# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Analytics date | Source language | Target language |
|---|---|---|---|
| HPLT.en-eu | 10/25/2023 | English (en) | Basque (eu) |

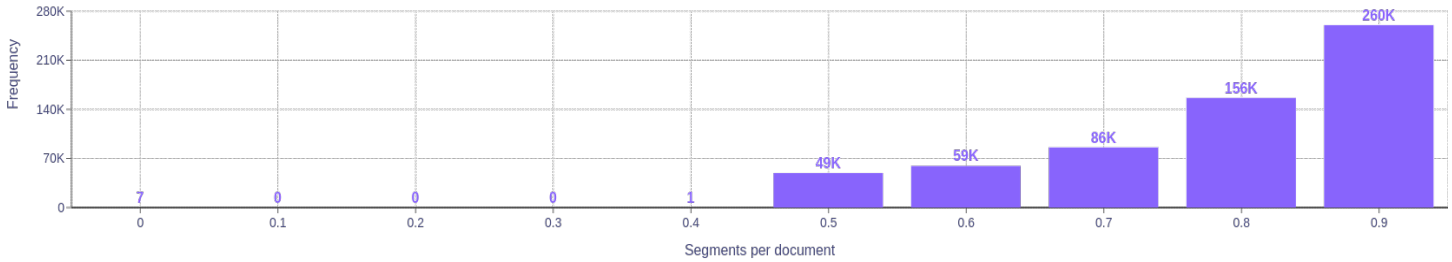
## Volumes

| Segments | Unique segments | Src tokens | Trg tokens | Src size | Trg size |
|---|---|---|---|---|---|
| 610,694 | 2,095 (0.34 %) | 12M | 9.7M | 59.02 MB | 61.91 MB |

## Type-Token Ratio

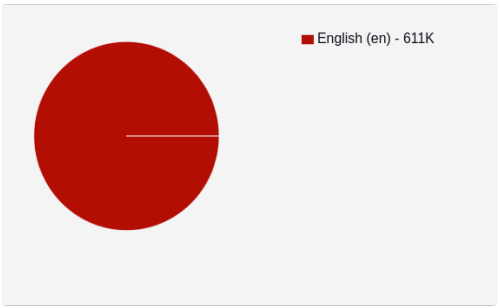
| Source | Target |
|---|---|
| 0.02 | 0.06 |

## Translation likelihood

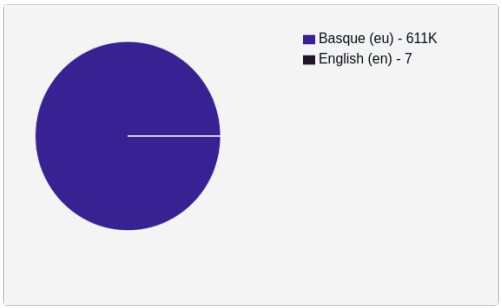



## Language Distribution

### Source



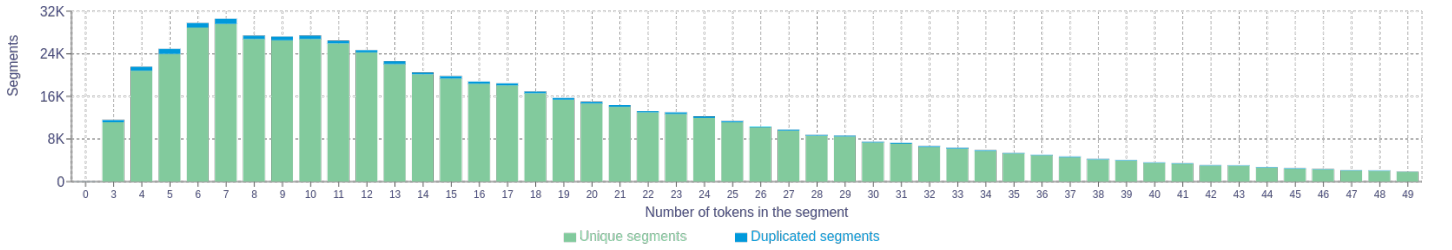

### Target




## Source segment length distribution by token

**<= 49** tokens = **571K** segments | **14K** duplicates

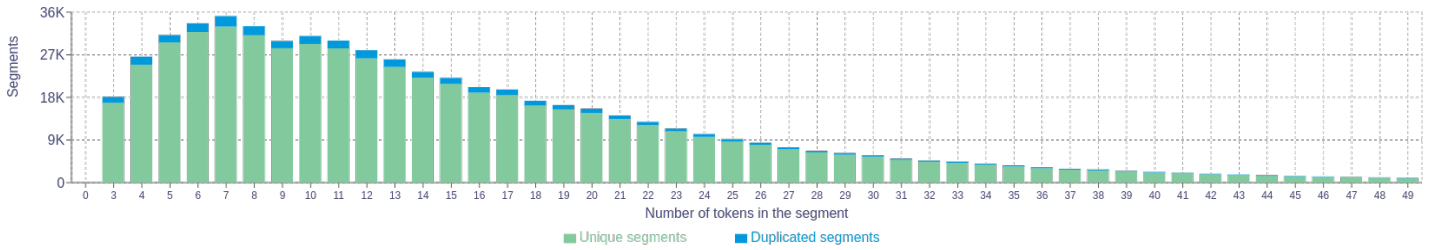
**> 50** tokens = **26K** segments | **501** duplicates



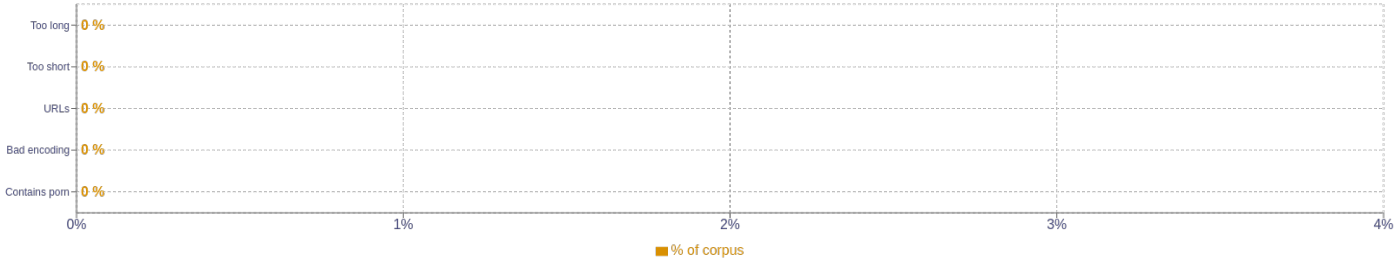

## Target segment length distribution by token

**<= 49** tokens = **564K** segments | **35K** duplicates

**> 50** tokens = **12K** segments | **850** duplicates



### Segment pair noise distribution




### Source n-grams

| Size | n-grams |
|---|---|
| 1 | information \| 18854 · one \| 18626 · basque \| 17897 · use \| 16812 · new \| 16674 |
| 2 | built surface \| 6667 · basque country \| 5721 · united states \| 4209 · personal data \| 4072 · get full \| 3546 |
| 3 | get full analysis \| 3544 · analysis of surname \| 2437 · name and surname \| 1610 · analysis of name \| 1103 · time genie timegenie \| 1020 |
| 4 | full analysis of surname \| 2437 · distance to the sea \| 1555 · full analysis of name \| 1103 · registered on our database \| 895 · male get full analysis \| 874 |
| 5 | get full analysis of surname \| 2437 · get full analysis of name \| 1103 · university of the basque country \| 544 · try one of these games \| 526 · wikimedia commons has media related \| 486 |

### Target n-grams

| Size | n-grams |
|---|---|
| 1 | behar \| 21573 · izango \| 18744 · egiten \| 17726 · nahi \| 16767 · duen \| 15944 |
| 2 | eraikitako azalera \| 6694 · ibilbide en \| 5812 · ahal izango \| 4916 · estatu batuak \| 3221 · entziklopedia askea \| 3131 |
| 3 | lortu abizenaren analisi \| 2445 · abizenaren analisi osoa \| 2441 · ameriketako estatu batuak \| 1895 · atzeko plano pertsonalizatua \| 1569 · bilaketarekin bat datozen \| 1349 |
| 4 | lortu abizenaren analisi osoa \| 2441 · bilaketarekin bat datozen emaitzak \| 1348 · bilatu filtros zure bilaketarekin \| 1342 · commonsen badira fitxategi gehiago \| 1105 · wikimedia commonsen badira fitxategi \| 1101 |
| 5 | filtros zure bilaketarekin bat datozen \| 1342 · wikimedia commonsen badira fitxategi gehiago \| 1101 · ohikoena eta ezohikoena den abizena \| 800 · gizonezkoa talde izenaren azterketa osoa \| 559 · artisten eta kultur arloko eragileen \| 450 |

## 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).

**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.

**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