

General overview

| Corpus           | Date      | Language   |
|------------------|-----------|------------|
| hplt-v3-lua_Latn | 9/18/2025 | Luba-Lulua |

Volumes

| Docs  | Segments | Unique segments  | Tokens | Characters | Size     |
|-------|----------|------------------|--------|------------|----------|
| 1,634 | 50,740   | 47,190 (93.00 %) | 2.3M   | 12,345,454 | 11.87 MB |

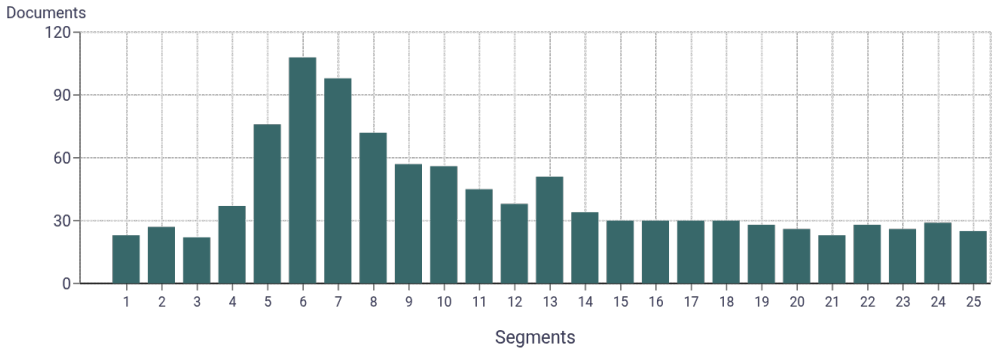
Top 10 domains

| Domain             | Docs | % of total |
|--------------------|------|------------|
| jw.org             | 1.1K | 67.75%     |
| editorial7.net     | 146  | 8.94%      |
| radiookapi.net     | 144  | 8.81%      |
| biblafrique.net    | 44   | 2.69%      |
| biblafrique.org    | 31   | 1.90%      |
| wikimedia.org      | 28   | 1.71%      |
| canalblog.com      | 11   | 0.67%      |
| thessage.com       | 10   | 0.61%      |
| gotquestions.org   | 10   | 0.61%      |
| legende-tshakap... | 9    | 0.55%      |

Top 10 TLDs

| Domain | Docs | % of total |
|--------|------|------------|
| org    | 1.2K | 74.17%     |
| net    | 339  | 20.75%     |
| com    | 58   | 3.55%      |
| cd     | 8    | 0.49%      |
| ca     | 5    | 0.31%      |
| be     | 5    | 0.31%      |
| info   | 3    | 0.18%      |
| gc.ca  | 2    | 0.12%      |
| ru     | 1    | 0.06%      |
| eu     | 1    | 0.06%      |

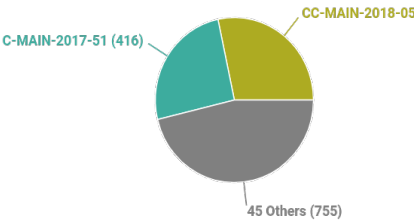
Documents size (in segments) ⓘ



≤ 25 segments **64.2%** (1K documents)  
> 25 segments **35.8%** (585 documents)

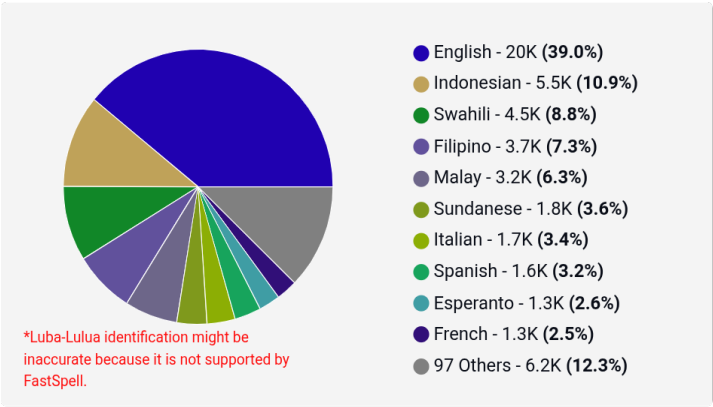
Document collections

CC = **90.76%**  
IA = **9.24%**

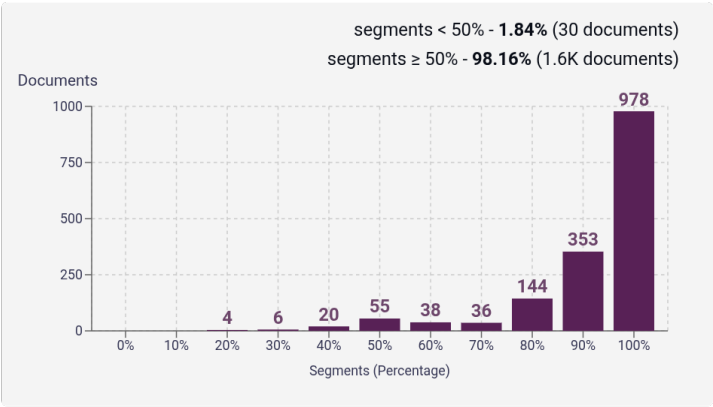


Language Distribution

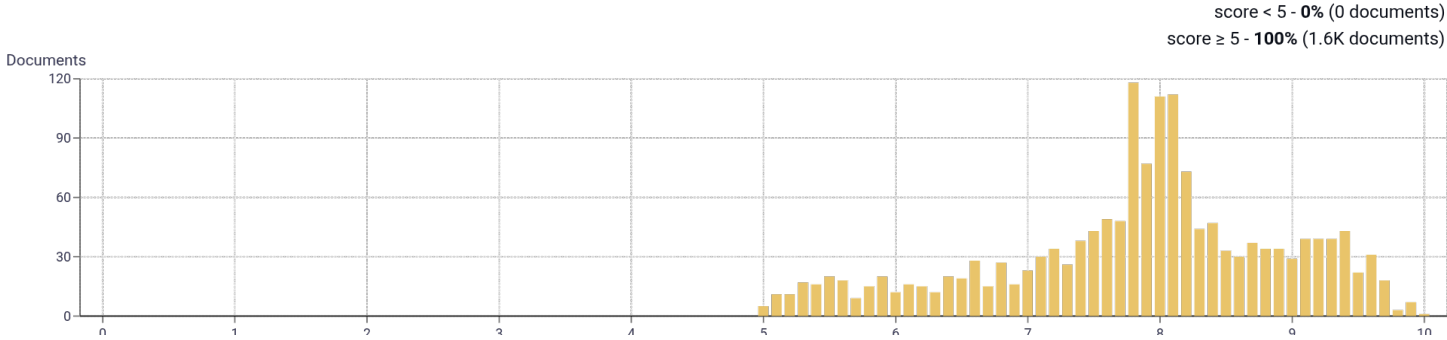
Number of segments in the Luba-Lulua corpus



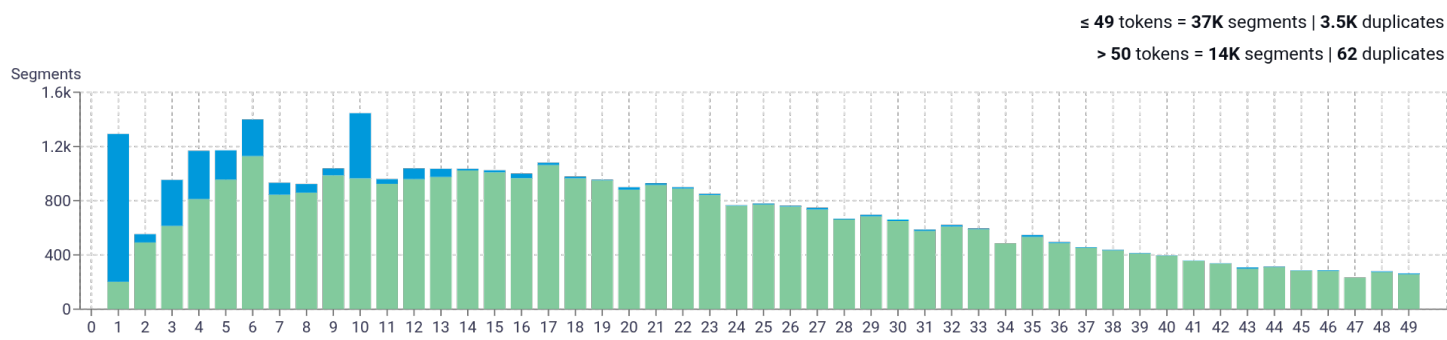
Percentage of segments in Luba-Lulua inside documents



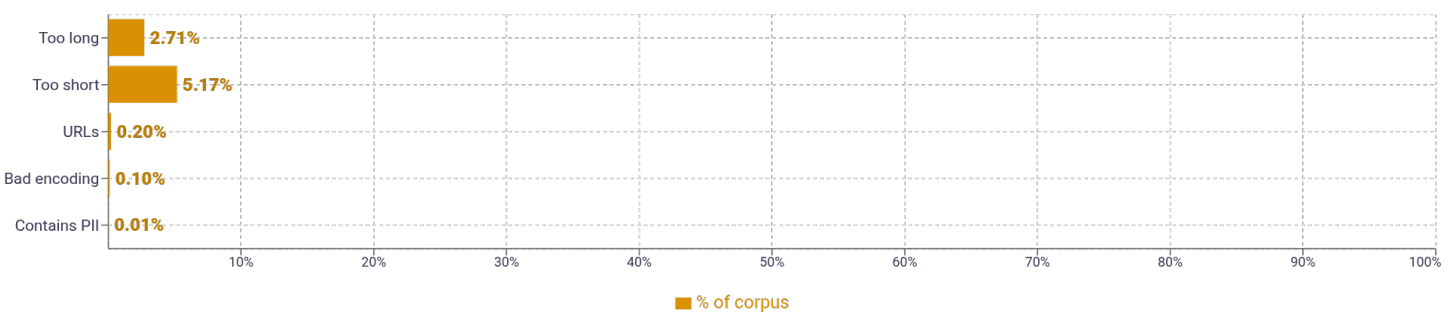
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

| SIZE | N-GRAMS  |  |
|------|--|--|
| 1    | ne   105,105    bua   55,574    mu   48,335    wa   37,918    udi   26,965   |  |
| 2    | bualu bua   5,870    ne bua   5,277    pa buloba   2,828    wa nzambi   2,586    udi ne   2,581  |  |
| 3    | udi wamba ne   1,157    bantu ba bungi   803    ba pa buloba   651    tudi ne bua   649    too ne ku   538   |  |
| 4    | nkudimuinu wa bulongolodi bupiabupia   461    ne mushinga wa bungi   258    mu mukanda wa nzambi   243    bible udi wamba ne   236<br>wa nzambi udi wamba   200                |  |
| 5    | wa nzambi udi wamba ne   176    ba balume ne ba bakaji   96    bua tshinyi tudi ne bua   91    mu bidimu lukama bia kumpala   85<br>mukalenge mutambe bunene yehowa wamba   82 |  |

# 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    |                                  |       |   |       |