

General overview

Corpus	Date	Language
hplt-v3-epo_Latn	9/17/2025	Esperanto (eo)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
715,290	23,245,746	17,835,713 (76.73 %)	707M	3,707,801,717	3.52 GB

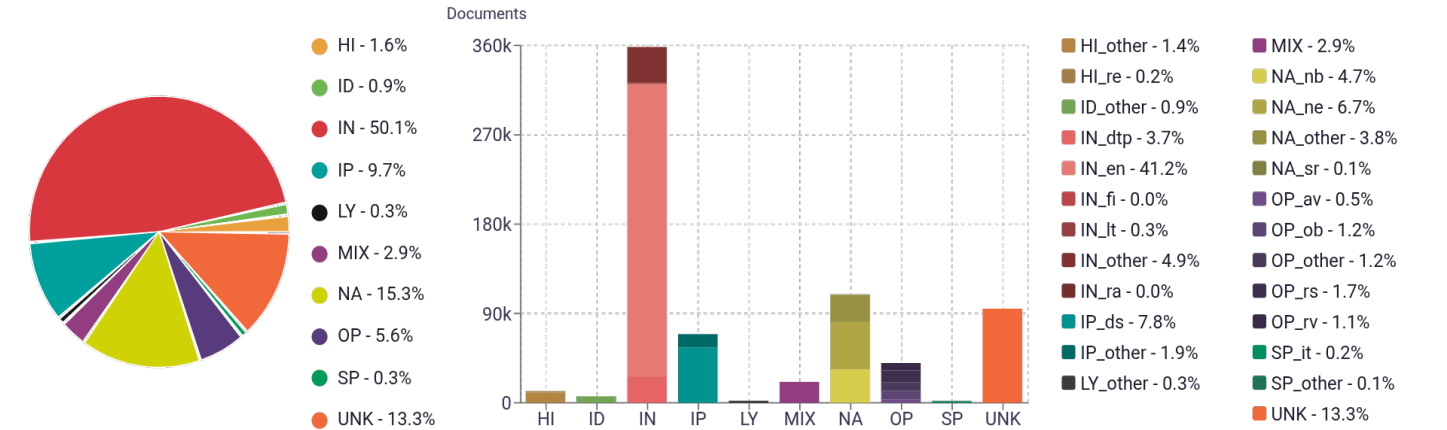
Top 10 domains

Domain	Docs	% of total
wikipedia.org	186K	25.95%
wikitrans.net	96K	13.39%
blogspot.com	11K	1.58%
cri.cn	10K	1.46%
martech.zone	7.5K	1.04%
wordpress.com	6.1K	0.86%
wikilingue.com	6K	0.85%
wikisource.org	5.7K	0.79%
pola-retradio.org	5.4K	0.75%
esperantio.net	5.1K	0.72%

Top 10 TLDs

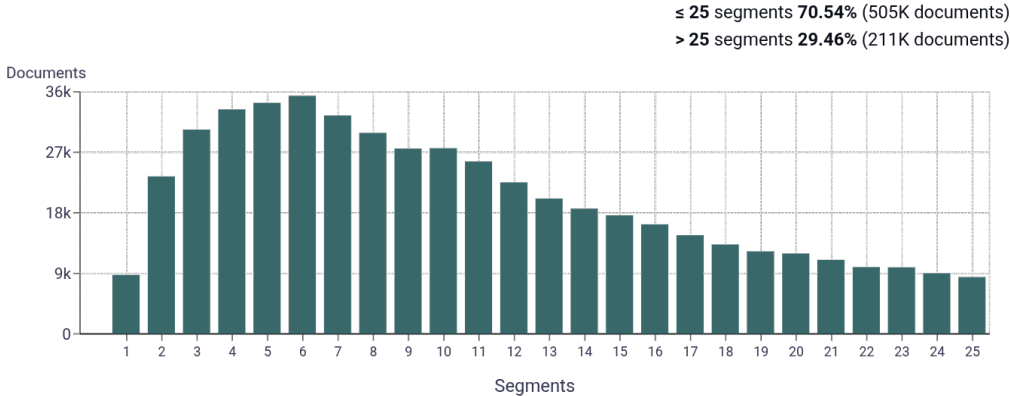
Domain	Docs	% of total
org	250K	34.92%
com	213K	29.80%
net	129K	18.01%
cn	13K	1.83%
ru	11K	1.53%
zone	7.5K	1.04%
de	5.3K	0.74%
com.br	4.8K	0.68%
org.cn	4.7K	0.66%
news	4K	0.56%

Register labels



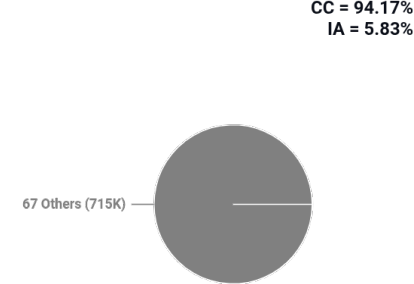
MT:10.0% | 72K Documents

Documents size (in segments)



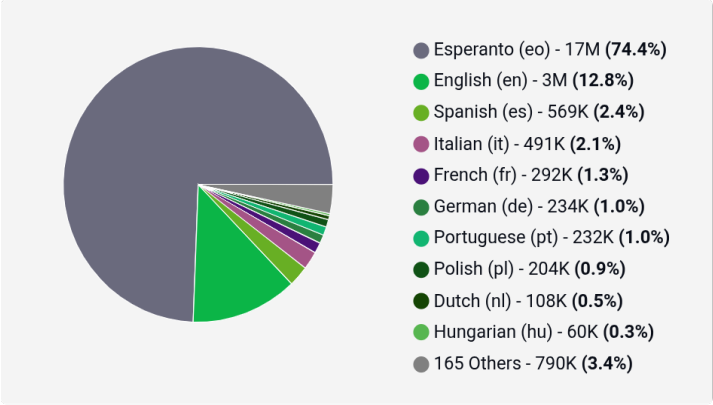
≤ 25 segments 70.54% (505K documents)  
> 25 segments 29.46% (211K documents)

Document collections

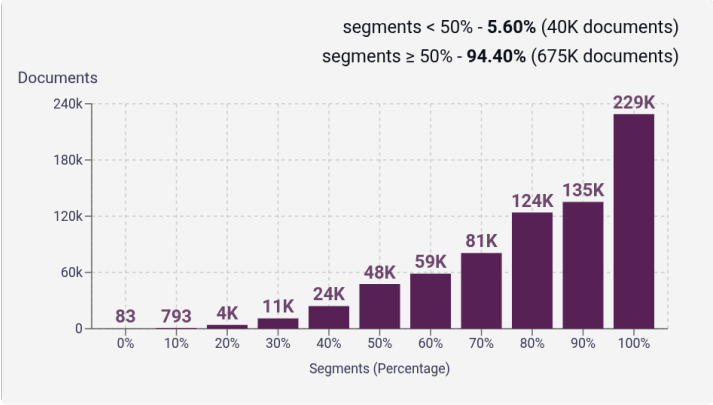


Language Distribution

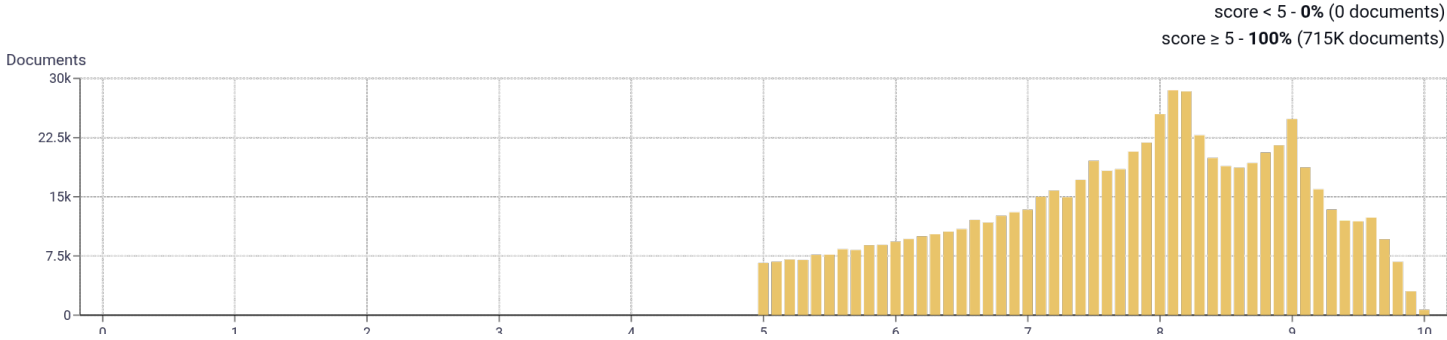
Number of segments in the Esperanto (eo) corpus



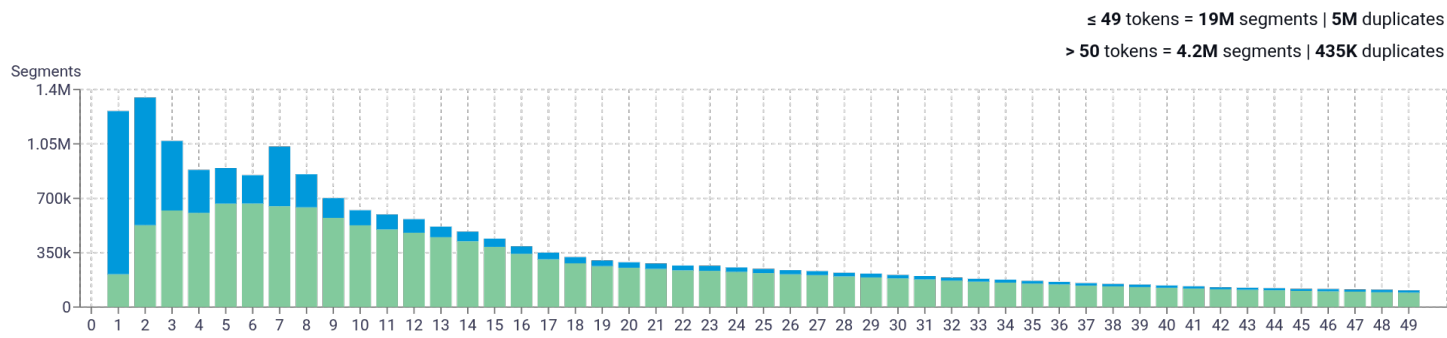
Percentage of segments in Esperanto (eo) inside documents



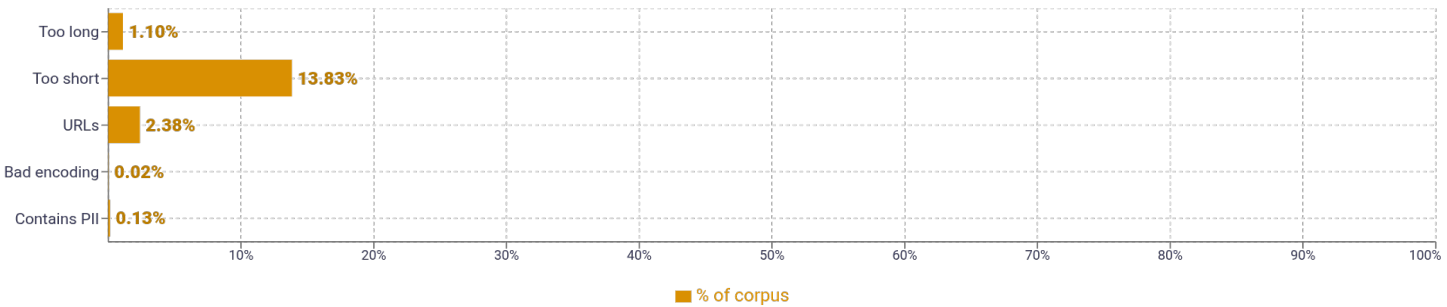
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>kiel   3,420,017</div> <div>kun   3,361,346</div> <div>pri   2,384,779</div> <div>el   1,948,902</div> <div>povas   1,840,911</div>	
2	<div>povas esti   442,215</div> <div>redakti fonton   417,721</div> <div>of the   251,663</div> <div>new york   148,745</div> <div>same kiel   137,161</div>	
3	<div>from the original   105,250</div> <div>archived from the   104,125</div> <div>the original on   97,373</div> <div>as translated by   94,566</div> <div>translated by gramtrans   94,554</div>	
4	<div>archived from the original   104,123</div> <div>from the original on   97,368</div> <div>as translated by gramtrans   94,554</div> <div>the new york times   26,864</div> <div>ĉe la wayback maŝino   19,034</div>	
5	<div>archived from the original on   96,427</div> <div>per la retarkivo wayback machine   8,705</div> <div>kiam por forigi tiun ŝablonmesaĝon   8,560</div> <div>kiel kaj kiam por forigi   8,467</div> <div>ĉe la interreta filma datenbazo   7,137</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  $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$ , 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				