# HPLT Analytics report

## General overview

| Corpus | Analytics date | Language |
|---|---|---|
| lmo_Latn.jsonl.tsv | 11/5/2024 | Lombard (lmo) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Size | Characters |
|---|---|---|---|---|---|
| 146,162 | 2,124,663 | 841,958 (39.63 %) | 76M | 340.5 MB | 343,386,720 |

### Top 10 domains

| Domain | Docs | % of total |
|---|---|---|
| wikipedia.org | 62K | 42.60 |
| rtr.ch | 53K | 36.38 |
| gr.ch | 2K | 1.33 |
| uslaval.it | 1.4K | 0.93 |
| provinzia.bz.it | 940 | 0.64 |
| admin.ch | 714 | 0.49 |
| wikisource.org | 702 | 0.48 |
| noeles.info | 612 | 0.42 |
| playsuisse.ch | 607 | 0.42 |
| blogspot.com | 603 | 0.41 |

### Top 10 TLDs

| Domain | Docs | % of total |
|---|---|---|
| ch | 69K | 46.99 |
| org | 64K | 44.02 |
| it | 5.1K | 3.47 |
| com | 3.2K | 2.19 |
| bz.it | 1.4K | 0.99 |
| info | 935 | 0.64 |
| net | 606 | 0.41 |
| is | 551 | 0.38 |
| de | 157 | 0.11 |
| eu | 123 | 0.08 |

## Documents size (in segments)

<= 25 segments **90.3%** (132K documents)
> 25 segments **9.7%** (14K documents)
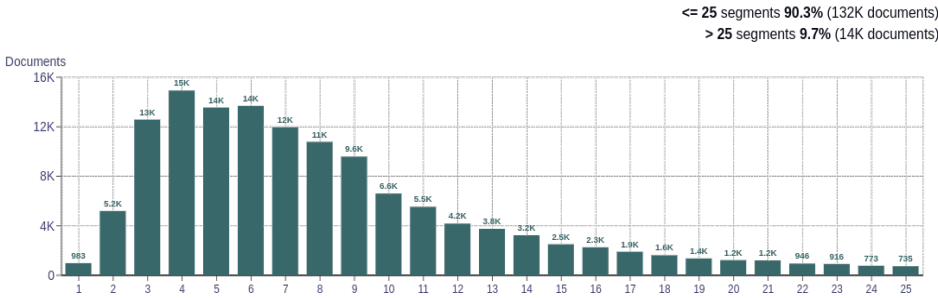


## Documents by collection

cc18 (27K) · cc22 (40K) · cc21 (26K) · 18 Others (54K)
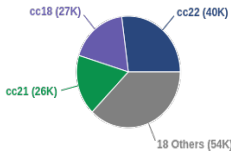


## Language Distribution

### Number of segments

- Italian (it) - 548K
- Romansh (rm) - 229K
- Lombard (lmo) - 221K
- Catalan (ca) - 163K
- eml (eml) - 134K
- English (en) - 132K
- German (de) - 104K
- Spanish (es) - 90K
- French (fr) - 87K
- Piedmontese (pms) - 40K
- 161 Others - 377K
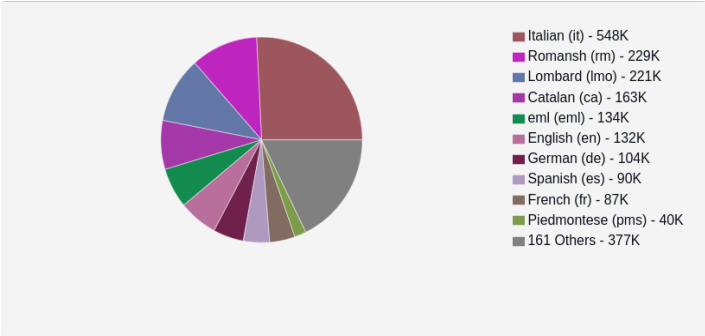


### Percentage of segments in Lombard (lmo) inside documents
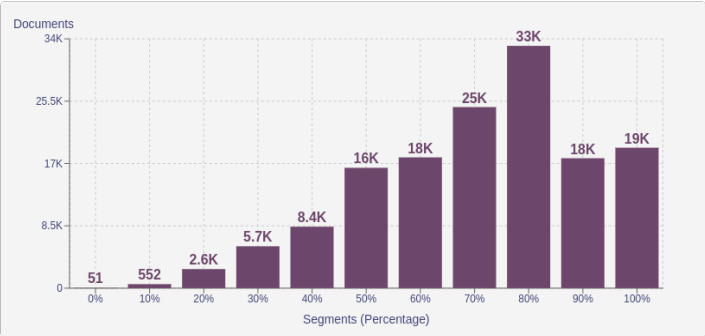


## Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (146K documents)



## Segment length distribution by token

<= 49 tokens = **605K** segments | **1M** duplicates
> 50 tokens = **494K** segments | **257K** duplicates



- Unique segments
- Duplicated segments

## Segment noise distribution

| | |
|---|---|
| Too long | 0.00 % |
| Too short | 5.50 % |
| URLs | 0.97 % |
| Bad encoding | 0.02 % |
| Contains PII | 0.16 % |
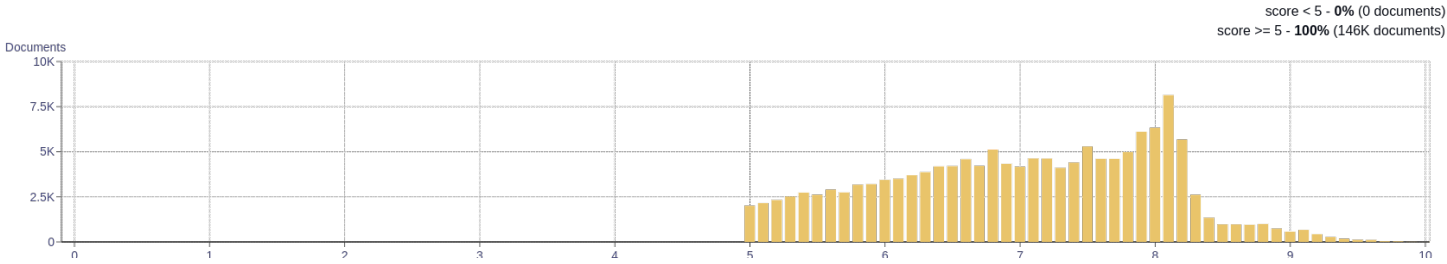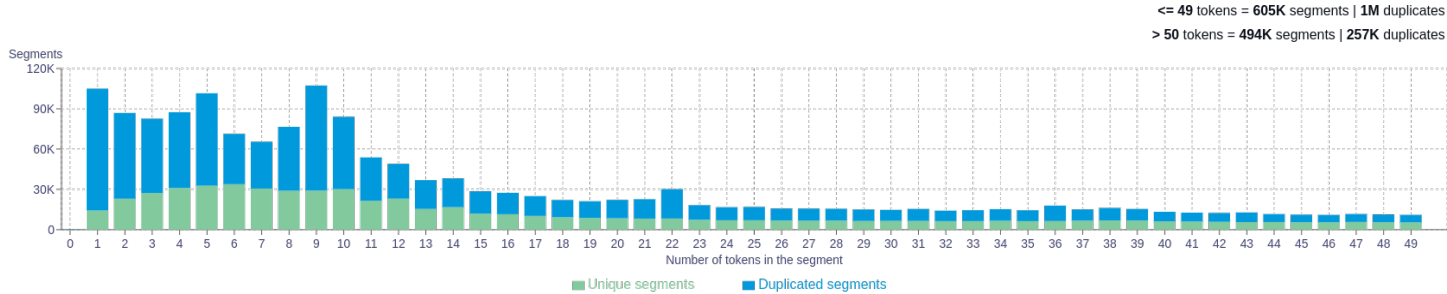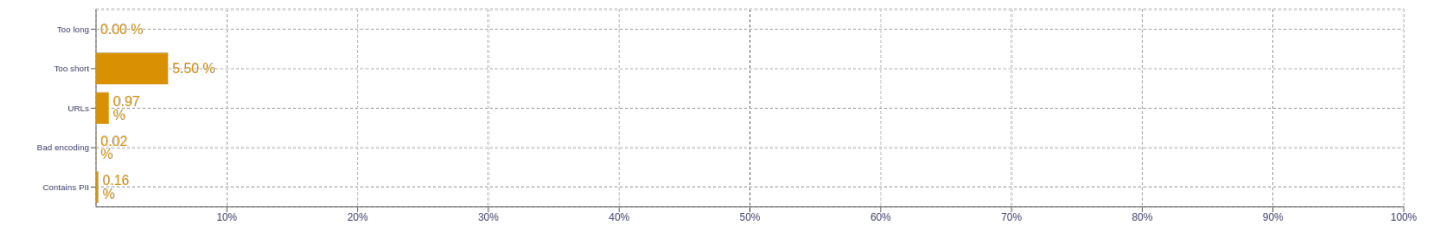
## About HPLT Analytics

**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