

General overview

| Corpus           | Date      | Language       |
|------------------|-----------|----------------|
| hplt-v3-zsm_Latn | 9/19/2025 | Standard Malay |

Volumes

| Docs       | Segments    | Unique segments       | Tokens | Characters     | Size     |
|------------|-------------|-----------------------|--------|----------------|----------|
| 17,365,290 | 502,955,646 | 295,458,173 (58.74 %) | 13B    | 70,741,014,024 | 66.18 GB |

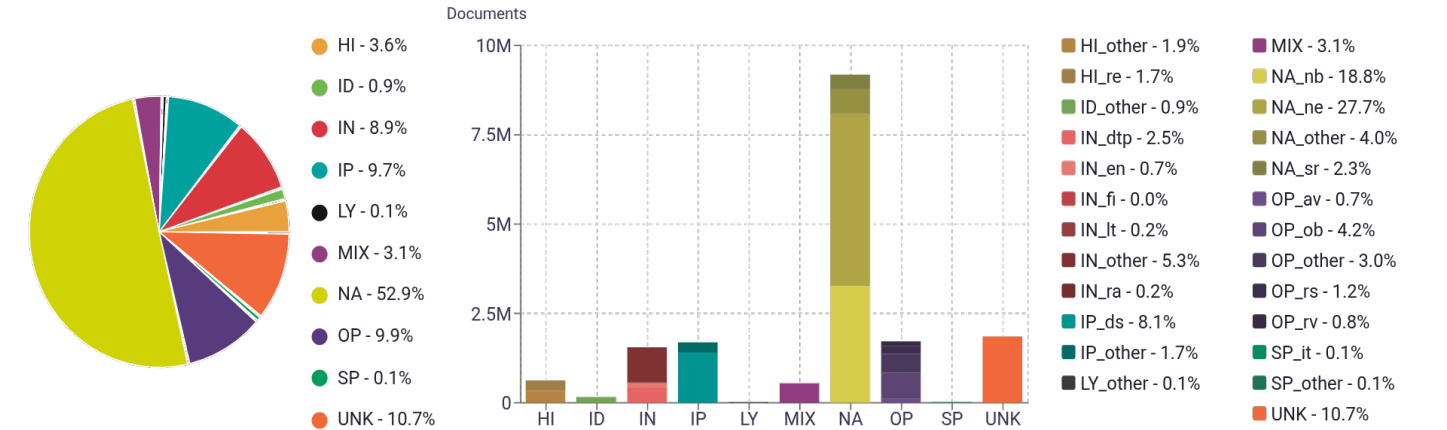
Top 10 domains

| Domain             | Docs | % of total |
|--------------------|------|------------|
| blogspot.com       | 5.1M | 29.55%     |
| wordpress.com      | 311K | 1.79%      |
| blogspot.my        | 274K | 1.58%      |
| sinarharian.com.my | 162K | 0.93%      |
| astroawani.com     | 144K | 0.83%      |
| mstar.com.my       | 118K | 0.68%      |
| utusan.com.my      | 113K | 0.65%      |
| hotels.com         | 97K  | 0.56%      |
| wikipedia.org      | 95K  | 0.55%      |
| hmetro.com.my      | 93K  | 0.53%      |

Top 10 TLDs

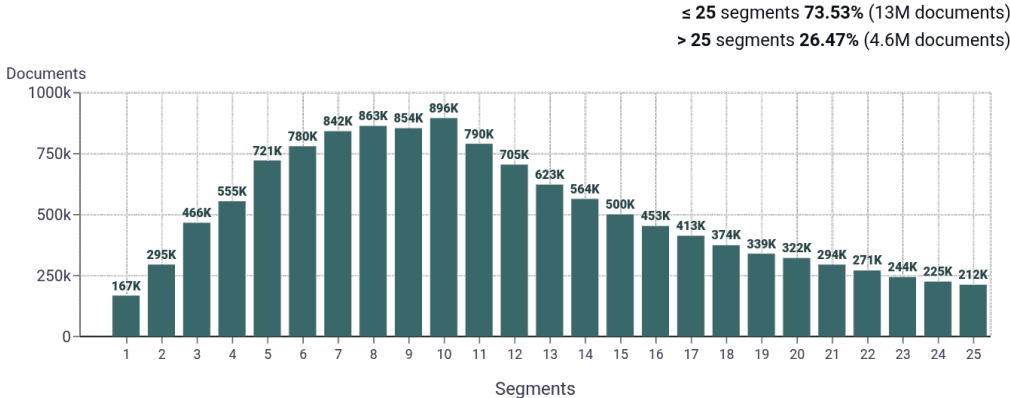
| Domain | Docs | % of total |
|--------|------|------------|
| com    | 12M  | 66.96%     |
| my     | 1.6M | 9.26%      |
| com.my | 1.3M | 7.74%      |
| net    | 559K | 3.22%      |
| org    | 350K | 2.01%      |
| gov.my | 223K | 1.28%      |
| sg     | 143K | 0.82%      |
| info   | 141K | 0.81%      |
| edu.my | 139K | 0.80%      |
| com.bn | 93K  | 0.53%      |

Register labels

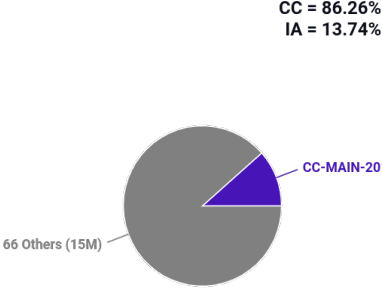


MT:6.7% | 1.2M Documents

Documents size (in segments)

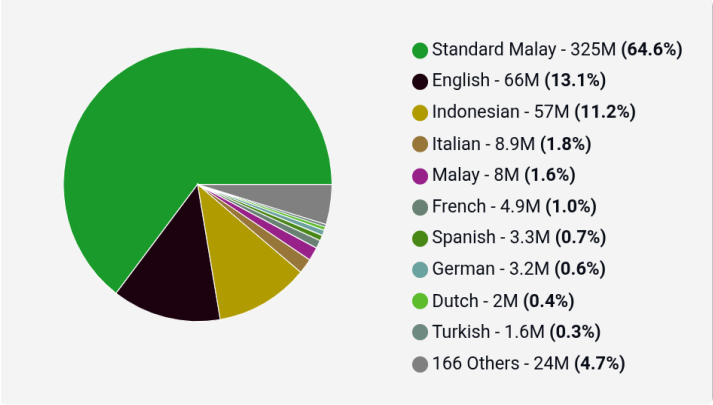


Document collections

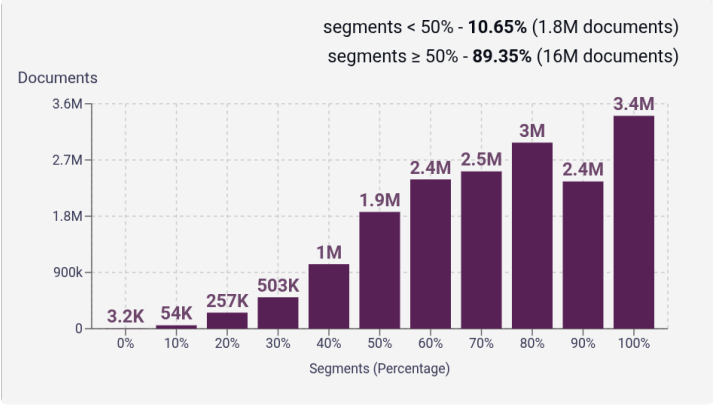


Language Distribution

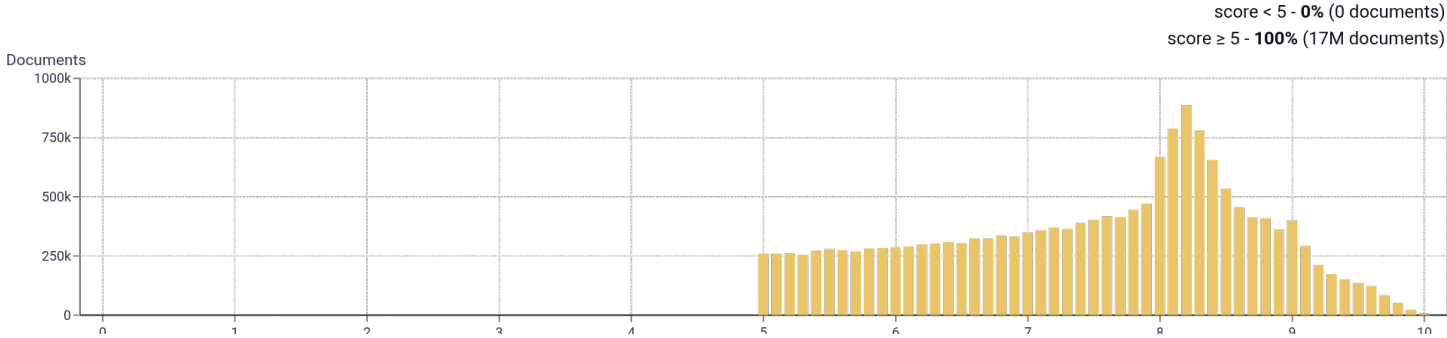
Number of segments in the Standard Malay corpus



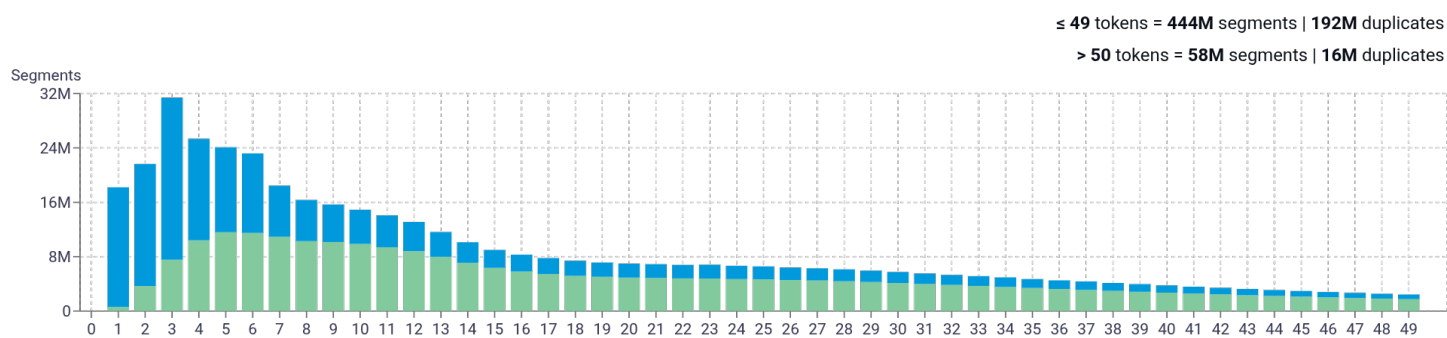
Percentage of segments in Standard Malay inside documents



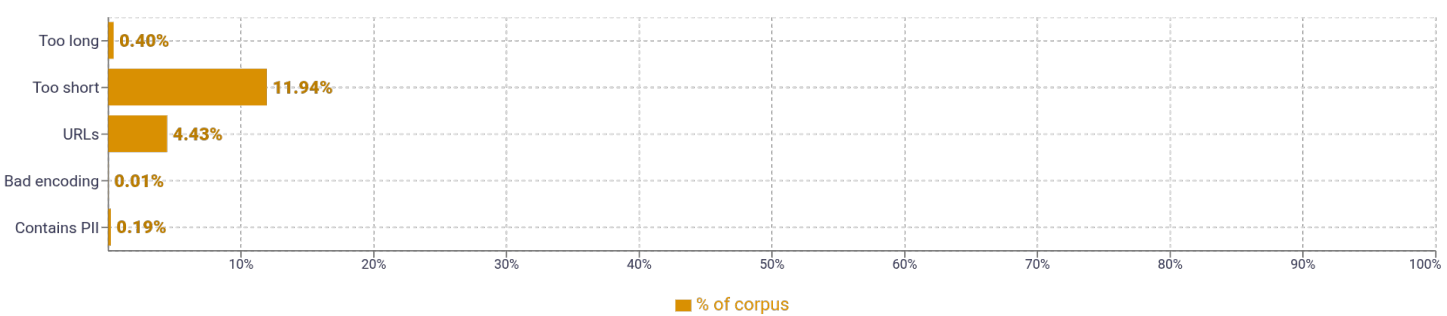
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

| SIZE | N-GRAMS  |  |
|------|--|--|
| 1    | aku   39,433,532   tak   35,808,314   nak   31,998,633   ni   31,796,139   dia   31,574,977  |  |
| 2    | a comment   3,041,912   post a   2,987,202   terima kasih   2,922,212   laman web   1,746,828   ibu bapa   1,603,209   |  |
| 3    | post a comment   2,985,169   kah kah kah   490,065   perintah kawalan pergerakan   248,309   tepat pada masanya   240,895<br>nabi muhammad saw   218,028                                       |  |
| 4    | assalamualaikum dan salam sejahtera   123,213   ibu pejabat polis daerah   93,727   tular di media sosial   89,077   komen dan share ya   83,922<br>bayi yang baru lahir   78,056              |  |
| 5    | lupa komen dan share ya   76,587   jangan lupa komen dan share   75,750   nik abdul aziz nik mat   66,294   view this post on instagram   56,082<br>komen yang diberikan oleh pembaca   53,385 |  |

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 of 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    |                                  |       |   |       |