

General overview

| Corpus | Date | Language |
|----------------------|-----------|------------|
| HPLT-v2-tam_TamI.tsv | 9/19/2024 | Tamil (ta) |

Volumes

| Docs | Segments | Characters | Size |
|-----------|-------------|----------------|----------|
| 6,105,834 | 168,591,779 | 26,076,202,009 | 64.51 GB |

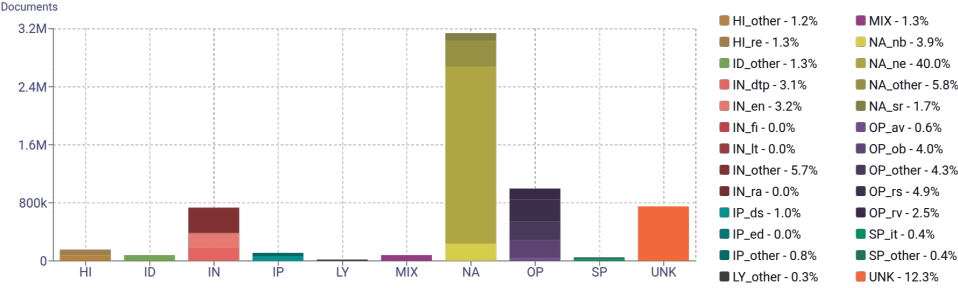
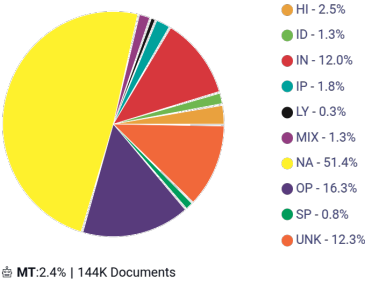
Top 10 domains

| Domain | Docs | % of total |
|---------------|------|------------|
| blogspot.com | 522K | 8.55% |
| blogspot.in | 264K | 4.32% |
| wikipedia.org | 202K | 3.30% |
| dinamalar.com | 146K | 2.39% |
| wordpress.com | 105K | 1.73% |
| vikatan.com | 104K | 1.70% |
| dinamani.com | 93K | 1.52% |
| blogspot.sg | 86K | 1.41% |
| dinakaran.com | 84K | 1.37% |
| blogspot.ch | 83K | 1.36% |

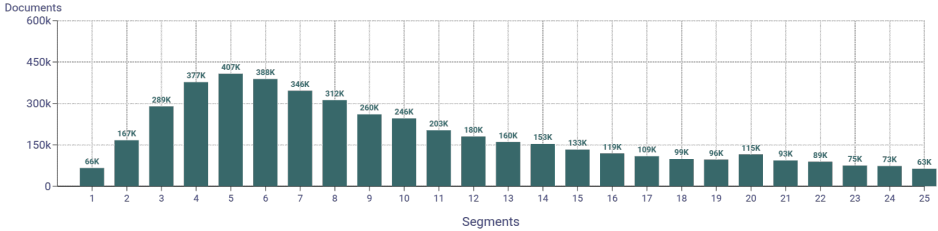
Top 10 TLDs

| Domain | Docs | % of total |
|--------|------|------------|
| com | 4.1M | 66.66% |
| in | 661K | 10.82% |
| org | 383K | 6.27% |
| net | 218K | 3.57% |
| lk | 119K | 1.95% |
| sg | 88K | 1.44% |
| ch | 86K | 1.40% |
| info | 45K | 0.74% |
| ca | 45K | 0.74% |
| ae | 39K | 0.64% |

Register labels

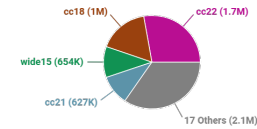


Documents size (in segments)



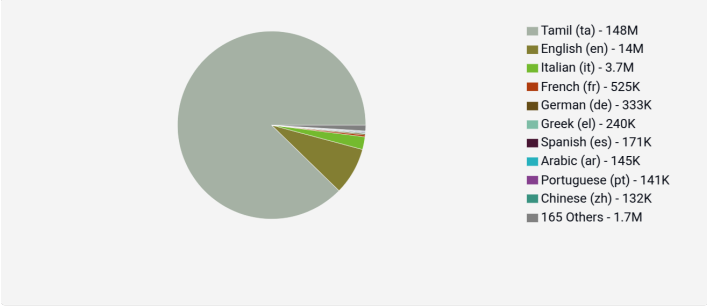
<= 25 segments 75.65% (4.6M documents)
> 25 segments 24.35% (1.5M documents)

Documents by collection

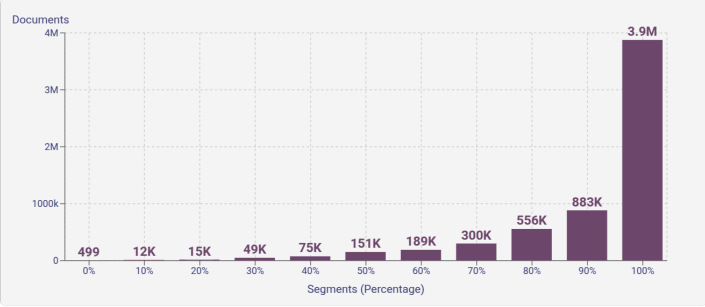


Language Distribution

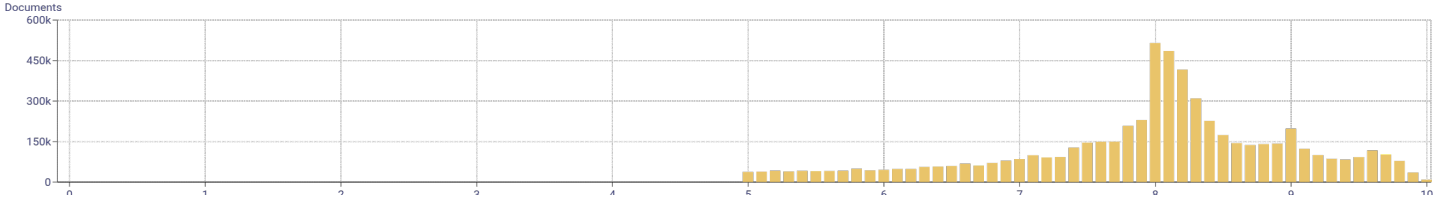
Number of segments in the Tamil (ta) corpus



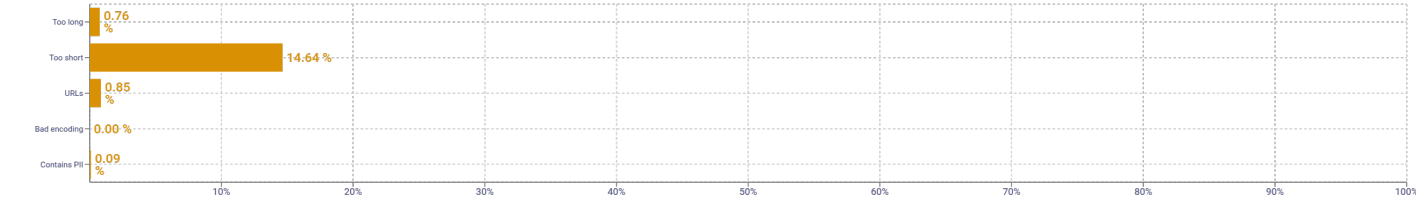
Percentage of segments in Tamil (ta) inside documents



Distribution of documents by document score



Segment noise distribution



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 *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>, , , 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 | | | | |