HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
fur_Latn.jsonl.tsv	11/27/2024	Friulian (fur)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
36,666	730,045	266,635 (36.52 %)	24M	114,043,358	112.39 MB

Top 10 domains

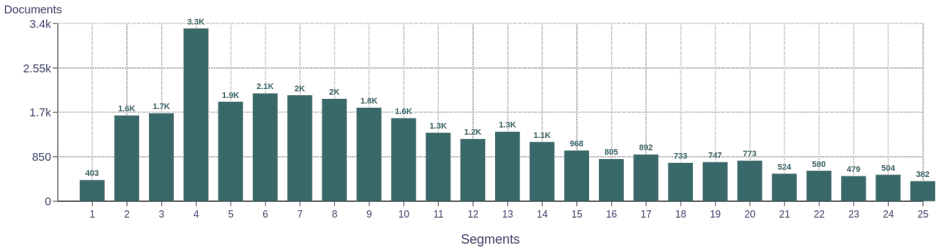
Domain	Docs	% of total
wikipedia.org	11K	28.96
lapatriedalfriul.org	7.8K	21.15
blogspot.com	3.3K	9.04
blogspot.it	2.9K	7.99
tuugo.at	1.7K	4.68
contecurte.eu	1.4K	3.80
wordpress.com	836	2.28
glesiefurlane.org	791	2.16
blogspot.ch	670	1.83
arlef.it	413	1.13

Top 10 TLDs

Domain	Docs	% of total
org	20K	54.64
com	5.5K	14.89
it	5.2K	14.12
eu	2.1K	5.83
at	1.9K	5.08
ch	702	1.91
ud.it	209	0.57
net	197	0.54
in	132	0.36
fvg.it	98	0.27

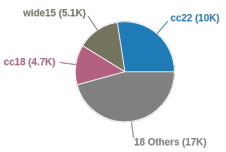
Documents size (in segments)

<= 25 segments **83.73%** (31K documents)
> 25 segments **16.27%** (6K documents)



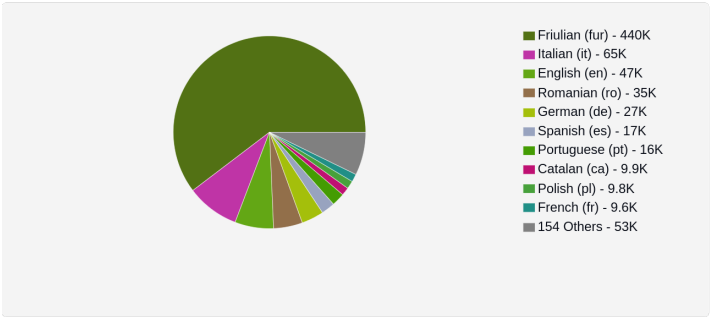
Documents by collection

CC = 59.29%
IA = 40.71%

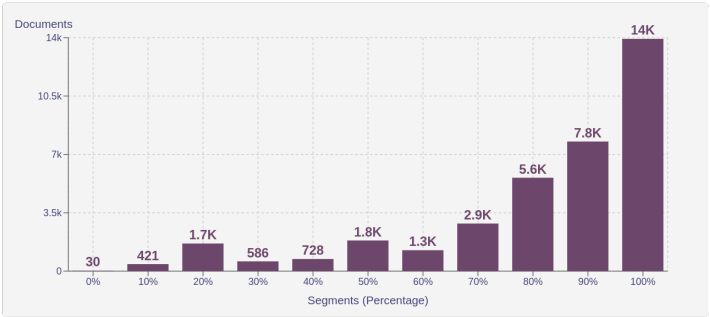


Language Distribution

Number of segments in the Friulian (fur) corpus

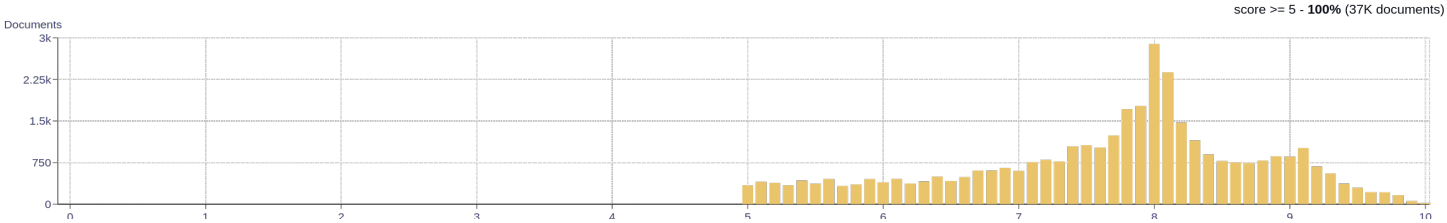


Percentage of segments in Friulian (fur) inside documents



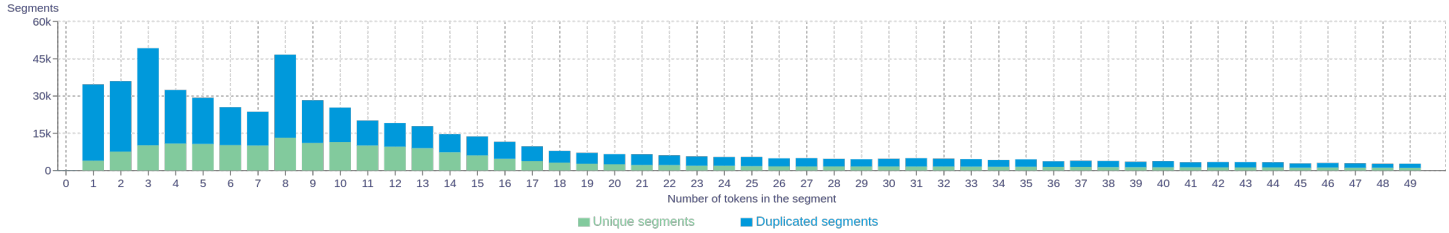
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (37K documents)

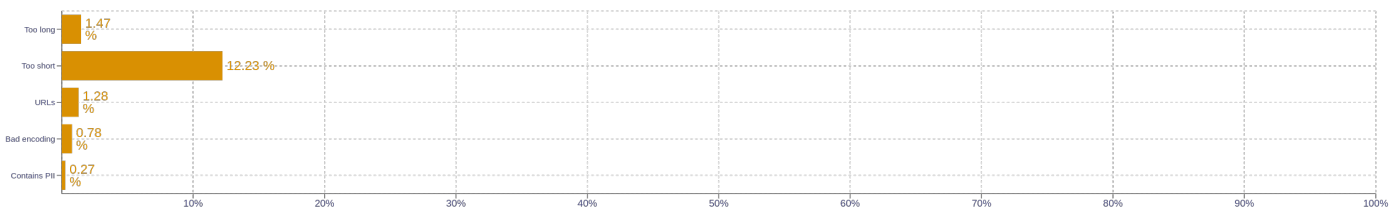


Segment length distribution by token

≤ 49 tokens = **202K** segments | **380K** duplicates
> 50 tokens = **148K** segments | **83K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	l | 78798 · ai | 70541 · fat | 46094 · friûl | 45132 · furlan | 39870
2	clap clap | 24950 · daj dajdaj | 23820 · dajdaj daj | 23816 · lenghe furlane | 13231 · fat fat | 12209
3	daj dajdaj daj | 23816 · dajdaj daj dajdaj | 23808 · modifiche il codiç | 23347 · clap clap clap | 23156 · fat fat fat | 12190
4	daj dajdaj daj dajdaj | 23808 · dajdaj daj dajdaj daj | 23804 · clap clap clap clap | 21366 · fat fat fat fat | 12176 · do re mi fa | 1886
5	daj dajdaj daj dajdaj daj | 23804 · dajdaj daj dajdaj daj dajdaj | 23796 · clap clap clap clap clap | 19725 · fat fat fat fat fat | 12163 · fa so la ti do | 2162

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