

General overview

Corpus	Analytics 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	112.39 MB	114,043,358

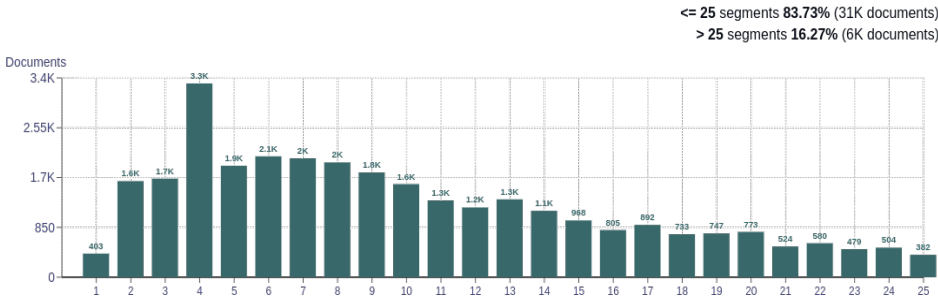
Top 10 domains

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
arief.it	413	1.13

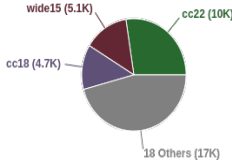
Top 10 TLDs

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)

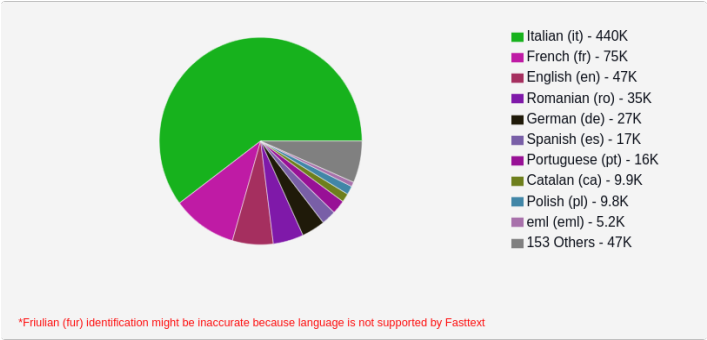


Documents by collection

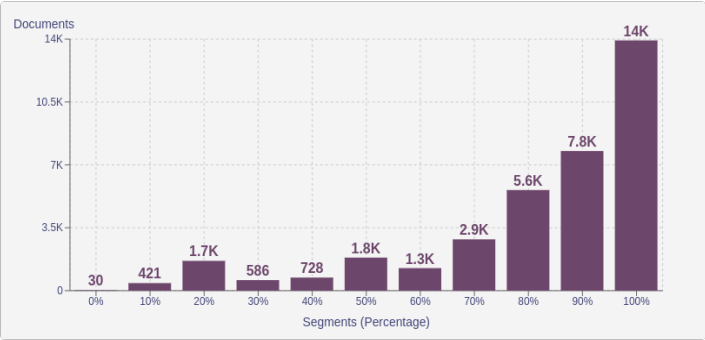


Language Distribution

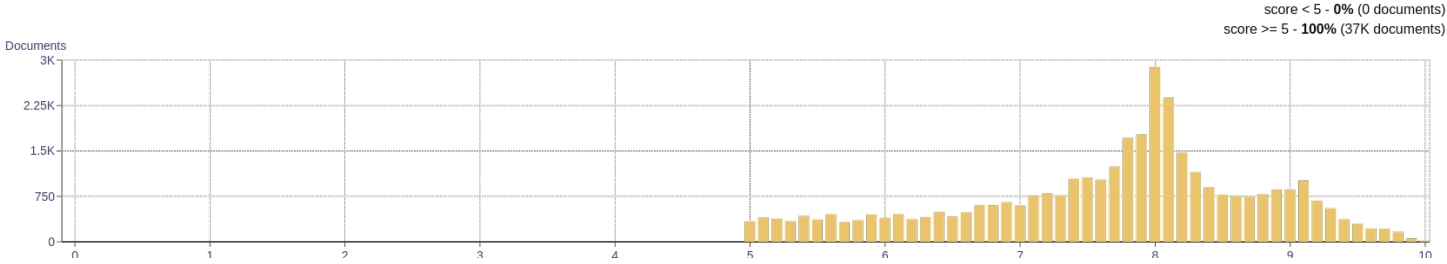
Number of segments



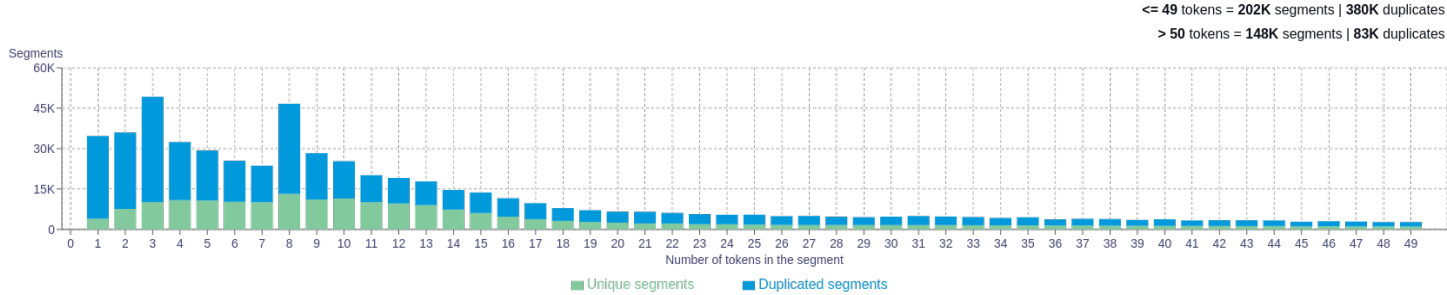
Percentage of segments in Friulian (fur) inside documents



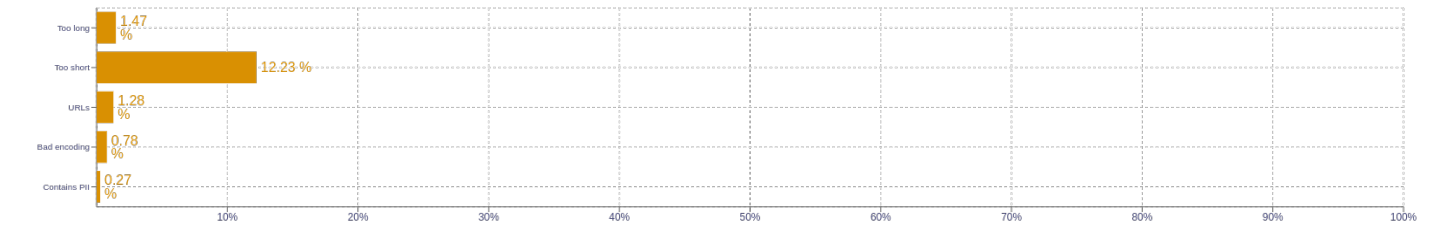
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

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

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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>, , , 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>