

General overview

Corpus	Date	Language
hplt-v3-hun_Latn	9/18/2025	Hungarian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
75,122,917	1,776,578,526	885,735,919 (49.86 %)	48B	294,005,482,561	300.3 GB

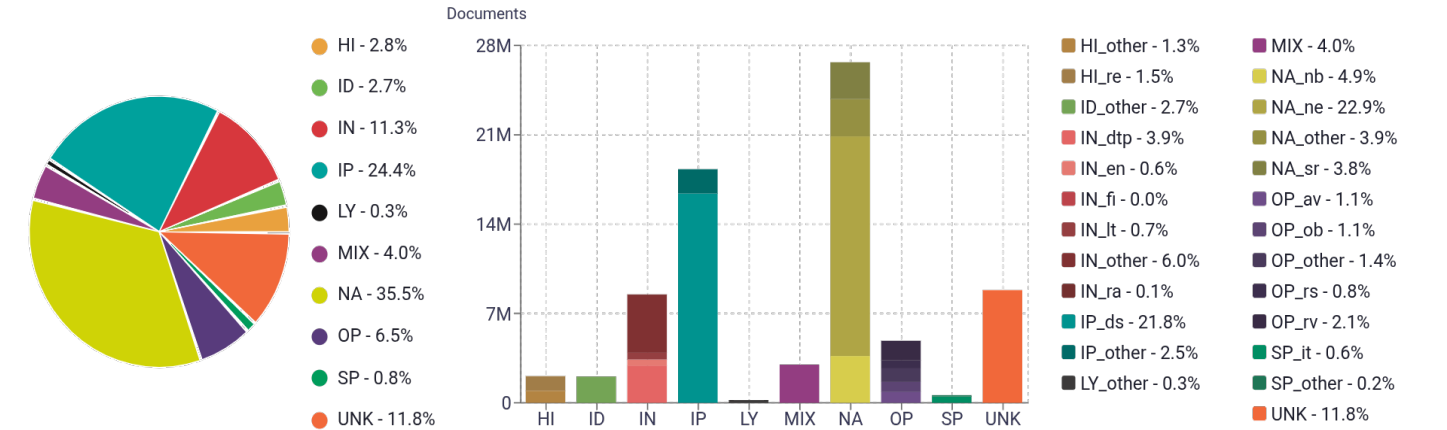
Top 10 domains

Domain	Docs	% of total
blogspot.com	1.2M	1.63%
blog.hu	857K	1.14%
index.hu	648K	0.86%
24.hu	631K	0.84%
docplayer.hu	623K	0.83%
origo.hu	609K	0.81%
blogspot.hu	564K	0.75%
blikk.hu	343K	0.46%
femina.hu	313K	0.42%
wikipedia.org	281K	0.37%

Top 10 TLDs

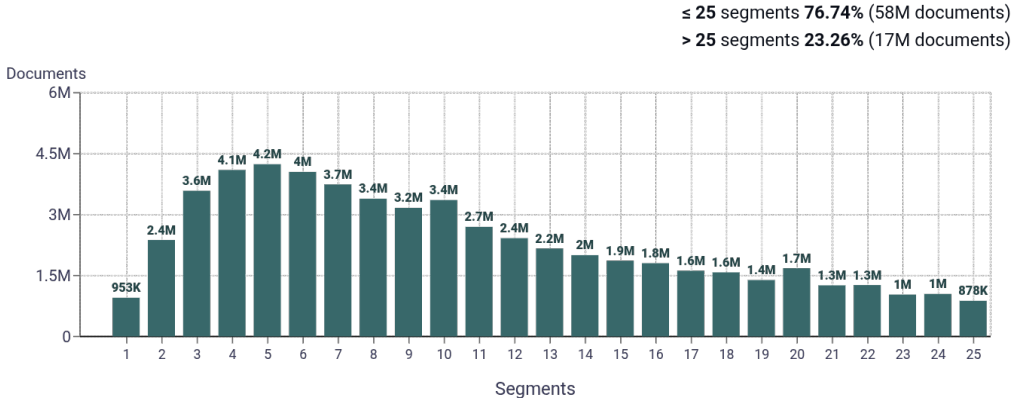
Domain	Docs	% of total
hu	57M	75.28%
com	9M	12.04%
org	1.5M	1.99%
ro	1.3M	1.72%
eu	1.1M	1.50%
net	1M	1.33%
info	720K	0.96%
sk	504K	0.67%
ma	302K	0.40%
news	174K	0.23%

Register labels

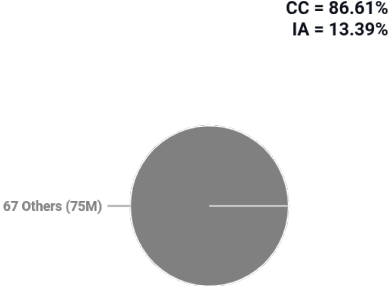


MT:8.1% | 6.1M Documents

Documents size (in segments) ⓘ

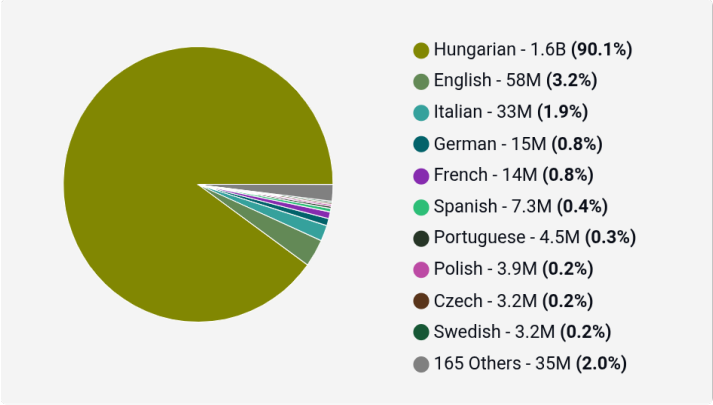


Document collections

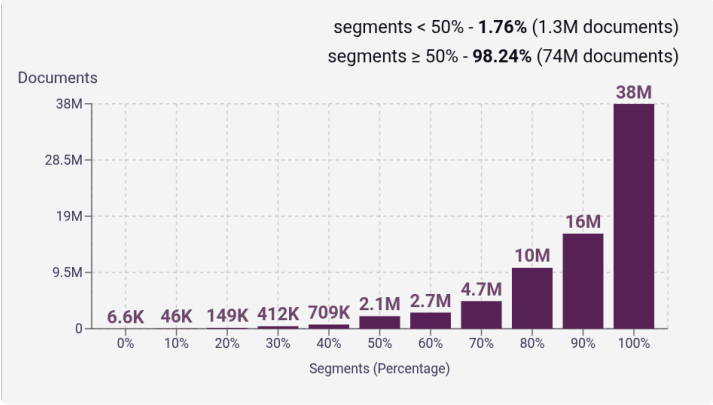


Language Distribution

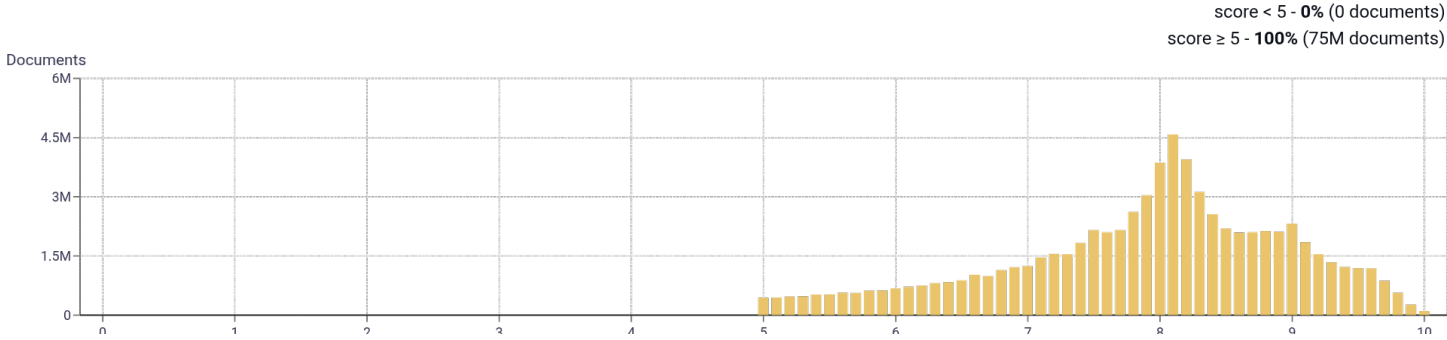
Number of segments in the Hungarian corpus



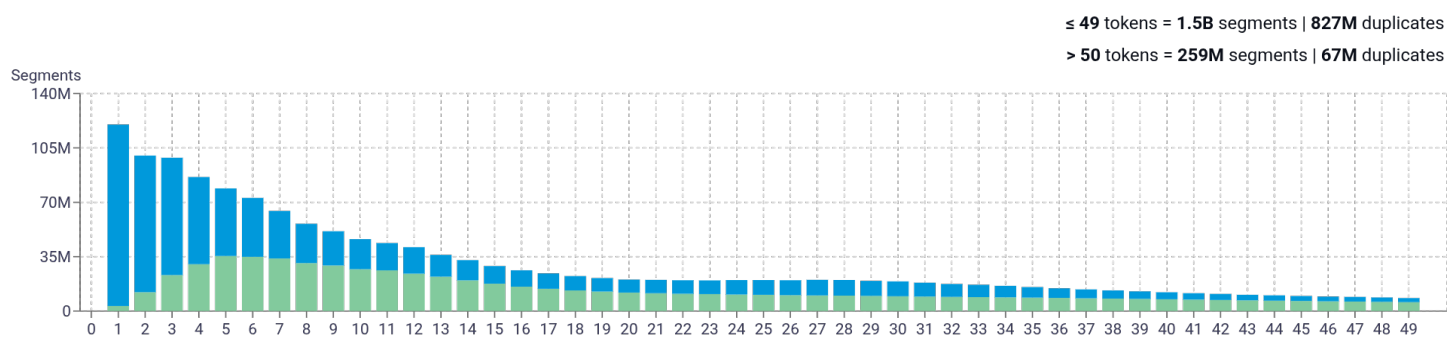
Percentage of segments in Hungarian inside documents



Distribution of documents by document score

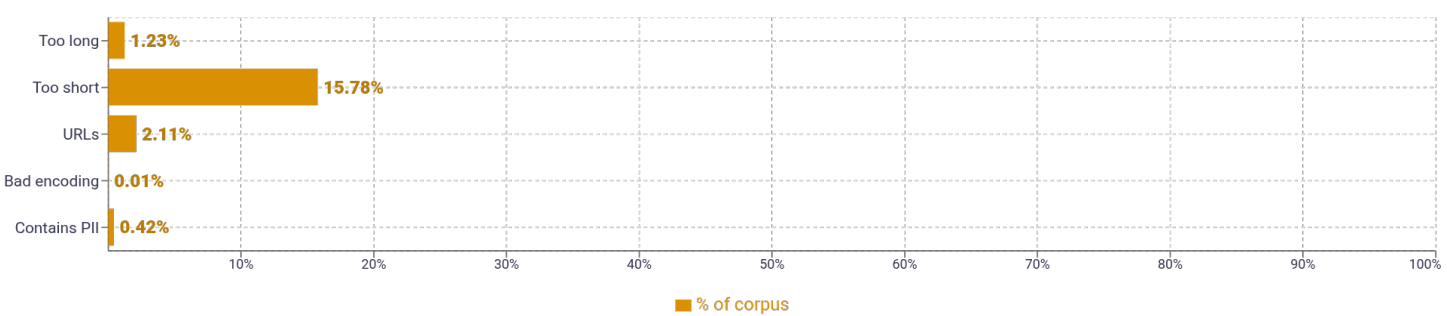


Segment length distribution by token



≤ 49 tokens = 1.5B segments | 827M duplicates  
> 50 tokens = 259M segments | 67M duplicates

Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>is   404,841,975</div> <div>ha   144,057,035</div> <div>magyar   52,818,235</div> <div>két   42,761,084</div> <div>első   42,061,566</div>	
2	<div>továbbra is   5,873,197</div> <div>magas vérnyomás   5,366,124</div> <div>lehetővé teszi   4,385,691</div> <div>társkereső oldal   3,815,207</div> <div>te is   2,821,427</div>	
3	<div>leszokni a dohányzásról   1,074,735</div> <div>ingyenes társkereső oldal   993,326</div> <div>megyei jogú város   853,017</div> <div>tünetei és kezelése   717,957</div> <div>ha nem is   708,690</div>	
4	<div>j e g y   451,832</div> <div>k ö n y   401,526</div> <div>ö n y v   399,881</div> <div>hívta fel a figyelmet   364,898</div> <div>g y z ő   359,099</div>	
5	<div>j e g y z   430,878</div> <div>k ö n y v   399,553</div> <div>g y z ő k   349,542</div> <div>y z ő k ő   347,673</div> <div>z ő k ő n   345,975</div>	

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  $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$ , 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 stopwords. 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				