# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Date | SL | TL |
|---|---|---|---|
| hplt-v2-en-gu.tsv | 1/21/2025 | English (en) | Gujarati (gu) |

## Volumes

| Segments | SL tokens | SL characters | SL size |
|---|---|---|---|
| 716,777 | 19M | 95,695,067 | 91.74 MB |

| TL tokens | TL characters | TL size |
|---|---|---|
| 19M | 99,124,140 | 239.04 MB |

## Dataset top 10 domains

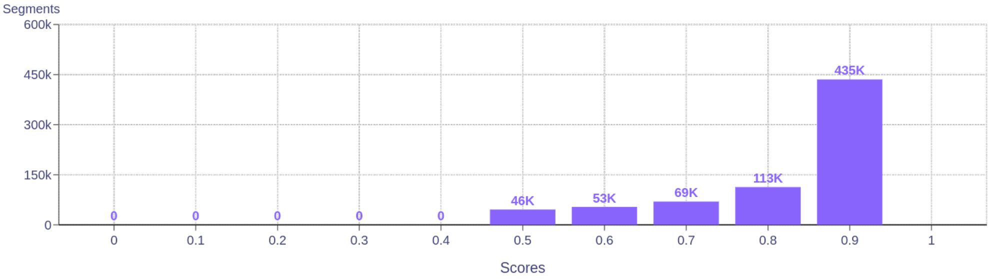
| SL domain | Segments | TL domain | Segments |
|---|---|---|---|
| wikipedia.org | 12.7% | wikipedia.org | 8.8% |
| educationbro.com | 5.0% | itsmygame.org | 3.8% |
| itsmygame.org | 4.9% | websiterating.com | 3.5% |
| websiterating.com | 3.5% | educationbro.com | 2.4% |
| vessoft.com | 2.1% | wondershare.com | 1.8% |
| wondershare.com | 1.6% | vessoft.in | 1.7% |
| worldwidescripts.net | 1.2% | news18.com | 1.2% |
| teesupport.com | 1.2% | worldwidescripts.net | 1.2% |
| angelone.in | 1.2% | angelone.in | 1.2% |
| whatsapp.com | 1.2% | whatsapp.com | 1.1% |

## Dataset top 10 TLDs

| SL domain | Segments | TL domain | Segments |
|---|---|---|---|
| com | 90.1% | com | 69.9% |
| org | 27.9% | org | 20.0% |
| in | 7.7% | in | 10.2% |
| net | 6.0% | net | 4.9% |
| top | 1.5% | top | 1.5% |
| io | 1.2% | co.in | 1.3% |
| gov.in | 1.1% | io | 1.1% |
| trade | 1.1% | gov.in | 0.9% |
| info | 0.9% | trade | 0.8% |
| plus | 0.9% | info | 0.8% |

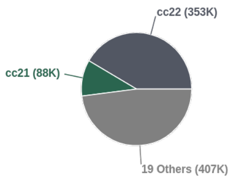
## Translation likelihood

≥ 5 = 717K segments | **100.0%**
≥ 8 = 548K segments | **76.5%**
< 5 = 0 segments | **0.0%**



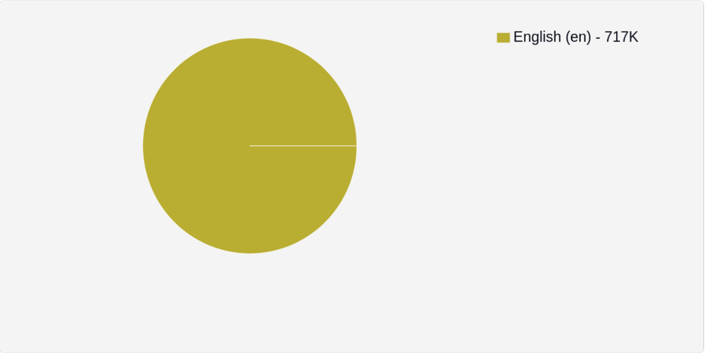
## Collections

**CC = 67.11%**
**IA = 32.89%**

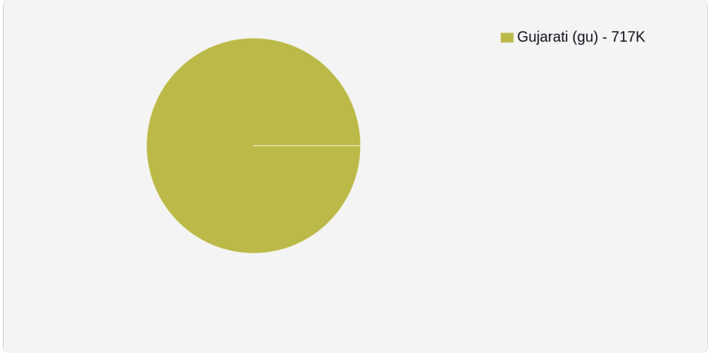


## Language Distribution

### Source

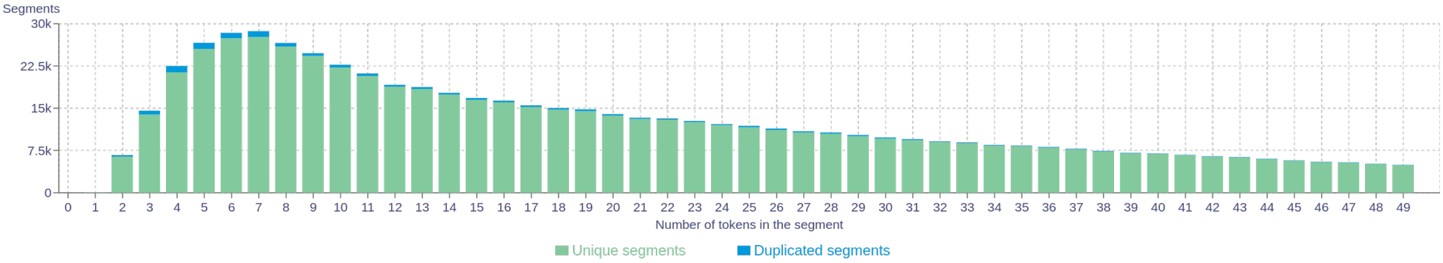


### Target



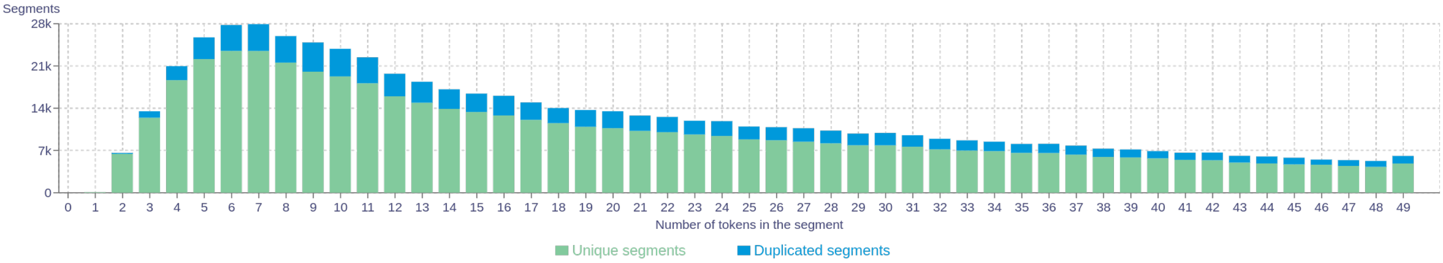
## Source segment length distribution by token

**<= 49** tokens = **607K** segments | **14K** duplicates
**> 50** tokens = **96K** segments | **1.1K** duplicates

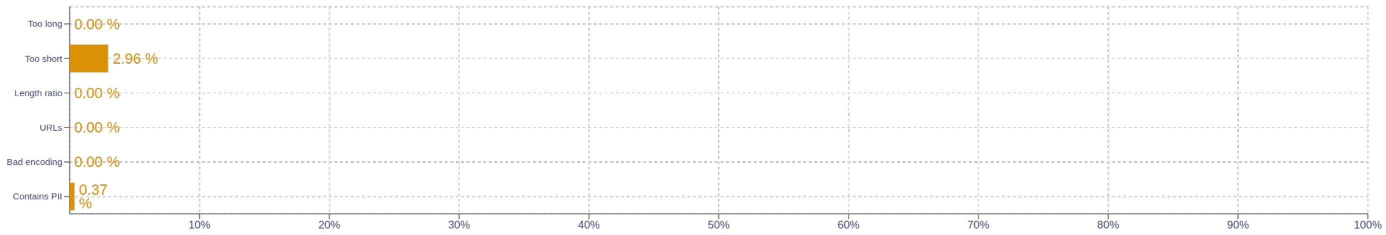


## Target segment length distribution by token

**<= 49** tokens = **500K** segments | **109K** duplicates
**> 50** tokens = **107K** segments | **20K** duplicates



## Segment pair noise distribution



## Source n-grams

| Size | n-grams |
|---|---|
| 1 | also \| 37968<br>game \| 32236<br>one \| 30495<br>new \| 29514<br>use \| 26958 |
| 2 | prime minister \| 4033<br>personal information \| 3339<br>html code \| 3284<br>new delhi \| 2905<br>united states \| 2853 |
| 3 | like the game \| 4593<br>send the link \| 3271<br>share the game \| 3270<br>copy and send \| 3269<br>copy the code \| 3266 |
| 4 | link to a friend \| 3269<br>game with the world \| 3269<br>paste in the html \| 3258<br>code of your site \| 3258<br>characteristics of the game \| 1986 |
| 5 | friend or all your friends \| 3269<br>copy and send the link \| 3269<br>copy the code and paste \| 3259<br>paste in the html code \| 3258<br>html code of your site \| 3258 |

## Target n-grams

| Size | n-grams |
|---|---|
| 1 | સાથે \| 106427<br>કરો \| 82743<br>કરવા \| 76711<br>તમારા \| 74698<br>દ્વારા \| 54206 |
| 2 | ક્લિક કરો \| 10193<br>ફેરફાર કરો \| 8216<br>તમારી વેબસાઇટ \| 6932<br>કરવામાં આવશે \| 5522<br>પસંદ કરો \| 5436 |
| 3 | તમારા બધા મિત્રો \| 3284<br>પેસ્ટ નકલ કરો \| 3273<br>સાઇટ ના html \| 3271<br>કોડ અને પેસ્ટ \| 3271<br>html કોડ કોડ \| 3271 |
| 4 | વેબસાઇટ પર આ રમત \| 5874<br>સાઇટ ના html કોડ \| 3271<br>તમારી સાઇટ ના html \| 3271<br>કોડ કોડ અને પેસ્ટ \| 3271<br>કોડ અને પેસ્ટ નકલ \| 3271 |
| 5 | તમારી વેબસાઇટ પર આ રમત \| 5874<br>સાઇટ ના html કોડ કોડ \| 3271<br>તમારી સાઇટ ના html કોડ \| 3271<br>કોડ કોડ અને પેસ્ટ નકલ \| 3271<br>કોડ અને પેસ્ટ નકલ કરો \| 3271 |

# 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