HPLT Analytics report

General overview

Corpus	Date	Language
est_Latn.jsonl.tsv	6/6/2025	Estonian (et)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
8,449,258	264,352,085	102,779,262 (38.88 %)	5.6B	35,759,962,808	34.35 GB

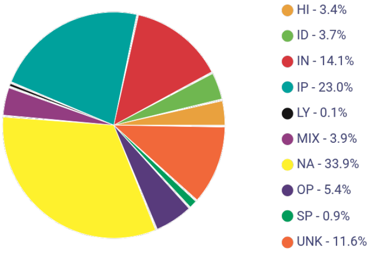
Top 10 domains

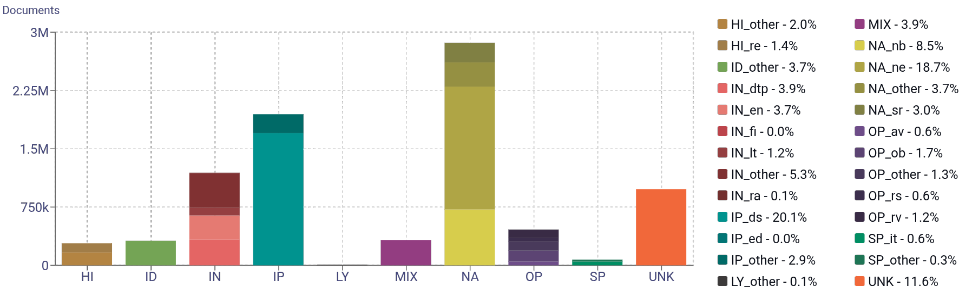
Domain	Docs	% of total
wikipedia.org	293K	3.47%
blogspot.com	290K	3.43%
postimees.ee	288K	3.40%
err.ee	249K	2.95%
delfi.ee	225K	2.67%
aripaev.ee	158K	1.87%
pilguheit.com	149K	1.76%
ohtuleht.ee	117K	1.39%
kliinik.ee	93K	1.10%
wordpress.com	85K	1.00%

Top 10 TLDs

Domain	Docs	% of total
ee	5.5M	65.36%
com	1.7M	19.75%
org	436K	5.16%
eu	222K	2.63%
net	116K	1.37%
fi	76K	0.90%
com.ee	53K	0.62%
edu.ee	49K	0.58%
info	42K	0.50%
pt	21K	0.25%

Register labels

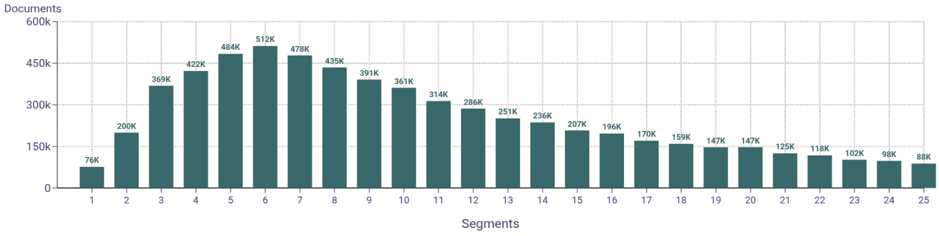




🤖 **MT**:8.5% | 722K Documents

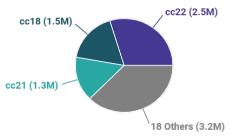
Documents size (in segments)

<= 25 segments **75.41%** (6.4M documents)
> 25 segments **24.59%** (2.1M documents)



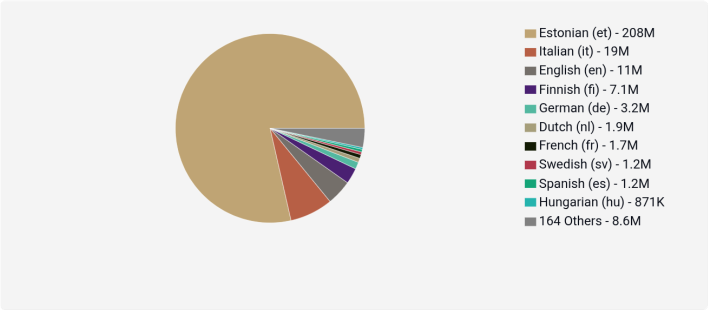
Documents by collection

CC = 73.12%
IA = 26.88%

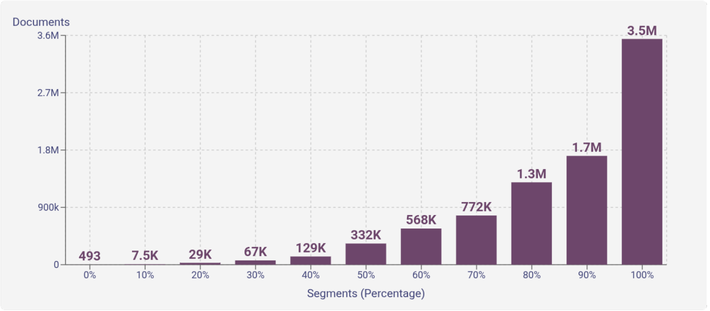


Language Distribution

Number of segments in the Estonian (et) corpus

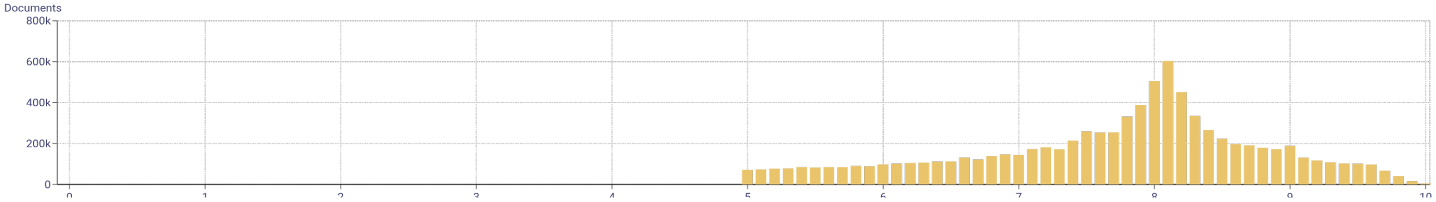


Percentage of segments in Estonian (et) inside documents



Distribution of documents by document score

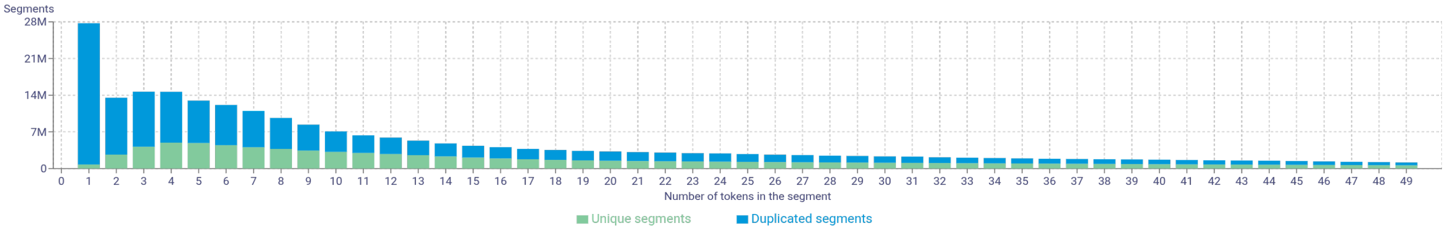
score < 5 - **0%** (0 documents)
score >= 5 - **100%** (8.4M documents)



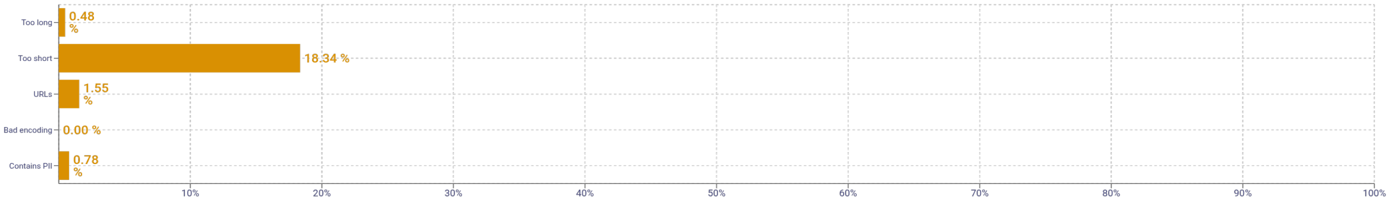
Segment length distribution by token

≤ **49** tokens = **86M** segments | **149M** duplicates

> **50** tokens = **29M** segments | **12M** duplicates



Segment noise distribution



Frequent n-grams

<table>
<tr><th>Size</th><th>n-grams</th></tr>
<tr><td>1</td><td>eesti | 14056503
võib | 10394869
kes | 9821756
kuid | 9028429
ole | 8052598</td></tr>
<tr><td>2</td><td>võib olla | 1415584
samal ajal | 918895
mitte ainult | 798046
ettevõttega ühendust | 739660
euroopa liidu | 664639</td></tr>
<tr><td>3</td><td>viimase tunni jooksul | 332931
vaadanud seda hotelli | 332641
külastajat on vaadanud | 332641
hotelli viimase tunni | 332641
olulisemate uudiste kokkuvõte | 276237</td></tr>
<tr><td>4</td><td>vaadanud seda hotelli viimase | 332641
hotelli viimase tunni jooksul | 332641
vastas dr jüri ennet | 93500
president toomas hendrik ilves | 87180
isamaa ja res publica | 73437</td></tr>
<tr><td>5</td><td>vaadanud seda hotelli viimase tunni | 332641
külastajat on vaadanud seda hotelli | 332641
vana ning kuulub väljaande digitaalsesse | 48927
aastat vana ning kuulub väljaande | 48927
võib olla vajalik kaasaegsete allikatega | 48926</td></tr>
</table>

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