

General overview

| Corpus | Date | Language |
|------------------|-----------|-----------------------|
| hplt-v3-pbt_Arab | 9/24/2025 | Southern Pashto (pbt) |

Volumes

| Docs | Segments | Unique segments | Tokens | Characters | Size |
|---------|------------|----------------------|--------|---------------|---------|
| 918,708 | 15,745,595 | 12,584,332 (79.92 %) | 577M | 2,431,797,440 | 3.99 GB |

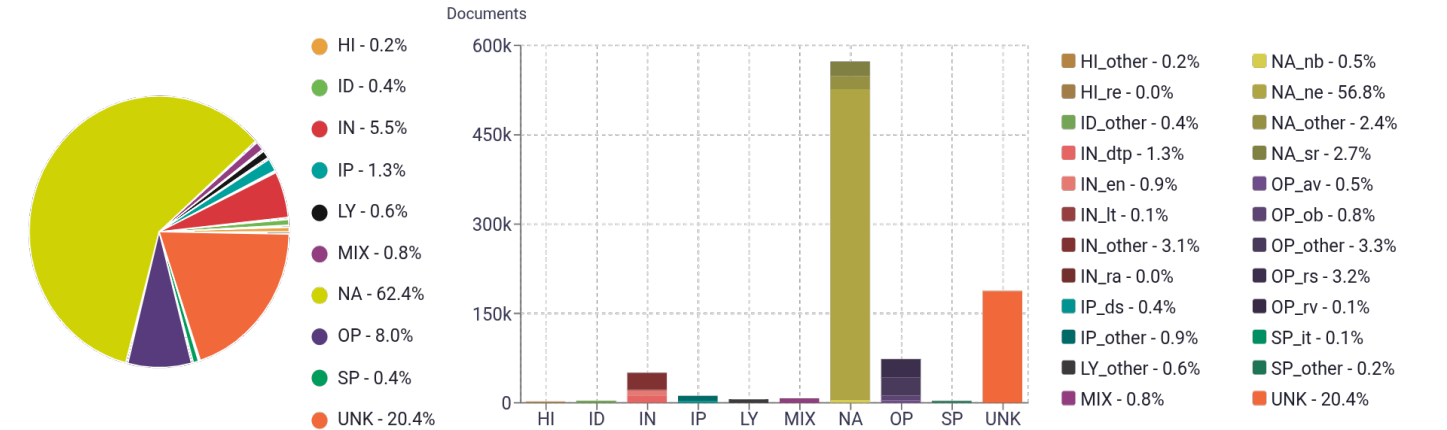
Top 10 domains

| Domain | Docs | % of total |
|------------------|------|------------|
| nunn.asia | 61K | 6.66% |
| pashtovoa.com | 44K | 4.74% |
| mashaalradio.com | 40K | 4.34% |
| azadiradio.com | 29K | 3.19% |
| voadeewanews.com | 26K | 2.84% |
| taand.com | 23K | 2.51% |
| bbc.com | 21K | 2.31% |
| bakhtarnews.af | 19K | 2.07% |
| tolafghan.com | 18K | 2.00% |
| dw.com | 18K | 1.95% |

Top 10 TLDs

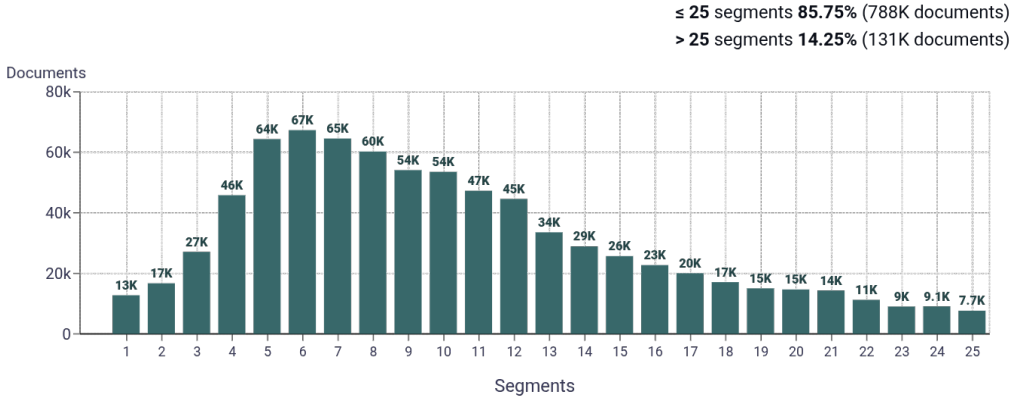
| Domain | Docs | % of total |
|--------|------|------------|
| com | 583K | 63.47% |
| af | 66K | 7.14% |
| asia | 62K | 6.72% |
| net | 45K | 4.91% |
| org | 28K | 3.04% |
| gov.af | 27K | 2.89% |
| cn | 15K | 1.61% |
| tv | 11K | 1.24% |
| com.af | 7.8K | 0.85% |
| co | 6.3K | 0.69% |

Register labels

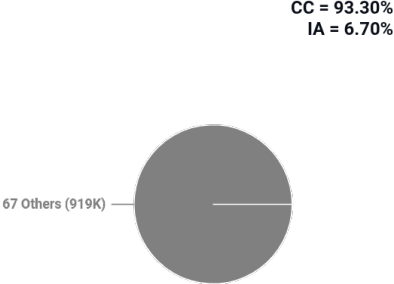


MT:15.8% | 145K Documents

Documents size (in segments) ⓘ

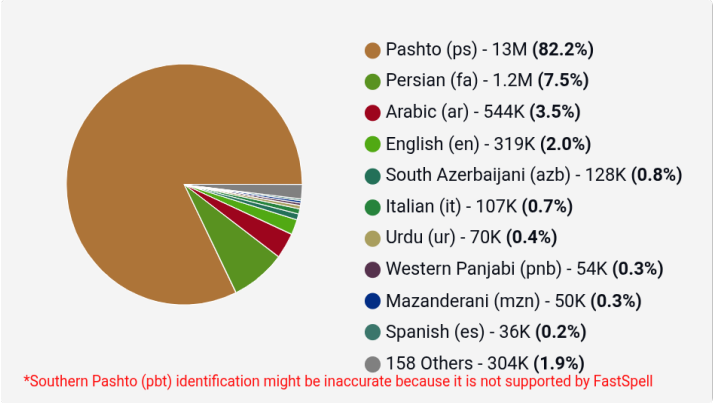


Document collections

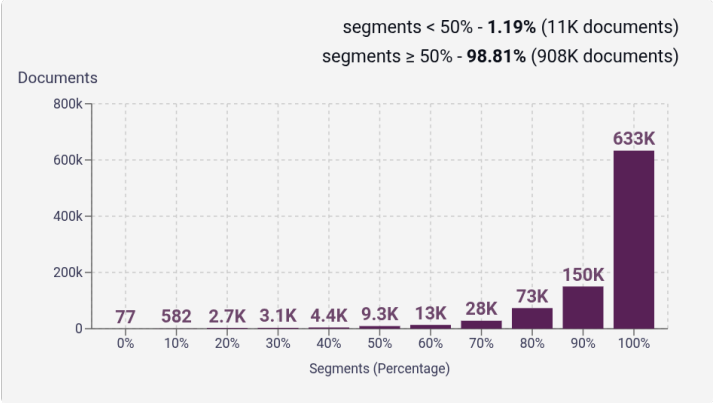


Language Distribution

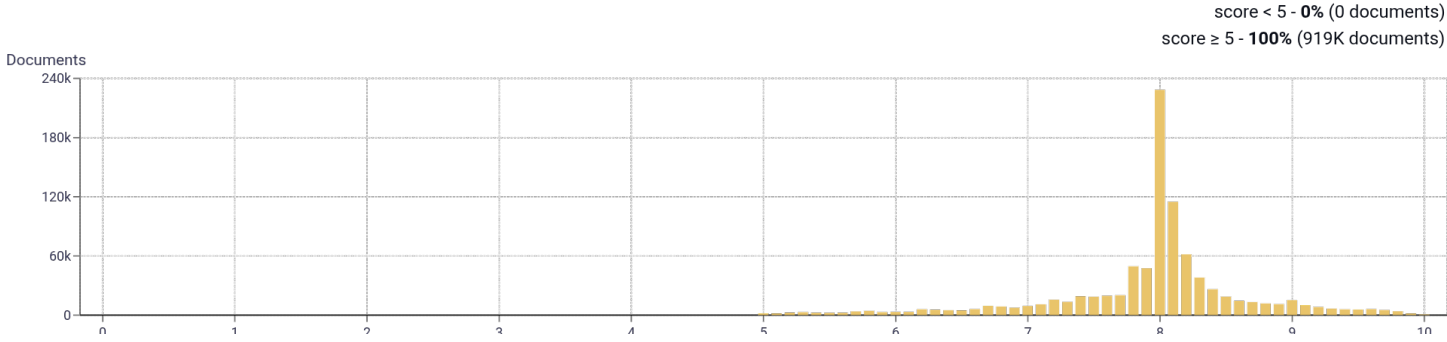
Number of segments in the Southern Pashto (pbt) corpus



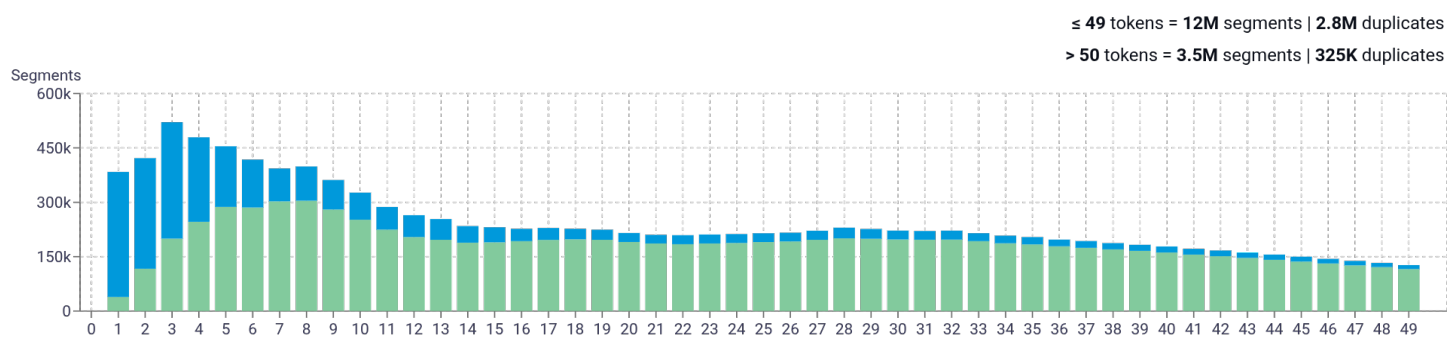
Percentage of segments in Southern Pashto (pbt) inside documents



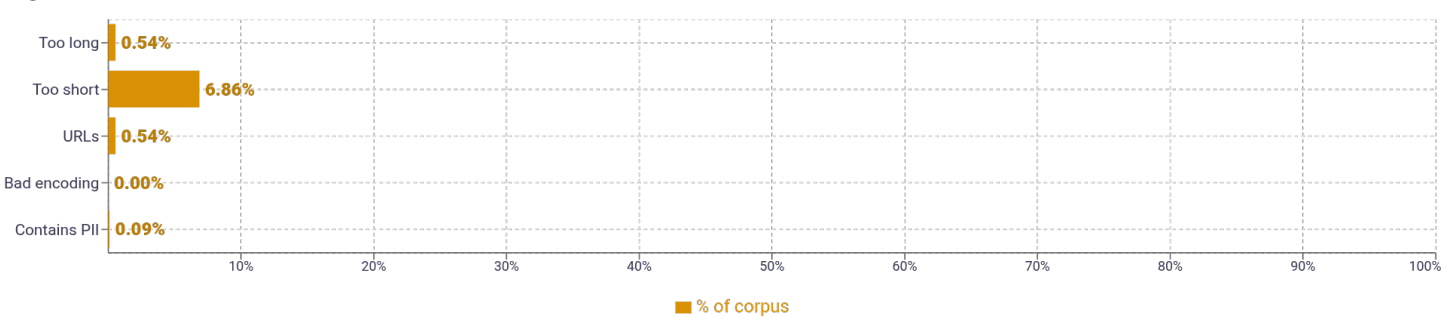
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

| SIZE | N-GRAMS | |
|------|--|---|
| 1 | 11,437,703 چې 10,860,973 کې 3,154,710 دي 3,086,039 دي 2,723,042 يې | 📄 |
| 2 | 365,914 افغانستان کې 356,822 دي چې 310,129 چې تاسو 245,306 حال کې 240,102 کې چې | 📄 |
| 3 | 121,664 تاسو کولی شئ 102,433 حال کې چې 100,235 چې په دي 90,824 چې د افغانستان 75,747 چې د دي | 📄 |
| 4 | 54,826 چې په افغانستان کې 49,411 پداسې حال کې چې 46,164 صلى الله عليه وسلم 27,568 حال کې ده چې 19,260 صلى الله عليه وسلم | 📄 |
| 5 | 26,219 رسول الله صلى الله عليه 11,407 رسول الله صلى الله عليه 10,201 آژانس د خبر له مخې 8,548 اړه څه نه دي ويلي 7,540 جمهور خبري آژانس د خبر | 📄 |

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 | ntp |
| 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 | | | | |