HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
tur_Latn.jsonl.tsv	8/2/2025	Turkish (tr)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
116,565,998	2,573,945,640	950,713,593 (36.94 %)	61B	387,178,346,456	394.5 GB

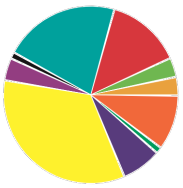
Top 10 domains

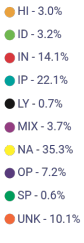
Domain	Docs	% of total
blogspot.com	1.7M	1.47%
blogspot.com.tr	1.5M	1.26%
hurriyet.com.tr	1.4M	1.17%
wikipedia.org	580K	0.50%
docplayer.biz.tr	529K	0.45%
sabah.com.tr	485K	0.42%
sikayetvar.com	454K	0.39%
haberler.com	422K	0.36%
eksisozluk.com	396K	0.34%
haberaktuel.com	393K	0.34%

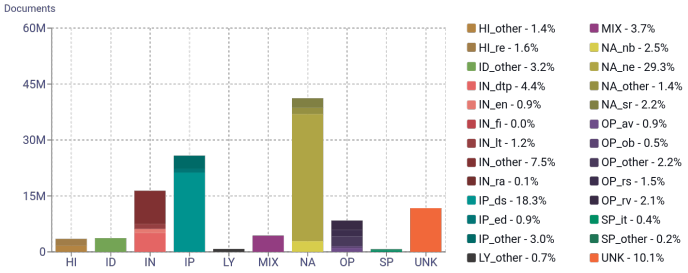
Top 10 TLDs

Domain	Docs	% of total
com	76M	64.79%
com.tr	13M	11.02%
net	10M	8.89%
org	6.1M	5.19%
org.tr	1.1M	0.91%
biz.tr	774K	0.66%
info	723K	0.62%
gen.tr	691K	0.59%
gov.tr	683K	0.59%
edu.tr	678K	0.58%

Register labels



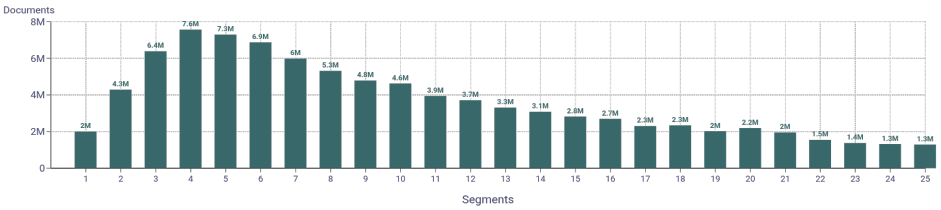




MT:5.1% | 6M Documents

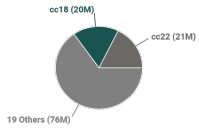
Documents size (in segments)

<= 25 segments **78.06%** (91M documents)
> 25 segments **21.94%** (26M documents)



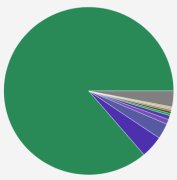
Documents by collection

CC = 57.62%
IA = 42.38%

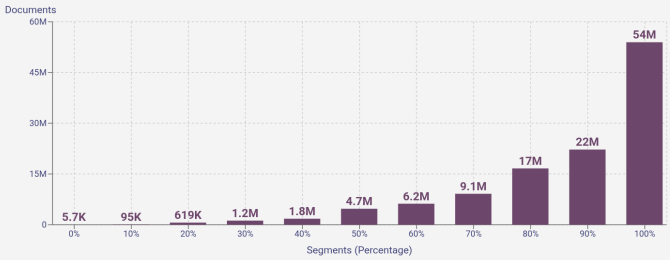


Language Distribution

Number of segments in the Turkish (tr) corpus

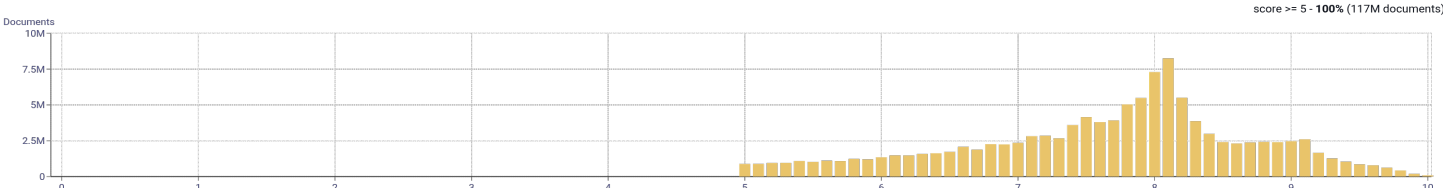


Percentage of segments in Turkish (tr) inside documents



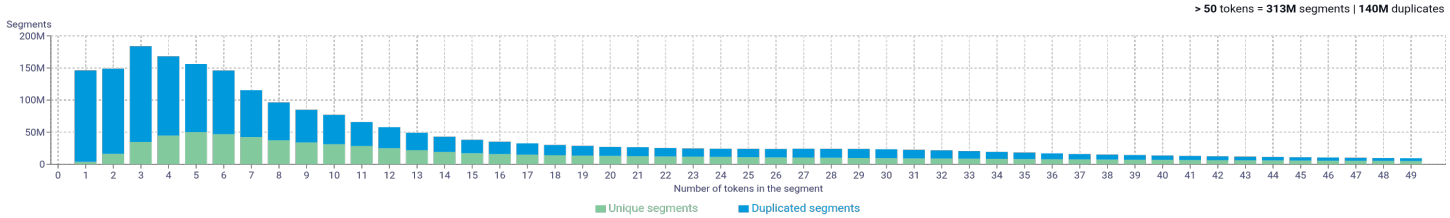
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (117M documents)

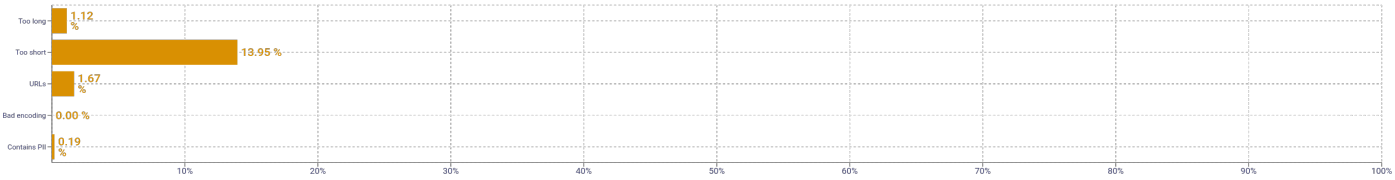


Segment length distribution by token

≤ 49 tokens = **2.3B** segments | **1.5B** duplicates
> 50 tokens = **313M** segments | **140M** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	kadar | 109545129 · sonra | 101920662 · yeni | 87121091 · büyük | 80820284 · a | 72290561
2	yer alan | 16033898 · olmak üzere | 13902631 · aynı zamanda | 13217097 · yanı sıra | 9806686 · söz konusu | 9643975
3	recep tayyip erdoğan | 2616712 · başta olmak üzere | 2336301 · yönetim kurulu başkanı | 2023839 · türkçe karakter kullanılmayan | 1906858 · büyük harflerle yazılmış | 1883535
4	büyük harflerle yazılmış yorumlar | 1880371 · karakter kullanılmayan ve büyük | 1821461 · kullanılmayan ve büyük harflerle | 1821455 · harflerle yazılmış yorumlar onaylanmamaktadır | 1813817 · cumhurbaşkanı recep tayyip erdoğan | 1006823

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 or 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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	**HI**
Recipe	re
Informational persuasion	**IP**
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	**IN**
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	**OP**
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av