# HPLT Analytics report

## General overview

| Corpus | Date | Language |
|---|---|---|
| HPLT-v2-hrv_Latn.tsv | 9/19/2024 | Croatian (hr) |

### Volumes

| Docs | Segments | Characters | Size |
|---|---|---|---|
| 12,303,820 | 297,130,317 | 47,710,541,358 | 45.7 GB |

### Top 10 domains

| Domain | Docs | % of total |
|---|---|---|
| wikipedia.org | 264K | 2.15% |
| dnevnik.hr | 170K | 1.38% |
| blogspot.com | 138K | 1.12% |
| tportal.hr | 95K | 0.77% |
| skole.hr | 88K | 0.72% |
| jutarnji.hr | 84K | 0.68% |
| index.hr | 82K | 0.67% |
| metro-portal.hr | 73K | 0.59% |
| 24sata.hr | 66K | 0.54% |
| hrt.hr | 59K | 0.48% |

### Top 10 TLDs

| Domain | Docs | % of total |
|---|---|---|
| hr | 5.5M | 45.05% |
| com | 3.4M | 27.95% |
| org | 592K | 4.81% |
| net | 551K | 4.48% |
| ba | 421K | 3.43% |
| rs | 340K | 2.76% |
| info | 288K | 2.34% |
| com.hr | 247K | 2.01% |
| eu | 219K | 1.78% |
| news | 140K | 1.14% |

## Register labels



HI - 6.2%
ID - 1.4%
IN - 16.7%
IP - 24.5%
LY - 0.1%
MIX - 5.5%
NA - 27.5%
OP - 8.5%
SP - 0.7%
UNK - 8.7%

**MT**:6.0% | 738K Documents

HI_other - 2.6% | MIX - 5.5%
HI_re - 3.6% | NA_nb - 3.3%
ID_other - 1.4% | NA_ne - 19.9%
IN_dtp - 5.6% | NA_other - 3.7%
IN_en - 2.4% | NA_sr - 0.6%
IN_fi - 0.0% | OP_av - 2.0%
IN_lt - 2.0% | OP_ob - 1.4%
IN_other - 6.5% | OP_other - 1.6%
IN_ra - 0.2% | OP_rs - 1.9%
IP_ds - 21.8% | OP_rv - 1.6%
IP_ed - 0.0% | SP_it - 0.6%
IP_other - 2.7% | SP_other - 0.2%
LY_other - 0.1% | UNK - 8.7%
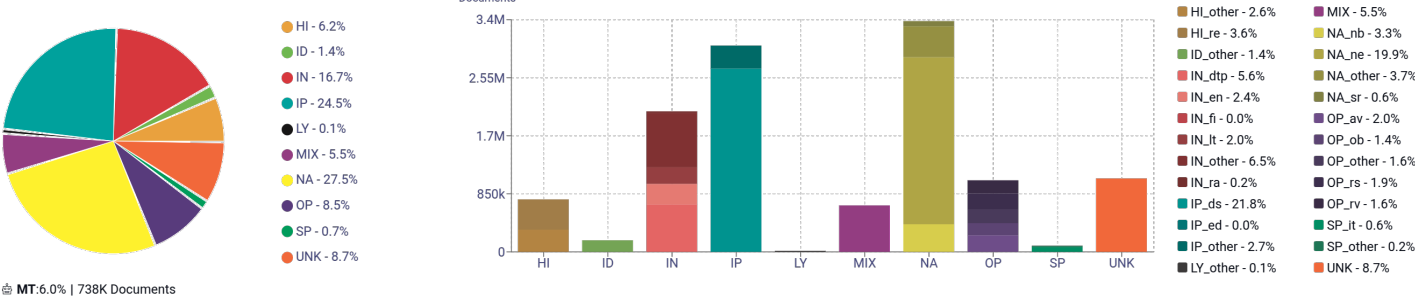
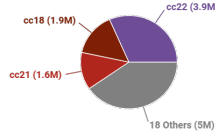## Documents size (in segments)
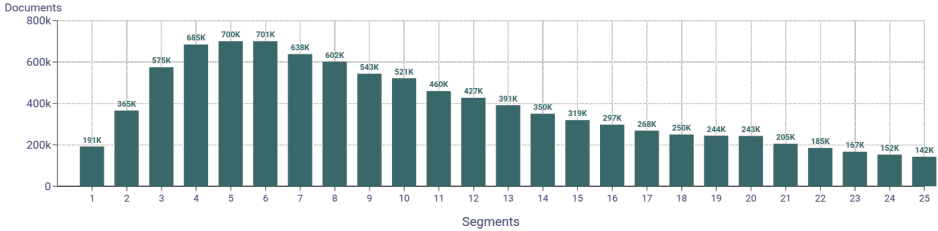
<= 25 segments **78.21%** (9.6M documents)
> 25 segments **21.79%** (2.7M documents)



## Documents by collection

CC = 69.33%
IA = 30.67%



cc22 (3.9M)
cc18 (1.9M)
cc21 (1.6M)
18 Others (5M)

## Language Distribution

### Number of segments in the Croatian (hr) corpus



Croatian (hr) - 232M
English (en) - 14M
Serbian (sr) - 12M
Italian (it) - 7.5M
Bosnian (bs) - 5M
Polish (pl) - 3.1M
French (fr) - 2.9M
German (de) - 2.6M
Czech (cs) - 2.2M
Portuguese (pt) - 1.3M
165 Others - 14M
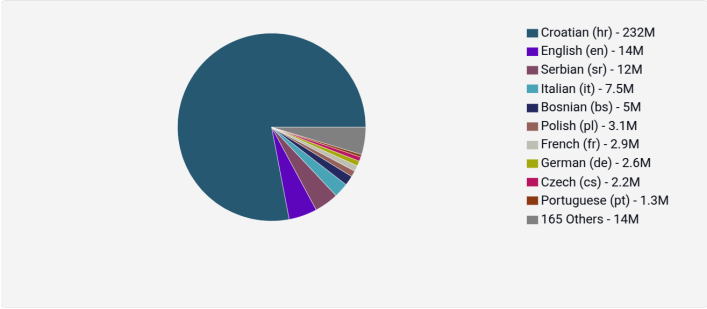
### Percentage of segments in Croatian (hr) inside documents
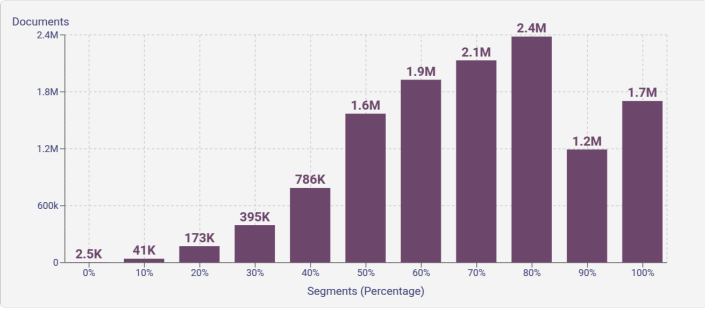


## Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (12M documents)



## Segment noise distribution



Too long - 1.06 %
Too short - 15.24 %
URLs - 1.82 %
Bad encoding - 0.00 %
Contains PII - 0.41 %
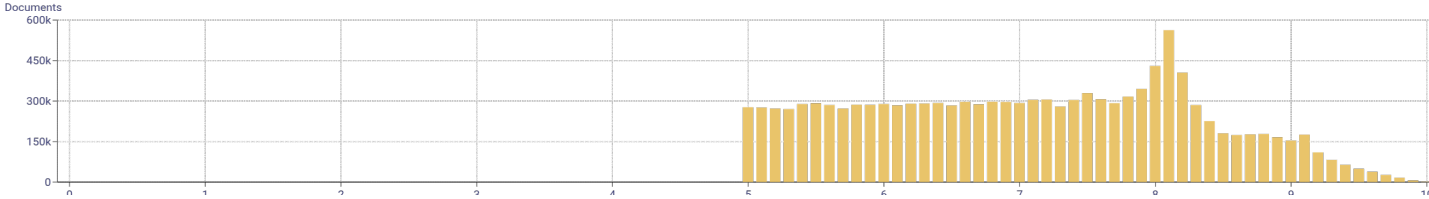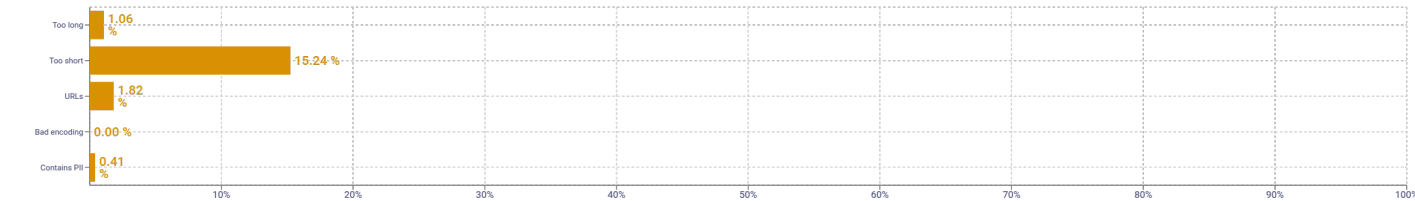
## About HPLT Analytics

**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 | | | | |