HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-snd_Arab	9/18/2025	Sindhi

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
363,829	6,274,353	5,162,379 (82.28 %)	261M	1,087,312,087	1.78 GB

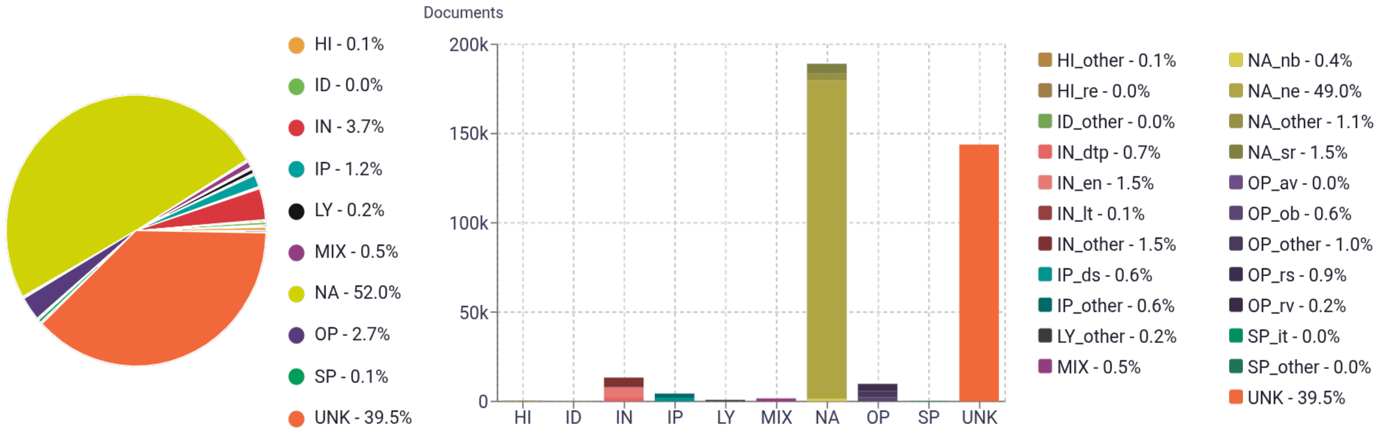
Top 10 domains

Domain	Docs	% of total
awamiawaz.pk	42K	11.47%
ktnnews.tv	17K	4.60%
awamiawaz.com	16K	4.42%
thetimenews.tv	16K	4.32%
dailysindhyar.com	13K	3.60%
pahenjiakhbar.com	13K	3.59%
sindhexpress.co...	12K	3.26%
dailysarwan.com	12K	3.17%
voiceofsindh.co...	8.4K	2.30%
androidsis.com	7.3K	2.00%

Top 10 TLDs

Domain	Docs	% of total
com	209K	57.37%
pk	47K	12.80%
tv	38K	10.35%
com.pk	29K	7.94%
org	14K	3.75%
net	9.5K	2.60%
zone	4.6K	1.27%
fr	1.3K	0.36%
ru	1.1K	0.31%
co.uk	776	0.21%

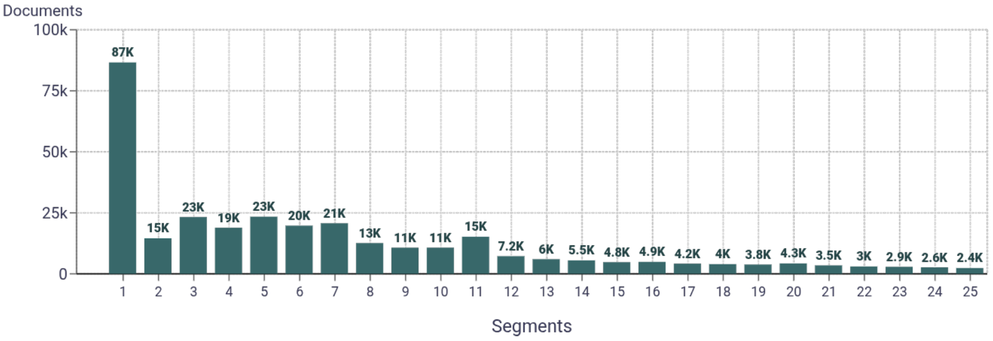
Register labels



MT:37.4% | 136K Documents

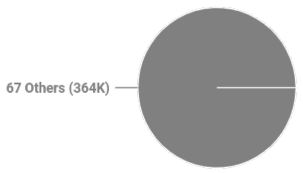
Documents size (in segments)

≤ 25 segments **86.73%** (316K documents)
> 25 segments **13.27%** (48K documents)



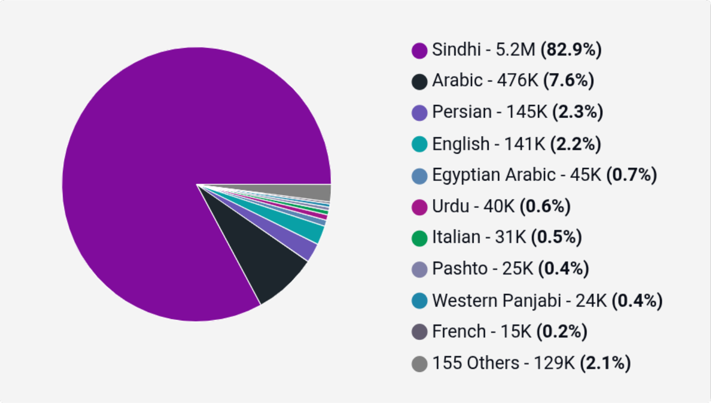
Document collections

CC = 97.65%
IA = 2.35%



Language Distribution

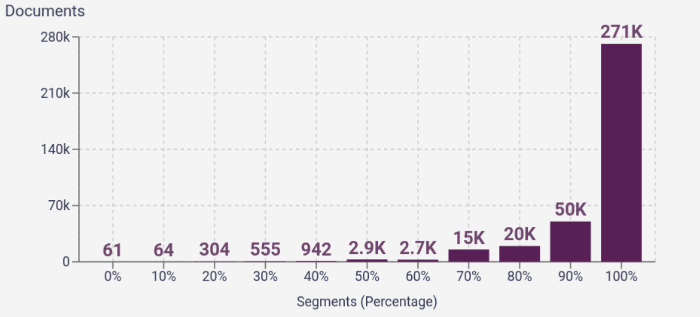
Number of segments in the Sindhi corpus



Percentage of segments in Sindhi inside documents

segments < 50% - **0.53%** (1.9K documents)

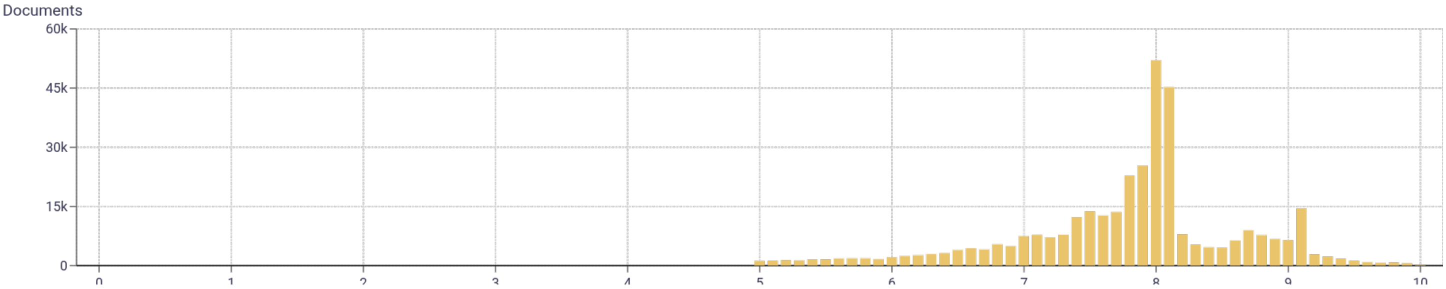
segments ≥ 50% - **99.47%** (362K documents)



Distribution of documents by document score

score < 5 - **0%** (0 documents)

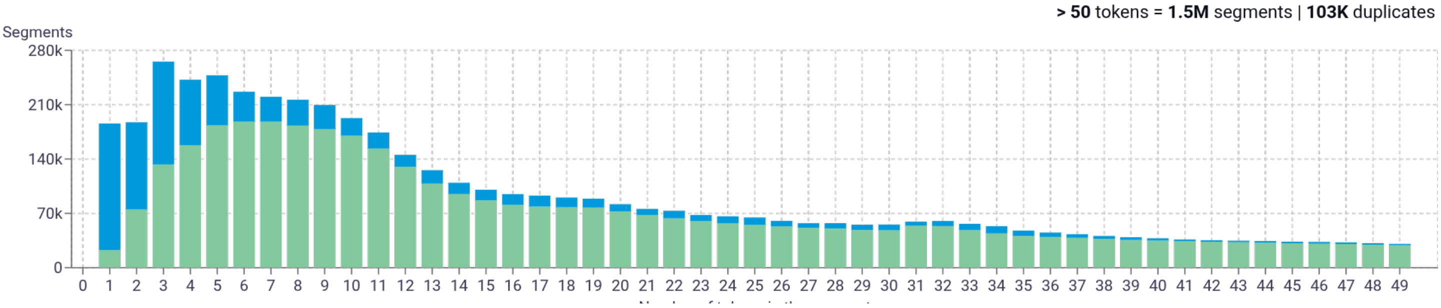
score ≥ 5 - **100%** (364K documents)



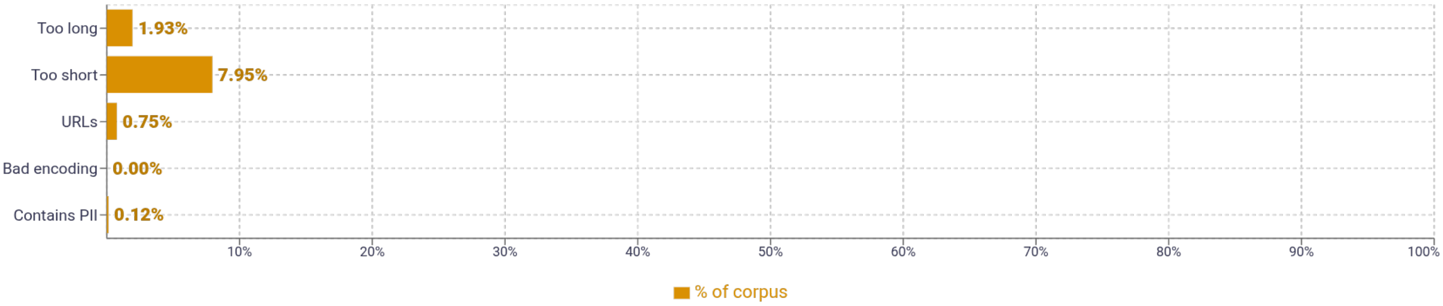
Segment length distribution by token

≤ **49** tokens = **4.7M** segments | **1M** duplicates

> 50 tokens = **1.5M** segments | **103K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	1,354,554 | نه 1,082,601 | ڪيو 859,316 | جا 629,232 | طور 510,207 | ڪئي
2	110,766 | ڪئي وئي 90,187 | اسلام آباد 85,569 | ڪيو ويندو 81,426 | ڪيو وڃي 78,105 | عمران خان
3	36,139 | پي ٽي آء 20,193 | استعمال ڪيو ويندو 16,392 | ايس ايس پي 15,426 | بلاول ڀٽو زرداري 14,641 | مراد علي شاهه
4	9,705 | سنڌ جي وڏي وزير 7,219 | سيد مراد علي شاهه 6,736 | چيئرمين بلاول ڀٽو زرداري 6,712 | وزير سيد مراد علي 6,199 | وڏي وزير سيد مراد
5	6,105 | وڏي وزير سيد مراد علي 5,767 | وزير سيد مراد علي شاهه 4,856 | ڊبليو ڊبليو ڊبليو ڊبليو ڊبليو 4,691 | سنڌ جي وڏي وزير سيد 3,830 | پ پ چيئرمين بلاول ڀٽو

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.	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				