

General overview

Corpus	Date	Language
hplt-v3-tur_Latn	9/19/2025	Turkish (tr)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
159,466,598	3,113,393,511	1,658,134,100 (53.26 %)	79B	509,655,706,808	518.56 GB

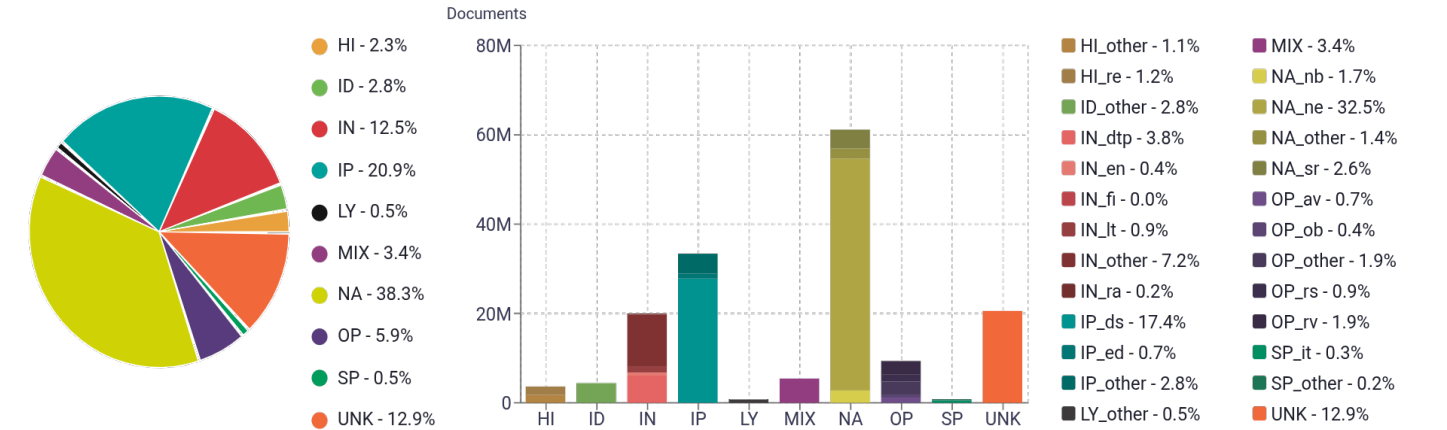
Top 10 domains

Domain	Docs	% of total
blogspot.com	2M	1.26%
sikayetvar.com	778K	0.49%
hurriyet.com.tr	768K	0.48%
blogspot.com.tr	734K	0.46%
sabah.com.tr	642K	0.40%
docplayer.biz.tr	607K	0.38%
haberler.com	546K	0.34%
milliyet.com.tr	541K	0.34%
kanalahaber.com	465K	0.29%
wordpress.com	382K	0.24%

Top 10 TLDs

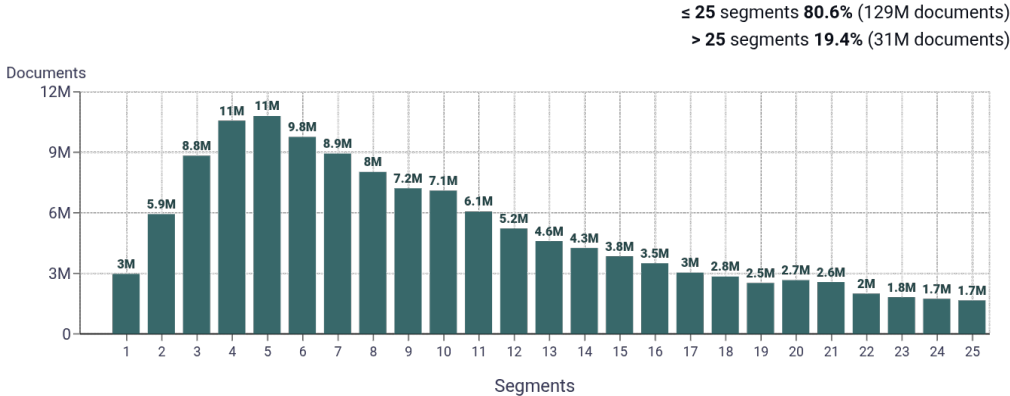
Domain	Docs	% of total
com	100M	62.56%
com.tr	21M	13.23%
net	13M	8.19%
org	7.7M	4.81%
org.tr	1.4M	0.90%
xyz	1.1M	0.68%
info	949K	0.60%
gen.tr	900K	0.56%
edu.tr	873K	0.55%
biz.tr	832K	0.52%

Register labels

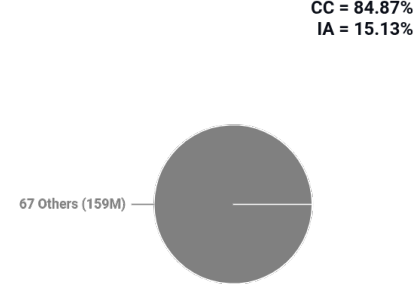


MT:8.5% | 14M Documents

Documents size (in segments) ⓘ

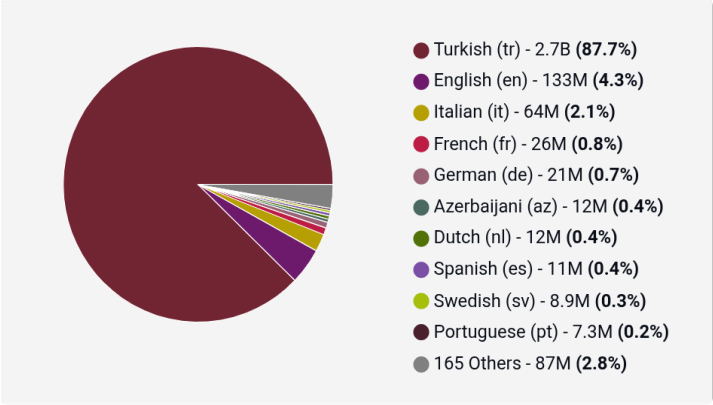


Document collections

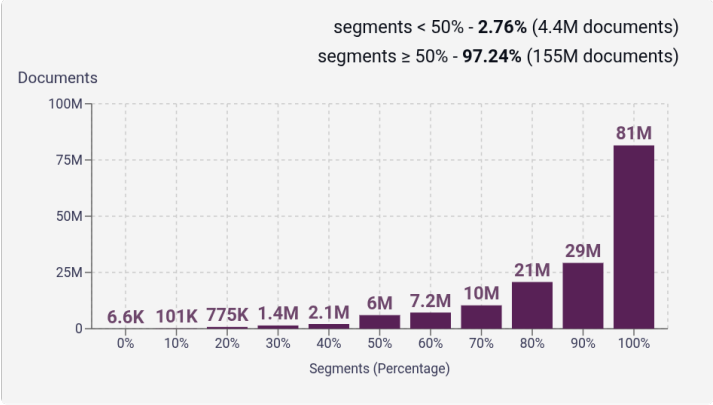


Language Distribution

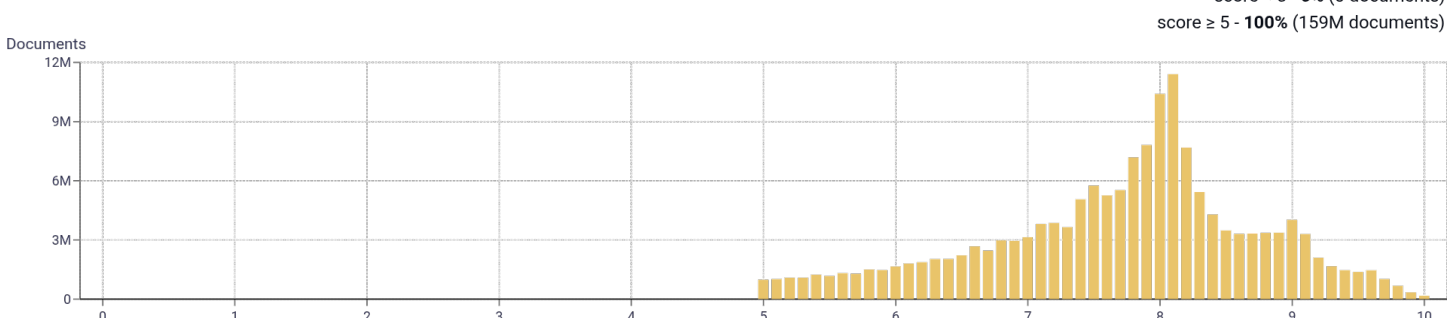
Number of segments in the Turkish (tr) corpus



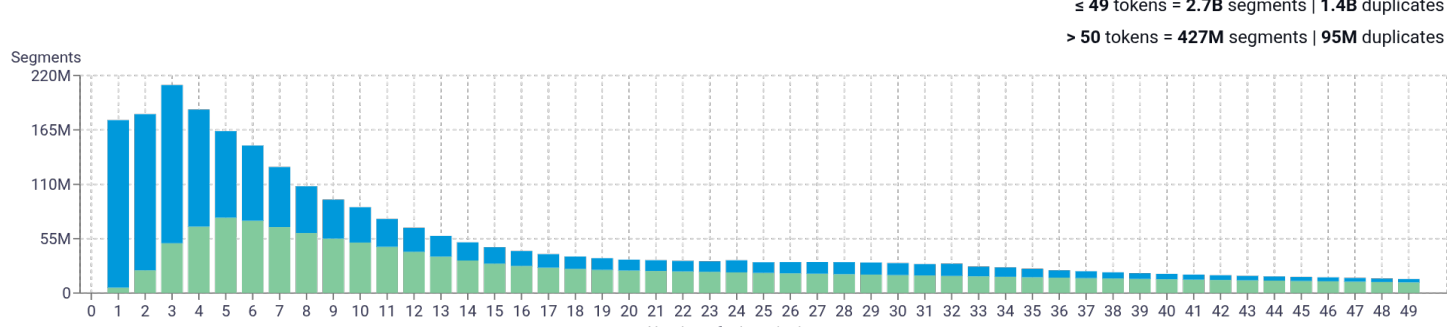
Percentage of segments in Turkish (tr) inside documents



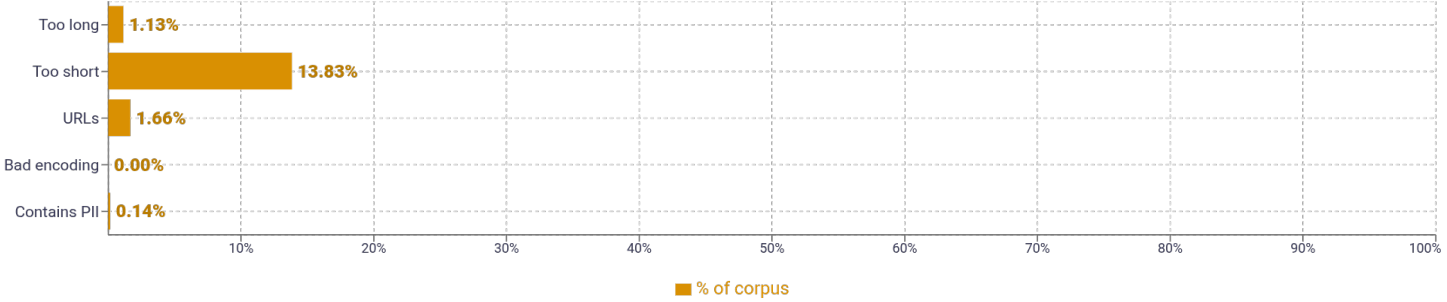
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>kadar   137,370,641</div> <div>yeni   127,473,329</div> <div>sonra   121,984,027</div> <div>büyük   108,721,923</div> <div>şekilde   100,267,052</div>	
2	<div>aynı zamanda   25,160,883</div> <div>yer alan   22,536,563</div> <div>olmak üzere   20,308,548</div> <div>yanı sıra   17,837,756</div> <div>söz konusu   12,533,051</div>	
3	<div>dahil olmak üzere   5,566,176</div> <div>hızlı bir şekilde   4,716,787</div> <div>hiçbir şekilde sorumlu   3,879,272</div> <div>şekilde sorumlu tutulamaz   3,851,584</div> <div>doğrudan veya dolaylı   3,786,738</div>	
4	<div>hiçbir şekilde sorumlu tutulamaz   3,850,051</div> <div>ilgili doğrudan veya dolaylı   3,646,573</div> <div>dolaylı tüm sorumluluğu tek   3,646,110</div> <div>sorumluluğu tek başınıza üstleniyorsunuz   3,646,089</div> <div>etmiş bulunuyor ve yorumunuzla   3,442,739</div>	
5	<div>doğrudan veya dolaylı tüm sorumluluğu   3,646,183</div> <div>dolaylı tüm sorumluluğu tek başınıza   3,646,076</div> <div>yorumunuzla ilgili doğrudan veya dolaylı   3,645,546</div> <div>kabul etmiş bulunuyor ve yorumunuzla   3,441,568</div> <div>etmiş bulunuyor ve yorumunuzla ilgili   3,440,802</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 \*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>, <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.	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				