

General overview

Corpus	Date	Language
hplt-v3-ekk_Latn	9/17/2025	Estonian (et)

Volumes

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

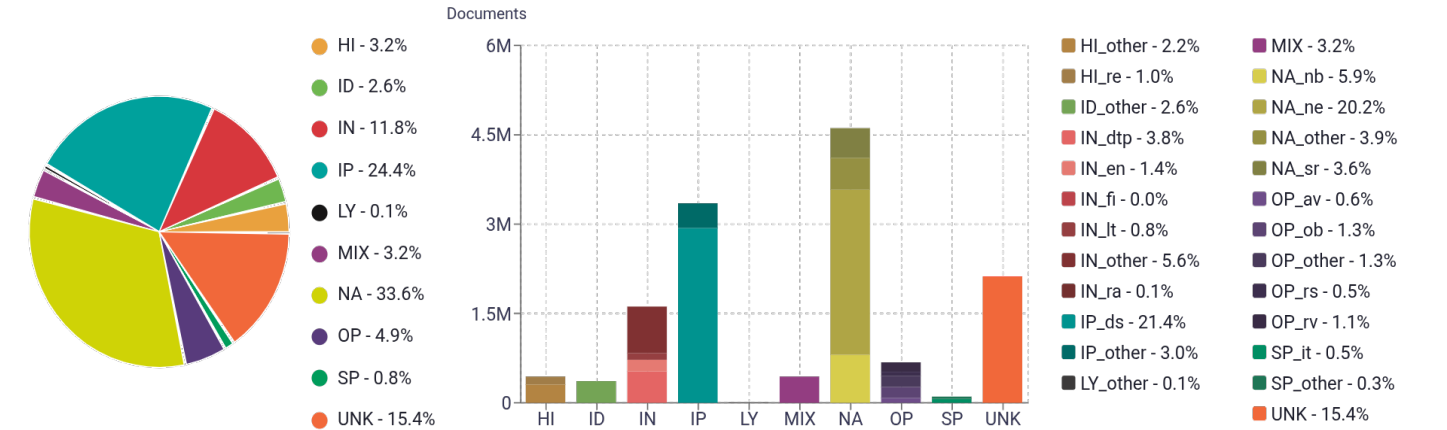
Top 10 domains

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

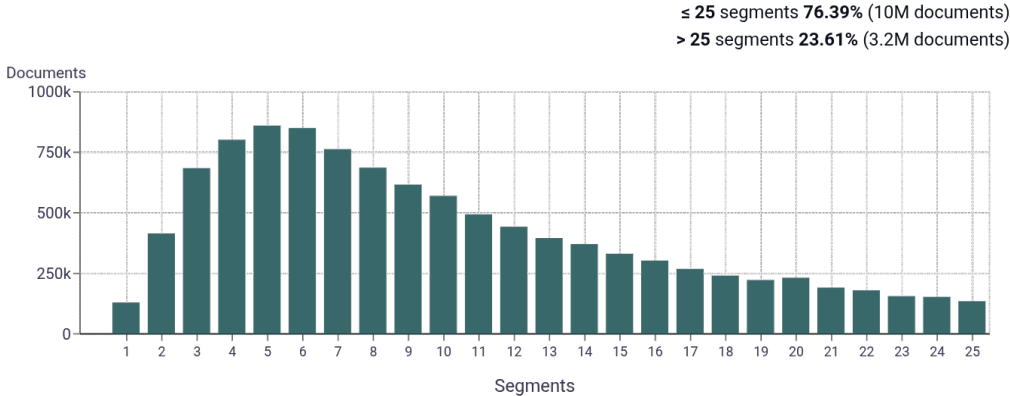
Domain	Docs	% of total
ee	9M	65.59%
com	2.9M	21.03%
org	438K	3.19%
eu	421K	3.06%
net	179K	1.30%
edu.ee	73K	0.53%
info	65K	0.47%
pt	58K	0.42%
fi	55K	0.40%
ru	32K	0.24%

Register labels

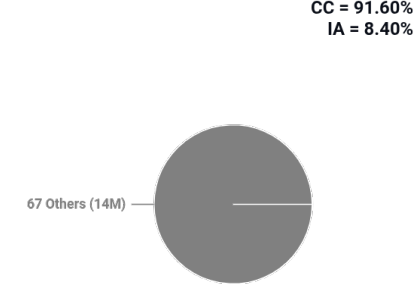


MT:11.7% | 1.6M Documents

Documents size (in segments) ⓘ

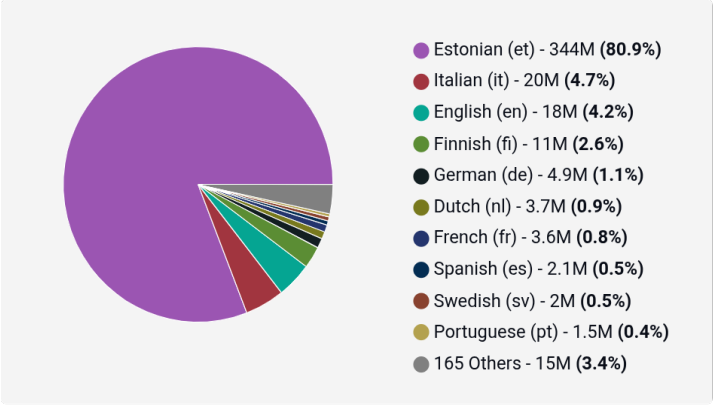


Document collections

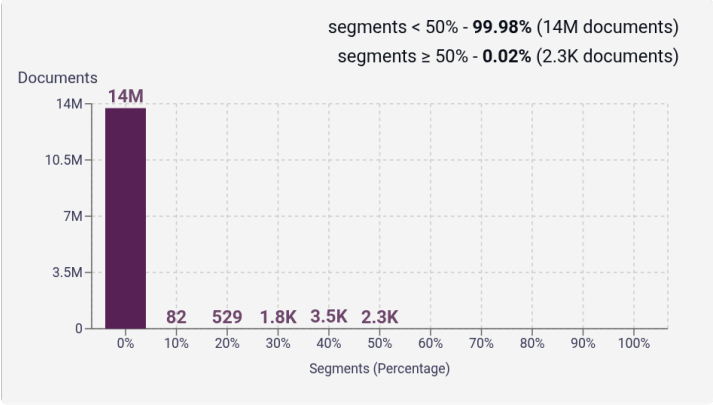


Language Distribution

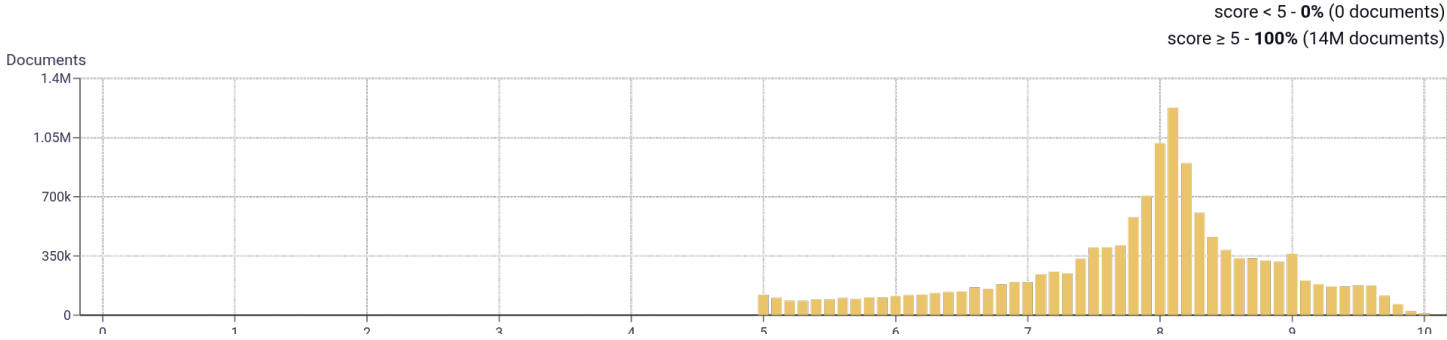
Number of segments in the Estonian (et) corpus



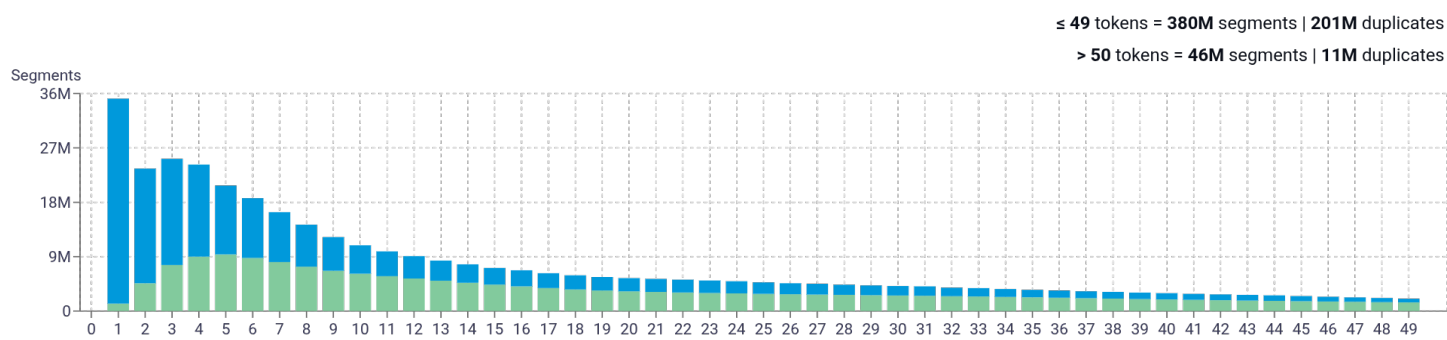
Percentage of segments in Estonian (et) inside documents



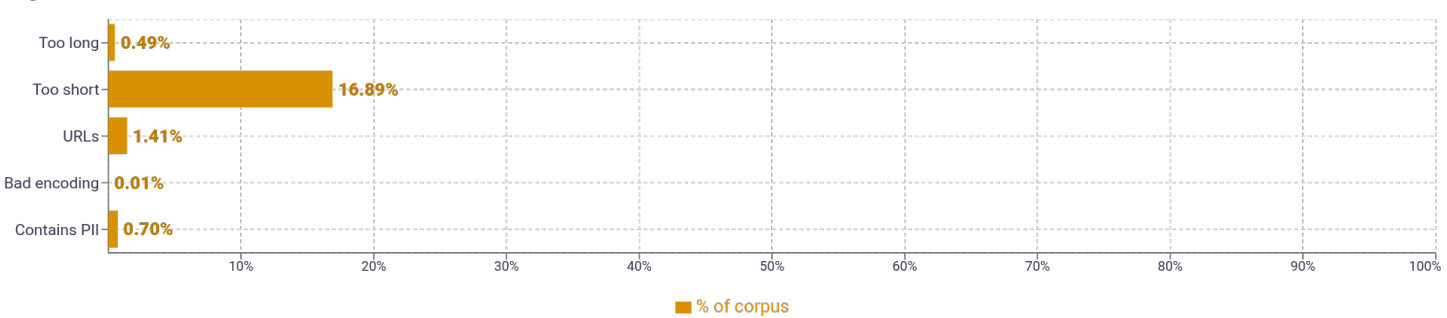
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>võib   20,927,794</div> <div>kuid   17,112,049</div> <div>eesti   17,011,109</div> <div>kes   15,403,647</div> <div>pärast   13,681,447</div>	
2	<div>võib olla   3,063,429</div> <div>mitte ainult   2,020,503</div> <div>samal ajal   1,886,004</div> <div>korda päevas   1,301,279</div> <div>ette nähtud   1,165,584</div>	
3	<div>olulisemate uudiste kokkuvõte   275,773</div> <div>võtke meiega ühendust   250,854</div> <div>kolm korda päevas   224,161</div> <div>teil on vaja   215,821</div> <div>kaks korda päevas   213,006</div>	
4	<div>palun võtke meiega ühendust   119,320</div> <div>kuvab err kommenteerija täisnime   100,124</div> <div>tooteülevaateid veel ei ole   89,898</div> <div>president toomas hendrik ilves   87,585</div> <div>aastat vana ning kuulub   85,163</div>	
5	<div>aastat vana ning kuulub väljaande   66,950</div> <div>vana ning kuulub väljaande digitaalsesse   66,949</div> <div>uuenda ega kaasajasta arhiveeritud sisu   66,945</div> <div>võib olla vajalik kaasaegsete allikatega   66,943</div> <div>olla vajalik kaasaegsete allikatega tutvumine   66,943</div>	

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 of 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	ntp
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				