HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-fur_Latn	9/17/2025	Friulian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
55,016	1,113,382	478,541 (42.98 %)	45M	213,125,456	211.19 MB

Top 10 domains

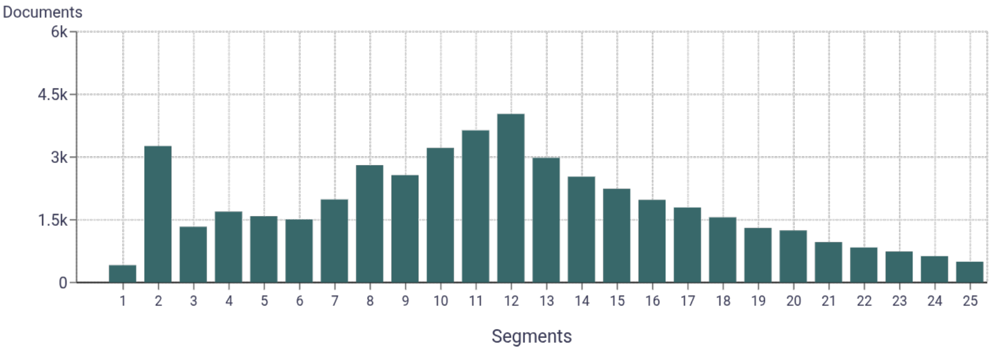
Domain	Docs	% of total
lapatriedalfriu...	28K	51.72%
blogspot.com	4.4K	7.95%
lavosdaifurlans...	3.4K	6.11%
contecurte.eu	2.4K	4.42%
blogspot.it	1.9K	3.40%
wikipedia.org	1.7K	3.06%
wordpress.com	1.6K	2.95%
bibie.org	1.1K	1.95%
claap.org	1K	1.89%
glesiefurlane.org	914	1.66%

Top 10 TLDs

Domain	Docs	% of total
org	34K	61.87%
com	11K	19.37%
it	5.7K	10.40%
eu	3.3K	5.99%
ud.it	393	0.71%
fvg.it	205	0.37%
net	161	0.29%
ch	78	0.14%
info	48	0.09%
in	45	0.08%

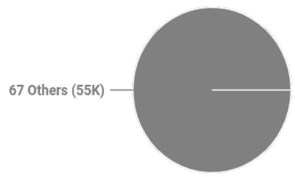
Documents size (in segments) ⓘ

≤ 25 segments **86.1%** (47K documents)
> 25 segments **13.9%** (7.6K documents)



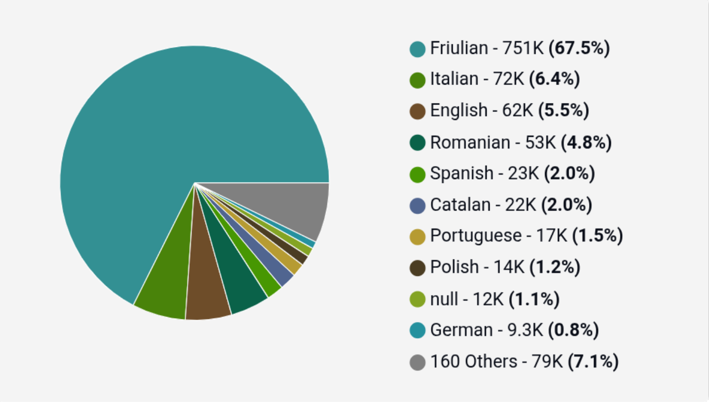
Document collections

CC = 93.85%
IA = 6.15%



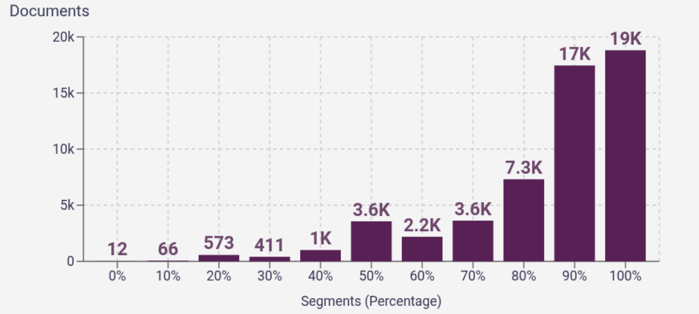
Language Distribution

Number of segments in the Friulian corpus



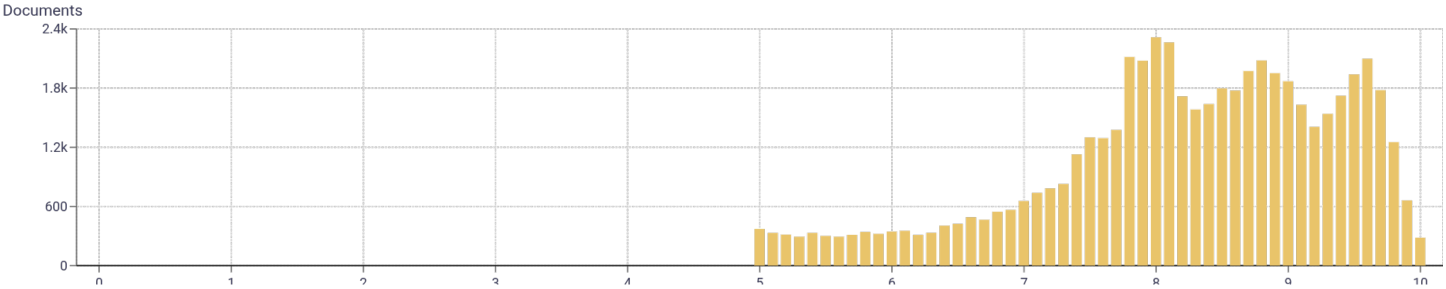
Percentage of segments in Friulian inside documents

segments < 50% - **3.76%** (2.1K documents)
segments ≥ 50% - **96.24%** (53K documents)



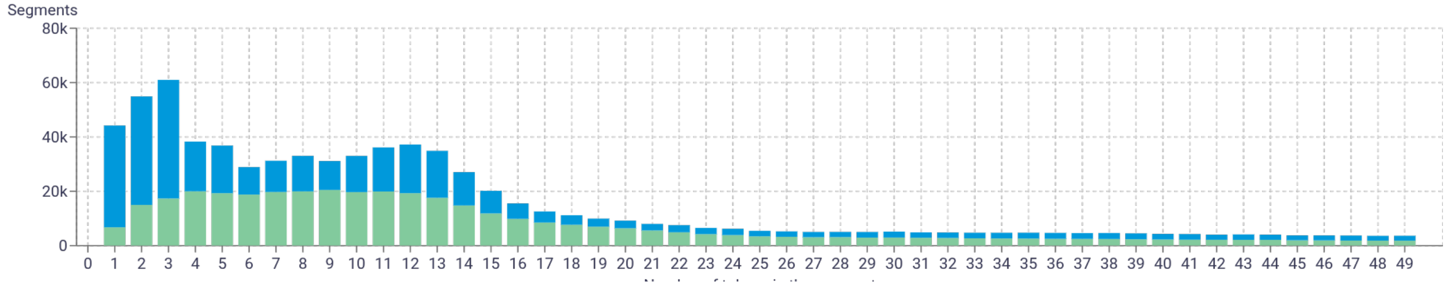
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (55K documents)

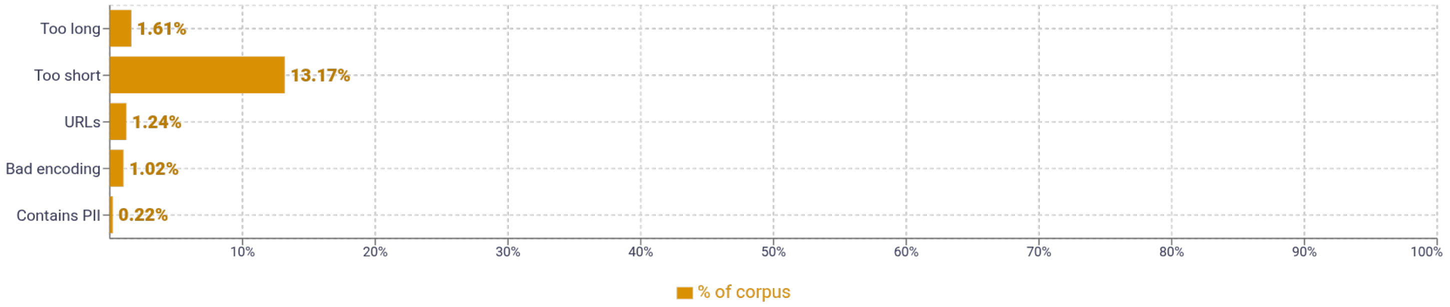


Segment length distribution by token

≤ 49 tokens = **750K** segments | **368K** duplicates
> 50 tokens = **363K** segments | **269K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	lei | 157,296 ai | 149,561 l | 144,626 friûl | 111,526 agns | 101,084
2	lenghe furlane | 28,073 chê volte | 12,644 erika adami | 12,577 feminis furlanis | 11,697 furlanis fuartis | 9,502
3	patrie dal friûl | 33,789 feminis furlanis fuartis | 9,502 cu la autore | 8,625 dì di vuê | 7,654 cu la storie | 7,600
4	storie di cheste tiere | 7,368 gnûf arcivescul di udin | 7,361 robe che o podìn | 7,350 lui e par nô | 7,346 jentrâ in sintonie cu | 7,346
5	robe che o podìn sperâ | 7,345 rivi a colp a jentrâ | 7,345 miôr robe che o podìn | 7,345 cu la storie di cheste | 7,345 colp a jentrâ in sintonie | 7,345

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