HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
slv_Latn.jsonl.tsv	6/7/2025	Slovenian (sl)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
10,277,172	238,597,185	107,466,653 (45.04 %)	6.4B	35,029,805,659	33.47 GB

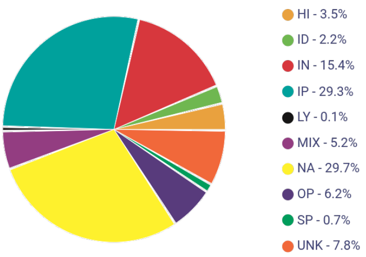
Top 10 domains

Domain	Docs	% of total
wikipedia.org	276K	2.69%
sta.si	226K	2.20%
siol.net	141K	1.37%
delo.si	130K	1.26%
blogspot.com	120K	1.16%
dnevnik.si	83K	0.81%
metropolitan.si	77K	0.75%
slo-tech.com	73K	0.71%
rtvslo.si	72K	0.70%
agoda.com	71K	0.69%

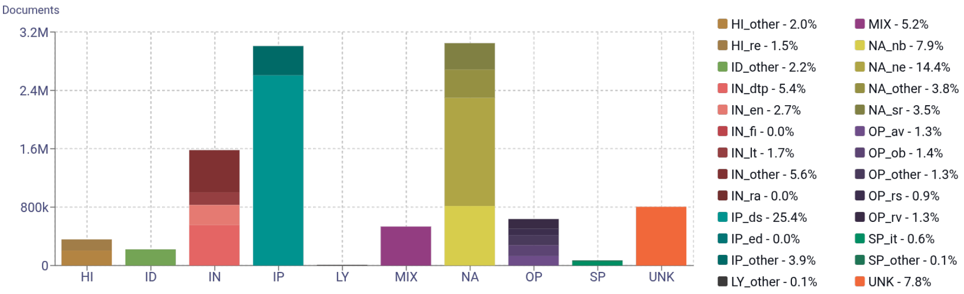
Top 10 TLDs

Domain	Docs	% of total
si	6.3M	61.21%
com	2.2M	21.01%
net	558K	5.43%
org	555K	5.40%
eu	252K	2.45%
info	120K	1.17%
se	28K	0.27%
je	24K	0.24%
tv	22K	0.21%
at	19K	0.19%

Register labels

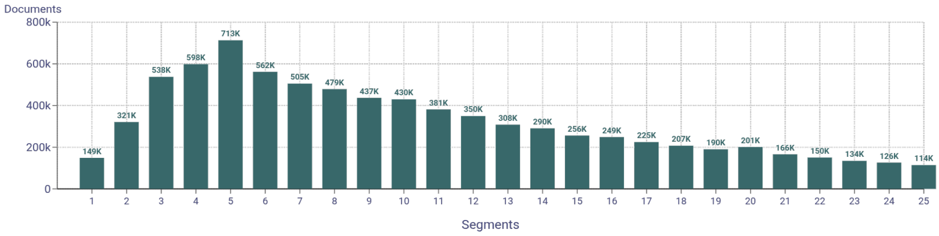


MT:5.0% | 514K Documents



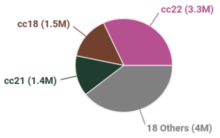
Documents size (in segments)

<= 25 segments **78.6%** (8.1M documents)
> 25 segments **21.4%** (2.2M documents)



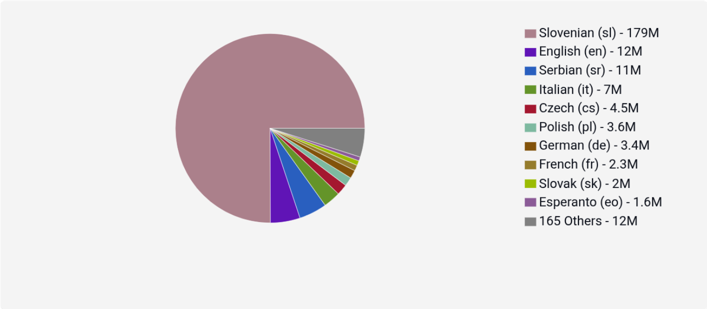
Documents by collection

CC = 70.36%
IA = 29.64%

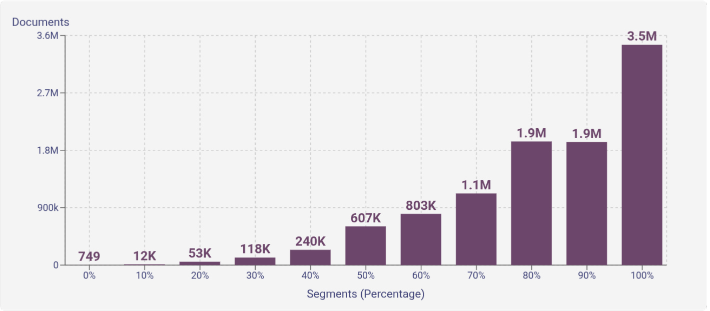


Language Distribution

Number of segments in the Slovenian (sl) corpus

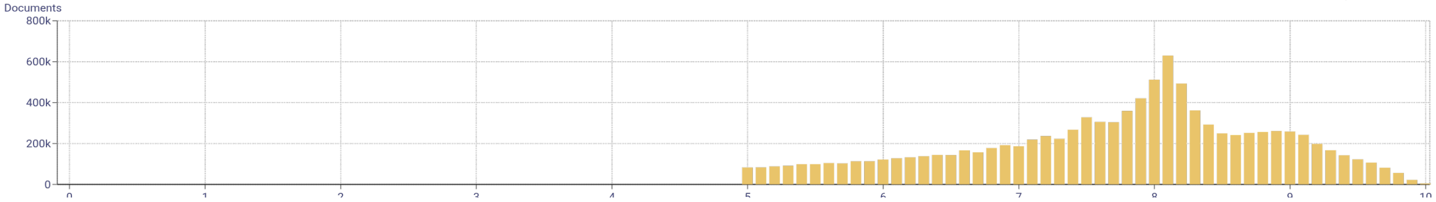


Percentage of segments in Slovenian (sl) inside documents



Distribution of documents by document score

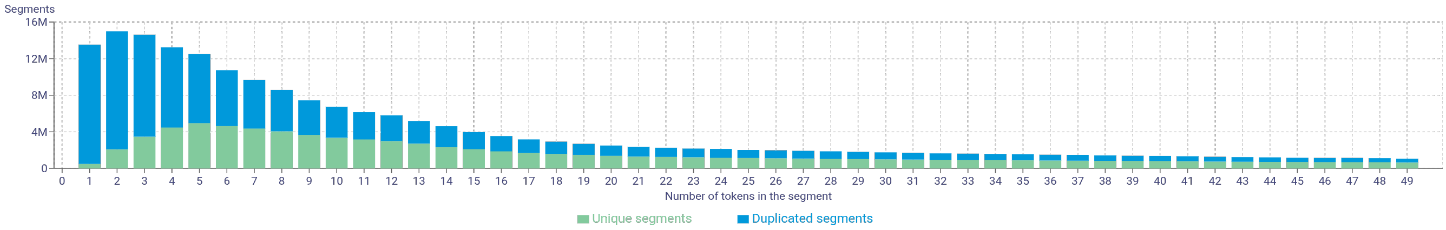
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (10M documents)



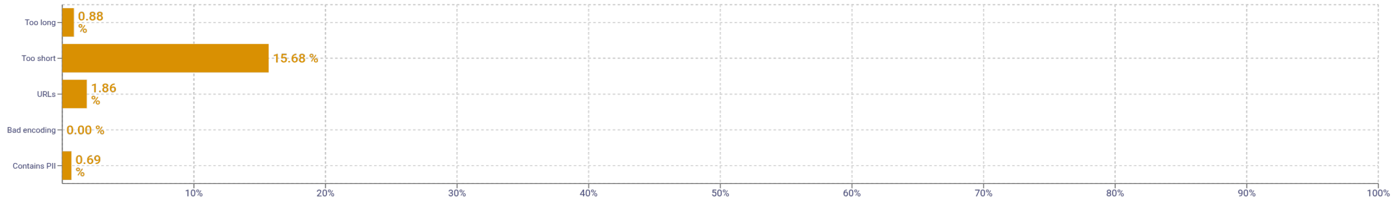
Segment length distribution by token

≤ 49 tokens = **83M** segments | **116M** duplicates

> 50 tokens = **40M** segments | **16M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	lahko | 25982227 zelo | 6691629 a | 6631583 leta | 6410881 zato | 6406868
2	uredi kodo | 974504 spletni strani | 834745 republike slovenije | 684698 spletne strani | 655489 osebnih podatkov | 540481
3	uradni list rs | 388741 člena tega zakona | 187426 državna revizijska komisija | 166992 dodaj v košarico | 159247 odstavka tega člena | 157674
4	celotna novica je dostopna | 136900 evropskega parlamenta in sveta | 64408 e-poštni naslov je zaščiten | 55487 obvezno pokojninsko in invalidsko | 48545 uradnem listu republike slovenije | 47721
5	novica je dostopna le naročnikom | 136880 naslov je zaščiten proti smetenju | 55436 kazensko odgovoren za javno spodbujanje | 44940 posameznik kazensko odgovoren za javno | 44926 odgovoren za javno spodbujanje sovraštva | 44788

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av