# HPLT Analytics report

HPLTAnalytics

## General overview

| Corpus | Date | Language |
|---|---|---|
| ron_Latn.jsonl.tsv | 7/2/2025 | Romanian (ro) |

## Volumes

| Docs | Segments | Unique segments | Tokens | Characters | Size |
|---|---|---|---|---|---|
| 65,876,360 | 1,696,435,521 | 613,954,867 (36.19 %) | 47B | 249,026,904,418 | 239.79 GB |

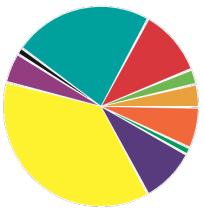
## Top 10 domains

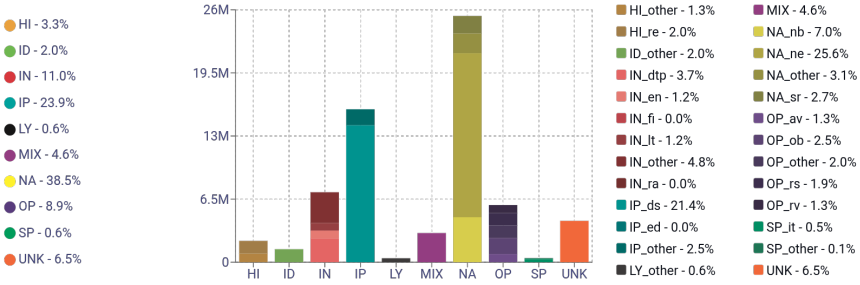
| Domain | Docs | % of total |
|---|---|---|
| blogspot.com | 1.9M | 2.83% |
| wordpress.com | 1.6M | 2.46% |
| blogspot.ro | 1.4M | 2.08% |
| 9am.ro | 810K | 1.23% |
| wikipedia.org | 473K | 0.72% |
| ziare.com | 462K | 0.70% |
| hotnews.ro | 423K | 0.64% |
| wall-street.ro | 363K | 0.55% |
| citatepedia.ro | 320K | 0.49% |
| ziuaconstanta.ro | 305K | 0.46% |

## Top 10 TLDs

| Domain | Docs | % of total |
|---|---|---|
| ro | 48M | 72.39% |
| com | 9.7M | 14.75% |
| md | 1.8M | 2.77% |
| net | 1.7M | 2.58% |
| org | 1.2M | 1.90% |
| eu | 837K | 1.27% |
| info | 774K | 1.17% |
| com.ro | 132K | 0.20% |
| tv | 124K | 0.19% |
| biz | 104K | 0.16% |

## Register labels

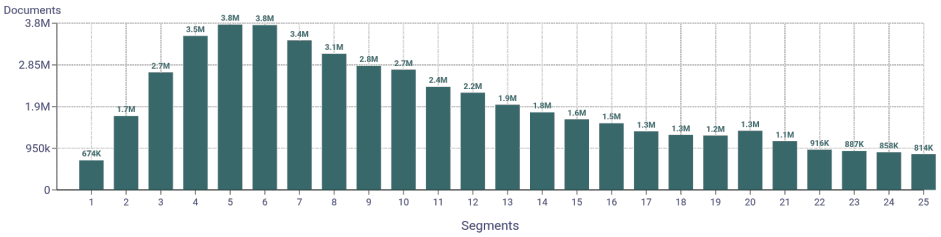





**MT**:2.8% | 1.8M Documents

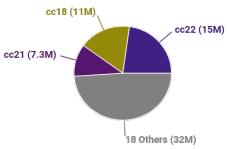
## Documents size (in segments)

**<= 25** segments **74.86%** (49M documents)
**> 25** segments **25.14%** (17M documents)



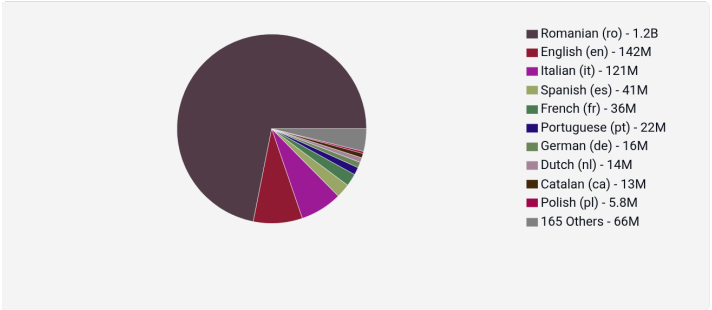

## Documents by collection

**CC = 61.30%**
**IA = 38.70%**

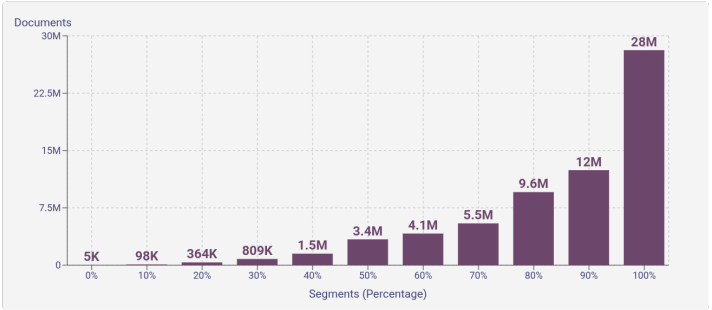



## Language Distribution

### Number of segments in the Romanian (ro) corpus

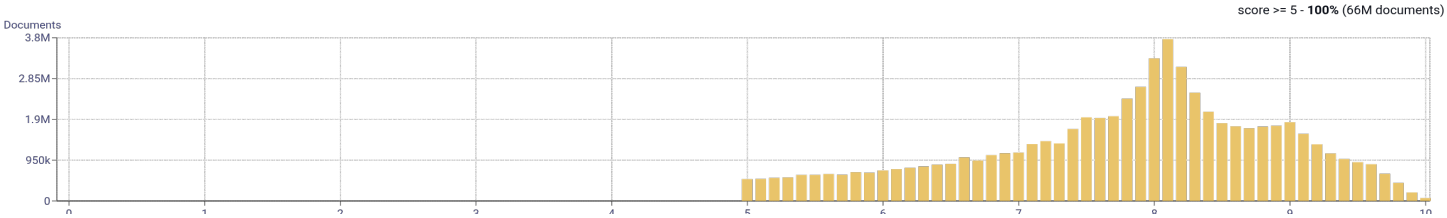



### Percentage of segments in Romanian (ro) inside documents



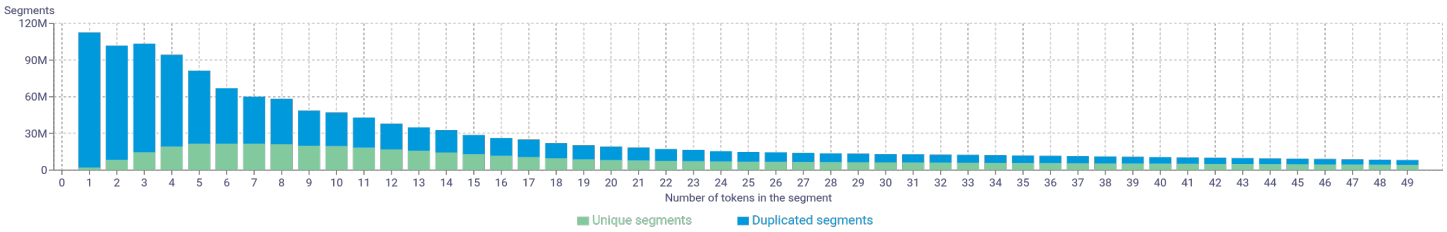

## Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (66M documents)

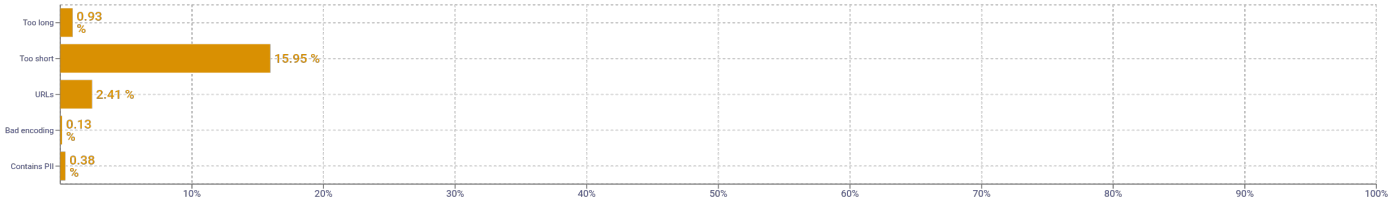


### Segment length distribution by token

**≤ 49** tokens = **1.4B** segments | **956M** duplicates
**> 50** tokens = **272M** segments | **131M** duplicates




### Segment noise distribution




### Frequent n-grams

| Size | n-grams |
|---|---|
| 1 | şi \| 319210101 · ani \| 55116609 · s-a \| 51432959 · trebuie \| 50996888 · mare \| 48168602 |
| 2 | precum şi \| 3323079 · de-a lungul \| 3266640 · anul trecut \| 3231450 · descrierea joc \| 3078139 · read more \| 2836339 |
| 3 | având în vedere \| 1460107 · citeste mai departe \| 1080082 · avand in vedere \| 1055181 · ani de zile \| 936920 · de-a lungul timpului \| 908283 |
| 4 | datelor cu caracter personal \| 659144 · acord scris din partea \| 614650 · caractere din acest articol \| 610316 · articol daca precizati sursa \| 517975 · adăugat de gheorghe culicovschi \| 446723 |
| 5 | aboneaza-te la 9am sau conecteaza-te \| 752785 · 9am sau conecteaza-te prin facebook \| 752785 · acord scris din partea internet \| 346751 · barbie reunite intr-o colectie unica \| 276638 · acestora revine integral autorului comentariului \| 230272 |

## 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. |
|---|---|
| **Machine-translated** | MT |
| **Lyrical** | LY |
| **Spoken** | SP |
| Interview | it |
| **Interactive discussion** | ID |
| **Narrative** | NA |
| News report | ne |
| Sports report | sr |
| Narrative blog | nb |

| Name | Abbr. |
|---|---|
| **How-to or instructions** | **HI** |
| Recipe | re |
| **Informational persuasion** | **IP** |
| Description with intent to sell | ds |
| News & opinion blog or editorial | ed |
| **Informational description** | **IN** |
| Enciclopedia article | en |
| Research article | ra |

| Name | Abbr. |
|---|---|
| Description of a thing or person | dtp |
| FAQ | fi |
| Legal terms & conditions | lt |
| **Opinion** | **OP** |
| Review | rv |
| Opinion blog | ob |
| Denominational religious blog or sermon | rs |
| Advice | av |