

General overview

| Corpus | Date | Language |
|------------------|-----------|----------|
| hplt-v3-nso_Latn | 9/18/2025 | Pedi |

Volumes

| Docs | Segments | Unique segments | Tokens | Characters | Size |
|-------|----------|-------------------|--------|------------|----------|
| 8,183 | 234,057 | 172,408 (73.66 %) | 9.3M | 42,025,444 | 41.23 MB |

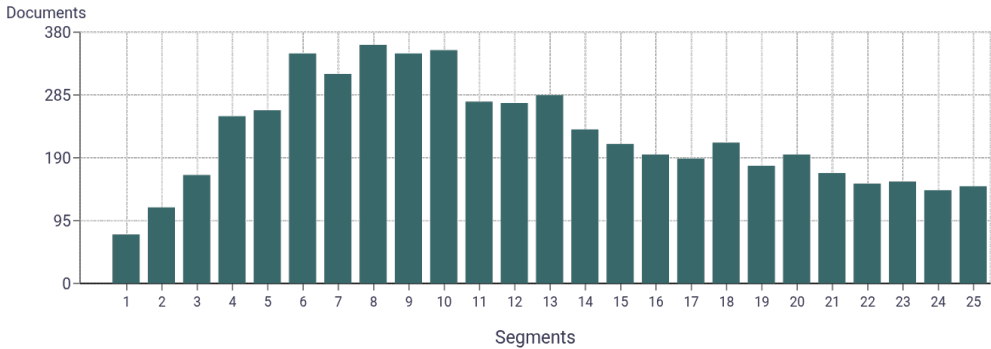
Top 10 domains

| Domain | Docs | % of total |
|--------------------|------|------------|
| jw.org | 3.7K | 45.35% |
| biblesa.co.za | 1.8K | 22.30% |
| southafrica.co.za | 806 | 9.85% |
| fundza.mobi | 185 | 2.26% |
| seiponemadireng... | 159 | 1.94% |
| nalibali.org | 115 | 1.41% |
| wikipedia.org | 93 | 1.14% |
| oxforddictionar... | 81 | 0.99% |
| sars.gov.za | 54 | 0.66% |
| sekhukhunetimes... | 47 | 0.57% |

Top 10 TLDs

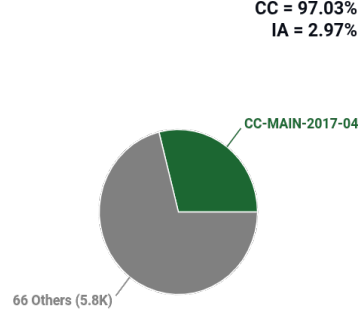
| Domain | Docs | % of total |
|--------|------|------------|
| org | 4K | 49.40% |
| co.za | 3.1K | 38.47% |
| com | 385 | 4.70% |
| mobi | 186 | 2.27% |
| gov.za | 88 | 1.08% |
| org.za | 48 | 0.59% |
| ru | 46 | 0.56% |
| net | 46 | 0.56% |
| fm | 41 | 0.50% |
| ac.za | 28 | 0.34% |

Documents size (in segments) ⓘ



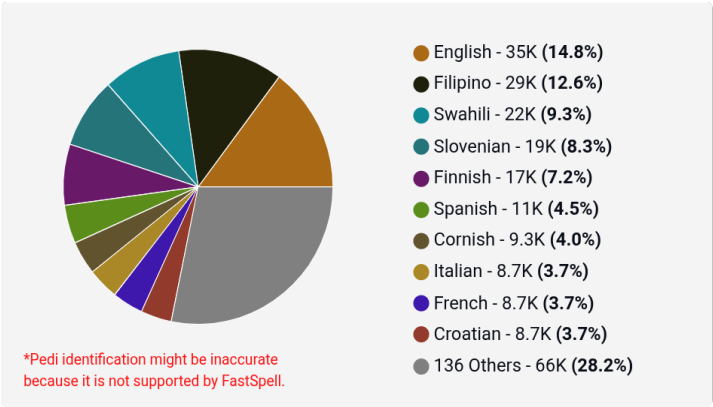
≤ 25 segments **68.46%** (5.6K documents)
> 25 segments **31.54%** (2.6K documents)

Document collections

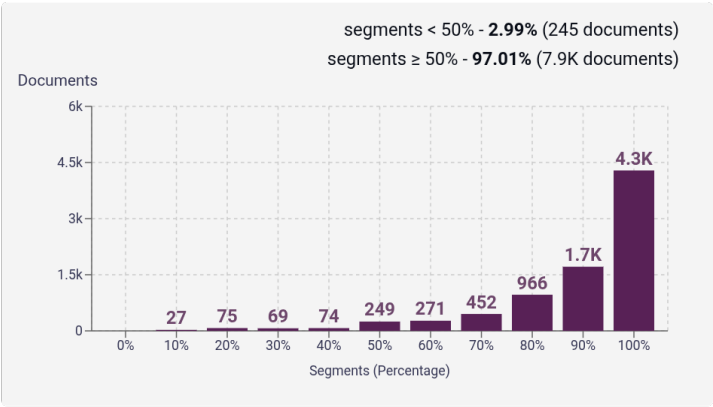


Language Distribution

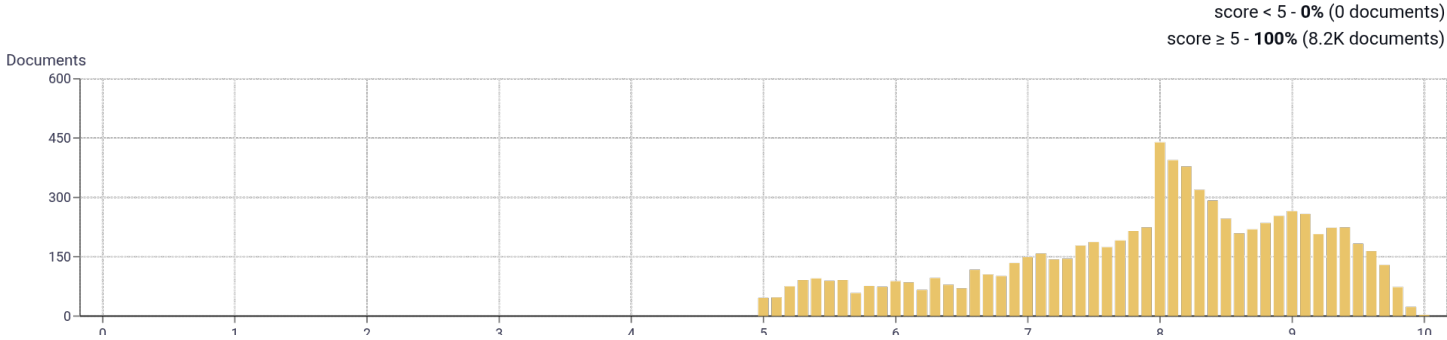
Number of segments in the Pedi corpus



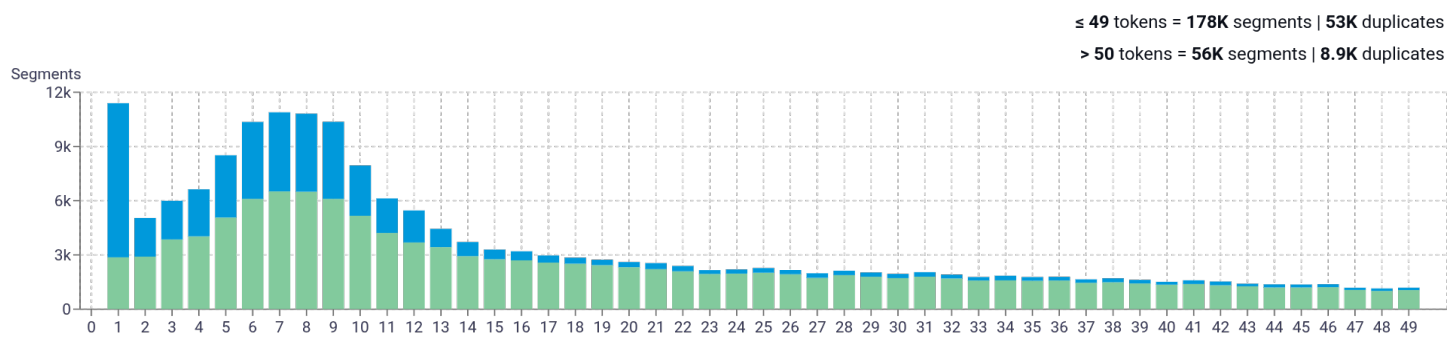
Percentage of segments in Pedi inside documents



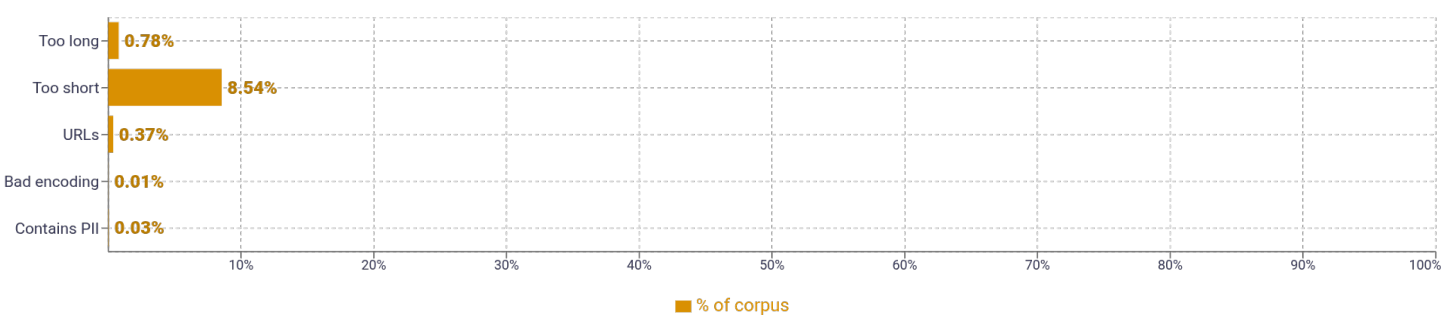
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

| SIZE | N-GRAMS | |
|------|---|---|
| 1 | ya 201,284wa 107,305re 105,061tša 88,829ge 79,792 | 📄 |
| 2 | yo mongwe 10,205wa gagwe 8,171ya gagwe 7,220tše dingwe 7,116baka la 6,459 | 📄 |
| 3 | dihlatse tša jehofa 2,771yo mongwe wa 2,523mongwe le yo 2,505bjalo ka ge 2,388bao ba bego 2,300 | 📄 |
| 4 | lega go le bjalo 3,486mongwe le yo mongwe 2,500yo mongwe le yo 2,453ge e le gabotse 1,411sengwe le se sengwe 1,259 | 📄 |
| 5 | yo mongwe le yo mongwe 2,448phetolelo ya lefase le lefsa 1,168ya lefase le lefsa ya 1,030lefase le lefsa ya mangwalo 1,017lefsa ya mangwalo a makgethwa 1,004 | 📄 |

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 | | | | |