# HPLT Analytics report

**HPLT**Analytics

## General overview

| Corpus | Date | Language |
|---|---|---|
| hplt-v3-ita_Latn | 9/18/2025 | Italian |

## Volumes

| Docs | Segments | Unique segments | Tokens | Characters | Size |
|---|---|---|---|---|---|
| 362,985,922 | 7,535,521,218 | 4,066,403,281 (53.96 %) | 238B | 1,294,420,645,275 | 1.19 TB |

## Top 10 domains

| Domain | Docs | % of total |
|---|---|---|
| blogspot.com | 6.8M | 1.88% |
| wordpress.com | 3.9M | 1.09% |
| blogspot.it | 1.8M | 0.51% |
| kijiji.it | 1.5M | 0.42% |
| altervista.org | 1.3M | 0.36% |
| docplayer.it | 1M | 0.28% |
| repubblica.it | 971K | 0.27% |
| corriere.it | 969K | 0.27% |
| tripadvisor.it | 856K | 0.24% |
| ilsole24ore.com | 817K | 0.23% |

## Top 10 TLDs

| Domain | Docs | % of total |
|---|---|---|
| it | 207M | 57.06% |
| com | 94M | 26.01% |
| org | 13M | 3.46% |
| net | 12M | 3.31% |
| eu | 6.3M | 1.74% |
| info | 4.4M | 1.21% |
| ch | 3.1M | 0.87% |
| tv | 1.7M | 0.47% |
| es | 1.2M | 0.33% |
| ru | 957K | 0.26% |

## Register labels



- HI - 2.7%
- ID - 2.0%
- IN - 11.6%
- IP - 25.6%
- LY - 0.2%
- MIX - 5.5%
- NA - 36.0%
- OP - 7.3%
- SP - 0.7%
- UNK - 8.5%

- HI_other - 1.4%
- HI_re - 1.2%
- ID_other - 2.0%
- IN_dtp - 5.0%
- IN_en - 0.6%
- IN_fi - 0.0%
- IN_lt - 0.7%
- IN_other - 5.2%
- IN_ra - 0.1%
- IP_ds - 21.3%
- IP_ed - 0.0%
- IP_other - 4.3%
- LY_other - 0.2%
- MIX - 5.5%
- NA_nb - 3.6%
- NA_ne - 22.5%
- NA_other - 3.9%
- NA_sr - 6.0%
- OP_av - 0.6%
- OP_ob - 2.1%
- OP_other - 1.5%
- OP_rs - 0.7%
- OP_rv - 2.3%
- SP_it - 0.5%
- SP_other - 0.1%
- UNK - 8.5%
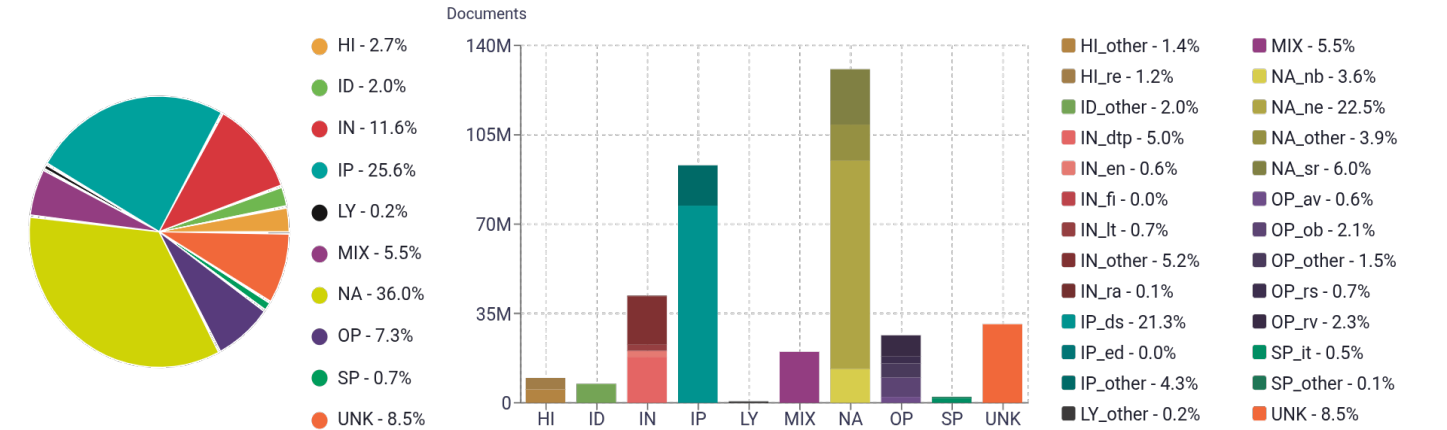
**MT**:5.3% | 19M Documents

## Documents size (in segments) ⓘ

≤ 25 segments **81.19%** (295M documents)
> 25 segments **18.81%** (68M documents)



## Document collections

CC = 90.55%
IA = 9.45%



67 Others (363M)
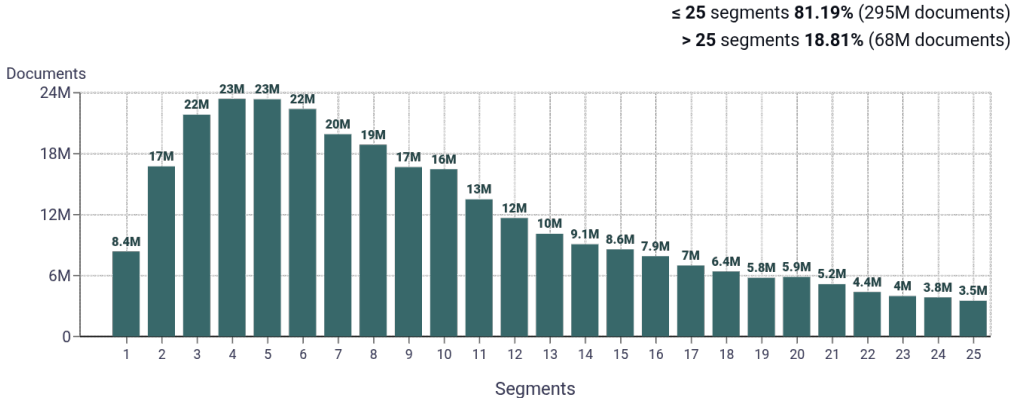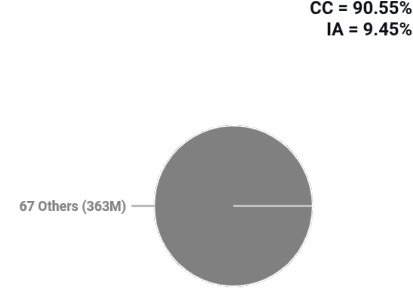
## Language Distribution

### Number of segments in the Italian corpus

- Italian - 6.6B **(88.0%)**
- English - 414M **(5.5%)**
- French - 116M **(1.5%)**
- Spanish - 85M **(1.1%)**
- German - 43M **(0.6%)**
- Portuguese - 40M **(0.5%)**
- Arabic - 28M **(0.4%)**
- Dutch - 20M **(0.3%)**
- Russian - 16M **(0.2%)**
- Catalan - 12M **(0.2%)**
- 165 Others - 131M **(1.7%)**
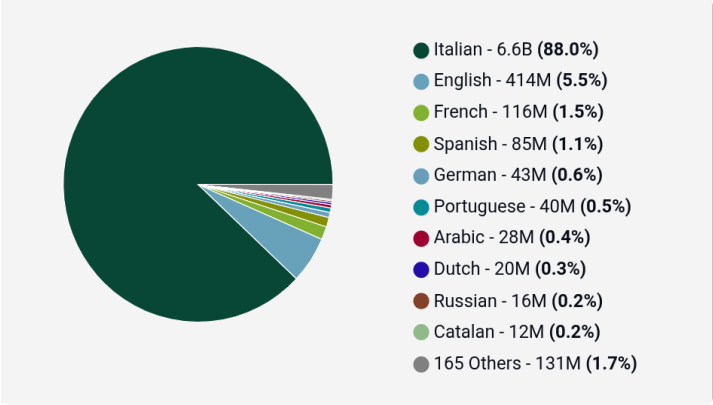
### Percentage of segments in Italian inside documents
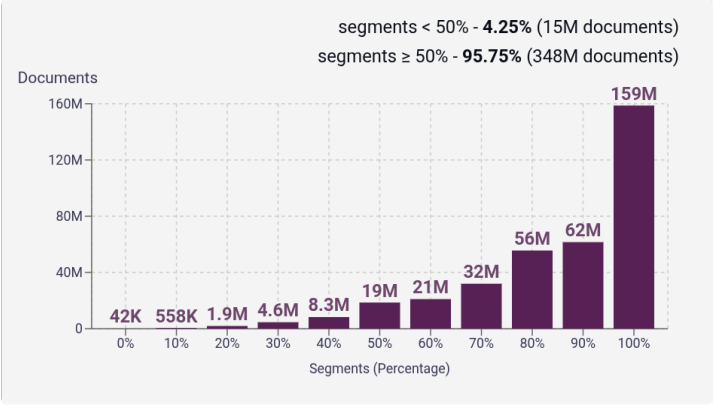
segments < 50% - **4.25%** (15M documents)
segments ≥ 50% - **95.75%** (348M documents)

Documents

| Segments (Percentage) | Documents |
|---|---|
| 0% | 42K |
| 10% | 558K |
| 20% | 1.9M |
| 30% | 4.6M |
| 40% | 8.3M |
| 50% | 19M |
| 60% | 21M |
| 70% | 32M |
| 80% | 56M |
| 90% | 62M |
| 100% | 159M |

### Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (363M documents)

Documents

### Segment length distribution by token

≤ **49** tokens = **6.1B** segments | **3.2B** duplicates
> **50** tokens = **1.4B** segments | **299M** duplicates

Segments

### Segment noise distribution

| | % of corpus |
|---|---|
| Too long | 1.51% |
| Too short | 15.39% |
| URLs | 2.10% |
| Bad encoding | 0.01% |
| Contains PII | 0.85% |

■ % of corpus
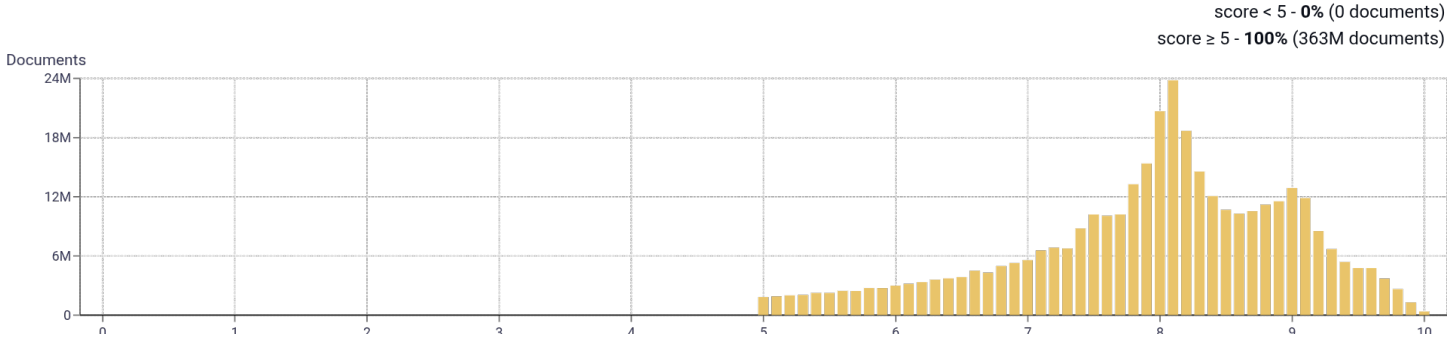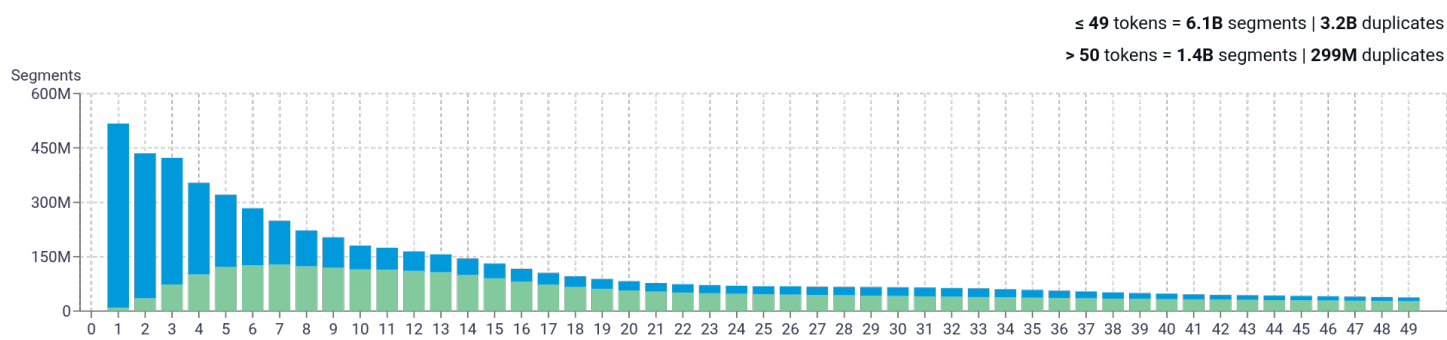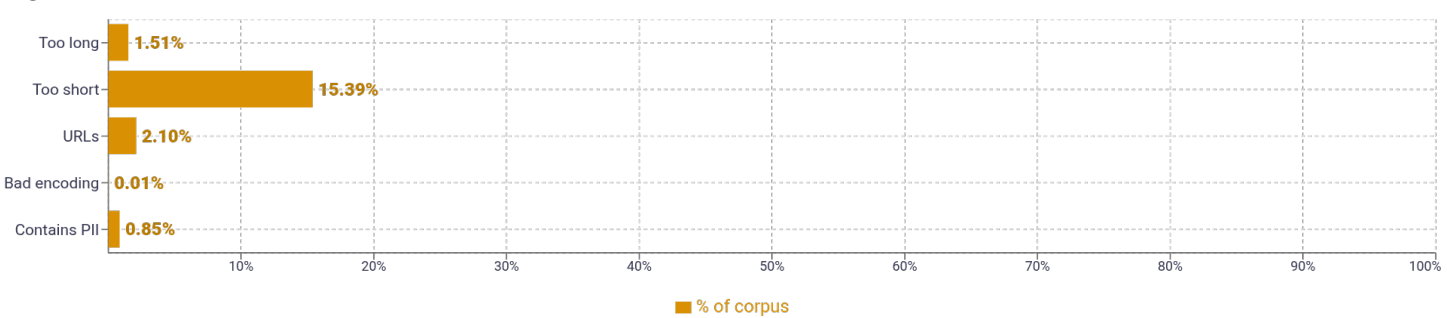
# Frequent n-grams

| SIZE | N-GRAMS | |
|---|---|---|
| 1 | essere \| 376,842,948   solo \| 277,436,536   anni \| 263,607,694   due \| 247,680,609   prima \| 242,934,848 | ⧉ |
| 2 | può essere \| 59,512,639   possono essere \| 31,827,562   dopo aver \| 25,119,554   deve essere \| 24,144,956   video porno \| 22,964,172 | ⧉ |
| 3 | continua a leggere \| 19,516,688   punto di vista \| 16,775,493   donna cerca uomo \| 12,954,258   milioni di euro \| 12,190,188   perdita di peso \| 11,118,312 | ⧉ |
| 4 | maggior parte dei casi \| 1,861,710   soggiorno con angolo cottura \| 1,711,764   donne che cercano uomini \| 1,588,728   due camere da letto \| 1,537,450   trattamento dei dati personali \| 1,330,810 | ⧉ |
| 5 | ancora non ci sono recensioni \| 1,682,214   donne in cerca di uomini \| 1,610,348   slots free spins no deposit \| 1,296,405   vita di tutti i giorni \| 1,113,824   stimato in base alle ricerche \| 1,061,137 | ⧉ |

# 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

**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 | | | | |