

General overview

Corpus	Date	Language
hplt-v3-ckb_Arab	9/17/2025	Central Kurdish

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
352,126	4,979,111	4,089,498 (82.13 %)	170M	952,173,905	1.63 GB

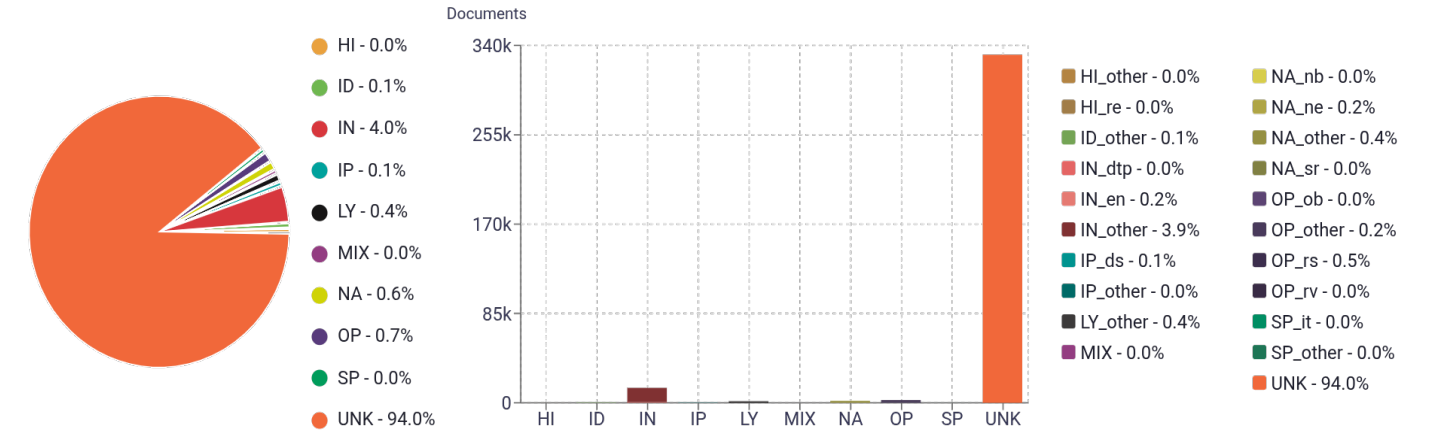
Top 10 domains

Domain	Docs	% of total
dengiamerika.com	29K	8.29%
kurdistan24.net	11K	3.15%
wishe.net	11K	3.13%
kurdistantv.net	11K	2.99%
komalah.org	10K	2.93%
dengekan.ca	7.3K	2.06%
hawlati.co	7.2K	2.04%
westganews.net	6.3K	1.78%
awene.com	5.3K	1.52%
payam.tv	5.1K	1.46%

Top 10 TLDs

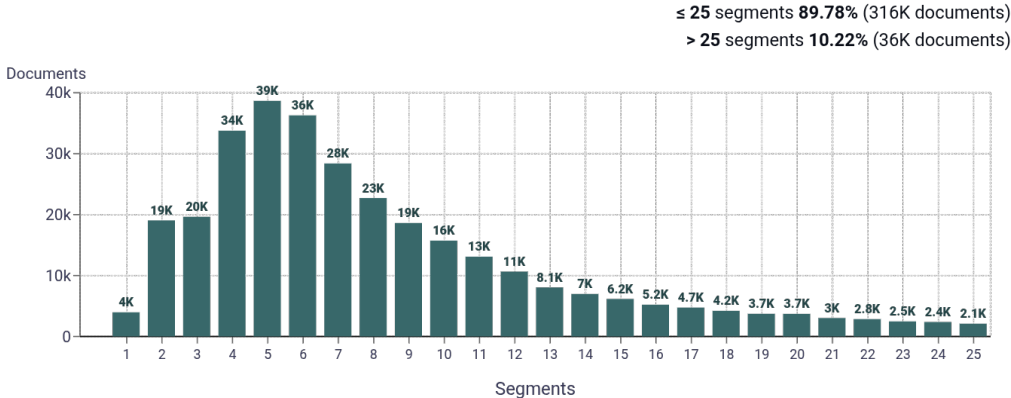
Domain	Docs	% of total
com	148K	42.09%
net	90K	25.47%
org	44K	12.63%
co	12K	3.48%
krd	9.8K	2.79%
tv	8.1K	2.30%
info	7.9K	2.25%
ca	7.3K	2.07%
live	4K	1.15%
ir	3.7K	1.06%

Register labels

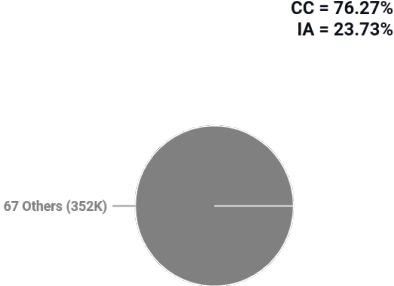


MT:69.9% | 246K Documents

Documents size (in segments)

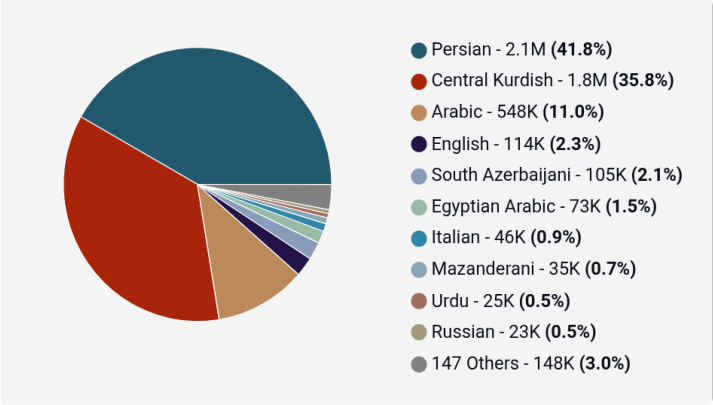


Document collections

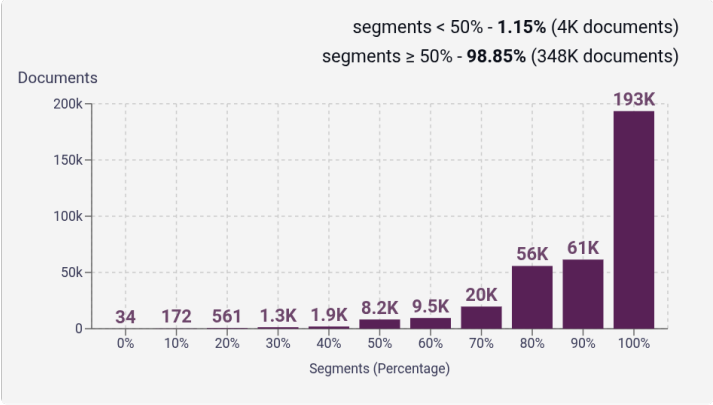


Language Distribution

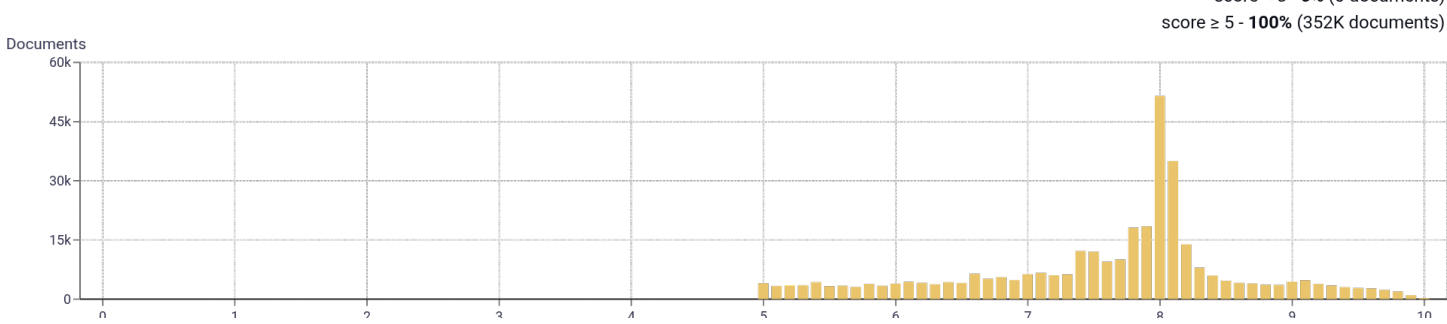
Number of segments in the Central Kurdish corpus



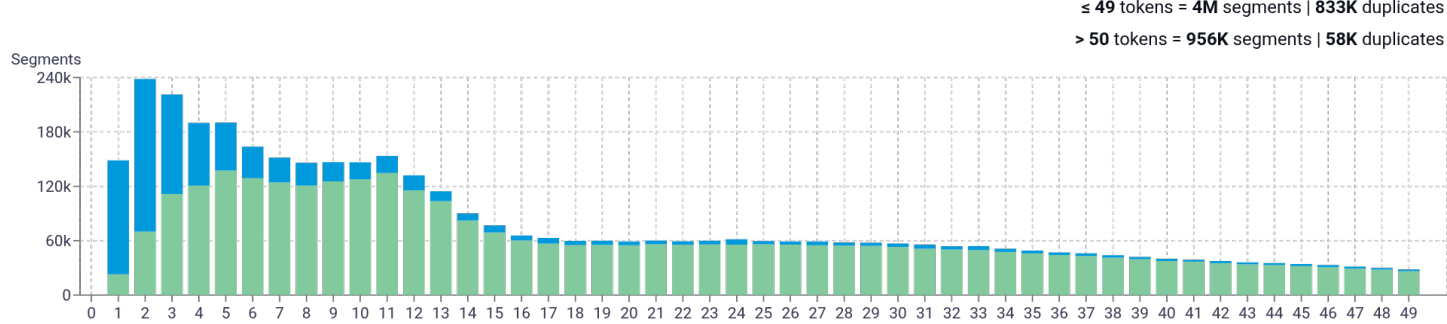
Percentage of segments in Central Kurdish inside documents



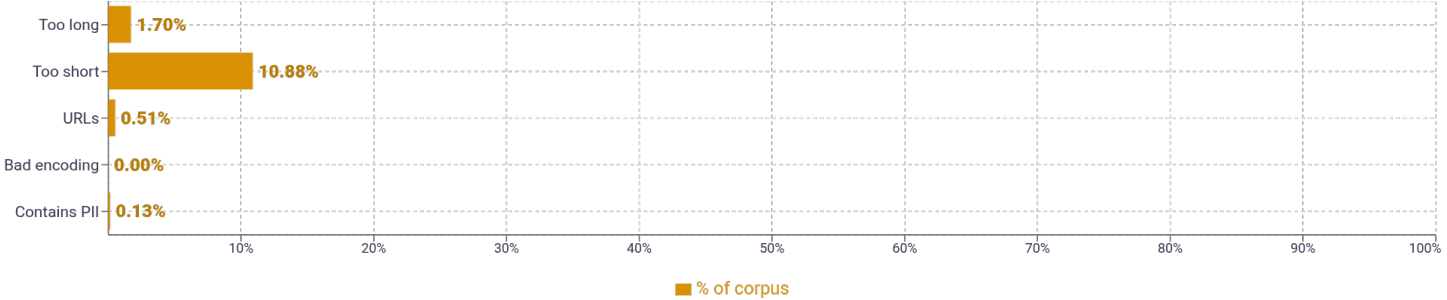
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	له 5,020,561 نههه 669,492 كه 810,884 نهو 1,314,667 كه 1,467,654 له 5,020,561	
2	له لايهن 57,756 له سهر 60,890 كه له 69,372 ههرتمى كوردستان 98,522 كه له 114,434	
3	صلى الله عليه 12,767 له ههرتمى كوردستان 13,745 حكومهتى ههرتمى كوردستان 15,337 صلى الله عليه 16,031 الله عليه وسلم 17,623	
4	برنگهى كليككردنى نهو فايهه 3,508 له برنگهى كليككردنى نهو 3,532 صلى الله عليه وسلم 12,113 صلى الله عليه وسلم 15,280 كۆمىنت بنوسه له فهىسيوك 3,405	
5	لترهوه كۆمىنت بنوسه له فهىسيوك 3,405 كۆمىنت بنوسه له فهىسيوك دهردهكهوئ 3,405 له برنگهى كليككردنى نهو فايهه 3,507 كوردستان و عىراق و ناوچهكه 2,250 برنگهى كليككردنى نهو فايهه دهنگىيانهى 2,259	

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				