

General overview

| Corpus           | Date      | Language  |
|------------------|-----------|-----------|
| hplt-v3-afr_Latn | 9/16/2025 | Afrikaans |

Volumes

| Docs      | Segments   | Unique segments      | Tokens | Characters    | Size    |
|-----------|------------|----------------------|--------|---------------|---------|
| 2,136,157 | 56,631,792 | 39,349,709 (69.48 %) | 1.7B   | 8,749,668,919 | 8.22 GB |

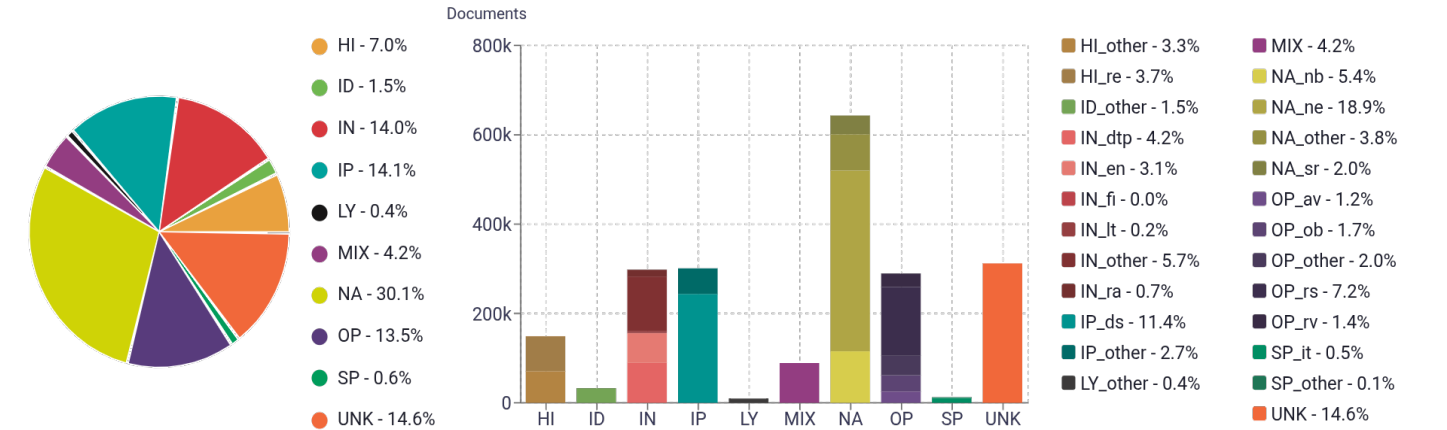
Top 10 domains

| Domain                             | Docs | % of total |
|------------------------------------|------|------------|
| <a href="#">maroelamedia.co.za</a> | 156K | 7.29%      |
| <a href="#">netwerk24.com</a>      | 60K  | 2.83%      |
| <a href="#">wikipedia.org</a>      | 53K  | 2.48%      |
| <a href="#">wordpress.com</a>      | 44K  | 2.06%      |
| <a href="#">litnet.co.za</a>       | 32K  | 1.49%      |
| <a href="#">landbou.com</a>        | 29K  | 1.35%      |
| <a href="#">lekkeslaap.co.za</a>   | 26K  | 1.20%      |
| <a href="#">dievyrburger.co.za</a> | 23K  | 1.09%      |
| <a href="#">sarie.com</a>          | 23K  | 1.08%      |
| <a href="#">software.net</a>       | 23K  | 1.08%      |

Top 10 TLDs

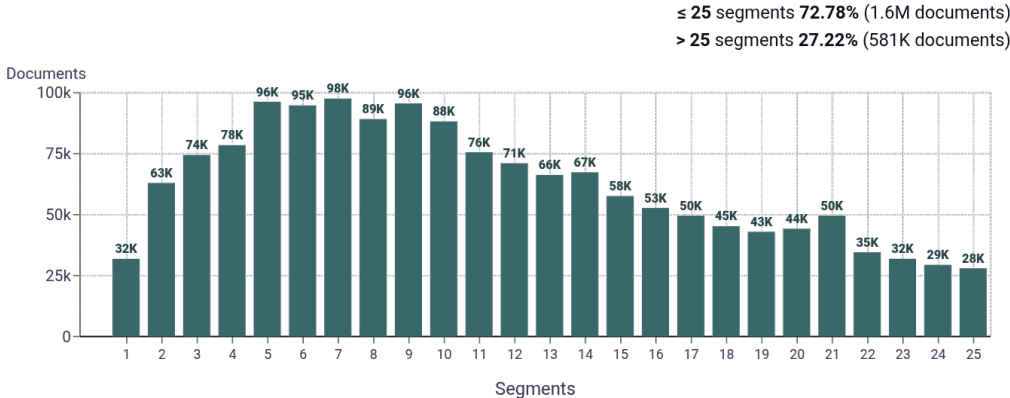
| Domain | Docs | % of total |
|--------|------|------------|
| com    | 840K | 39.35%     |
| co.za  | 733K | 34.33%     |
| org    | 172K | 8.04%      |
| net    | 94K  | 4.41%      |
| org.za | 43K  | 2.00%      |
| com.na | 35K  | 1.65%      |
| ac.za  | 27K  | 1.26%      |
| info   | 13K  | 0.61%      |
| ru     | 13K  | 0.61%      |
| pt     | 13K  | 0.61%      |

Register labels

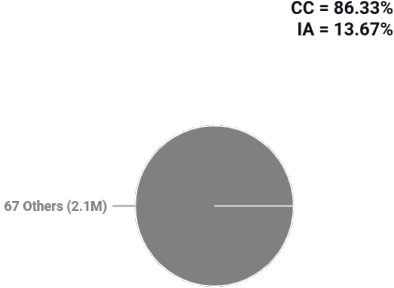


MT:10.8% | 230K Documents

Documents size (in segments) ⓘ

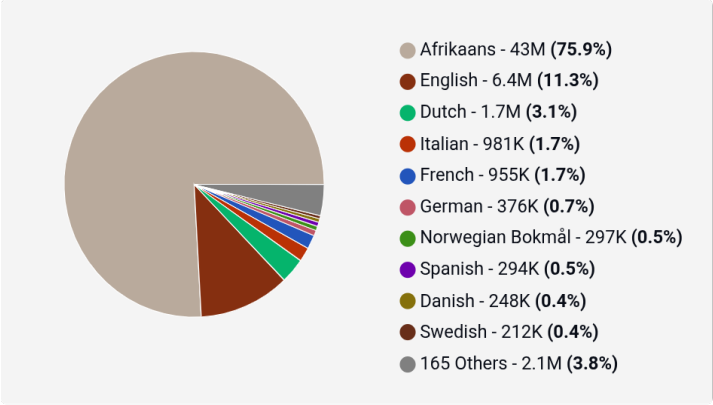


Document collections

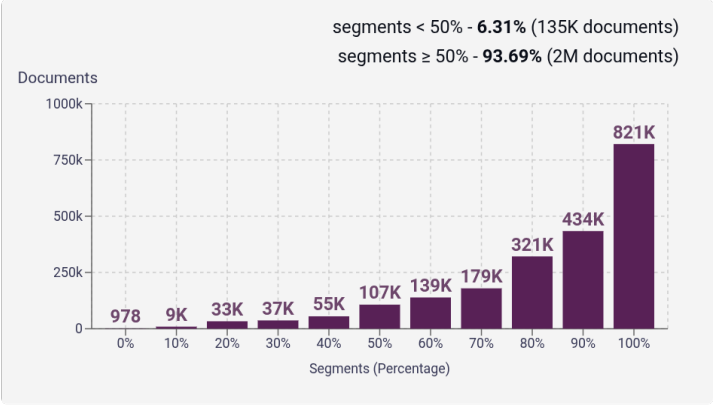


Language Distribution

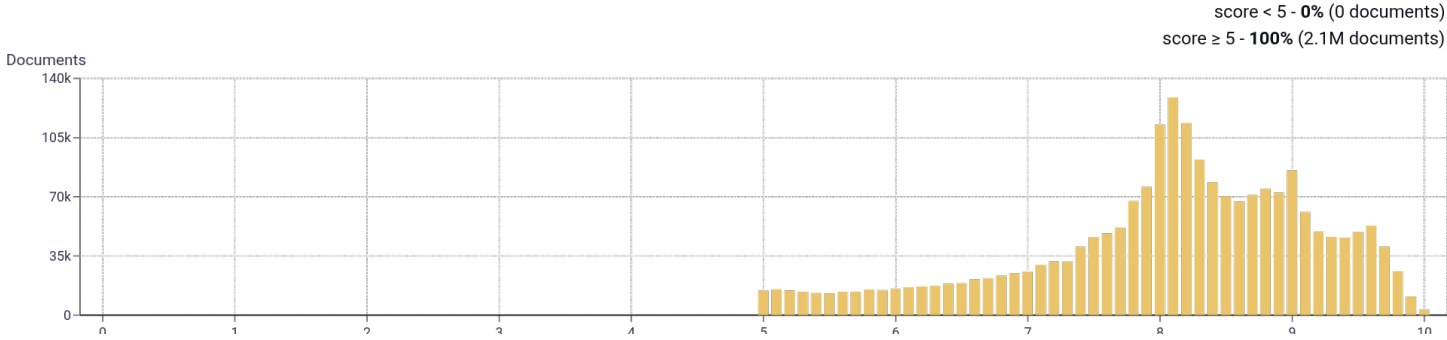
Number of segments in the Afrikaans corpus



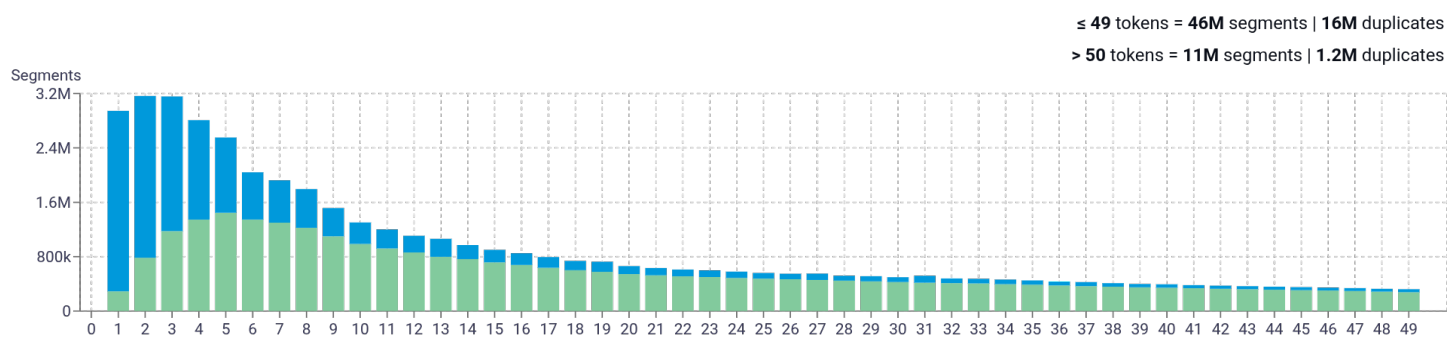
Percentage of segments in Afrikaans inside documents



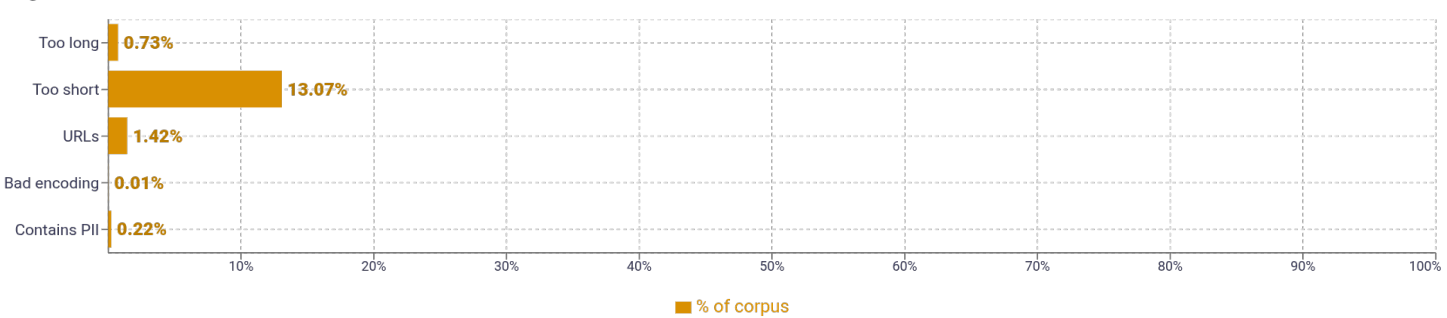
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

| SIZE | N-GRAMS  |  |
|------|--|--|
| 1    | oor   4,389,503the   3,980,924moet   3,962,365soos   3,373,295meer   3,268,062   |  |
| 2    | wil h    178,154to the   161,808eerste keer   158,511ten minste   145,202bin re opsies   139,170   |  |
| 3    | oor die algemeen   88,379voor te berei   65,244seker te maak   58,157data for your   57,950we are searching   57,948   |  |
| 4    | we are searching data   57,939searching data for your   57,939data for your request   57,939are searching data for   57,939will appear to access   53,216                              |  |
| 5    | we are searching data for   57,939searching data for your request   57,939are searching data for your   57,939will appear to access the   53,216to access the found materials   53,216 |  |

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    |                                  |       |   |       |