

General overview

Corpus	Date	Language
hplt-v3-slk_Latn	9/18/2025	Slovak (sk)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
36,370,822	768,114,978	410,671,047 (53.46 %)	21B	115,519,352,504	116.85 GB

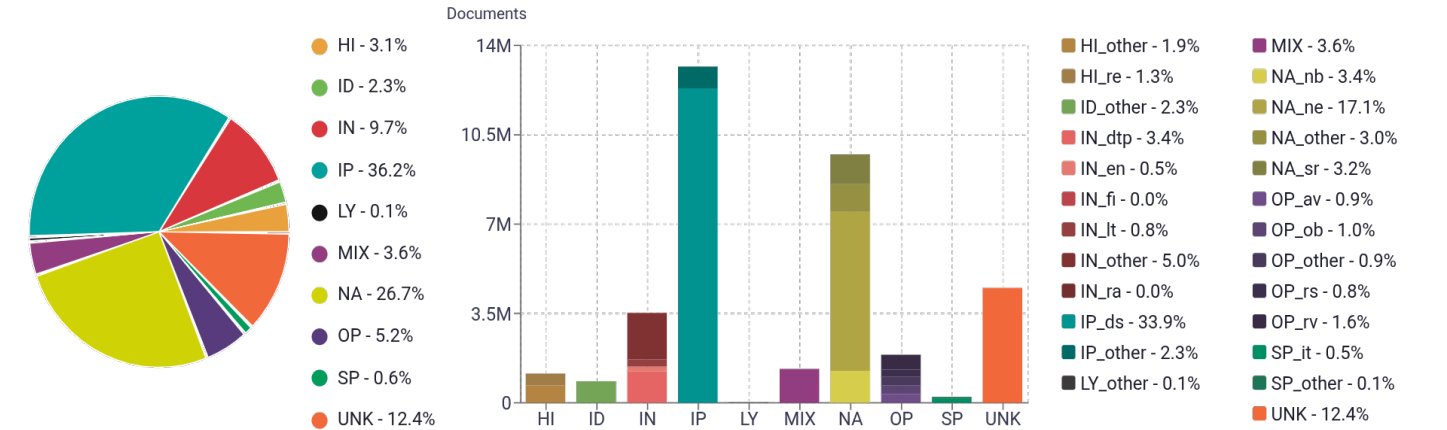
Top 10 domains

Domain	Docs	% of total
sme.sk	2.1M	5.87%
firebaseapp.com	932K	2.56%
web.app	878K	2.42%
pravda.sk	555K	1.53%
aktuality.sk	442K	1.21%
zoznam.sk	403K	1.11%
hnonline.sk	228K	0.63%
dnas24.sk	204K	0.56%
inzercia.sk	191K	0.53%
hlavnespravy.sk	169K	0.46%

Top 10 TLDs

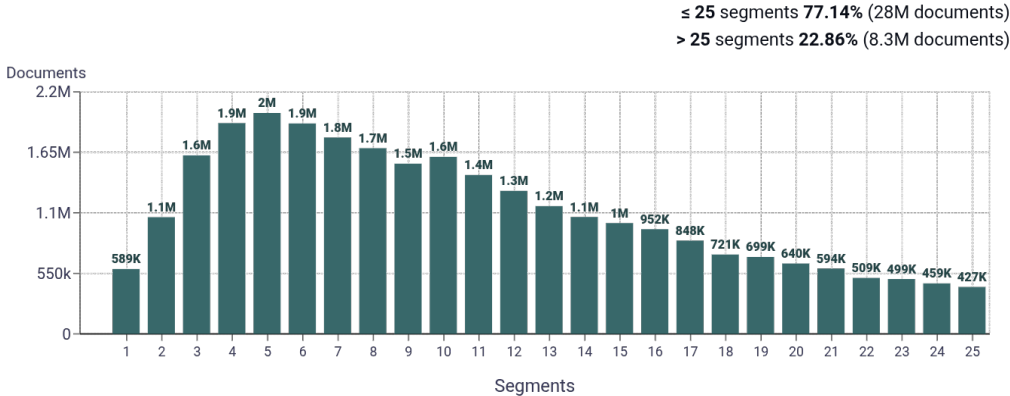
Domain	Docs	% of total
sk	27M	75.53%
com	4.1M	11.21%
cz	1.1M	3.06%
eu	893K	2.46%
app	885K	2.43%
org	370K	1.02%
net	317K	0.87%
info	185K	0.51%
xyz	63K	0.17%
ru	57K	0.16%

Register labels

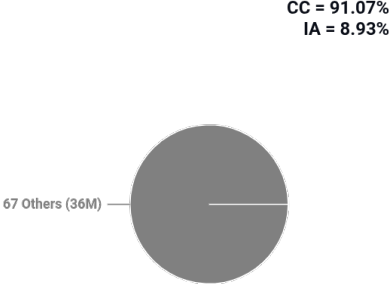


MT:9.6% | 3.5M Documents

Documents size (in segments) ⓘ

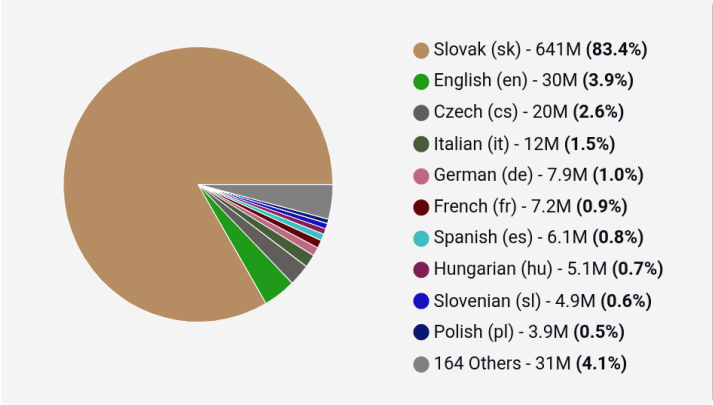


Document collections

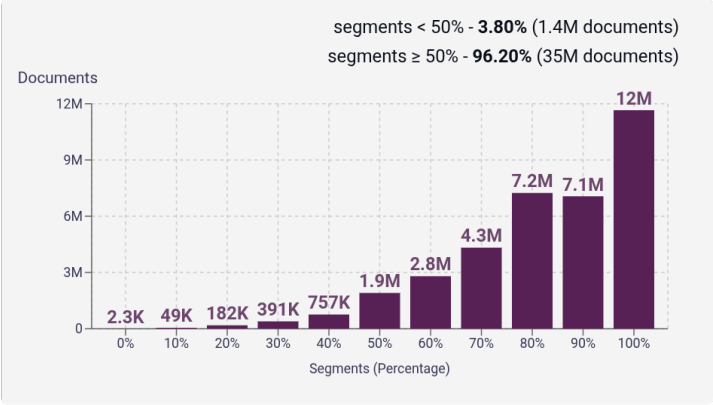


Language Distribution

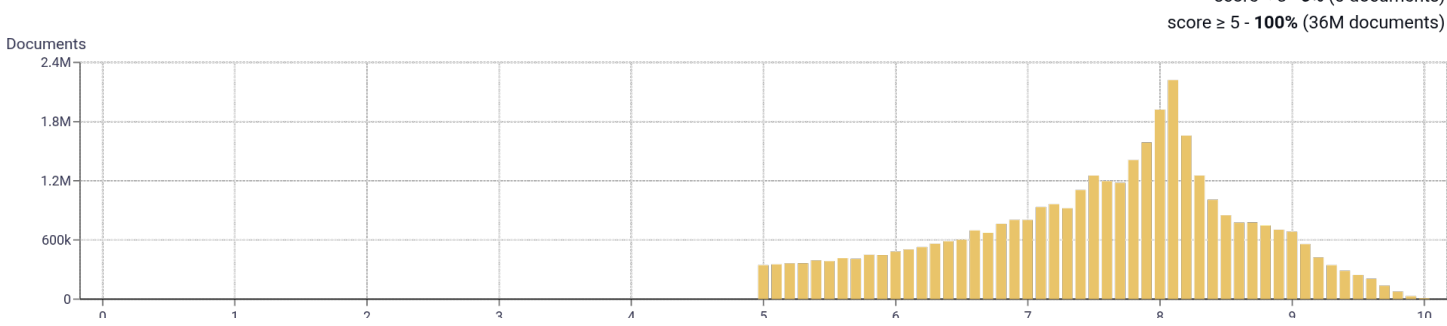
Number of segments in the Slovak (sk) corpus



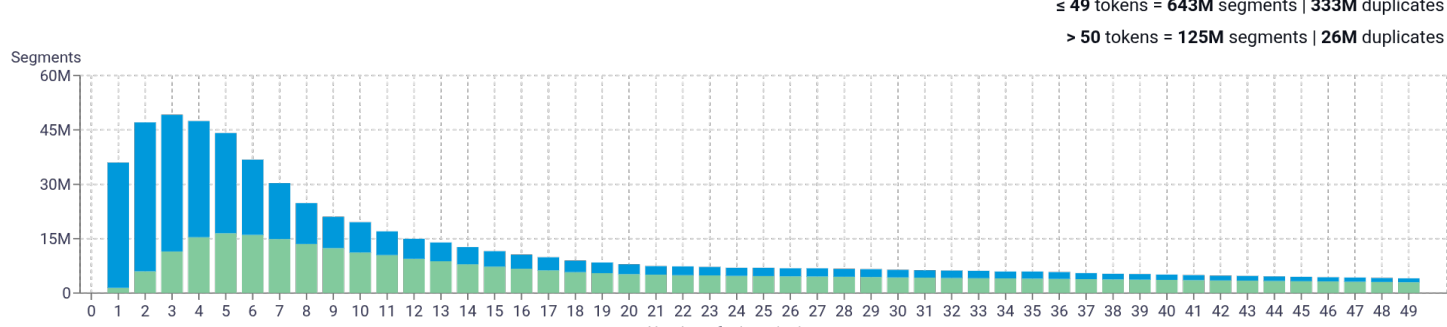
Percentage of segments in Slovak (sk) inside documents



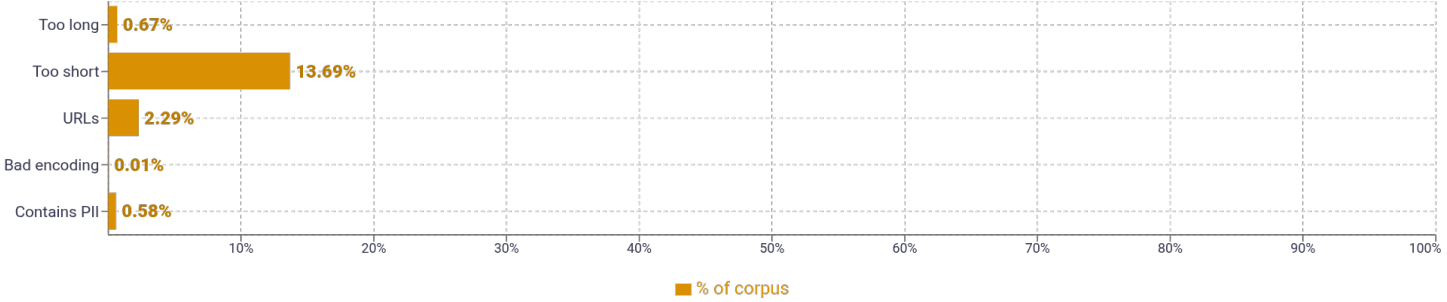
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	veľmi 20,386,560 roku 18,982,960 vás 15,289,491 všetky 15,018,136 u 14,149,847	
2	u nás 2,339,182 napriek tomu 2,302,121 slovenskej republiky 1,917,236 osobných údajov 1,797,176 celom svete 1,251,414	
3	zdarma v demoverzii 1,000,545 demoverzii a recenzia 1,000,544 hrací automat online 984,114 ponúkame na predaj 781,209 príspevok k tejto 578,776	
4	hranie zdarma v demoverzii 1,000,545 demoverzii a recenzia hry 1,000,544 napíše príspevok k tejto 578,443 príspevok k tejto položke 578,434 zmena a doplnení niektorých 340,638	
5	zdarma v demoverzii a recenzia 1,000,544 napíše príspevok k tejto položke 578,434 najdôležitejšie správy z východu slovenska 339,549 správy z východu slovenska čítajte 339,539 východu slovenska čítajte na korzar.sme.sk. 339,230	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 of 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>, , , 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				