HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-slv_Latn	9/18/2025	Slovenian (sl)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
16,812,095	402,087,512	226,735,448 (56.39 %)	11B	62,072,495,644	59.29 GB

Top 10 domains

Domain	Docs	% of total
delo.si	264K	1.57%
rtvslo.si	227K	1.35%
sta.si	194K	1.16%
metropolitan.si	181K	1.08%
dnevnik.si	176K	1.05%
zurnal24.si	148K	0.88%
siol.net	148K	0.88%
blogspot.com	135K	0.80%
arnes.si	107K	0.63%
ognjisce.si	97K	0.58%

Top 10 TLDs

Domain	Docs	% of total
si	10M	60.50%
com	3.8M	22.60%
net	776K	4.62%
org	537K	3.20%
eu	407K	2.42%
info	188K	1.12%
cz	126K	0.75%
tv	50K	0.30%
pl	45K	0.27%
sk	41K	0.24%

Register labels

- HI - 3.6%
- ID - 2.0%
- IN - 11.9%
- IP - 27.6%
- LY - 0.1%
- MIX - 4.6%
- NA - 31.8%
- OP - 5.5%
- SP - 0.6%
- UNK - 12.3%

- HI_other - 2.2%
- HI_re - 1.4%
- ID_other - 2.0%
- IN_dtp - 4.5%
- IN_en - 0.7%
- IN_fi - 0.0%
- IN_lt - 1.3%
- IN_other - 5.4%
- IN_ra - 0.1%
- IP_ds - 23.7%
- IP_other - 3.8%
- LY_other - 0.1%
- MIX - 4.6%
- NA_nb - 6.6%
- NA_ne - 17.1%
- NA_other - 3.9%
- NA_sr - 4.2%
- OP_av - 1.1%
- OP_ob - 1.2%
- OP_other - 1.3%
- OP_rs - 0.8%
- OP_rv - 1.1%
- SP_it - 0.5%
- SP_other - 0.1%
- UNK - 12.3%
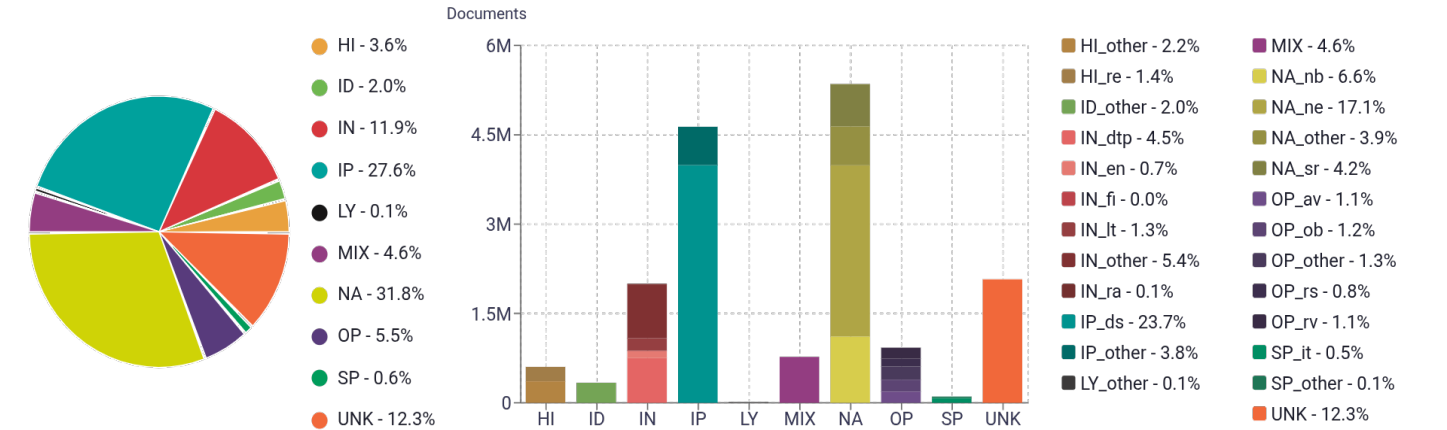
MT:9.0% | 1.5M Documents

Documents size (in segments) ⓘ

≤ **25** segments **80.59%** (14M documents)
> **25** segments **19.41%** (3.3M documents)

Document collections

CC = 91.38%
IA = 8.62%

67 Others (17M)
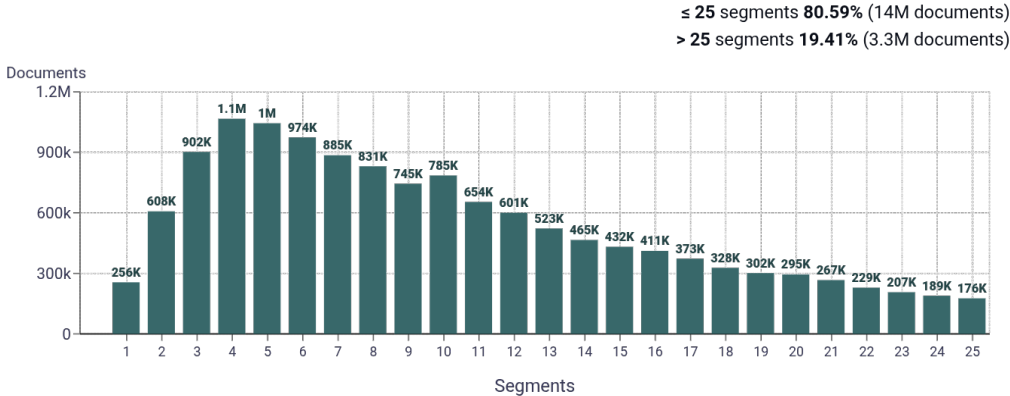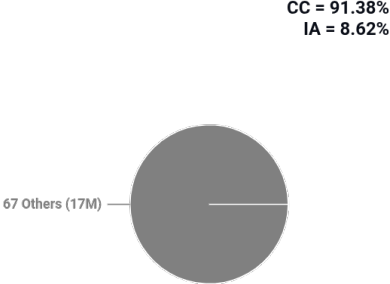
Language Distribution

Number of segments in the Slovenian (sl) corpus

- Slovenian (sl) - 313M **(77.9%)**
- English (en) - 19M **(4.7%)**
- Serbian (sr) - 18M **(4.5%)**
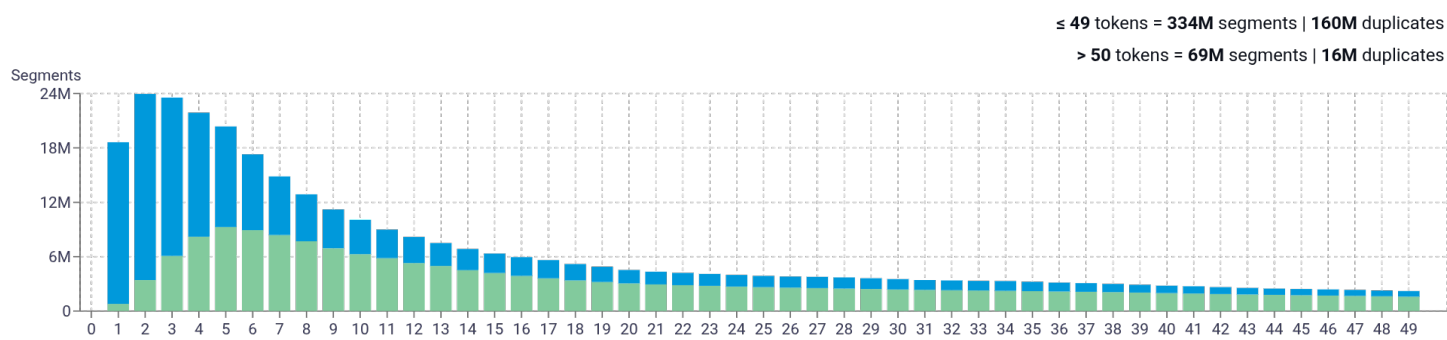- Italian (it) - 7.7M **(1.9%)**
- Czech (cs) - 6.2M **(1.5%)**
- Polish (pl) - 5.9M **(1.5%)**
- German (de) - 4.3M **(1.1%)**
- French (fr) - 3.4M **(0.8%)**
- Esperanto (eo) - 2.9M **(0.7%)**
- Slovak (sk) - 2.5M **(0.6%)**
- 165 Others - 19M **(4.7%)**
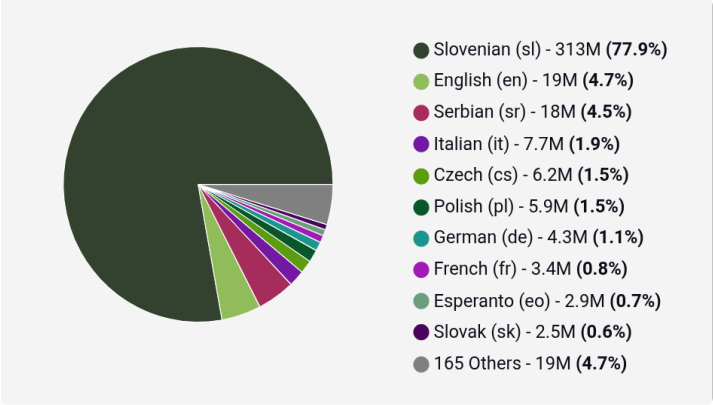
Percentage of segments in Slovenian (sl) inside documents
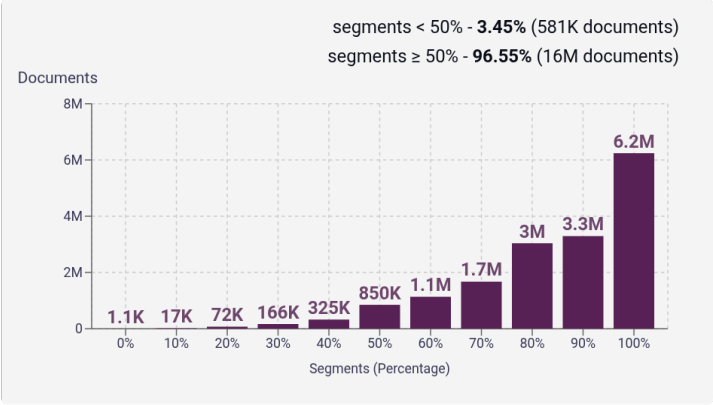
segments < 50% - **3.45%** (581K documents)
segments ≥ 50% - **96.55%** (16M documents)

Documents

- 0%: 1.1K
- 10%: 17K
- 20%: 72K
- 30%: 166K
- 40%: 325K
- 50%: 850K
- 60%: 1.1M
- 70%: 1.7M
- 80%: 3M
- 90%: 3.3M
- 100%: 6.2M

Segments (Percentage)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (17M documents)

Segment length distribution by token

≤ 49 tokens = **334M** segments | **160M** duplicates
> 50 tokens = **69M** segments | **16M** duplicates

Segments

Number of tokens in the segment

Segment noise distribution

- Too long: **0.74%**
- Too short: **14.28%**
- URLs: **1.58%**
- Bad encoding: **0.00%**
- Contains PII: **0.50%**

■ % of corpus
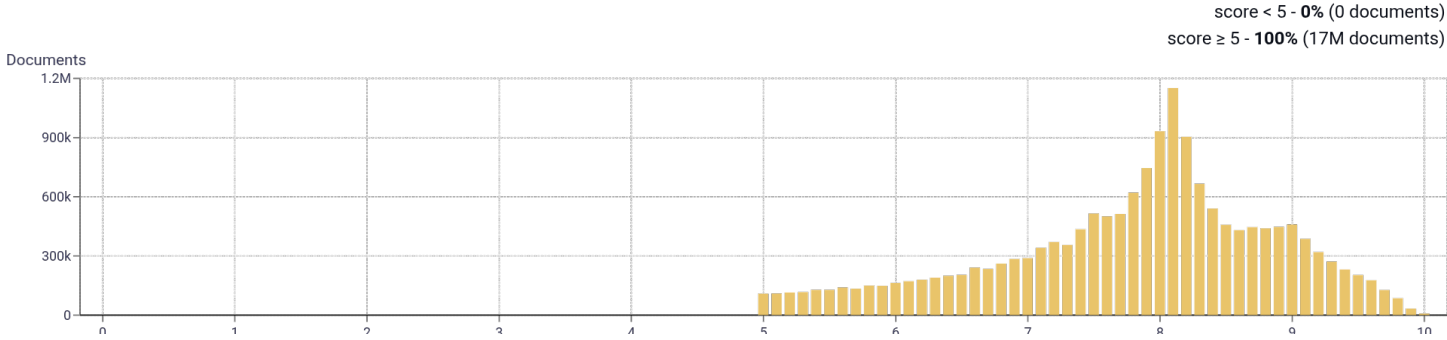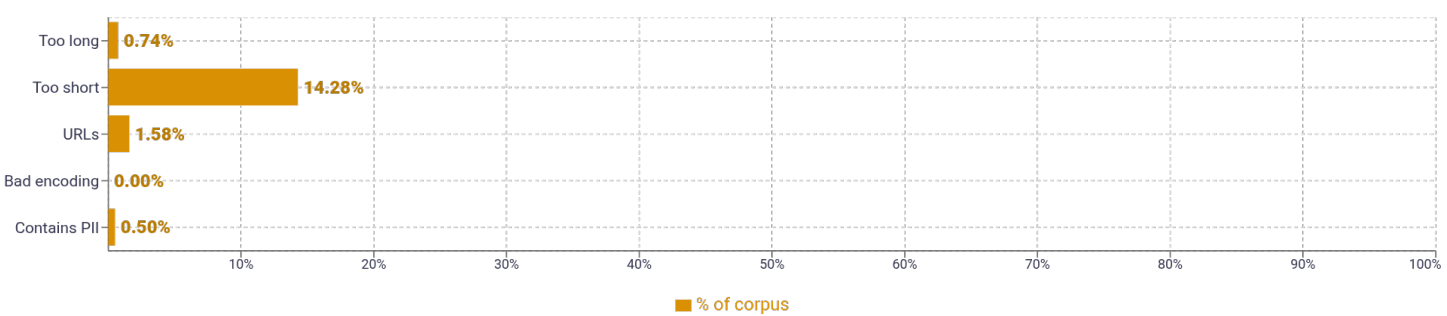
Frequent n-grams

SIZE	N-GRAMS
1	lahko | 53,169,347 zelo | 12,235,899 zato | 11,475,857 ima | 10,779,122 a | 10,310,684
2	spletni strani | 1,254,060 igralni avtomat | 1,026,230 free play | 883,423 igraj igralni | 843,597 play demo | 836,406
3	igraj igralni avtomat | 841,336 free play demo | 836,392 igre na srečo | 494,237 no deposit bonus | 442,458 iger na srečo | 439,968
4	free spins no deposit | 238,033 spins no deposit bonus | 237,314 casino free spins no | 235,692 casino no deposit bonus | 194,863 bonus and promo code | 192,492
5	free spins no deposit bonus | 237,309 casino free spins no deposit | 235,690 casino bonus and promo code | 190,471 odgovoren za javno spodbujanje sovraštva | 95,872 kazensko odgovoren za javno spodbujanje | 95,239

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				