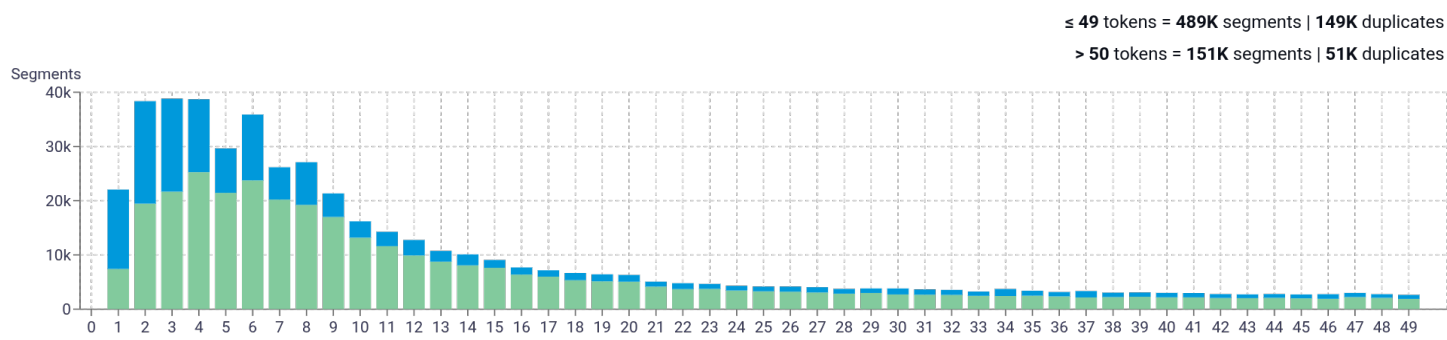
Percentage of segments in Silesian (szl) inside documents



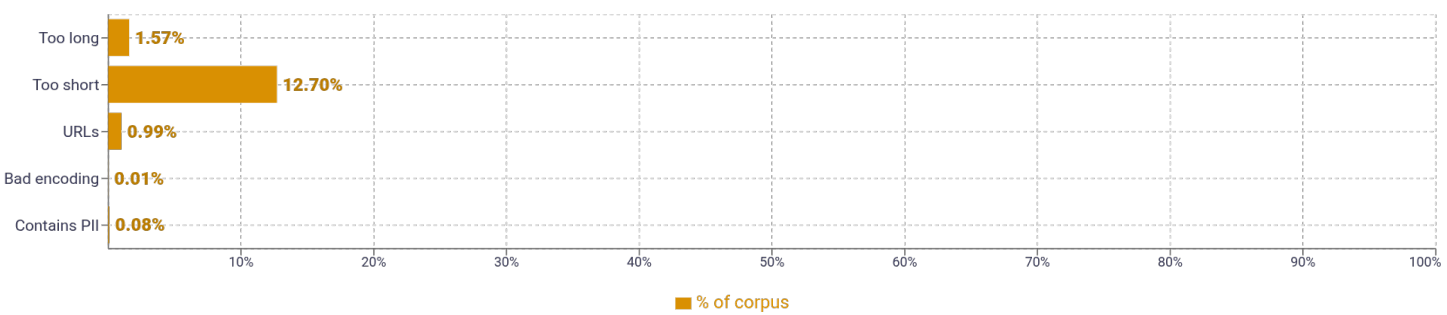
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	so 130,055 tak 53,836 tež 53,021 zo 50,658 su 45,948	
2	kaž tež 9,474 mjez druhim 5,902 k tomu 4,018 dr hab 3,999 wjace hač 3,620	
3	ginekolog i položnik 12,503 žórlowy tekst wobdžělać 2,474 založby za serbski 1,119 hač do kónca 1,026 hač do lěta 951	
4	založby za serbski lud 1,115 dzyń dzyń dzyń dzyń 730 založba za serbski lud 636 cordula ratajczakowa je so 454 wudaću nańdžeće mjez druhim 426	
5	dzyń dzyń dzyń dzyń dzyń 726 wudaću nańdžeće mjez druhim tole 426 aktualnym wudaću nańdžeće mjez druhim 426 serbske wotpowědniki za němske słowa 371 gdo sie za swój jynzyk 329	

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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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	ntp
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				