

General overview

| Corpus           | Date      | Language |
|------------------|-----------|----------|
| hplt-v3-shn_Mymr | 9/18/2025 | Shan     |

Volumes

| Docs   | Segments | Unique segments   | Tokens | Characters | Size      |
|--------|----------|-------------------|--------|------------|-----------|
| 12,287 | 157,377  | 130,406 (82.86 %) | 2.8M   | 38,351,084 | 100.42 MB |

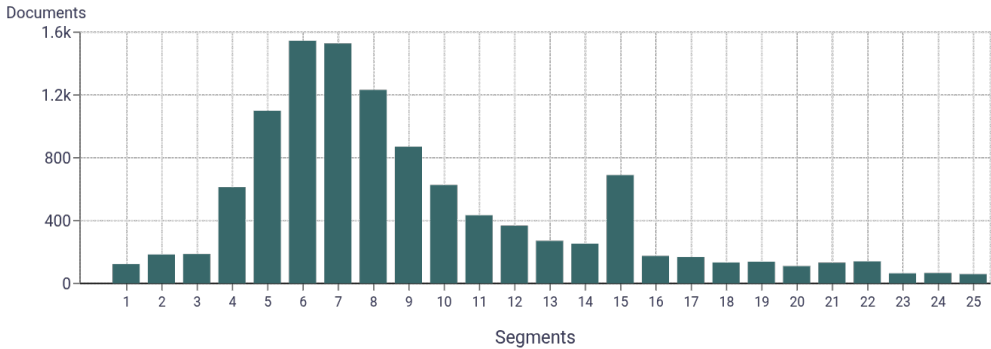
Top 10 domains

| Domain             | Docs | % of total |
|--------------------|------|------------|
| shannews.org       | 5.4K | 44.33%     |
| taifreedom.com     | 1.9K | 15.20%     |
| wikipedia.org      | 1.2K | 9.83%      |
| ssppssa.org        | 787  | 6.41%      |
| panglong.org       | 407  | 3.31%      |
| wikimedia.org      | 393  | 3.20%      |
| taiimages.com      | 273  | 2.22%      |
| bible.is           | 197  | 1.60%      |
| taifreedom.net     | 133  | 1.08%      |
| shanhumanrights... | 122  | 0.99%      |

Top 10 TLDs

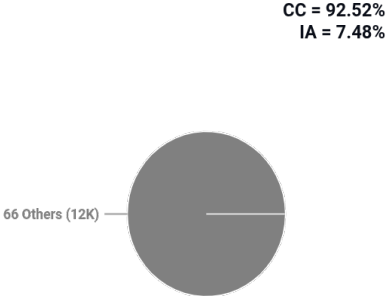
| Domain | Docs | % of total |
|--------|------|------------|
| org    | 8.8K | 71.99%     |
| com    | 2.9K | 23.45%     |
| is     | 197  | 1.60%      |
| net    | 171  | 1.39%      |
| info   | 57   | 0.46%      |
| media  | 50   | 0.41%      |
| org.mm | 29   | 0.24%      |
| blog   | 25   | 0.20%      |
| com.mm | 6    | 0.05%      |
| space  | 5    | 0.04%      |

Documents size (in segments) ⓘ



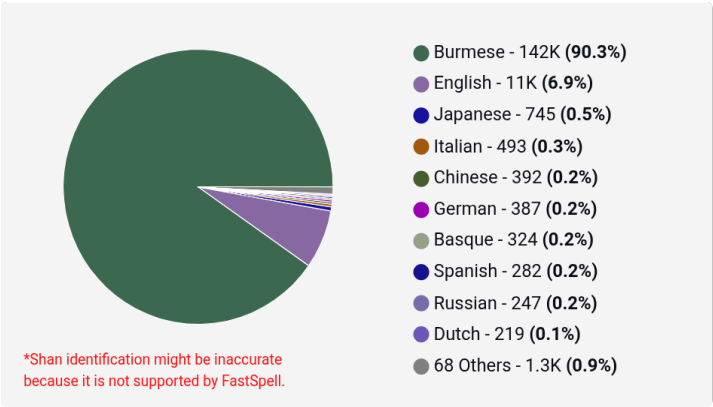
≤ 25 segments **91.27%** (11K documents)  
> 25 segments **8.73%** (1.1K documents)

Document collections

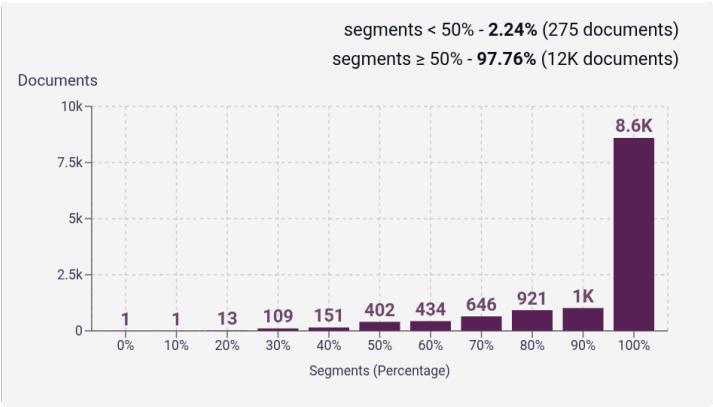


Language Distribution

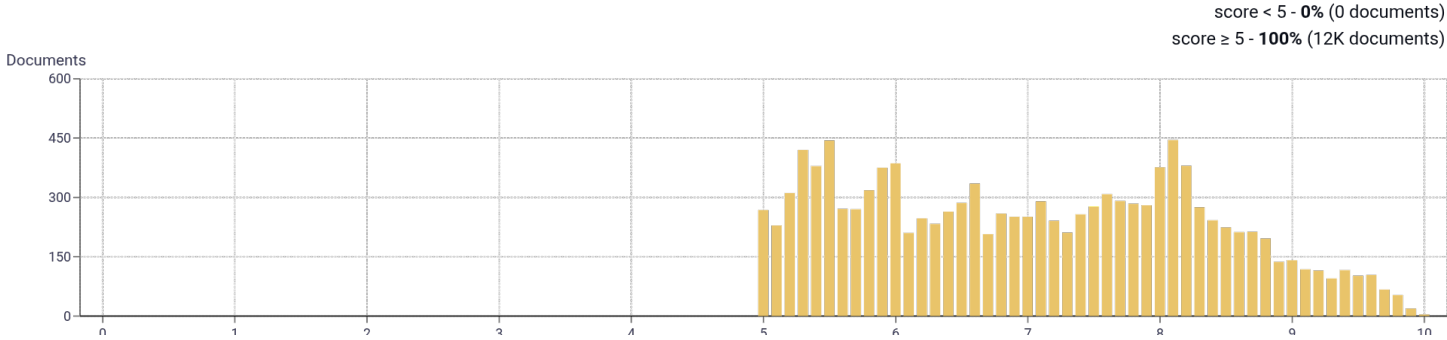
Number of segments in the Shan corpus



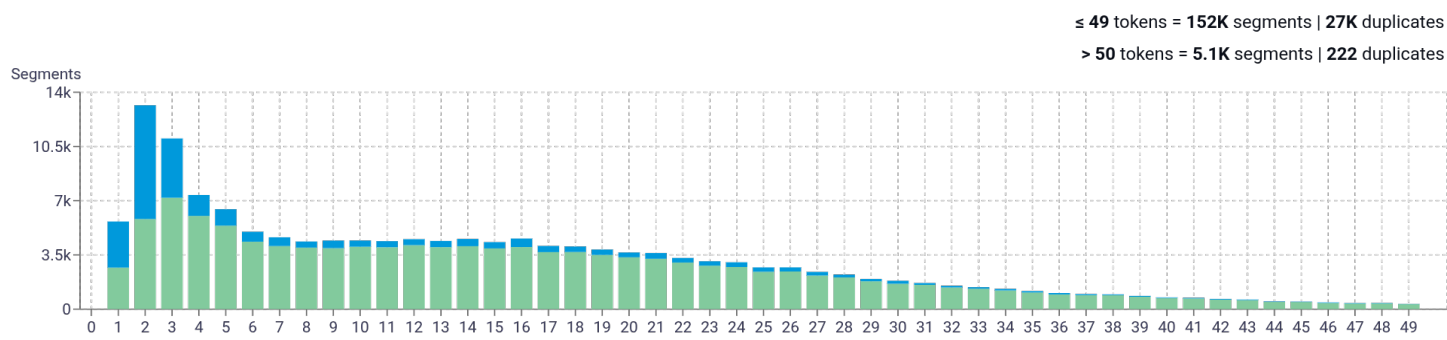
Percentage of segments in Shan inside documents



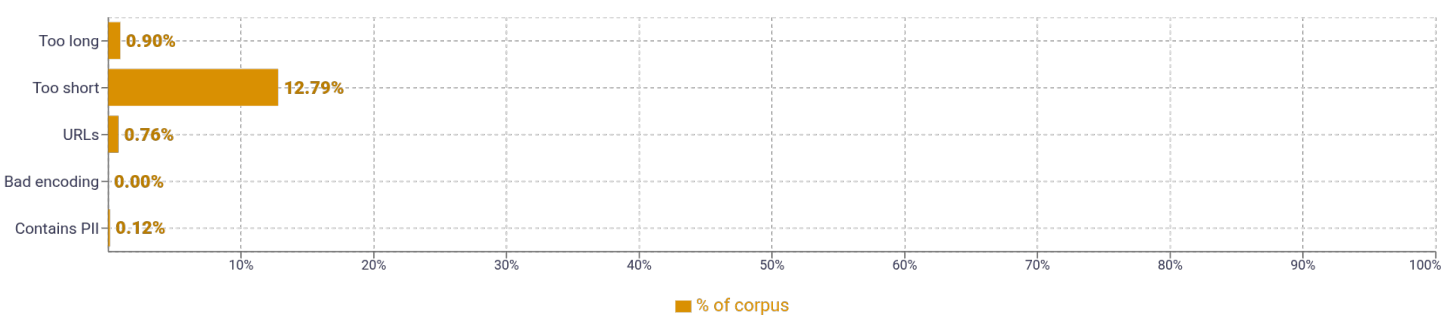
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

| SIZE | N-GRAMS  |  |
|------|--|--|
| 1    | <div>သိုင်း   37,853</div> <div>မိုင်း   26,896</div> <div>လှိုင်   15,361</div> <div>သေ   13,108</div> <div>ရှင်   10,512</div>   |  |
| 2    | <div>တပ်, သိုင်း   3,336</div> <div>cov id   2,754</div> <div>သိုင်း လိုင်းတီး   2,361</div> <div>လှိုင် လီး   1,993</div> <div>ာ,သီ, သိုင်း   1,699</div>   |  |
| 3    | <div>တပ်, သိုင်း လိုင်းတီး   1,470</div> <div>လှိုင် လီး မိုင်း   881</div> <div>ာ,သီ,ကပ်ခိုအ်း လိုင်းတီး/တပ်, သိုင်း   659</div> <div>ခွင် ဝ,သီ, သိုင်း   590</div> <div>တိတ်းလပ်းလှိုင်,မိုင်းတင်းပီအ် cov id   564</div>  |  |
| 4    | <div>ာ,သီ,ကပ်ခိုအ်း လိုင်းတီး/တပ်, သိုင်း လိုင်းတီး   340</div> <div>သိုင်း တပ်, သိုင်း လိုင်းတီး   298</div> <div>ခွင် ဝ,သီ,ကပ်ခိုအ်း လိုင်းတီး/တပ်, သိုင်း   262</div> <div>သိုင်း ကျေ့အ်းယှို,တပ်,ယှို,ကပ်,ပိုင်,ဟု,တီ,မို,သိုင်လိုင်းတီး တပ်, သိုင်း   207</div> <div>ကျေ့အ်းယှို,တပ်,ယှို,ကပ်,ပိုင်,ဟု,တီ,မို,သိုင်လိုင်းတီး တပ်, သိုင်း လိုင်းတီး   205</div>                          |  |
| 5    | <div>သိုင်း ကျေ့အ်းယှို,တပ်,ယှို,ကပ်,ပိုင်,ဟု,တီ,မို,သိုင်လိုင်းတီး တပ်, သိုင်း လိုင်းတီး   205</div> <div>ခွင် ဝ,သီ,ကပ်ခိုအ်း လိုင်းတီး/တပ်, သိုင်း လိုင်းတီး   147</div> <div>ကျေ့, သိုင်း ကျေ့အ်းယှို,တပ်,ယှို,ကပ်,ပိုင်,ဟု,တီ,မို,သိုင်လိုင်းတီး တပ်, သိုင်း   107</div> <div>ပေ ဝ,သေ, တွအ်းကျွင်,သေ လိုင်းတီးပွတ်း ရှင်   106</div> <div>ကျေ့, သိုင်း တပ်, သိုင်း လိုင်းတီး   105</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  $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$ , 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>, <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        | 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    |                                  |       |   |       |