

General overview

Corpus	Analytics date	Language
szl_Latn.jsonl.tsv	12/6/2024	Silesian (szl)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
40,934	636,571	283,641 (44.56 %)	18M	104.72 MB	103,242,084

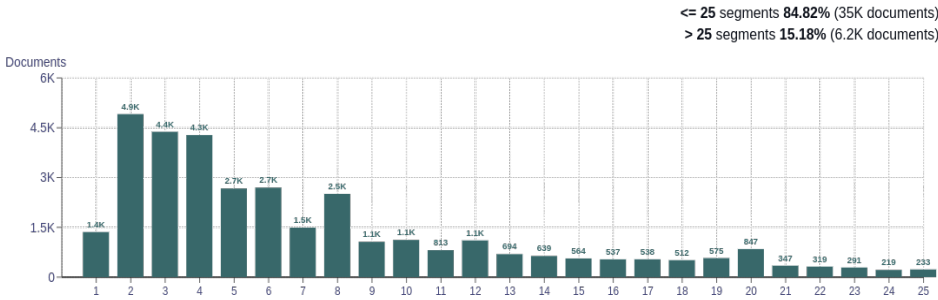
Top 10 domains

Domain	Docs	% of total
wikipedia.org	16K	38.03
serbske-nowiny.de	5.4K	13.13
slonskogodka.com	3.4K	8.21
mapa-kodow-pocztowych.pl	1.8K	4.35
wachtyrz.eu	1.2K	3.03
chopwkuchni.pl	1K	2.46
rozhlad.de	938	2.29
rymy.eu	680	1.66
domowina-verlag.de	449	1.10
gryfnie.com	355	0.87

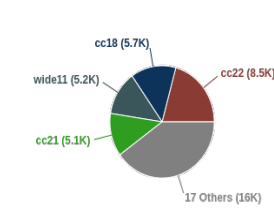
Top 10 TLDs

Domain	Docs	% of total
org	16K	39.62
de	8.7K	21.25
pl	6.6K	16.08
com	5.1K	12.58
eu	2.6K	6.32
com.pl	480	1.17
edu.pl	209	0.51
info	163	0.40
cz	159	0.39
net	126	0.31

Documents size (in segments)

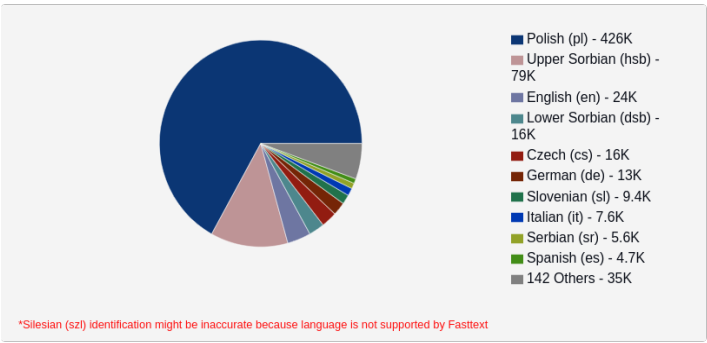


Documents by collection

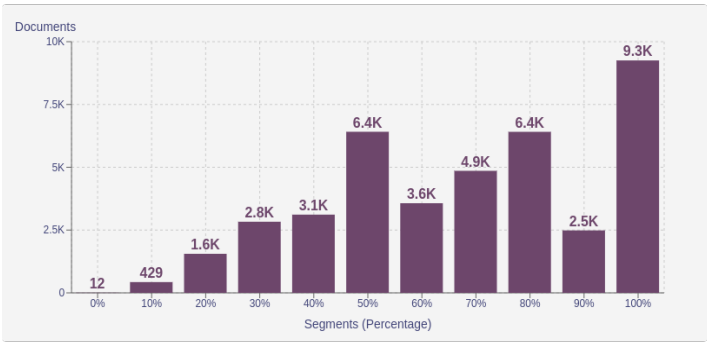


Language Distribution

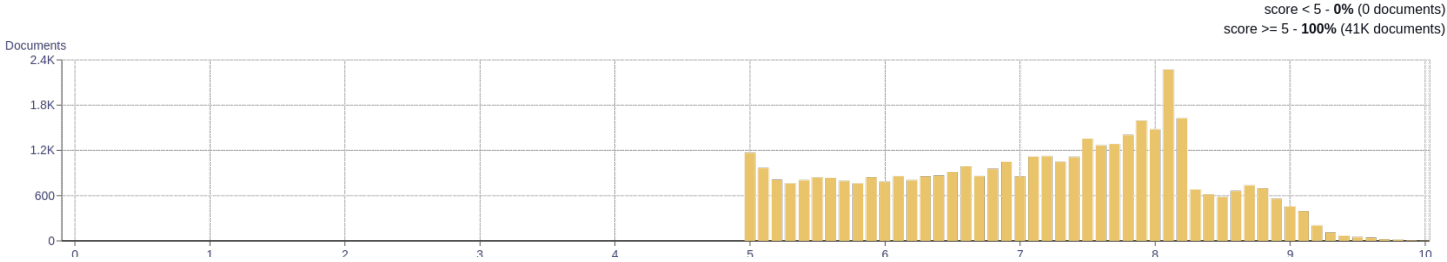
Number of segments



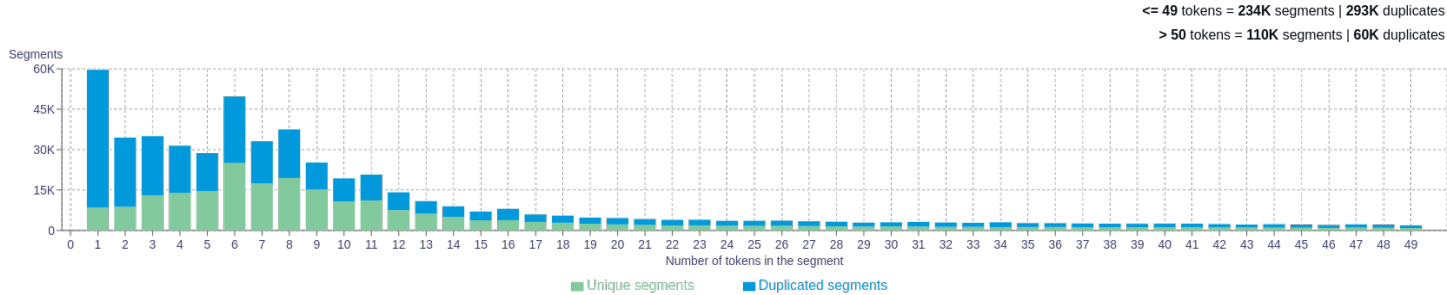
Percentage of segments in Silesian (szl) inside documents



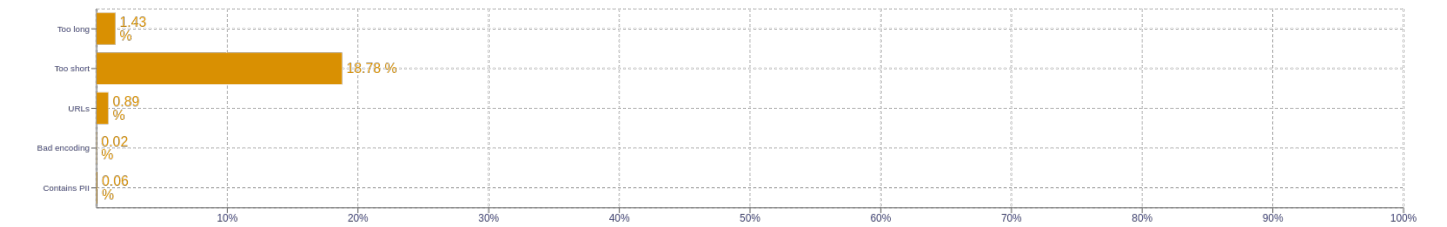
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	<div>warszawa 87038</div> <div>kod 77441</div> <div>pocztowy 76999</div> <div>so 66013</div> <div>jo 46665</div>
2	<div>kod pocztowy 76941</div> <div>warszawa kod 23807</div> <div>warszawa warszawa 11667</div> <div>ruda śląska 9550</div> <div>śląska kod 8043</div>
3	<div>warszawa kod pocztowy 23807</div> <div>śląska kod pocztowy 8043</div> <div>ruda śląska kod 8040</div> <div>warszawa mazowieckie m 6338</div> <div>bydgoszcz kod pocztowy 5684</div>
4	<div>ruda śląska kod pocztowy 8040</div> <div>tynf tynf tynf tynf 5373</div> <div>cycki cycki cycki cycki 3444</div> <div>grodzisk mazowiecki kod pocztowy 1499</div> <div>pocztowe w innych miejscowościach 1047</div>
5	<div>tynf tynf tynf tynf tynf 5372</div> <div>cycki cycki cycki cycki cycki 3431</div> <div>kody pocztowe w innych miejscowościach 1047</div> <div>k stawiznam a kulturje serbow 460</div> <div>wojtek jagielski na żywowojtek jagielski 447</div>

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.stinfo.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>