HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hun_Latn.jsonl.tsv	6/29/2025	Hungarian (hu)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
51,870,435	1,418,162,109	502,407,912 (35.43 %)	37B	223,833,111,798	228.8 GB

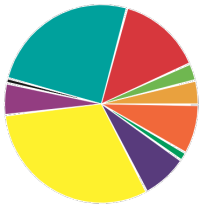
Top 10 domains

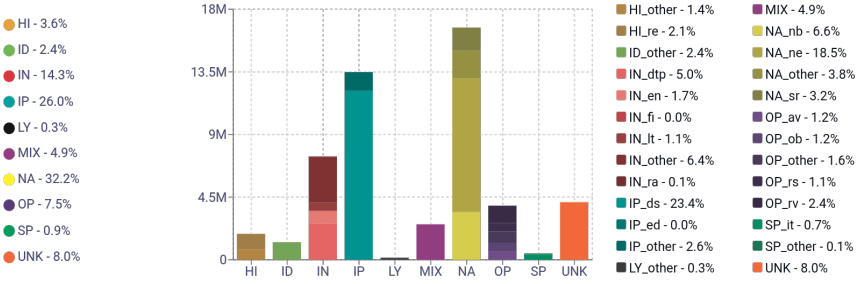
Domain	Docs	% of total
blogspot.com	1M	2.02%
blogspot.hu	1M	1.99%
wikipedia.org	766K	1.48%
index.hu	635K	1.22%
docplayer.hu	556K	1.07%
blog.hu	488K	0.94%
24.hu	421K	0.81%
origo.hu	231K	0.44%
blogspot.ro	189K	0.37%
wordpress.com	178K	0.34%

Top 10 TLDs

Domain	Docs	% of total
hu	39M	75.81%
com	6M	11.62%
org	1.4M	2.60%
ro	827K	1.59%
net	795K	1.53%
eu	692K	1.33%
info	640K	1.23%
sk	344K	0.66%
co.hu	176K	0.34%
ma	171K	0.33%

Register labels

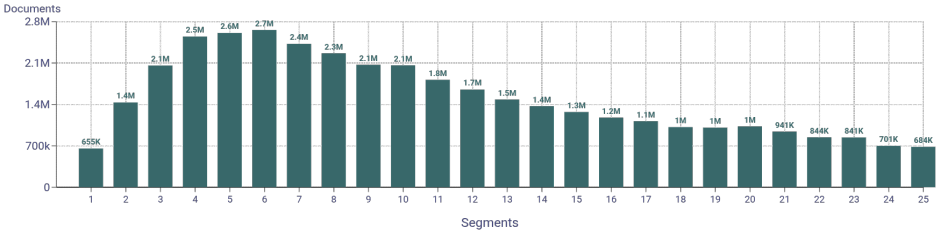




MT:4.7% | 2.4M Documents

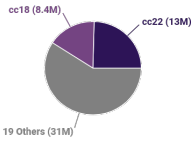
Documents size (in segments)

<= 25 segments **72.75%** (38M documents)
> 25 segments **27.25%** (14M documents)



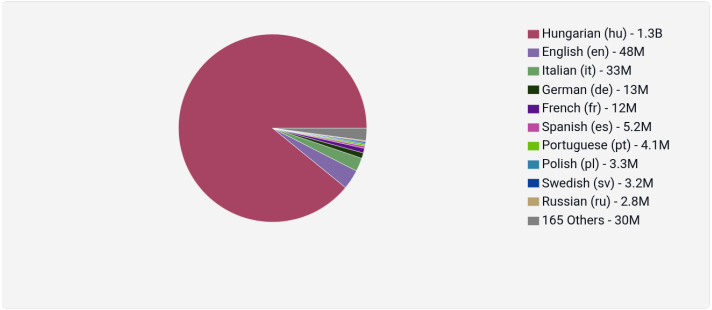
Documents by collection

CC = 60.10%
IA = 39.90%

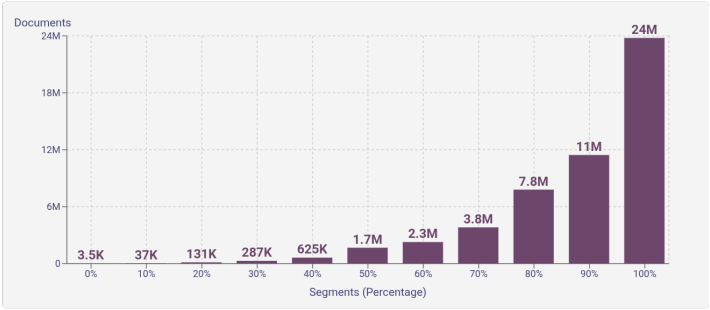


Language Distribution

Number of segments in the Hungarian (hu) corpus

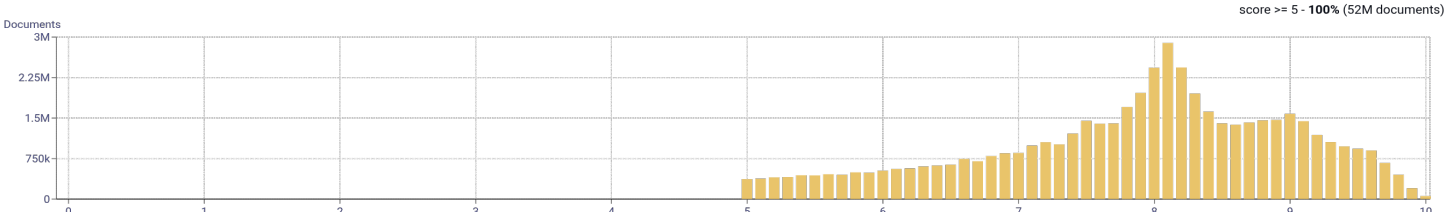


Percentage of segments in Hungarian (hu) inside documents



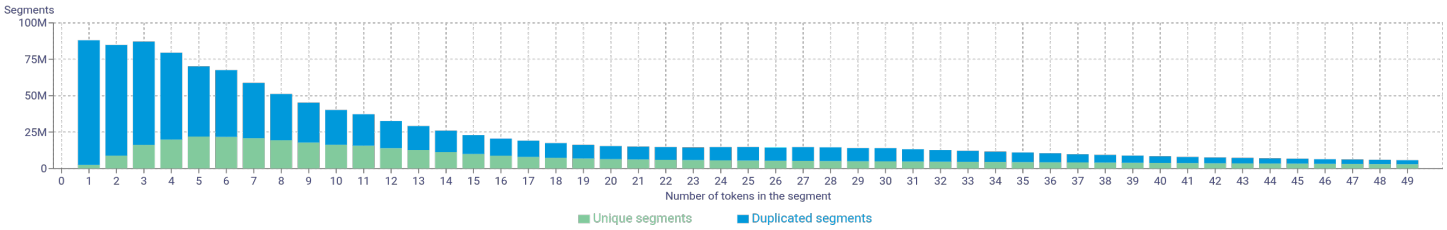
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (52M documents)

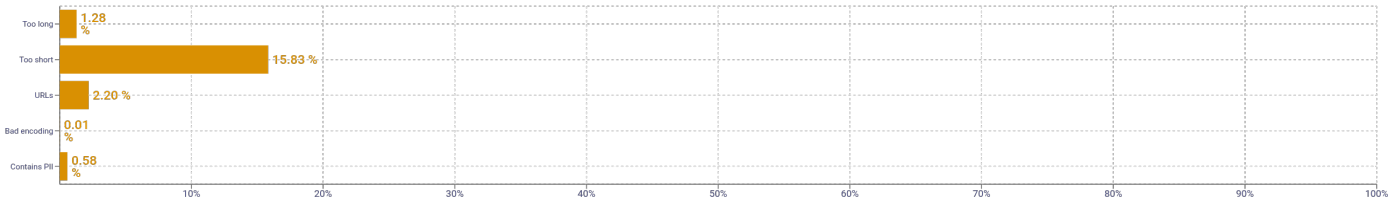


Segment length distribution by token

≤ 49 tokens = **1.2B** segments | **831M** duplicates
> 50 tokens = **191M** segments | **87M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	is | 304599096 · ha | 105143970 · magyar | 42237100 · két | 32692884 · első | 31630640
2	forrásszöveg szerkesztése | 2188876 · g y | 2078625 · lehetővé teszi | 2053368 · község önkormányzata | 1971967 · k ö | 1814283
3	d e l | 383515 · figyelembe kell venni | 330129 · előre is köszönöm | 322424 · alapfokú művészeti iskola | 294910 · emberi erőforrások minisztériuma | 287042
4	csatlakozz te is közösségünkhöz | 212028 · adja meg a kívánt | 170660 · elleni védelem alatt áll | 164046 · engedélyeznie kell a javascript | 161317 · budapesti műszaki és gazdaságtudományi | 156003
5	g y z ő k | 1342786 · elmúlt órában ezt a hotelt | 351197 · d e l e t | 345973 · adja meg a kívánt dátumokat | 164561 · európai parlament és a tanács | 155151

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