HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-ces_Latn	9/18/2025	Czech

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
107,802,248	2,467,616,974	1,166,958,008 (47.29 %)	66B	365,482,306,497	377.31 GB

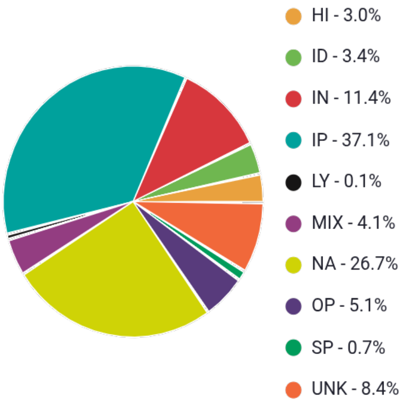
Top 10 domains

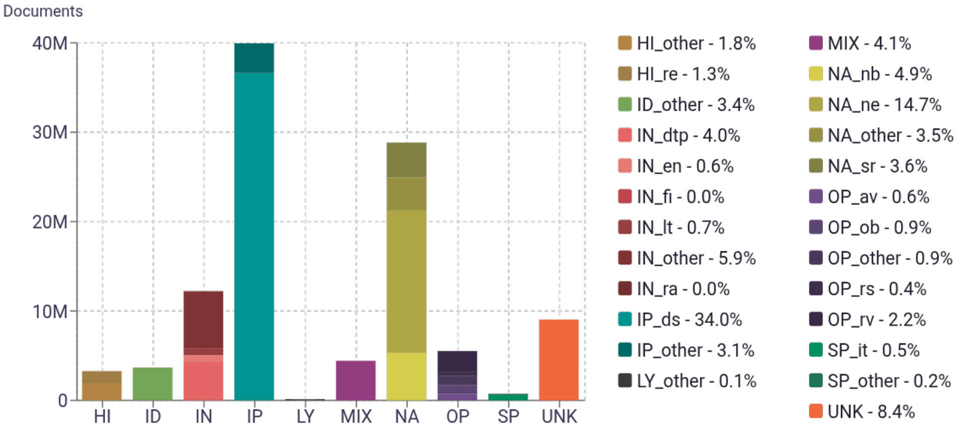
Domain	Docs	% of total
denik.cz	2.5M	2.31%
idnes.cz	1.6M	1.46%
novinky.cz	1.1M	1.02%
web.app	1M	0.97%
blog.cz	972K	0.90%
docplayer.cz	785K	0.73%
ceskatelevize.cz	647K	0.60%
blogspot.com	592K	0.55%
rozhlas.cz	538K	0.50%
firebaseapp.com	496K	0.46%

Top 10 TLDs

Domain	Docs	% of total
cz	90M	83.16%
com	8.3M	7.74%
eu	2.3M	2.17%
org	1.3M	1.23%
net	1.1M	1.06%
app	1.1M	0.98%
sk	912K	0.85%
info	683K	0.63%
ru	254K	0.24%
tv	133K	0.12%

Register labels

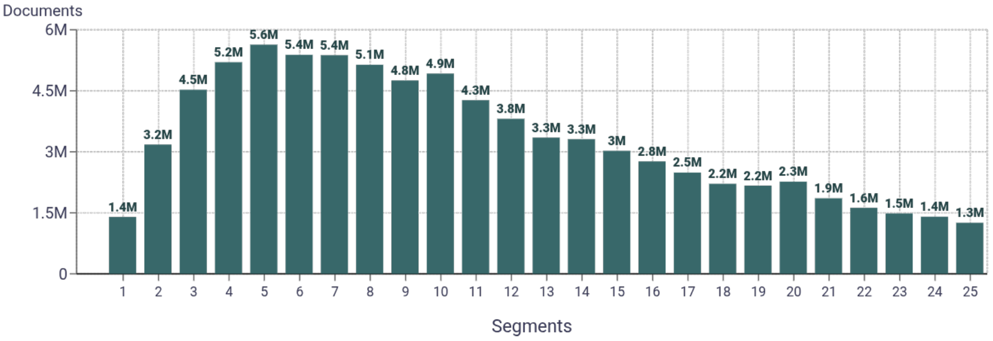




MT:4.8% | 5.2M Documents

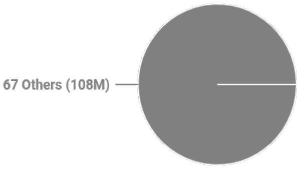
Documents size (in segments)

≤ **25** segments **76.81%** (83M documents)
> **25** segments **23.19%** (25M documents)



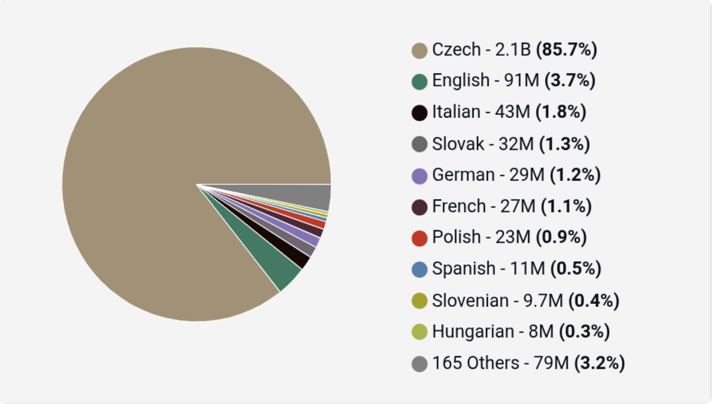
Document collections

CC = 90.67%
IA = 9.33%



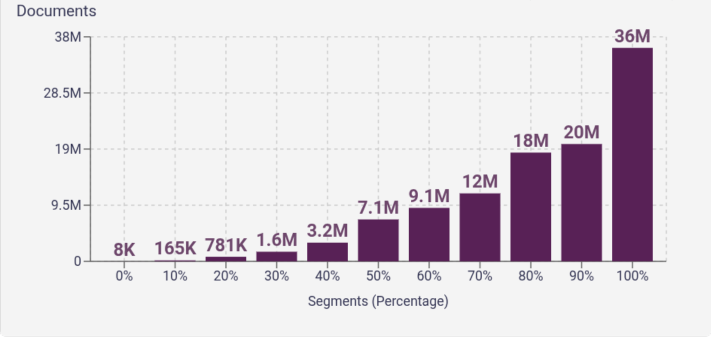
Language Distribution

Number of segments in the Czech corpus



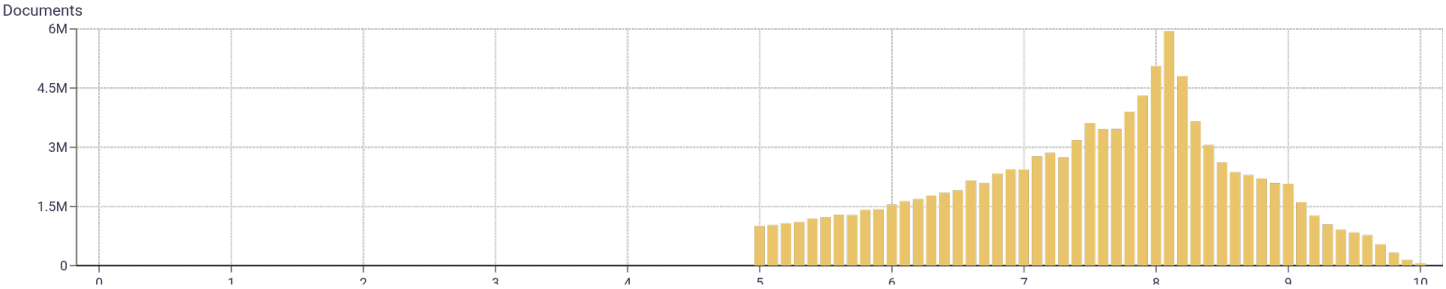
Percentage of segments in Czech inside documents

segments < 50% - **5.33%** (5.7M documents)
segments ≥ 50% - **94.67%** (102M documents)



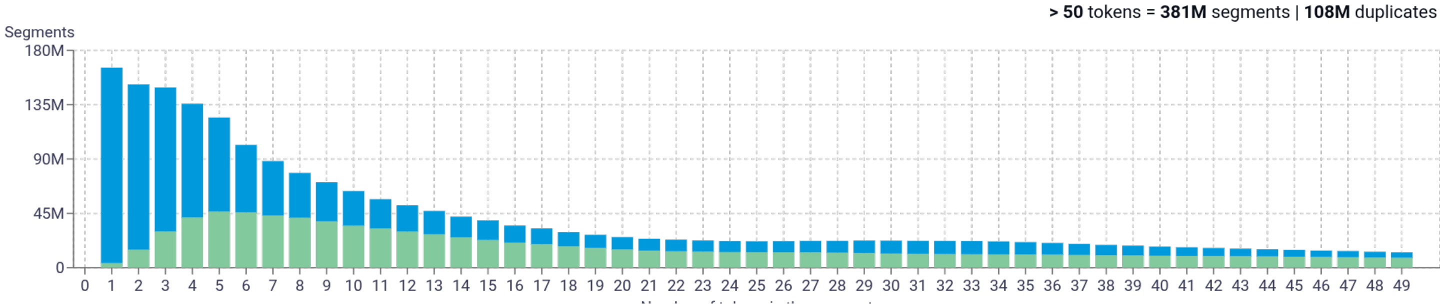
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (108M documents)

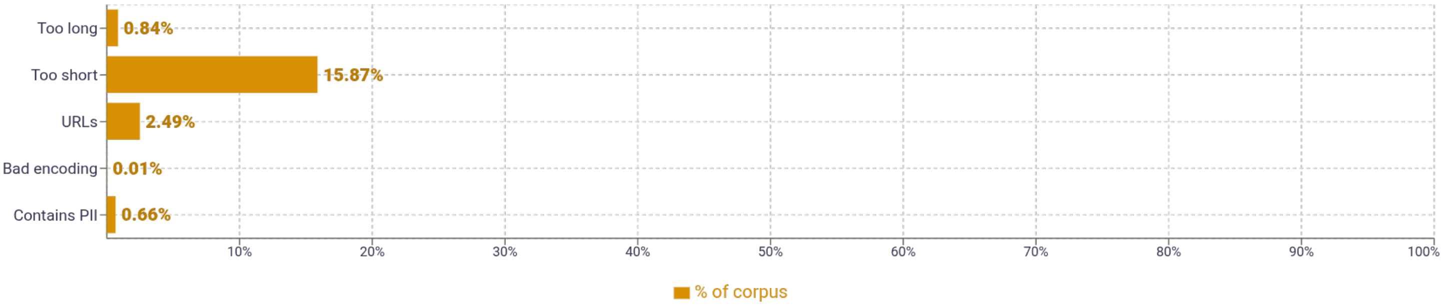


Segment length distribution by token

≤ 49 tokens = **2.1B** segments | **1.2B** duplicates
> 50 tokens = **381M** segments | **108M** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	a | 1,712,557,567 v | 1,031,555,076 se | 936,203,781 s | 487,047,348 z | 342,408,726
2	v roce | 29,393,125 že se | 25,101,426 se v | 22,491,588 a v | 20,701,832 v případě | 18,678,974
3	jedná se o | 7,888,261 v české republice | 6,198,254 se jedná o | 4,485,230 v současné době | 4,324,382 v souladu s | 3,638,486
4	napíše příspěvek k této | 1,792,064 příspěvek k této položce | 1,792,019 vltava labe media a.s. | 1,567,148 že se jedná o | 1,081,330 zdarma a bez rizika | 914,590
5	napíše příspěvek k této položce | 1,792,019 obsahu denik.cz je bez písemného | 795,236 publikování nebo šíření obsahu denik.cz | 795,235 denik.cz je bez písemného souhlasu | 795,224 základní škola a mateřská škola | 760,058

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				