# HPLT Analytics report

**HPLT**Analytics

## General overview

| Corpus | Date | Language |
|---|---|---|
| hplt-v3-ekk_Latn | 10/3/2025 | Estonian (ekk) |

## Volumes

| Docs | Segments | Unique segments | Duplication ratio | Tokens | Characters | Size |
|---|---|---|---|---|---|---|
| 13,735,919 | 425,813,766 | 214,496,716 (50.37 %) | 49.63% | 9.3B | 60,253,470,754 | 57.84 GB |

### Top 10 domains

| Domain | Docs | % of total |
|---|---|---|
| postimees.ee | 729K | 5.31% |
| delfi.ee | 608K | 4.43% |
| err.ee | 416K | 3.03% |
| blogspot.com | 316K | 2.30% |
| aripaev.ee | 315K | 2.30% |
| ohtuleht.ee | 194K | 1.41% |
| pilguheit.com | 149K | 1.08% |
| wikipedia.org | 139K | 1.01% |
| wordpress.com | 116K | 0.85% |
| kliinik.ee | 88K | 0.64% |

### Top 10 TLDs

| Domain | Docs | % of total |
|---|---|---|
| ee | 9M | 65.59% |
| com | 2.9M | 21.03% |
| org | 438K | 3.19% |
| eu | 421K | 3.06% |
| net | 179K | 1.30% |
| edu.ee | 73K | 0.53% |
| info | 65K | 0.47% |
| pt | 58K | 0.42% |
| fi | 55K | 0.40% |
| ru | 32K | 0.24% |

## Register labels



- HI - 3.2%
- ID - 2.6%
- IN - 11.8%
- IP - 24.4%
- LY - 0.1%
- MIX - 3.2%
- NA - 33.6%
- OP - 4.9%
- SP - 0.8%
- UNK - 15.4%

- HI_other - 2.2%
- HI_re - 1.0%
- ID_other - 2.6%
- IN_dtp - 3.8%
- IN_en - 1.4%
- IN_fi - 0.0%
- IN_lt - 0.8%
- IN_other - 5.6%
- IN_ra - 0.1%
- IP_ds - 21.4%
- IP_other - 3.0%
- LY_other - 0.1%
- MIX - 3.2%
- NA_nb - 5.9%
- NA_ne - 20.2%
- NA_other - 3.9%
- NA_sr - 3.6%
- OP_av - 0.6%
- OP_ob - 1.3%
- OP_other - 1.3%
- OP_rs - 0.5%
- OP_rv - 1.1%
- SP_it - 0.5%
- SP_other - 0.3%
- UNK - 15.4%
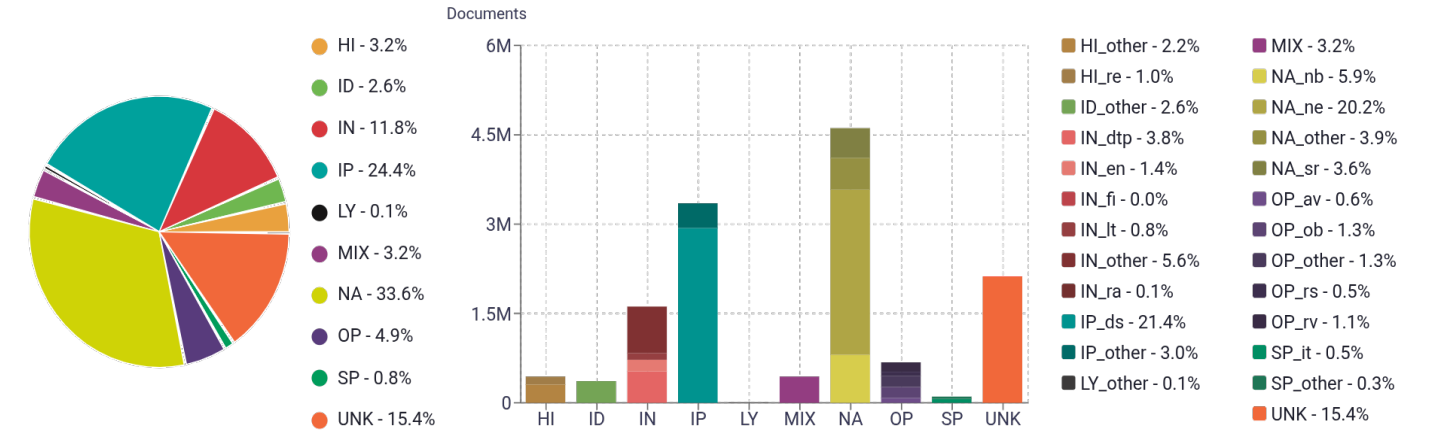
🤖 **MT**:11.7% | 1.6M Documents

## Documents size (in segments) ⓘ

≤ 25 segments **76.39%** (10M documents)
> 25 segments **23.61%** (3.2M documents)



## Document collections

**CC = 91.60%**
**IA = 8.40%**



67 Others (14M)
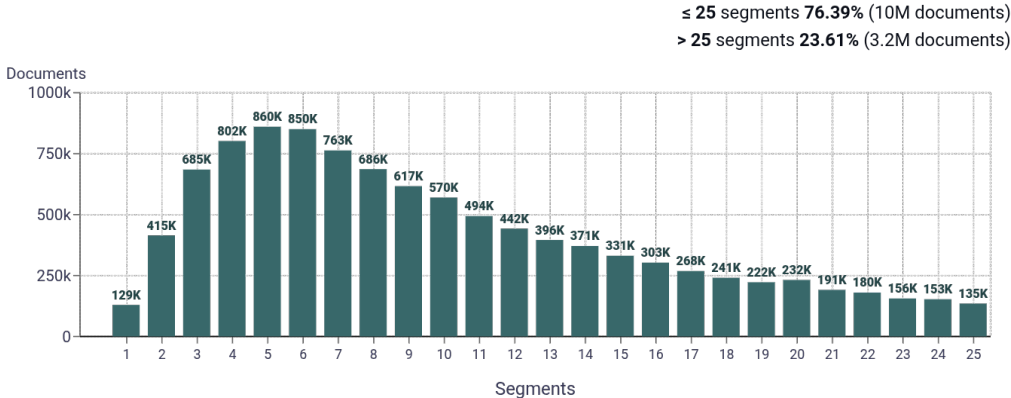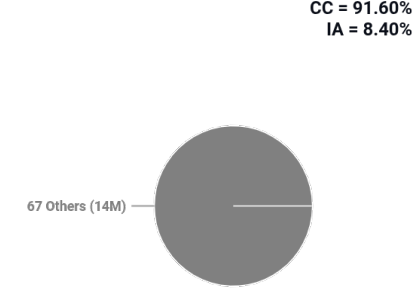
## Language Distribution

### Number of segments in the Estonian (ekk) corpus

- Estonian (et) - 344M **(80.9%)**
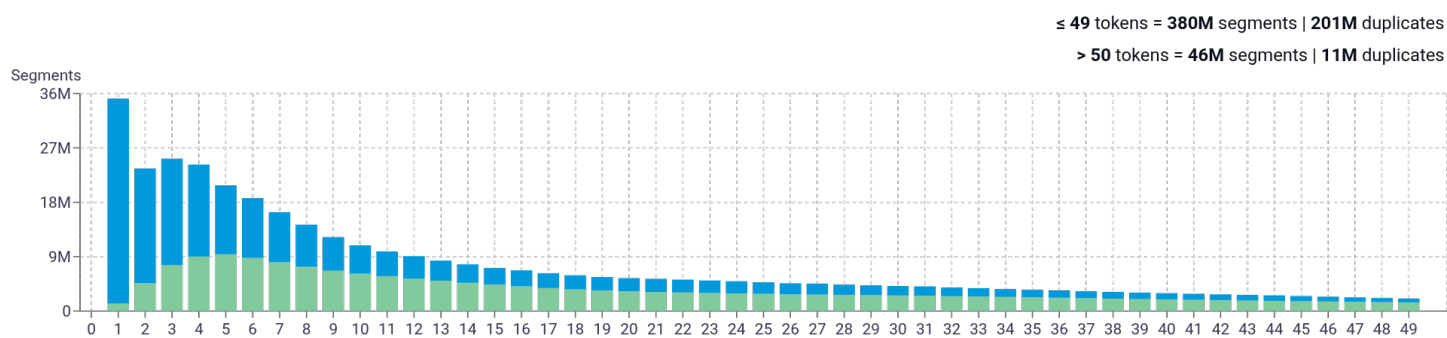- Italian (it) - 20M **(4.7%)**
- English (en) - 18M **(4.2%)**
- Finnish (fi) - 11M **(2.6%)**
- German (de) - 4.9M **(1.1%)**
- Dutch (nl) - 3.7M **(0.9%)**
- French (fr) - 3.6M **(0.8%)**
- Spanish (es) - 2.1M **(0.5%)**
- Swedish (sv) - 2M **(0.5%)**
- Portuguese (pt) - 1.5M **(0.4%)**
- 165 Others - 15M **(3.4%)**

*Estonian (ekk) identification might be inaccurate because it is not supported by FastSpell.
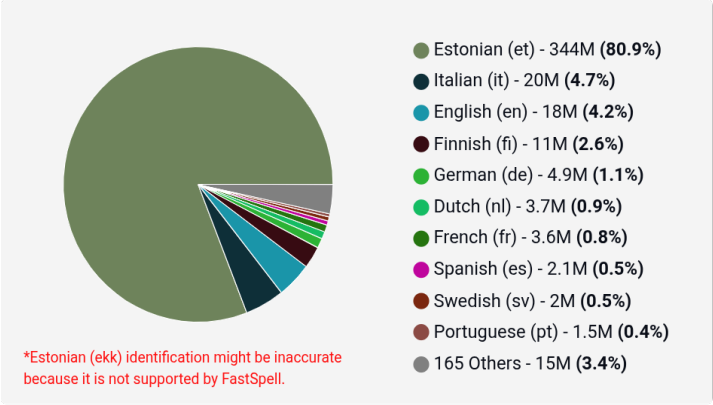
### Percentage of segments in Estonian (ekk) inside documents
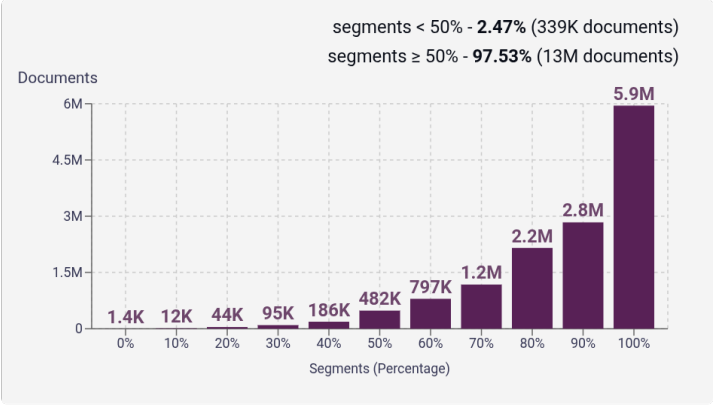
segments < 50% - **2.47%** (339K documents)
segments ≥ 50% - **97.53%** (13M documents)

Documents

| Segments (Percentage) | Documents |
|---|---|
| 0% | 1.4K |
| 10% | 12K |
| 20% | 44K |
| 30% | 95K |
| 40% | 186K |
| 50% | 482K |
| 60% | 797K |
| 70% | 1.2M |
| 80% | 2.2M |
| 90% | 2.8M |
| 100% | 5.9M |

## Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (14M documents)

Documents

## Segment length distribution by token

≤ 49 tokens = **380M** segments | **201M** duplicates
> 50 tokens = **46M** segments | **11M** duplicates

Segments

## Segment noise distribution

- Too long — **0.49%**
- Too short — **16.89%**
- URLs — **1.41%**
- Bad encoding — **0.01%**
- Contains PII — **0.23%**

■ % of corpus
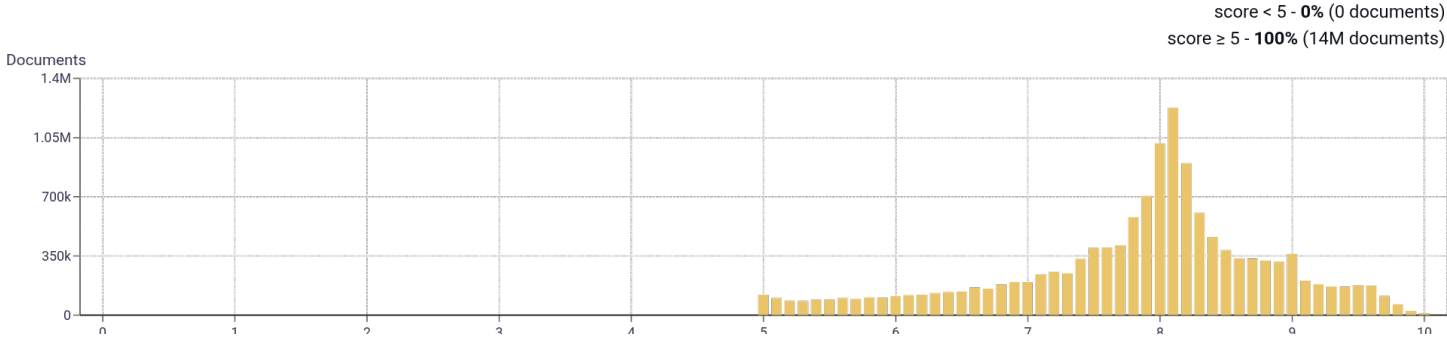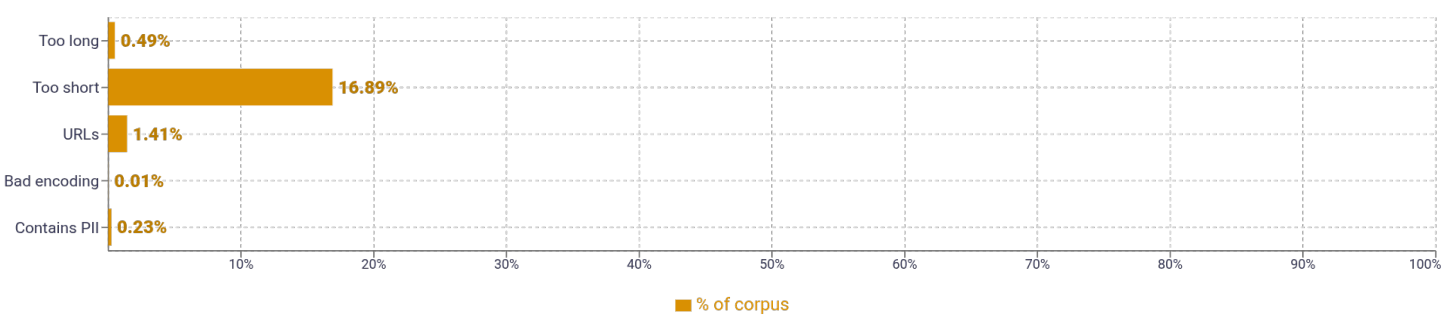
## Frequent n-grams

| SIZE | N-GRAMS | |
|---|---|---|
| 1 | ka \| 41,959,193   või \| 40,964,074   ning \| 31,602,929   võib \| 20,927,794   kuid \| 17,112,049 | ⧉ |
| 2 | võib olla \| 3,063,429   mitte ainult \| 2,020,503   samal ajal \| 1,886,004   korda päevas \| 1,301,279   vaid ka \| 1,276,030 | ⧉ |
| 3 | olulisemate uudiste kokkuvõte \| 275,773   võtke meiega ühendust \| 250,854   kolm korda päevas \| 224,161   teil on vaja \| 215,821   kaks korda päevas \| 213,006 | ⧉ |
| 4 | palun võtke meiega ühendust \| 119,320   kuvab err kommenteerija täisnime \| 100,124   tooteülevaateid veel ei ole \| 89,898   president toomas hendrik ilves \| 87,585   aastat vana ning kuulub \| 85,163 | ⧉ |
| 5 | aastat vana ning kuulub väljaande \| 66,950   vana ning kuulub väljaande digitaalsesse \| 66,949   ning kuulub väljaande digitaalsesse arhiivi \| 66,948   uuenda ega kaasajasta arhiveeritud sisu \| 66,945   võib olla vajalik kaasaegsete allikatega \| 66,943 | ⧉ |

# 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 | | | | |