

General overview

Corpus	Date	Language
hplt-v3-kmr_Latn	9/18/2025	Northern Kurdish

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
693,889	12,058,818	9,657,818 (80.09 %)	394M	1,951,981,773	2 GB

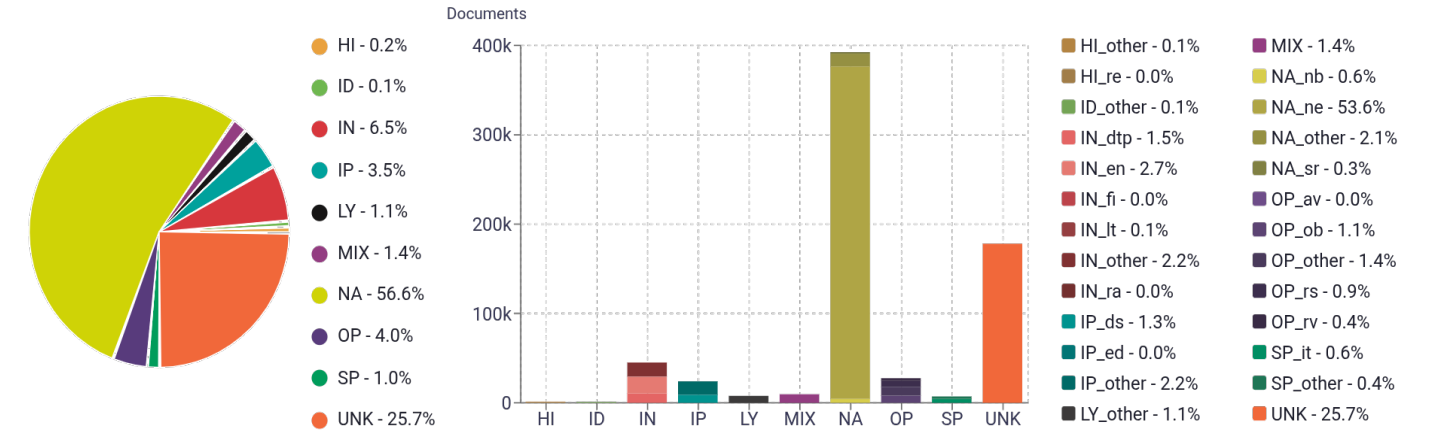
Top 10 domains

Domain	Docs	% of total
dengeamerika.com	31K	4.45%
ronahi.tv	31K	4.41%
rojevakurd.com	23K	3.31%
trtnuce.com	20K	2.93%
anfkurdi.com	20K	2.83%
kurdistan24.net	16K	2.32%
hawarnews.com	14K	2.08%
wikipedia.org	13K	1.94%
diyarname.com	11K	1.55%
denge-welat.org	11K	1.53%

Top 10 TLDs

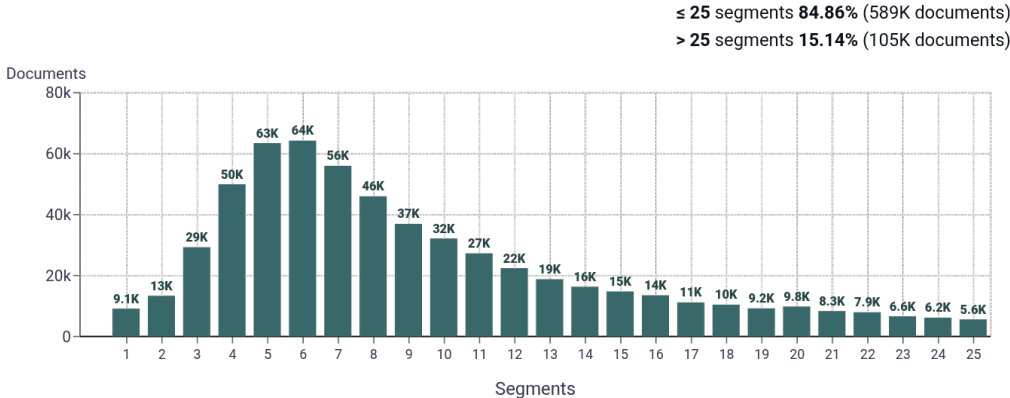
Domain	Docs	% of total
com	421K	60.69%
org	83K	11.96%
net	67K	9.60%
tv	35K	4.99%
info	13K	1.93%
am	11K	1.54%
com.tr	9.8K	1.42%
zone	5.2K	0.74%
ir	3.8K	0.54%
de	3.7K	0.54%

Register labels

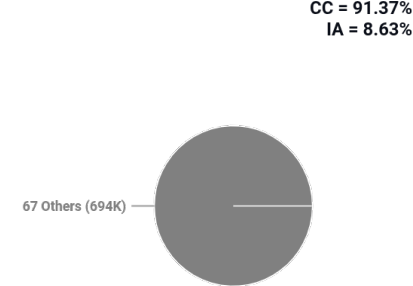


MT:20.3% | 141K Documents

Documents size (in segments)

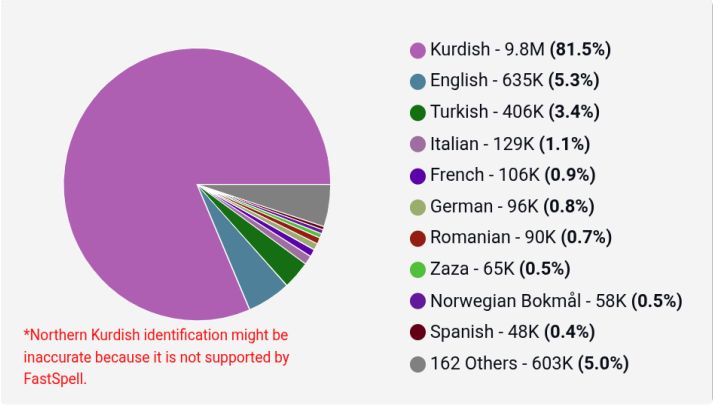


Document collections

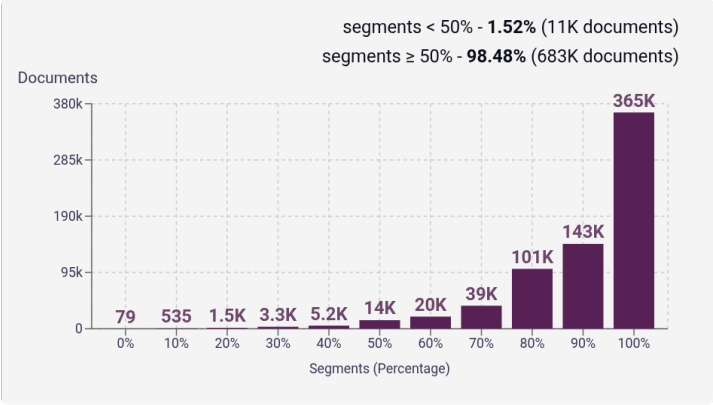


Language Distribution

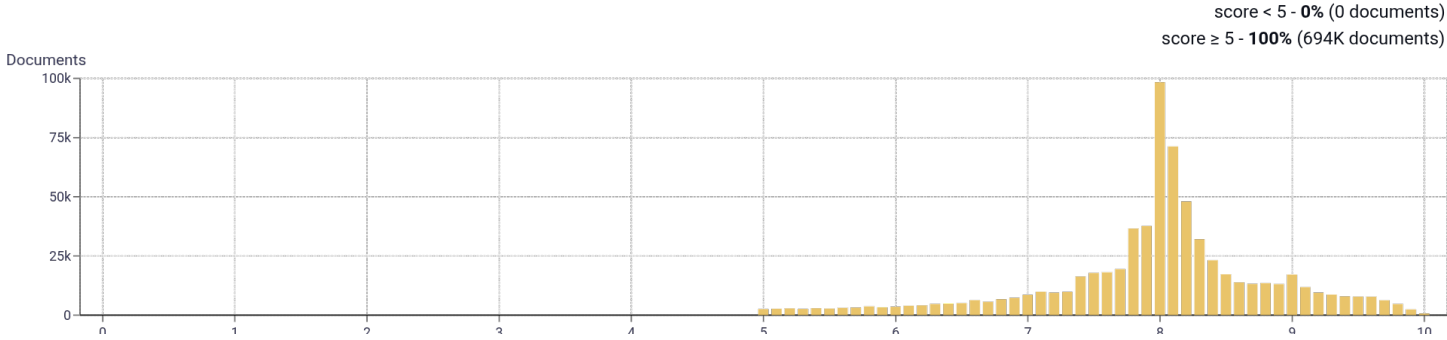
Number of segments in the Northern Kurdish corpus



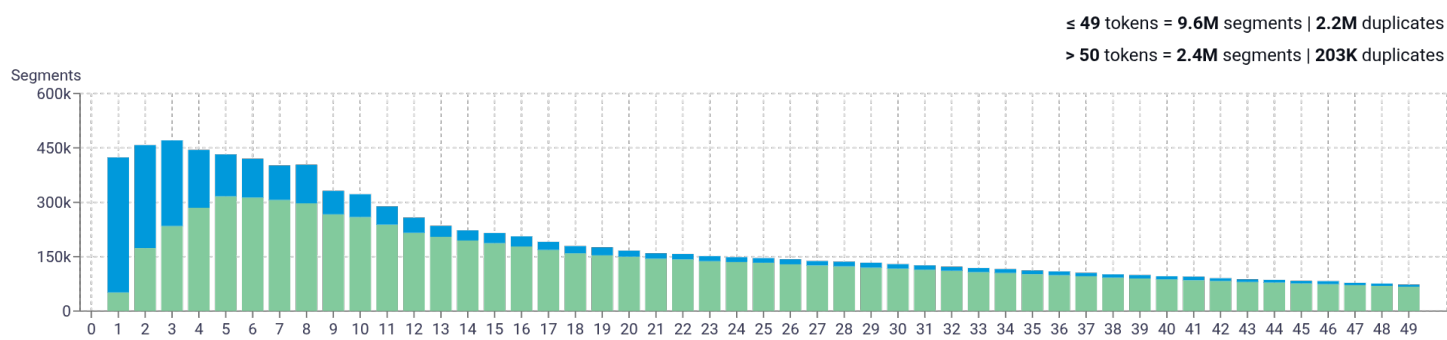
Percentage of segments in Northern Kurdish inside documents



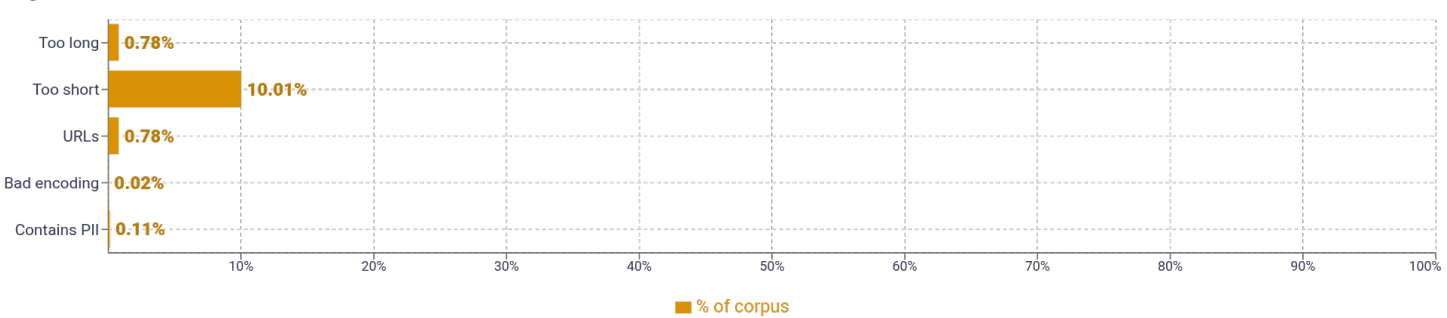
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	bikin 1,222,649 kirin 1,104,768 dike 1,098,050 dikin 795,893 bike 794,737	
2	dewleta tirk 159,035 herêma kurdistanê 155,283 bikar anîn 103,886 bikar binin 80,510 zimanê kurdî 69,425	
3	tirk a dagirker 39,468 bakur û rojhilatê 38,883 şert û mercên 29,565 rû bi rû 28,048 dest pê dike 26,106	
4	jiyana xwe ji dest 44,004 dewleta tirk a dagirker 24,972 bakur û rojhilatê sûriyê 23,037 hêzên çekdar ên ukraynayê 22,031 serfermandariya giştî ya hêzên 20,557	
5	serfermandariya giştî ya hêzên çekdar 20,555 rêberê gelê kurd abdullah ocalan 17,428 jiyana xwe ji dest dan 11,425 yekem be ku şîrove bike 10,766 beşa yekem be ku şîrove 10,764	

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