

General overview

Corpus	Date	Language
hplt-v3-lij_Latn	9/18/2025	Ligurian (lij)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
8,607	243,635	188,071 (77.19 %)	9.4M	35,342,899	41.65 MB

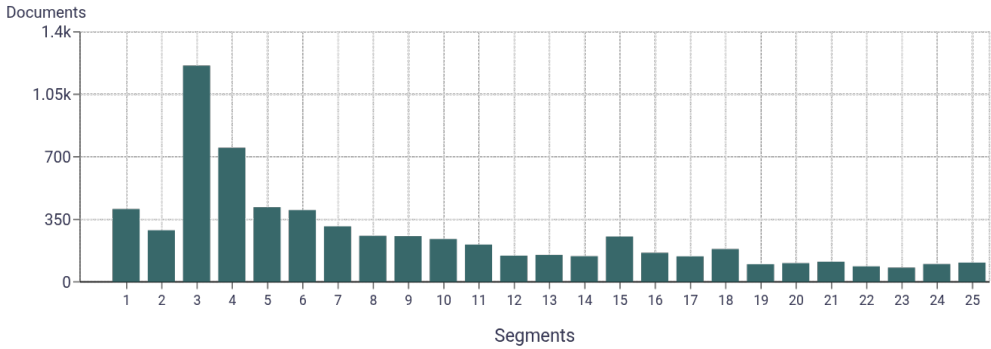
Top 10 domains

Domain	Docs	% of total
wikisource.org	2.4K	28.45%
wikipedia.org	2.3K	26.22%
eodishasamachar...	1.9K	22.47%
prameyanews.com	263	3.06%
ilsecoloxix.it	121	1.41%
biswabijayeew...	106	1.23%
hindustantimes.com	102	1.19%
primocanale.it	84	0.98%
francobampi.it	78	0.91%
conseggio-figur...	73	0.85%

Top 10 TLDs

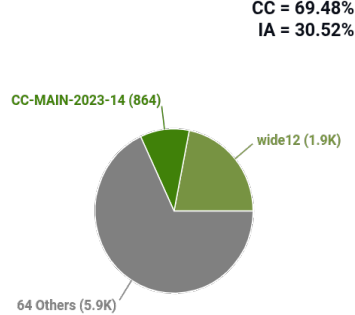
Domain	Docs	% of total
org	5K	58.06%
com	2.9K	33.44%
it	437	5.08%
net	52	0.60%
ru	26	0.30%
eu	25	0.29%
vn	18	0.21%
info	16	0.19%
click	15	0.17%
com.br	13	0.15%

Documents size (in segments) ⓘ



≤ 25 segments 77.04% (6.6K documents)  
> 25 segments 22.96% (2K documents)

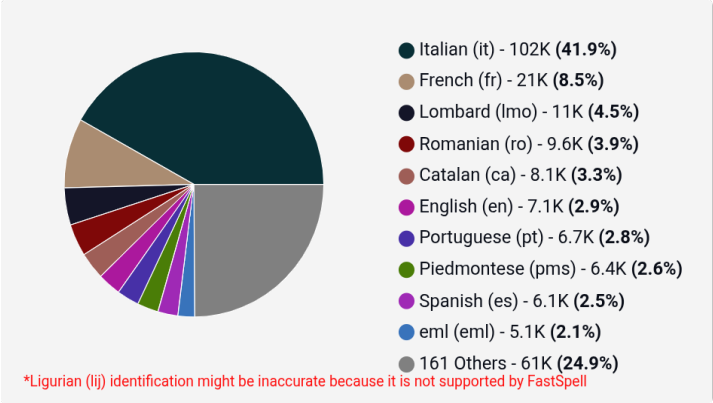
Document collections



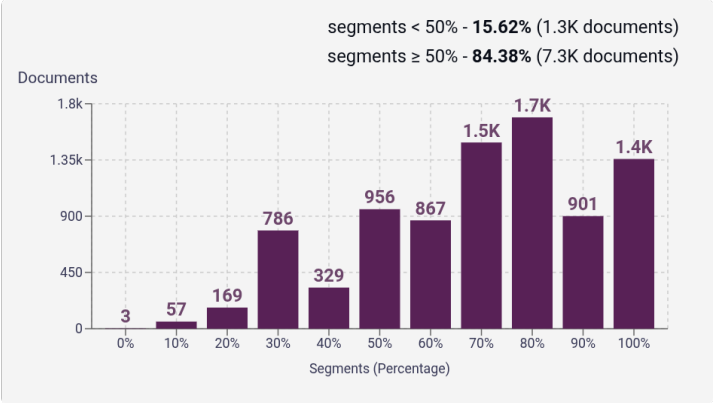
CC = 69.48%  
IA = 30.52%

Language Distribution

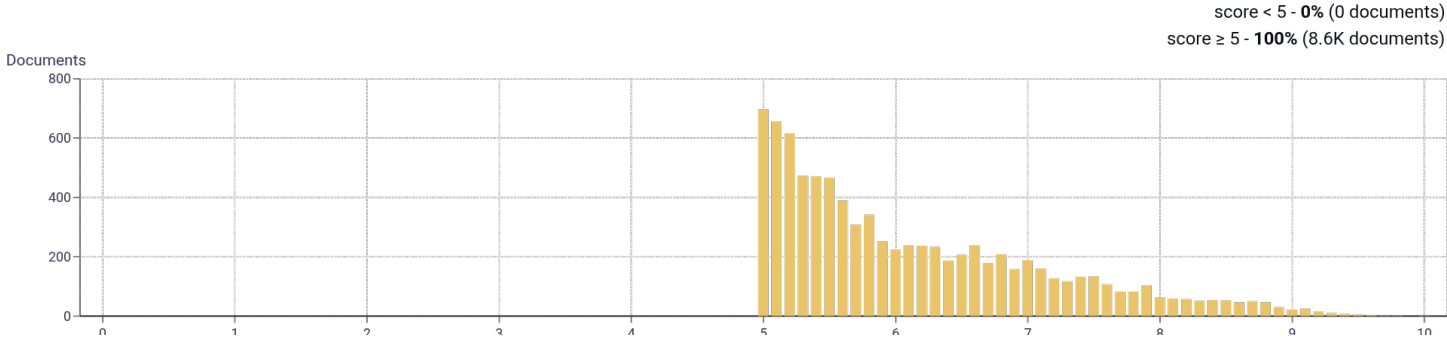
Number of segments in the Ligurian (lij) corpus



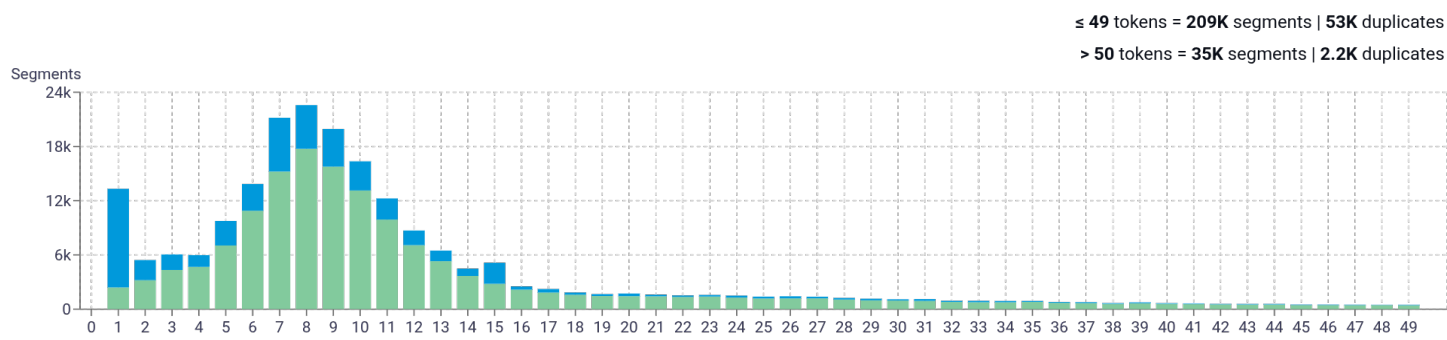
Percentage of segments in Ligurian (lij) inside documents



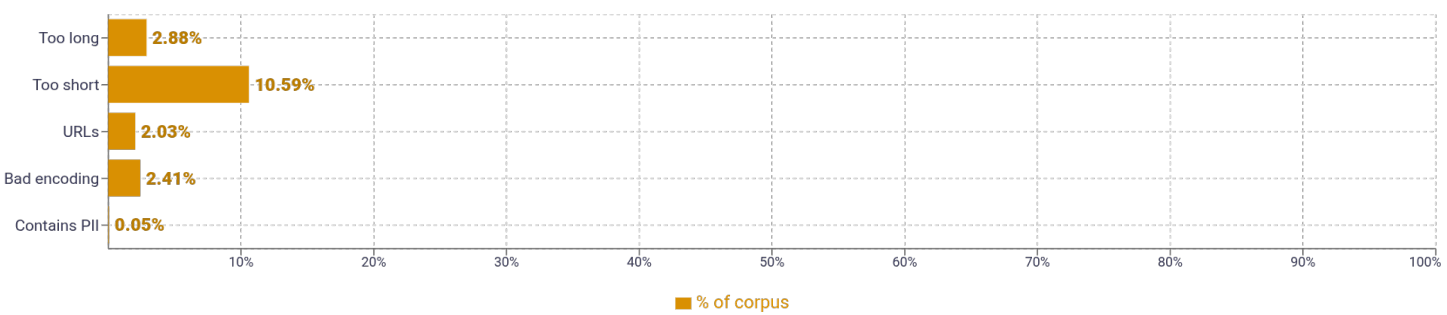
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>I   112,280</div> <div>ö   100,180</div> <div>ø   61,324</div> <div>è   56,181</div> <div>ð   52,943</div>	
2	<div>ö ø   7,860</div> <div>ø ö   6,981</div> <div>ä ä   6,875</div> <div>û ö   6,462</div> <div>ö ö   5,658</div>	
3	<div>ä ä ä   6,627</div> <div>ï ä ð   1,872</div> <div>ä ð ä   1,848</div> <div>ooo   1,714</div> <div>ïï ä   1,638</div>	
4	<div>ä ä ä ä   6,430</div> <div>ï ä ð ä   1,790</div> <div>oooo   1,558</div> <div>ä ð ä ä   449</div> <div>ö è í í   420</div>	
5	<div>ä ä ä ä ä   6,238</div> <div>ooooo   1,453</div> <div>ï ä ð ä ä   439</div> <div>ï ä ð ä ä   338</div> <div>è ä à ò û   307</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	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				