# HPLT Analytics report

HPLTAnalytics

## General overview

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

## Volumes

| Segments | SL tokens | SL characters | SL size |
|---|---|---|---|
| 317,120 | 8.8M | 45,532,138 | 43.67 MB |

| TL tokens | TL characters | TL size |
|---|---|---|
| 8.3M | 46,279,124 | 114.41 MB |

## Dataset top 10 domains

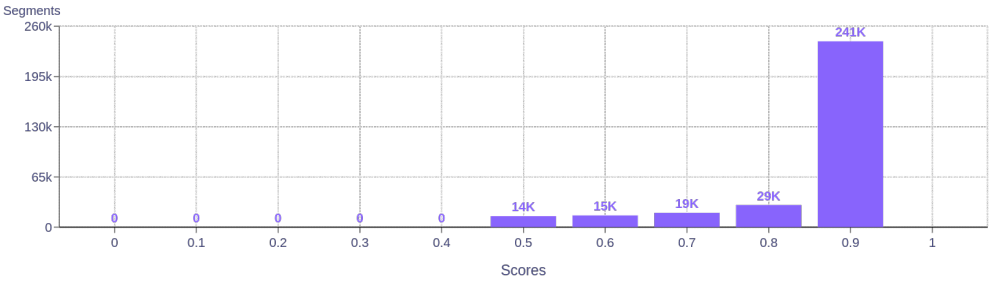
| SL domain | Segments | TL domain | Segments |
|---|---|---|---|
| biblegateway.com | 16.5% | biblegateway.com | 14.8% |
| educationbro.com | 7.4% | vessoft.com | 6.0% |
| vessoft.com | 7.0% | ofunnygames.com | 4.1% |
| ofunnygames.com | 4.6% | educationbro.com | 3.4% |
| manuals.plus | 2.1% | wikipedia.org | 1.8% |
| ustraveldocs.com | 2.0% | manuals.plus | 1.2% |
| wikipedia.org | 1.9% | jw.org | 1.1% |
| jw.org | 1.3% | pandaily.com | 1.0% |
| teesupport.com | 1.2% | eturbonews.com | 0.9% |
| pandaily.com | 1.0% | aoxactuator.com | 0.9% |

## Dataset top 10 TLDs

| SL domain | Segments | TL domain | Segments |
|---|---|---|---|
| com | 114.5% | com | 90.9% |
| org | 15.6% | org | 13.8% |
| net | 4.1% | net | 3.3% |
| plus | 2.1% | plus | 1.2% |
| eu | 1.0% | online | 0.8% |
| in | 0.9% | ru | 0.8% |
| online | 0.8% | zone | 0.7% |
| ru | 0.8% | gov.np | 0.6% |
| zone | 0.8% | eu | 0.6% |
| co.uk | 0.8% | com.np | 0.5% |

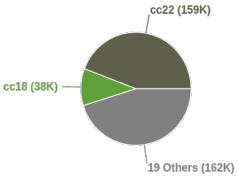
## Translation likelihood

≥ 5 = 317K segments | **100.0%**
≥ 8 = 269K segments | **84.9%**
< 5 = 0 segments | **0.0%**



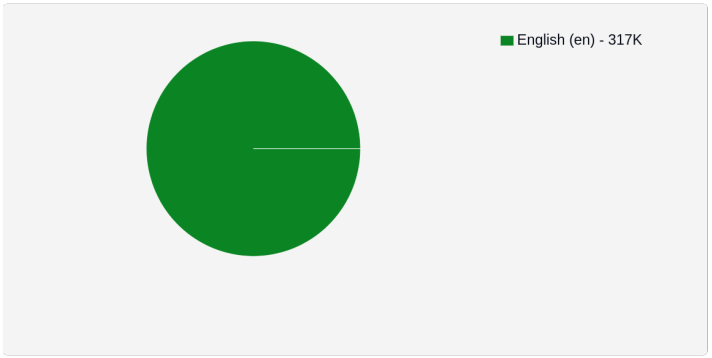

## Collections

**CC = 73.07%**
**IA = 26.93%**

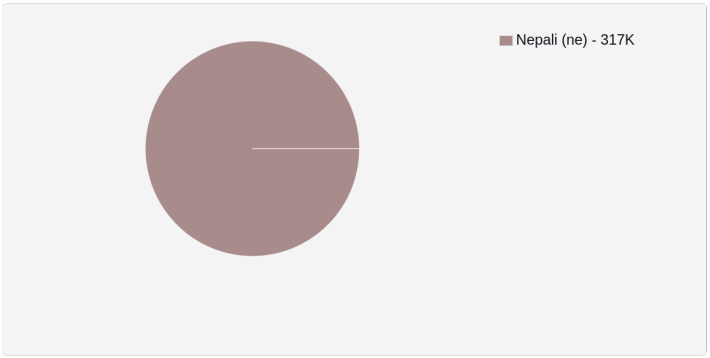



## Language Distribution

### Source

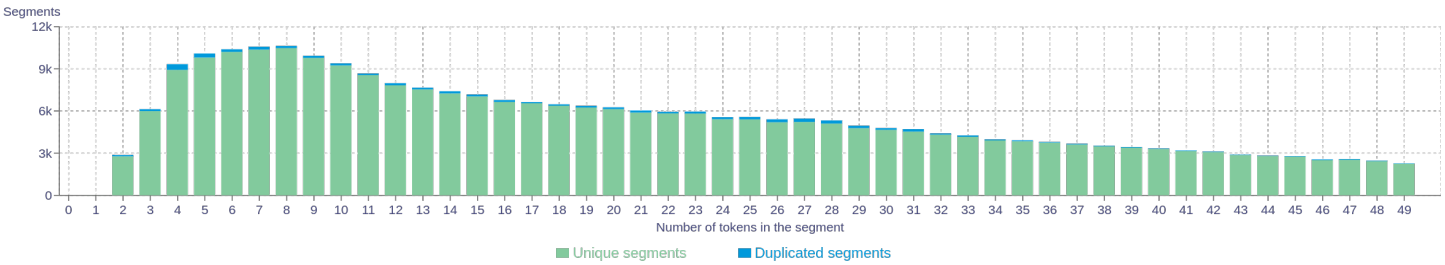



### Target



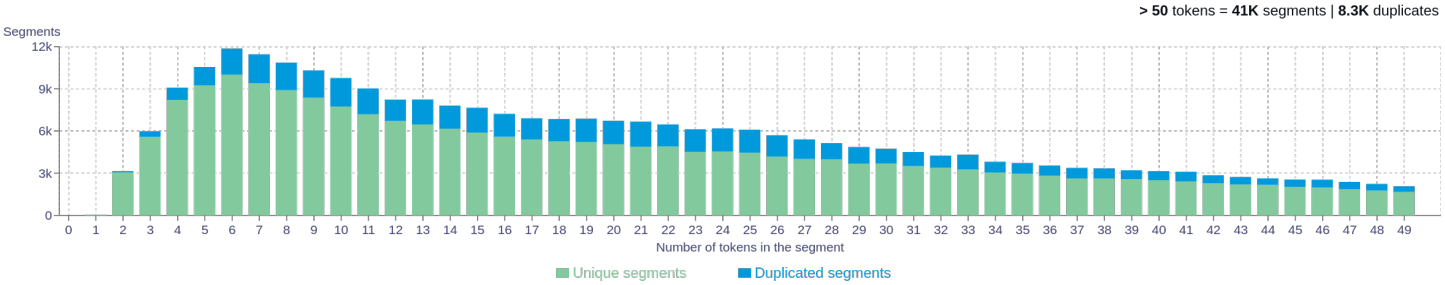

## Source segment length distribution by token

**<= 49** tokens = **264K** segments | **5.8K** duplicates
**> 50** tokens = **47K** segments | **523** duplicates

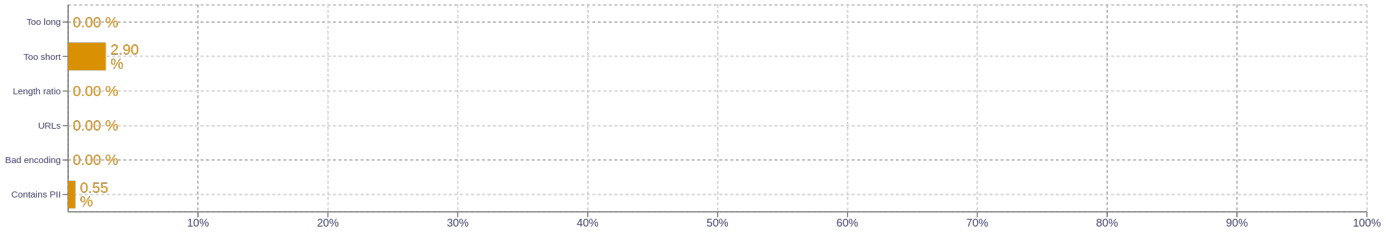



## Target segment length distribution by token

**<= 49** tokens = **220K** segments | **56K** duplicates
**> 50** tokens = **41K** segments | **8.3K** duplicates



### Segment pair noise distribution




### Source n-grams

| Size | n-grams |
|---|---|
| 1 | said \| 15602 · lord \| 15276 · also \| 14738 · one \| 14627 · use \| 14309 |
| 2 | personal information \| 2286 · united states \| 1910 · privacy policy \| 1867 · personal data \| 1736 · prime minister \| 1406 |
| 3 | name of jesus \| 1100 · around the world \| 808 · lord your god \| 701 · king of israel \| 640 · terms and conditions \| 585 |
| 4 | word of the lord \| 549 · north korean leader kim \| 519 · house of the lord \| 417 · offer you the best \| 388 · after-sale service and timely \| 378 |
| 5 | offer you the best after-sale \| 378 · best after-sale service and timely \| 378 · after-sale service and timely delivery \| 368 · north korean leader kim jong-un \| 271 · word of the lord came \| 233 |

### Target n-grams

| Size | n-grams |
|---|---|
| 1 | वा \| 40862 · तपाई \| 30834 · हामी \| 28735 · प्रयोग \| 26491 · रूपमा \| 23836 |
| 2 | प्रदान गर्दछ \| 3812 · अनुमति दिन्छ \| 3474 · प्रयोग गरिन्छ \| 2349 · संयुक्त राज्य \| 2343 · हामी तपाईलाई \| 2133 |
| 3 | प्रयोग गर्न सकिन्छ \| 1441 · संयुक्त राज्य अमेरिका \| 1361 · थप पढ्नुहोस्जाँच पठाउनुहोस् \| 700 · प्रयोग गर्न सक्नुहुन्छ \| 609 · हामीलाई सम्पर्क गर्नुहोस् \| 559 |
| 4 | उत्तर कोरियाली नेता किम \| 373 · सेवा र समयमै डेलिभरी \| 332 · समयमै डेलिभरी प्रदान गर्नेछौं \| 320 · व्यापक रूपमा प्रयोग गरिन्छ \| 285 · सक्नुहुन्छ र हामी तपाईलाई \| 223 |
| 5 | सेवा र समयमै डेलिभरी प्रदान \| 320 · बिक्री पछि सेवा र समयमै \| 223 · आश्वस्त हुन सक्नुहुन्छ र हामी \| 160 · किन्नको लागि आश्वस्त हुन सक्नुहुन्छ \| 158 · ट्रम्प र उत्तर कोरियाली नेता \| 152 |

## 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