

General overview

Corpus	Analytics date	Language
fao_Latn.jsonl.tsv	9/24/2024	Faroese (fo)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
239,923	4,526,061	2,174,019 (48.03 %)	112M	592.33 MB	577,511,299

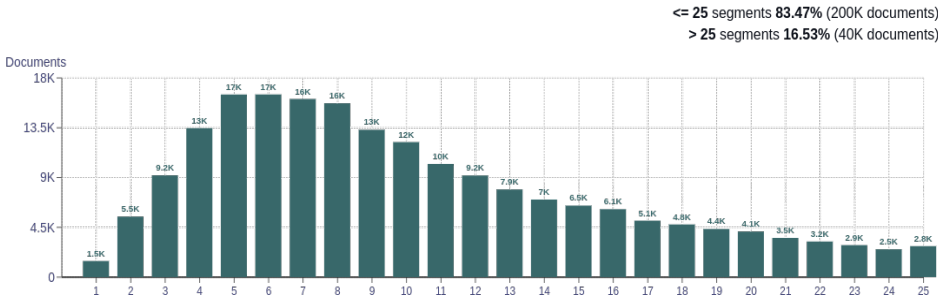
Top 10 domains

Domain	Docs	% of total
wikipedia.org	25K	10.43
snar.fo	8K	3.32
kvf.fo	6.6K	2.76
portal.fo	5.6K	2.31
vp.fo	5.4K	2.26
in.fo	5.3K	2.20
blogspot.com	3.8K	1.59
sangtekstir.com	3.6K	1.51
fsf.fo	3.4K	1.41
dimma.fo	3.1K	1.30

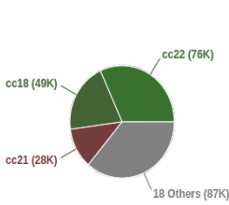
Top 10 TLDs

Domain	Docs	% of total
fo	176K	73.27
org	27K	11.42
com	24K	10.04
net	5K	2.07
dk	2.9K	1.22
info	1.1K	0.46
no	810	0.34
be	695	0.29
hk	509	0.21
is	504	0.21

Documents size (in segments)

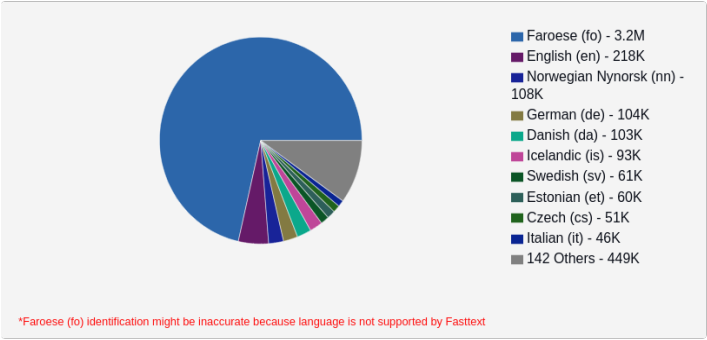


Documents by collection

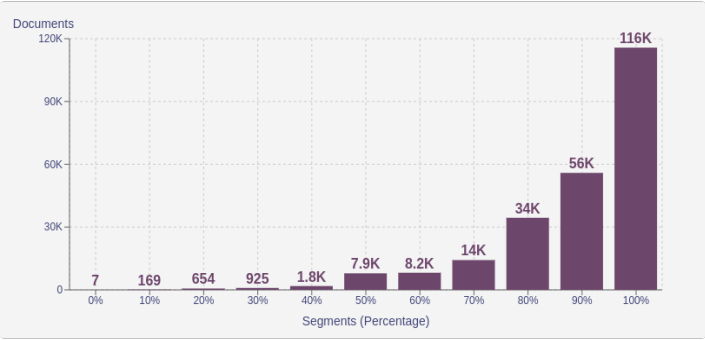


Language Distribution

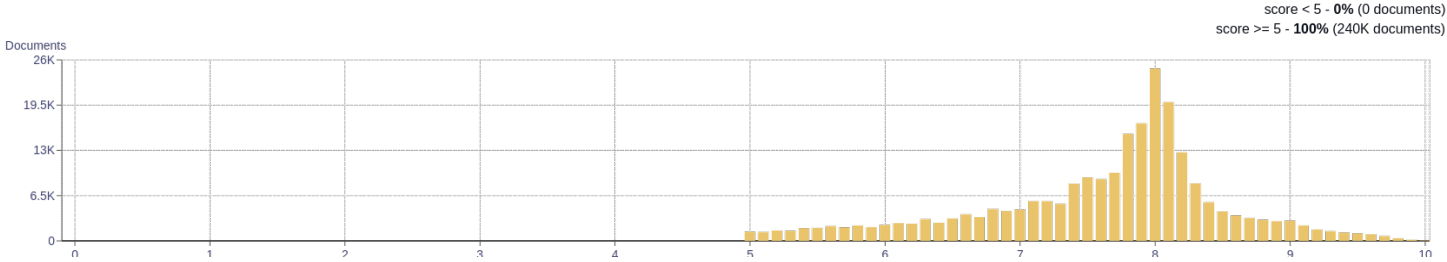
Number of segments



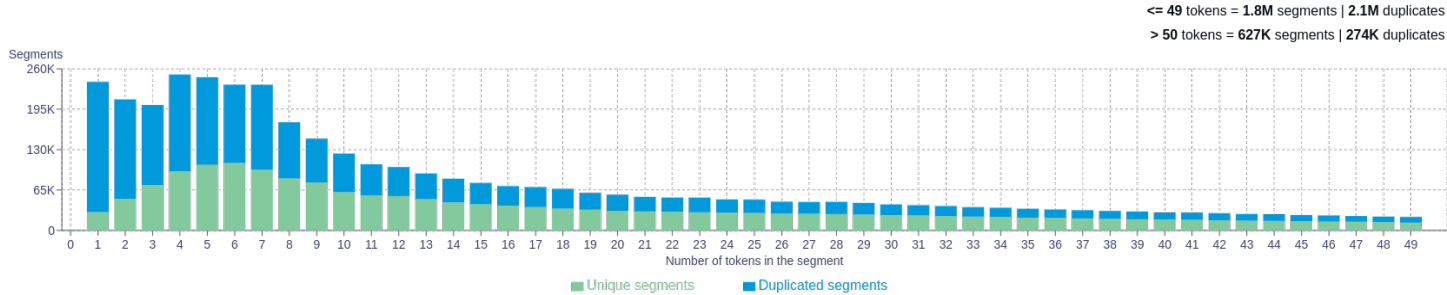
Percentage of segments in Faroese (fo) inside documents



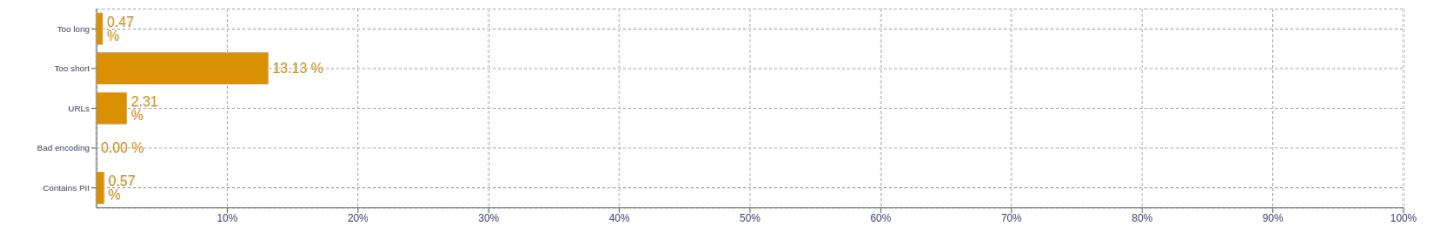
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	og 3387444 í 3249989 at 2768467 er 1627231 á 1349045
2	tá ið 102655 í færoyum 90913 er ein 79544 at fáa 76951 í dag 72703
3	at tað er 17509 og tað er 16845 ber til at 16276 av tí at 10303 í minsta lagi 10041
4	skriva so til vp 5324 varð fyrstu ferð løgd 3854 greinin varð fyrstu ferð 3853 fyrstu ferð løgd út 3853 posted by listin at 3789
5	varð fyrstu ferð løgd út 3853 greinin varð fyrstu ferð løgd 3853 partur av bókaheild til undirvísingina 3302 av bókaheild til undirvísingina í 3302 bókaheild til undirvísingina í støddfrøði 2820

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>