HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
ces_Latn.jsonl.tsv	7/9/2025	Czech (cs)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
75,287,957	1,926,002,523	657,228,256 (34.12 %)	50B	272,085,196,889	280.95 GB

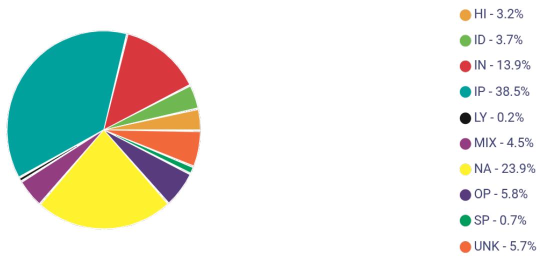
Top 10 domains

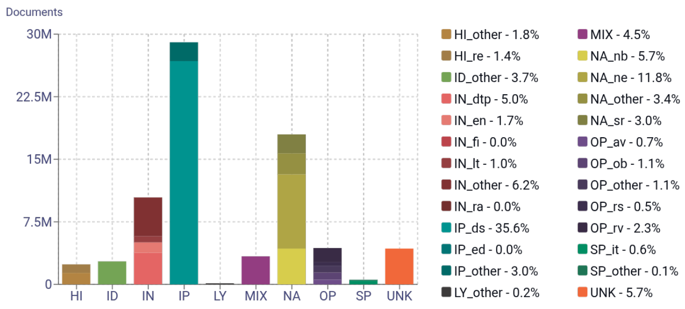
Domain	Docs	% of total
idnes.cz	1.3M	1.69%
wikipedia.org	952K	1.26%
denik.cz	840K	1.12%
blog.cz	716K	0.95%
docplayer.cz	697K	0.93%
novinky.cz	524K	0.70%
blogspot.com	493K	0.65%
ceskatelevize.cz	390K	0.52%
web.app	386K	0.51%
blogspot.cz	353K	0.47%

Top 10 TLDs

Domain	Docs	% of total
cz	63M	83.93%
com	5.2M	6.95%
org	1.6M	2.17%
eu	1.6M	2.16%
net	910K	1.21%
sk	625K	0.83%
info	562K	0.75%
app	390K	0.52%
ru	89K	0.12%
de	73K	0.10%

Register labels

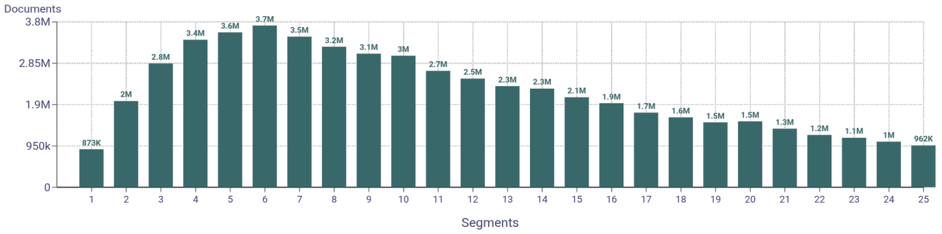




MT:2.5% | 1.8M Documents

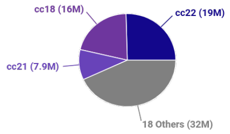
Documents size (in segments)

<= 25 segments **72.95%** (55M documents)
> 25 segments **27.05%** (20M documents)



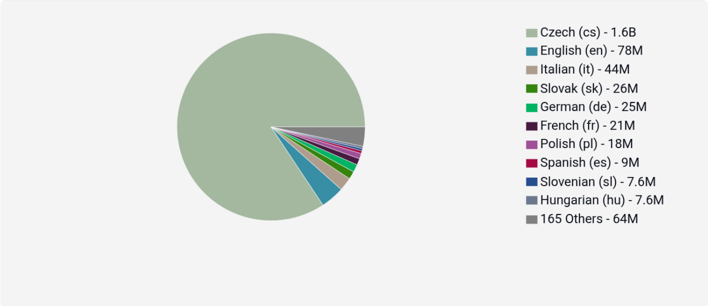
Documents by collection

CC = 70.74%
IA = 29.26%

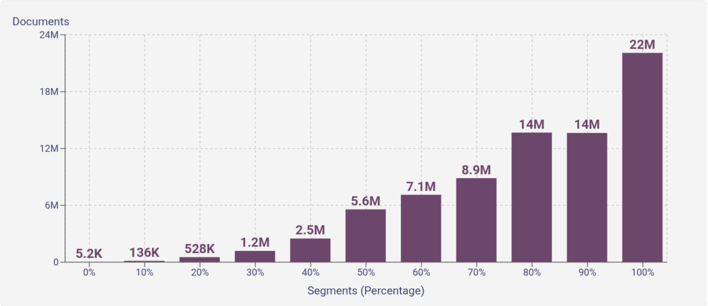


Language Distribution

Number of segments in the Czech (cs) corpus

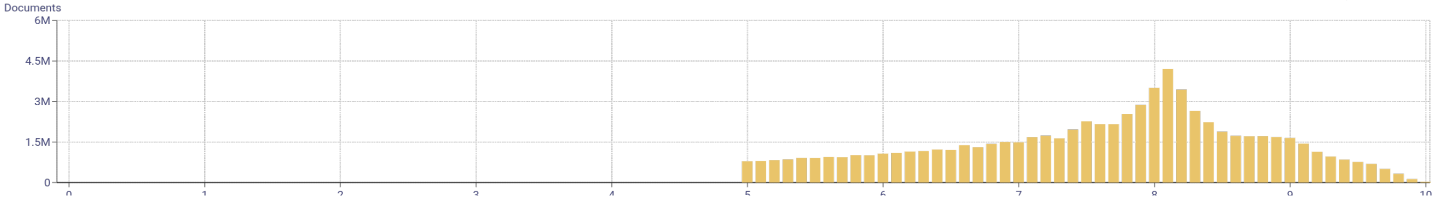


Percentage of segments in Czech (cs) inside documents



Distribution of documents by document score

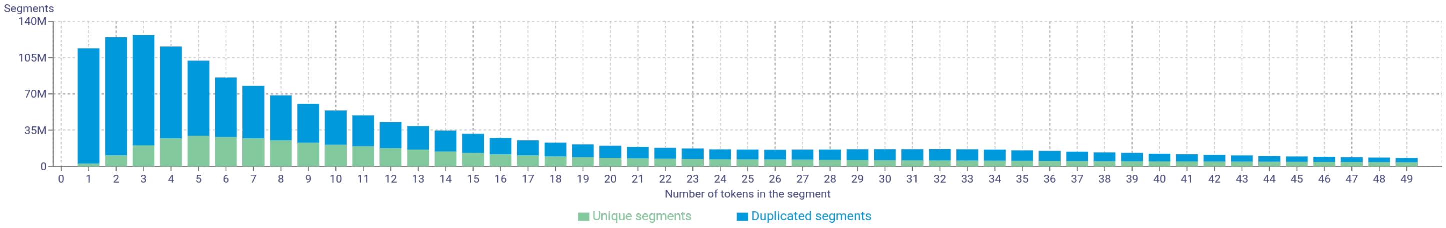
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (75M documents)



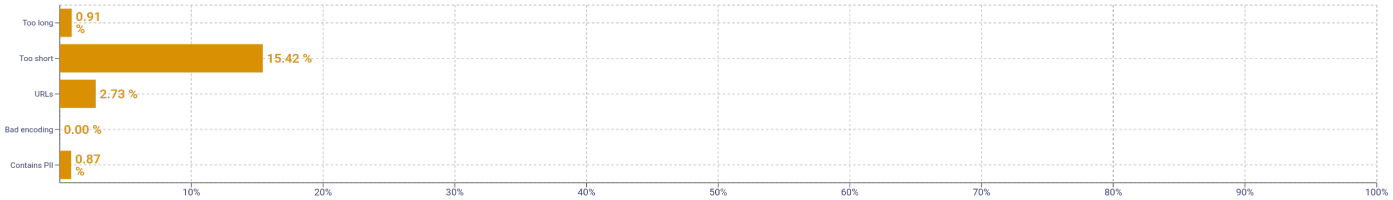
Segment length distribution by token

≤ 49 tokens = **1.7B** segments | **1.1B** duplicates

> 50 tokens = **269M** segments | **126M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	a | 1278867994 v | 756841972 se | 676811466 s | 350406834 z | 257525969
2	a v | 16122935 které se | 11901499 k tomu | 10777871 i v | 10502903 může být | 9624684
3	a mateřská škola | 1544173 a tak se | 1426461 a to i | 1387590 a to v | 1032725 a jak se | 819343
4	a v neposlední řadě | 585482 a zkvalitnění výuky prostřednictvím | 548364 a o změně některých | 479240 další články z rubriky | 468265 a od té doby | 326626
5	a zkvalitnění výuky prostřednictvím ict | 525254 a o změně některých zákonů | 311596 a služby neuvedené v přílohách | 276899 adresa je chráněna před spamboty | 215196 ať už se jedná o | 209365

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