HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
pbt_Arab.jsonl.tsv	9/20/2024	Southern Pashto (pbt)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
466,472	8,454,662	5,506,825 (65.13 %)	306M	1,295,601,474	2.13 GB

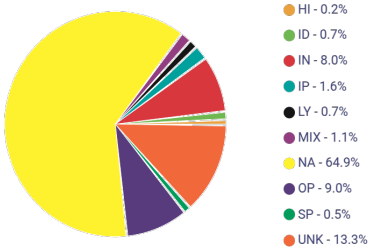
Top 10 domains

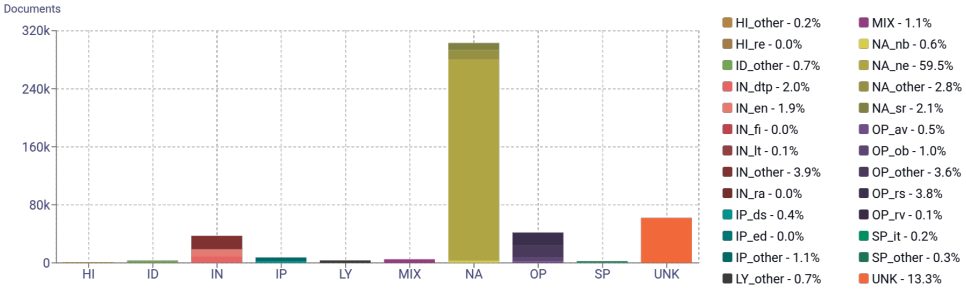
Domain	Docs	% of total
pashtovoa.com	38K	8.10%
mashaalradio.com	32K	6.81%
bakhtarnews.af	28K	6.01%
tolonews.com	24K	5.23%
nunn.asia	15K	3.21%
tolafghan.com	13K	2.89%
wikipedia.org	12K	2.63%
larawbar.net	12K	2.48%
taand.com	11K	2.44%
dw.com	10K	2.25%

Top 10 TLDs

Domain	Docs	% of total
com	288K	61.68%
af	49K	10.45%
gov.af	25K	5.33%
net	24K	5.11%
org	23K	4.93%
asia	15K	3.23%
cn	6.8K	1.46%
website	4.6K	0.98%
info	3.5K	0.74%
ir	2.7K	0.57%

Register labels

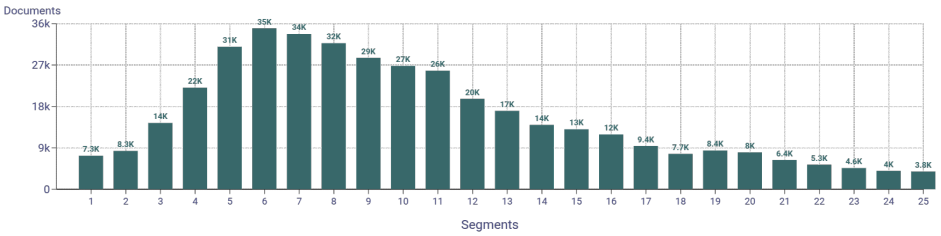




MT:7.3% | 34K Documents

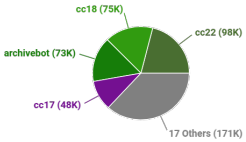
Documents size (in segments)

<= 25 segments **85.31%** (398K documents)
> 25 segments **14.69%** (69K documents)



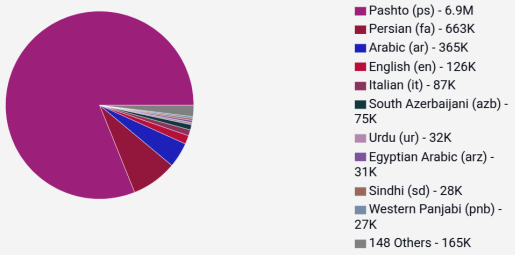
Documents by collection

CC = 60.73%
IA = 39.27%



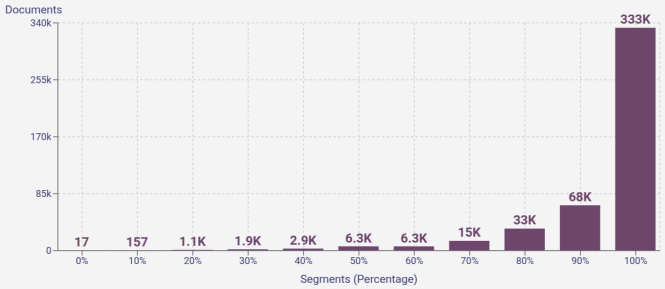
Language Distribution

Number of segments in the Southern Pashto (pbt) corpus



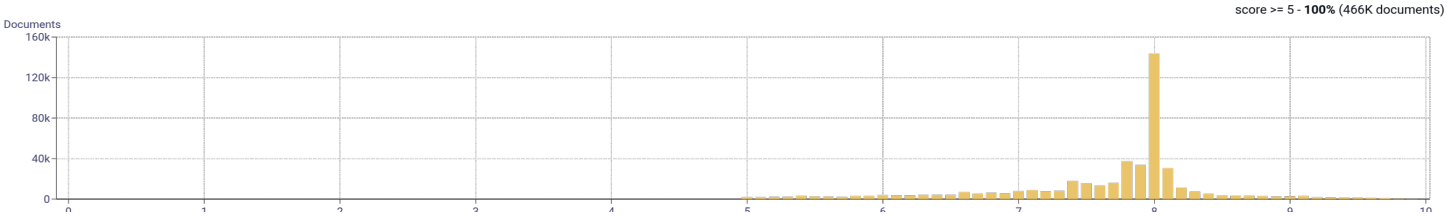
*Southern Pashto (pbt) identification might be inaccurate because it is not supported by FastSpell

Percentage of segments in Southern Pashto (pbt) inside documents



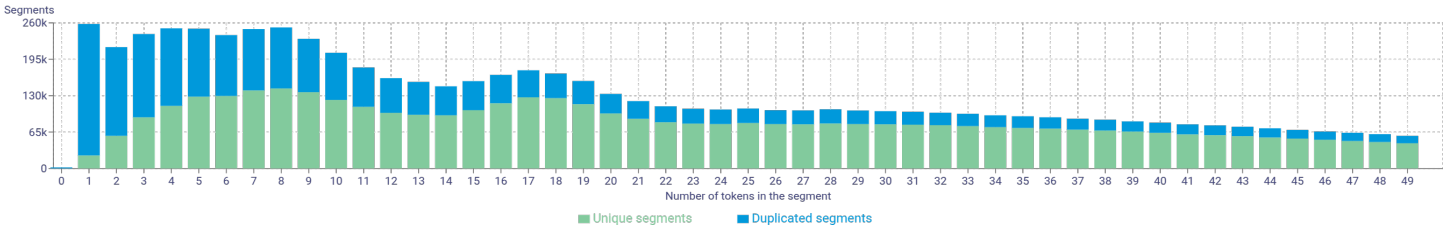
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (466K documents)

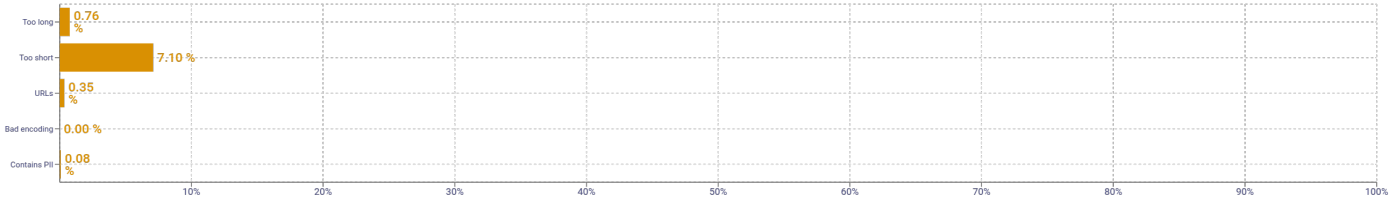


Segment length distribution by token

≤ 49 tokens = **4.2M** segments | **2.4M** duplicates
> 50 tokens = **1.9M** segments | **520K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	5754718 | چې 5332453 | کې 1706720 | دی 1624373 | دې 1341005 | يې
2	193226 | افغانستان کې 188791 | دې چې 130557 | برخه کې 116629 | کې چې 110752 | حال کې
3	56303 | چې په دې 51729 | چې د افغانستان 39227 | حال کې چې 36106 | کې د افغانستان 31550 | چې د دې
4	28146 | چې په افغانستان کې 25413 | صلی الله علیه وسلم 16442 | صلی الله علیه وسلم 14219 | افغانستان د اسلامي جمهوریت 13877 | صلی الله علیه وسلم
5	12440 | رسول الله صلی الله علیه 8720 | رسول الله صلی الله علیه 7995 | رسول الله صلی الله علیه 4799 | جمهور رئیس محمد اشرف غني 4178 | کې د پښتو او بلوڅو

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