# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Date | Language |
|---|---|---|
| hplt-v3-hau_Latn | 9/18/2025 | Hausa (ha) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Characters | Size |
|---|---|---|---|---|---|
| 743,843 | 15,098,921 | 11,570,766 (76.63 %) | 495M | 2,341,460,938 | 2.21 GB |

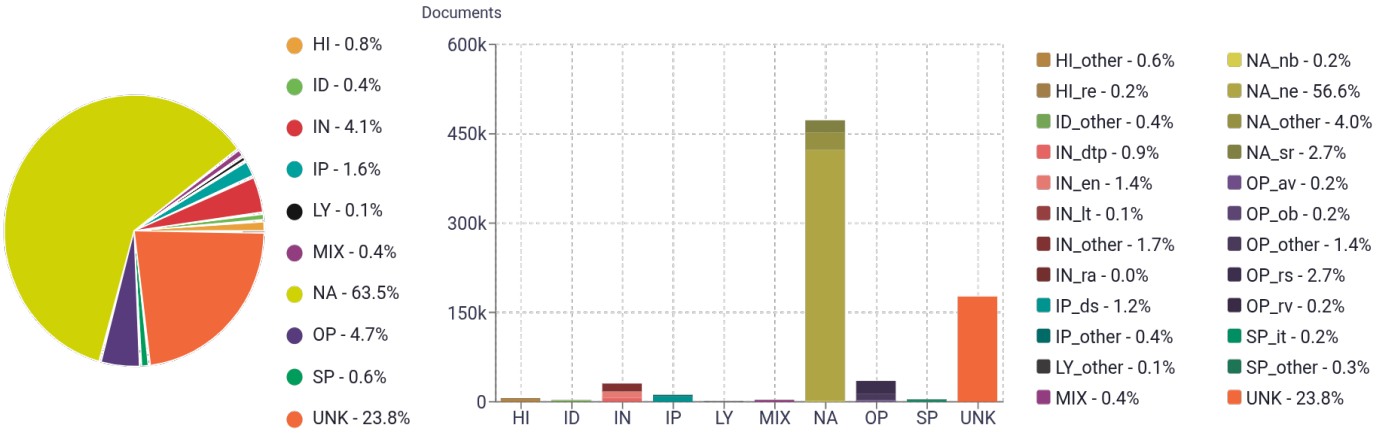
## Top 10 domains

| Domain | Docs | % of total |
|---|---|---|
| legit.ng | 73K | 9.76% |
| leadership.ng | 59K | 7.95% |
| voahausa.com | 32K | 4.32% |
| bbc.com | 25K | 3.43% |
| rfi.fr | 17K | 2.29% |
| premiumtimesng.com | 15K | 2.01% |
| cri.cn | 15K | 1.98% |
| naija.ng | 12K | 1.62% |
| dw.com | 11K | 1.52% |
| wikipedia.org | 11K | 1.44% |

## Top 10 TLDs

| Domain | Docs | % of total |
|---|---|---|
| com | 411K | 55.28% |
| ng | 171K | 22.94% |
| com.ng | 51K | 6.91% |
| org | 25K | 3.39% |
| fr | 19K | 2.50% |
| cn | 15K | 2.02% |
| net | 11K | 1.46% |
| zone | 6.8K | 0.92% |
| ir | 6K | 0.81% |
| eu | 3.9K | 0.52% |

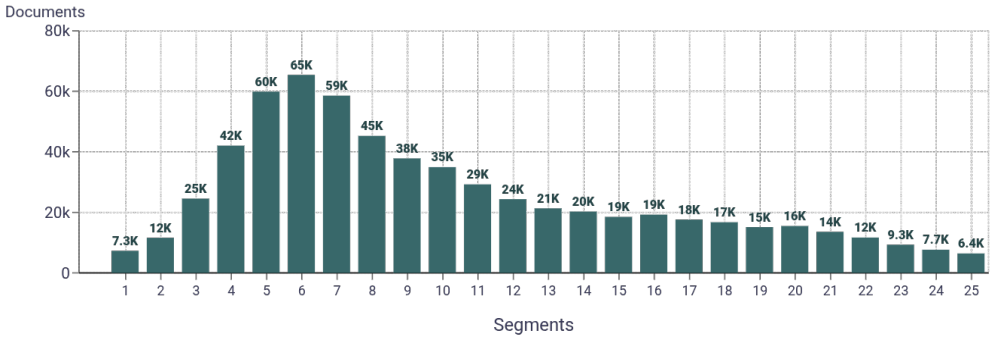
## Register labels




**MT**:21.4% | 159K Documents

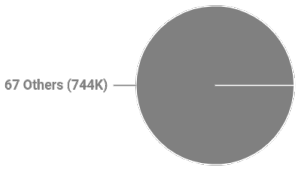
## Documents size (in segments)

**≤ 25** segments **85.27%** (634K documents)
**> 25** segments **14.73%** (110K documents)



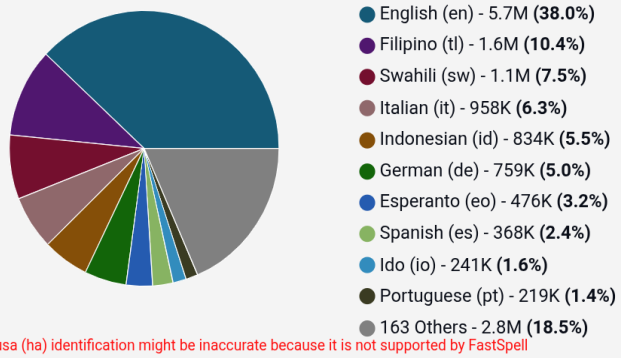

## Document collections

**CC = 97.44%**
**IA = 2.56%**



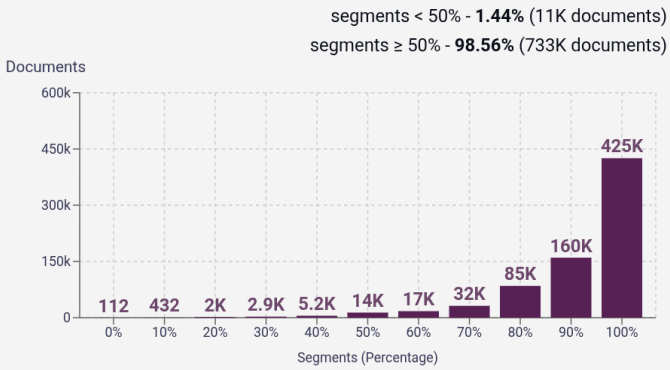
## Language Distribution

### Number of segments in the Hausa (ha) corpus

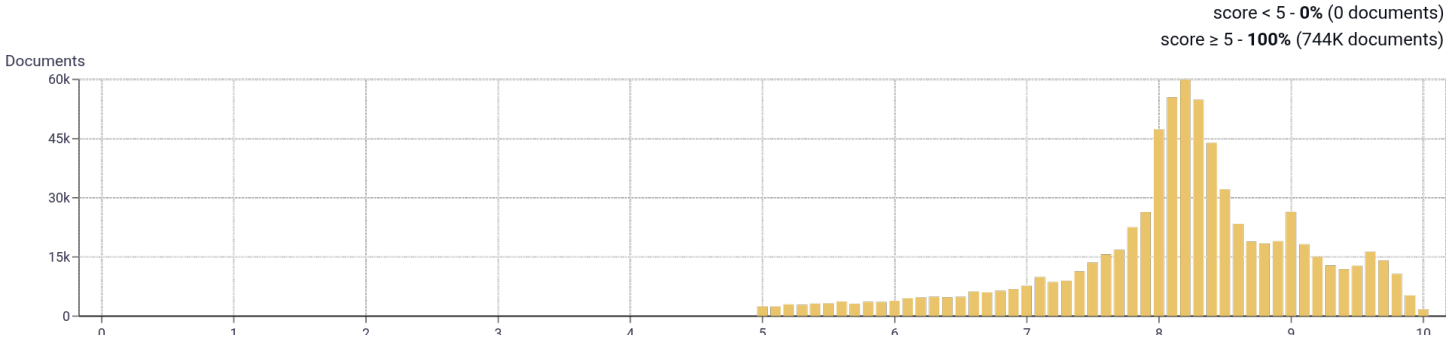



*Hausa (ha) identification might be inaccurate because it is not supported by FastSpell

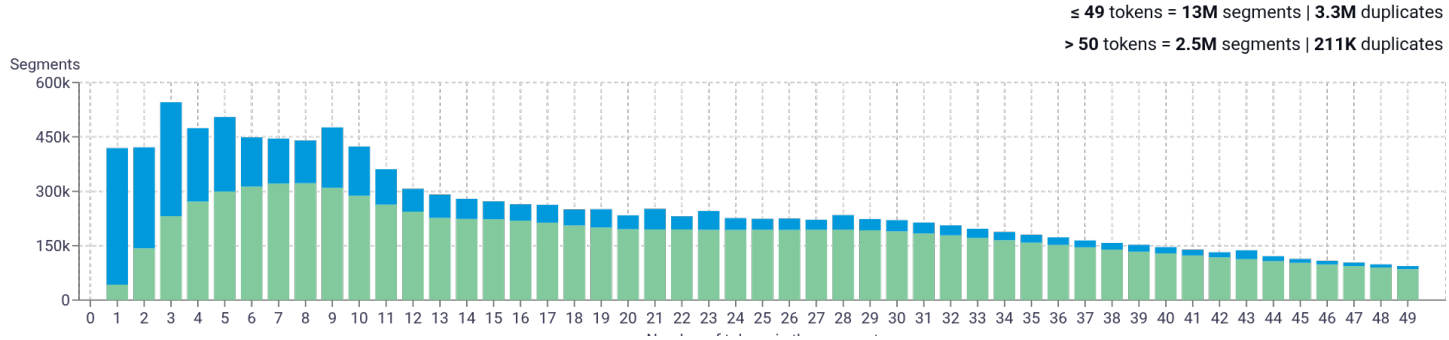
### Percentage of segments in Hausa (ha) inside documents



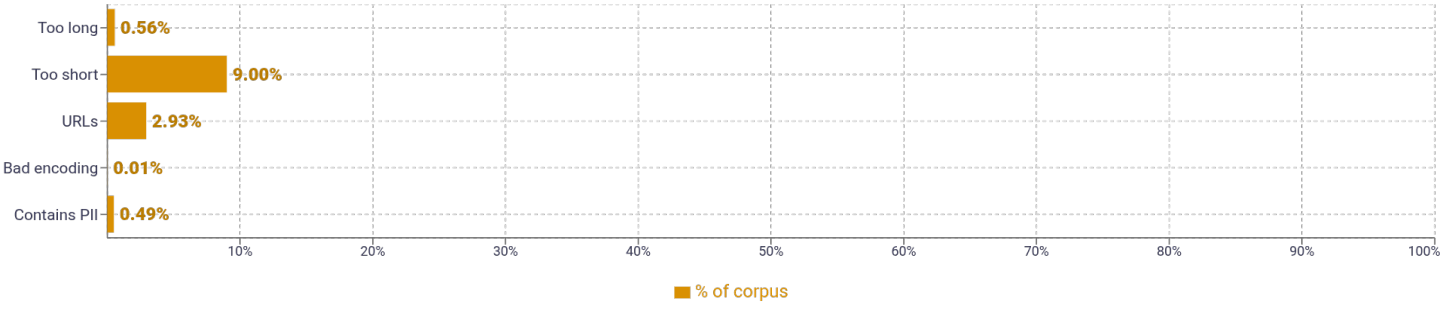

## Distribution of documents by document score




## Segment length distribution by token




## Segment noise distribution



### Frequent n-grams

| SIZE | N-GRAMS |
|---|---|
| 1 | tare \| 1,404,828 · allah \| 899,219 · amfani \| 891,132 · kasar \| 819,776 · jihar \| 797,644 |
| 2 | ci gaba \| 368,531 · shugaban kasa \| 176,469 · kasar sin \| 97,573 · jihar kano \| 97,019 · gwamnan jihar \| 93,215 |
| 3 | kasa da kasa \| 37,932 · dandalin sada zumunta \| 33,749 · shawara ko bukatar \| 33,696 · post a comment \| 32,888 · majalisar dinkin duniya \| 31,362 |
| 4 | wayar ku ta hannu \| 32,086 · shugaban kasa muhammadu buhari \| 30,078 · shafukanmu na dandalin sada \| 29,440 · shawara ko bukatar bamu \| 29,317 · latsa wannan domin samun \| 23,258 |
| 5 | shafukanmu na dandalin sada zumunta \| 29,439 · shawara ko bukatar bamu labari \| 29,316 · latsa wannan domin samun sabuwar \| 20,841 · kwararrun kwararrun kwararrun kwararrun kwararrun \| 20,550 · sabuwar manhajar labarai ta legit \| 14,678 |

## 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 or 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 | | | | |