

General overview

| Corpus | Date | Language |
|------------------|-----------|--------------|
| hplt-v3-kea_Latn | 9/18/2025 | Kabuverdianu |

Volumes

| Docs | Segments | Unique segments | Tokens | Characters | Size |
|-------|----------|------------------|--------|------------|---------|
| 3,080 | 50,811 | 44,938 (88.44 %) | 1.6M | 7,226,646 | 7.06 MB |

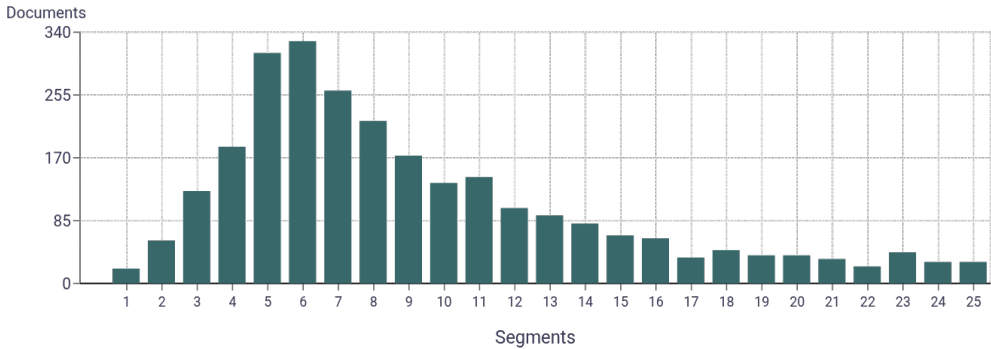
Top 10 domains

| Domain | Docs | % of total |
|--------------------|------|------------|
| dexamsabi.com | 1.2K | 38.15% |
| jw.org | 441 | 14.32% |
| dypk-portal.com | 353 | 11.46% |
| blogspot.com | 252 | 8.18% |
| deltacultura.org | 240 | 7.79% |
| santiagomagazin... | 109 | 3.54% |
| anacao.cv | 72 | 2.34% |
| blogspot.pt | 60 | 1.95% |
| kriolita.com | 18 | 0.58% |
| sullivaninjuryl... | 17 | 0.55% |

Top 10 TLDs

| Domain | Docs | % of total |
|---------|------|------------|
| com | 1.9K | 63.25% |
| org | 731 | 23.73% |
| cv | 223 | 7.24% |
| pt | 84 | 2.73% |
| com.br | 17 | 0.55% |
| gov | 16 | 0.52% |
| info | 14 | 0.45% |
| net | 13 | 0.42% |
| mus.br | 5 | 0.16% |
| publ.cv | 4 | 0.13% |

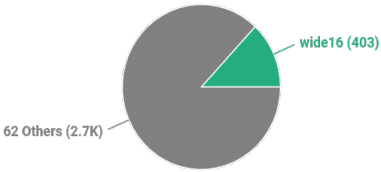
Documents size (in segments) ⓘ



≤ 25 segments **86.85%** (2.7K documents)
> 25 segments **13.15%** (405 documents)

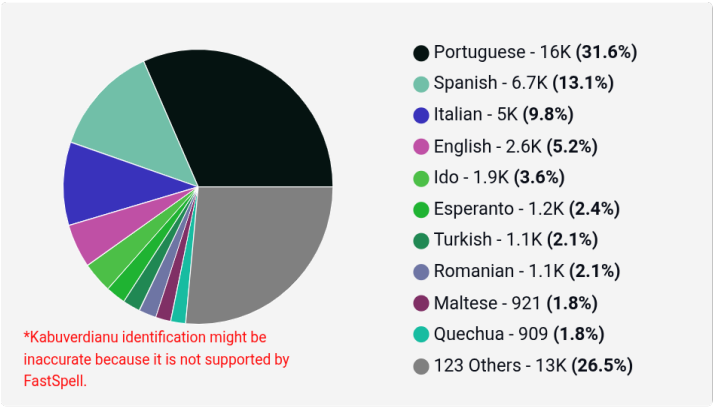
Document collections

CC = **78.31%**
IA = **21.69%**

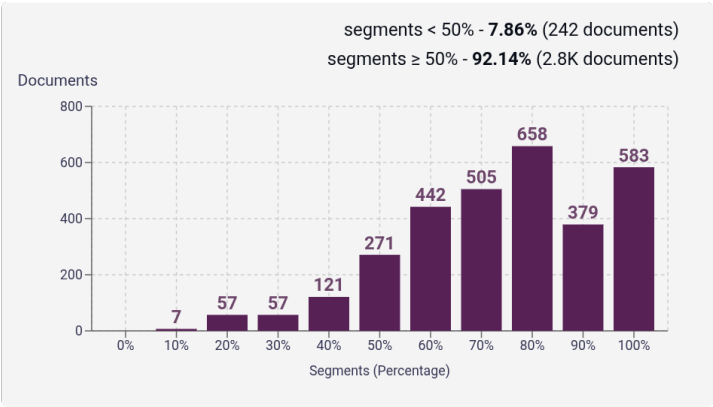


Language Distribution

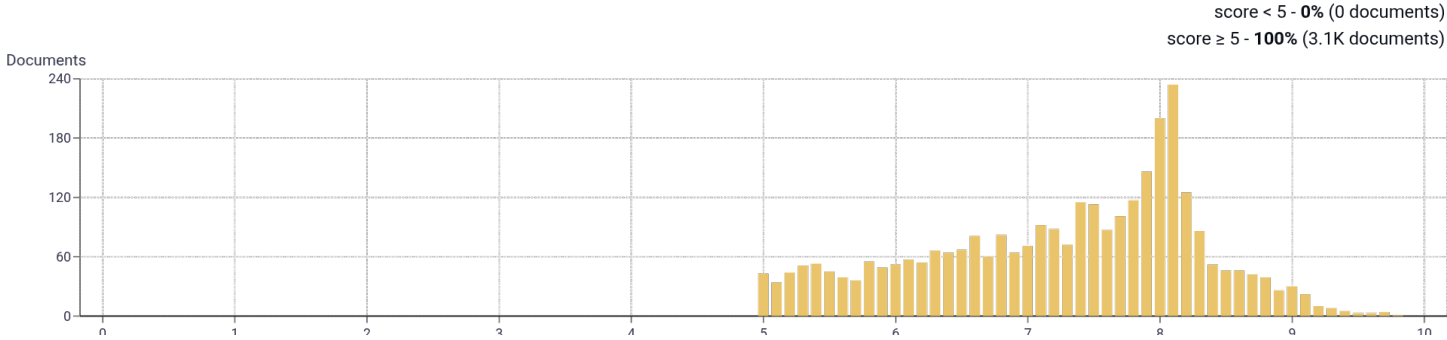
Number of segments in the Kabuverdianu corpus



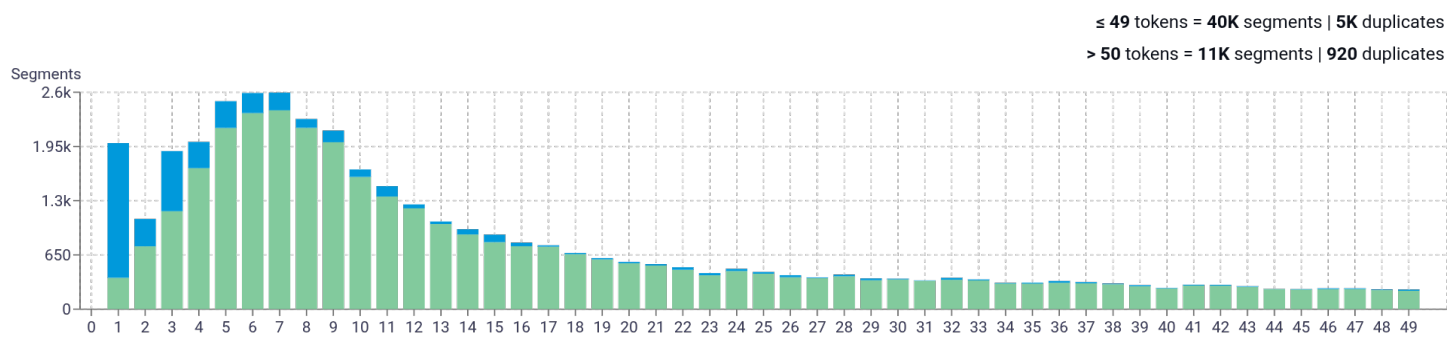
Percentage of segments in Kabuverdianu inside documents



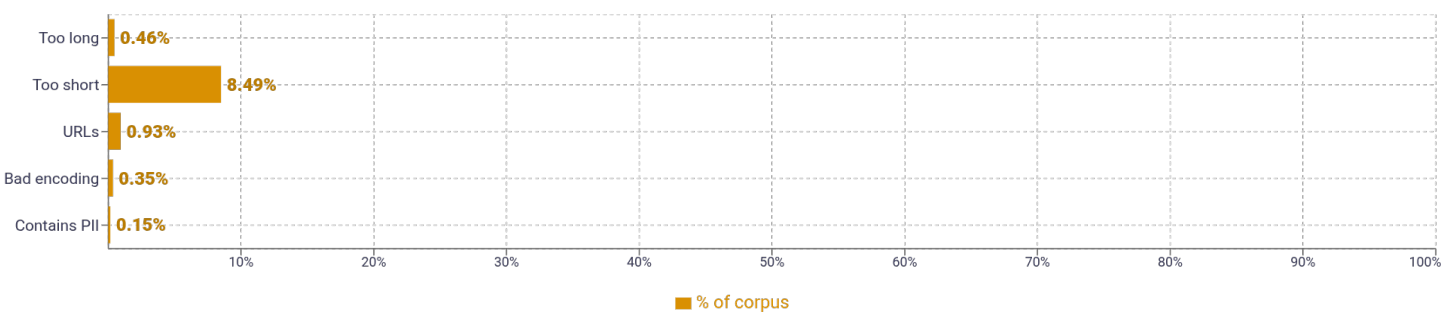
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

| SIZE | N-GRAMS | |
|------|--|--|
| 1 | ki 40,546pa 29,960ku 22,357un 17,747ka 17,067 | |
| 2 | ki nu 2,330fla ma 1,850ki sta 1,811cabo verde 1,483kuzê ki 1,324 | |
| 3 | un di kes 385kes algen ki 317óras ki nu 293kuzê ki nu 254modi ki nu 251 | |
| 4 | skodjedu ku spritu santu 102nha ida padri nikulau 96ida padri nikulau ferera 96marsianu nha ida padri 95ki nu ta faze 89 | |
| 5 | nha ida padri nikulau ferera 96marsianu nha ida padri nikulau 95sentru di idukason delta kultura 58leitura di biblia di kel 56biblia di kel simana li 56 | |

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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>, , , 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 | | | | |