HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-szl_Latn	10/27/2025	Silesian

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
30,839	485,408	357,655 (73.68 %)	26.32%	15M	83,554,565	84.25 MB

Top 10 domains

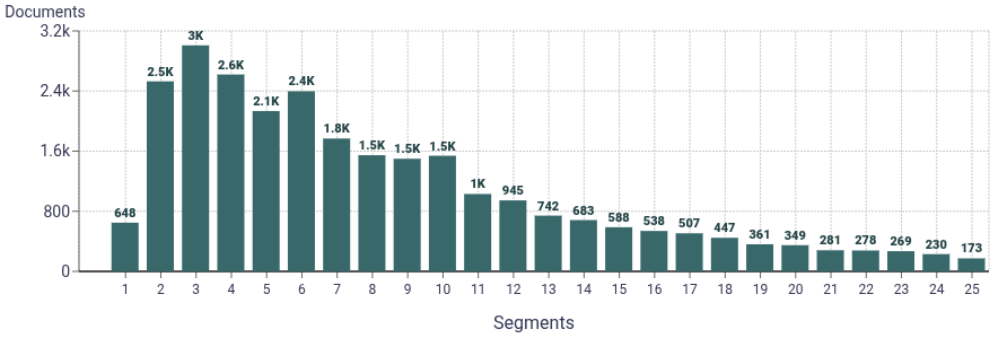
Domain	Docs	% of total
wikipedia.org	4.5K	14.50%
slonskogodka.com	2.8K	9.12%
wachtyrz.eu	1.3K	4.22%
chopwkuchni.pl	779	2.53%
mdr.de	693	2.25%
wordpress.com	530	1.72%
uj.edu.pl	434	1.41%
zycienaniebiesk...	382	1.24%
diesachsen.de	379	1.23%
viamedica.pl	369	1.20%

Top 10 TLDs

Domain	Docs	% of total
pl	9.2K	29.82%
org	5.3K	17.33%
com	5.2K	16.83%
de	4.3K	14.02%
eu	2.4K	7.71%
edu.pl	854	2.77%
com.pl	545	1.77%
cz	452	1.47%
net.pl	390	1.26%
info	365	1.18%

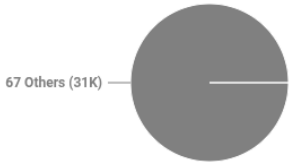
Documents size (in segments)

≤ 25 segments **87.9%** (27K documents)
> 25 segments **12.1%** (3.7K documents)



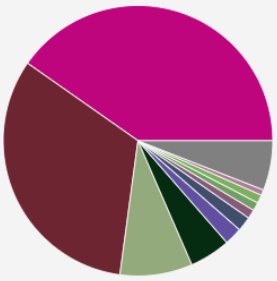
Document collections

CC = 82.34%
IA = 17.66%



Language Distribution

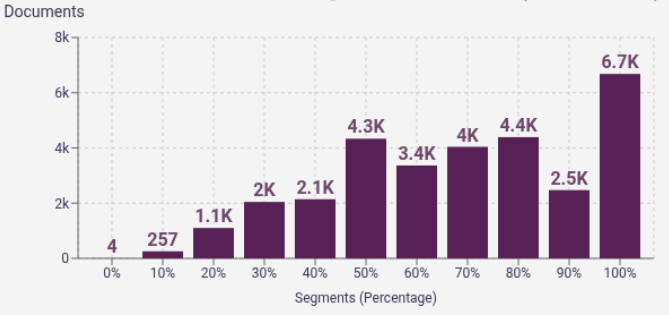
Number of segments in the Silesian corpus



- Silesian - 196K **(40.3%)**
- Polish - 158K **(32.6%)**
- Upper Sorbian - 42K **(8.7%)**
- English - 24K **(4.9%)**
- German - 10K **(2.1%)**
- Czech - 8.5K **(1.7%)**
- Lower Sorbian - 5.3K **(1.1%)**
- Italian - 4.9K **(1.0%)**
- Slovenian - 4.8K **(1.0%)**
- Spanish - 3.2K **(0.7%)**
- 144 Others - 29K **(5.9%)**

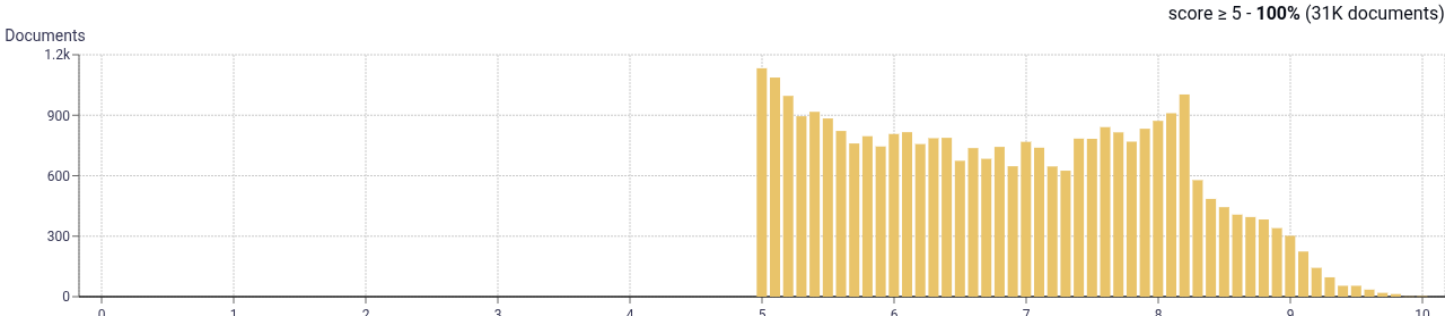
Percentage of segments in Silesian inside documents

segments < 50% - **18.00%** (5.6K documents)
segments ≥ 50% - **82.00%** (25K documents)



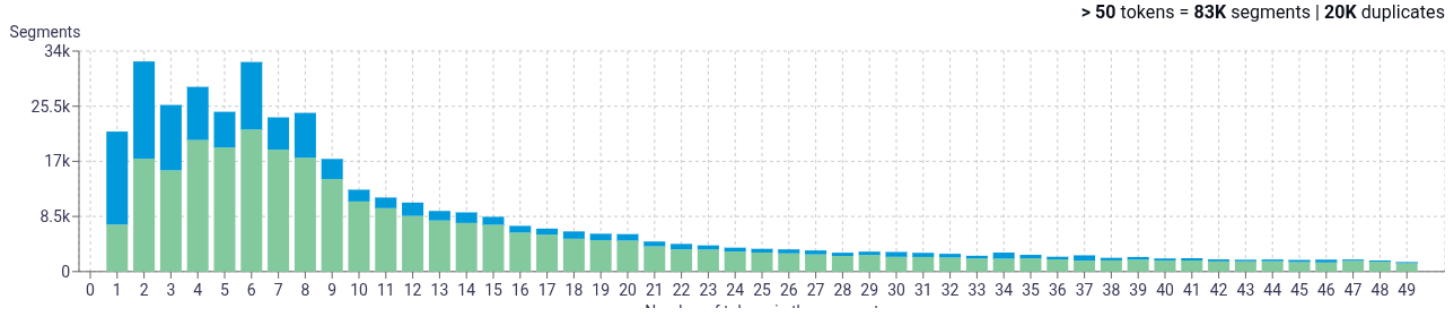
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (31K documents)

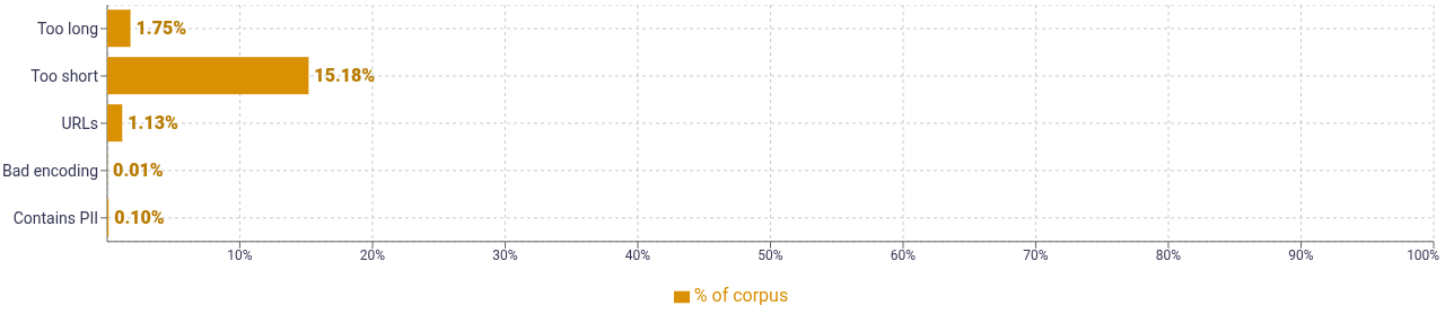


Segment length distribution by token

≤ 49 tokens = **402K** segments | **108K** duplicates
> 50 tokens = **83K** segments | **20K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	so | 36,455 · tak | 34,207 · ale | 32,928 · jan | 32,303 · że | 26,920
2	dr hab | 3,999 · žórłowy tekst | 2,476 · tekst wobdźěłać | 2,475 · teatr im | 2,240 · ni ma | 2,228
3	ginekolog i położnik | 12,503 · žórłowy tekst wobdźěłać | 2,474 · dzyń dzyń dzyń | 745 · pujcka bez uroku | 605 · ludowe nakładnistwo domowina | 527
4	dzyń dzyń dzyń dzyń | 730 · wudaću nańdźeće mjez druhim | 426 · nańdźeće mjez druhim tole | 426 · aktualnym wudaću nańdźeće mjez | 426 · założby za serbski lud | 419
5	dzyń dzyń dzyń dzyń dzyń | 726 · wudaću nańdźeće mjez druhim tole | 426 · aktualnym wudaću nańdźeće 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>, <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				