

General overview

Corpus	Date	Language
hplt-v3-fao_Latn	9/17/2025	Faroese (fo)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
323,746	5,361,617	3,702,614 (69.06 %)	135M	701,463,362	719.84 MB

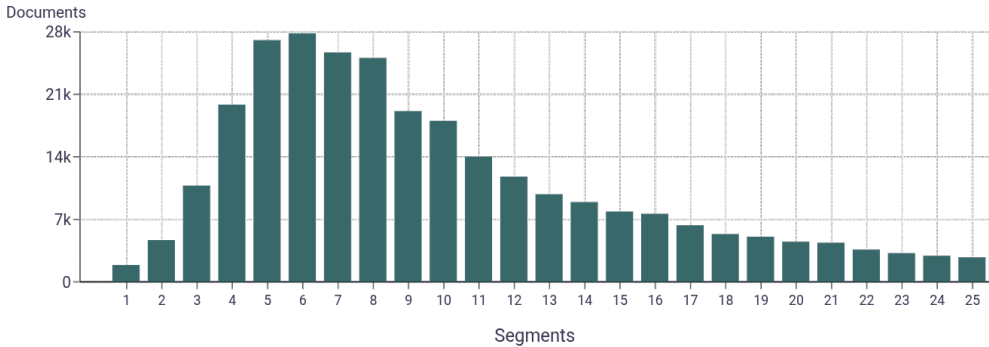
Top 10 domains

Domain	Docs	% of total
in.fo	19K	6.01%
vp.fo	13K	4.16%
kvf.fo	12K	3.84%
dagur.fo	7.5K	2.33%
portal.fo	7.3K	2.27%
r7.fo	7.2K	2.24%
nordlysid.fo	7.2K	2.21%
roysni.fo	7.1K	2.19%
jn.fo	6.9K	2.14%
setur.fo	6.1K	1.89%

Top 10 TLDs

Domain	Docs	% of total
fo	281K	86.69%
com	25K	7.86%
org	7.7K	2.37%
net	3.6K	1.10%
dk	3.2K	0.99%
info	964	0.30%
news	459	0.14%
be	311	0.10%
no	253	0.08%
is	207	0.06%

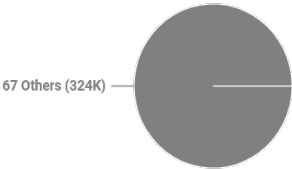
Documents size (in segments) ⓘ



≤ 25 segments **85.94%** (278K documents)
> 25 segments **14.06%** (46K documents)

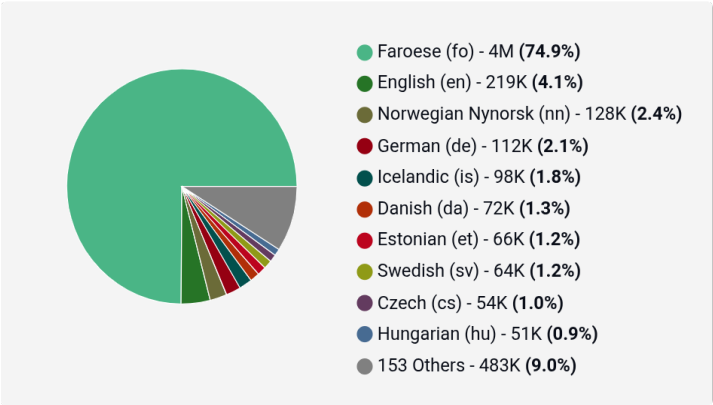
Document collections

CC = 92.70%
IA = 7.30%

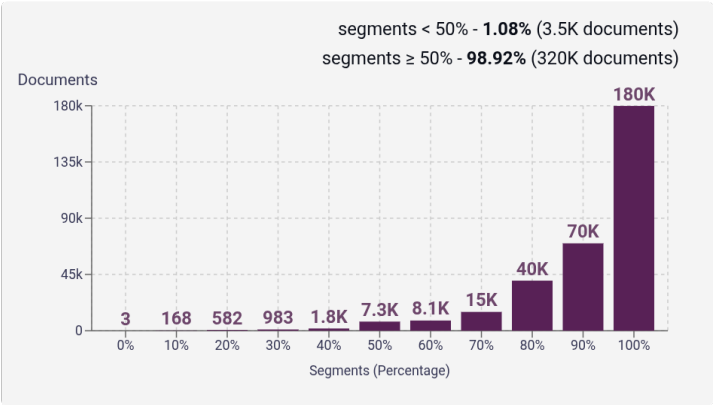


Language Distribution

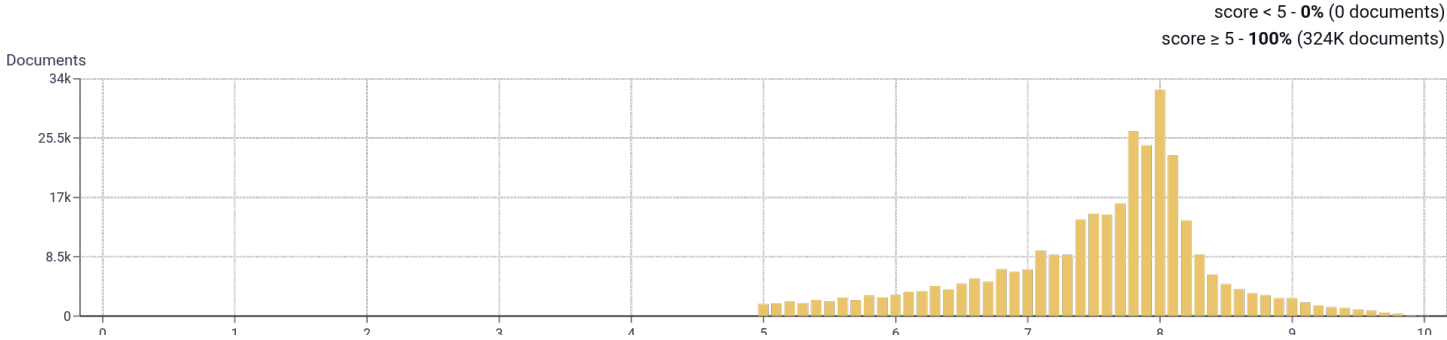
Number of segments in the Faroese (fo) corpus



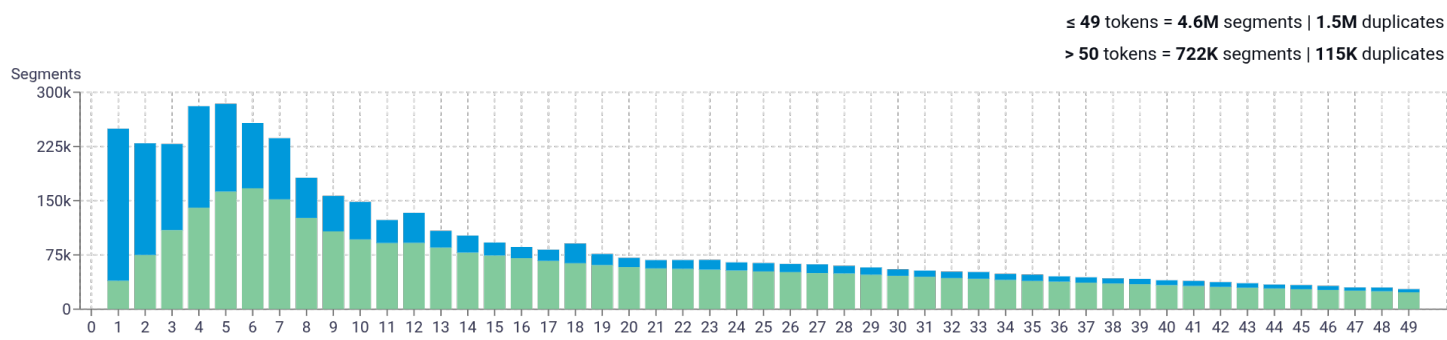
Percentage of segments in Faroese (fo) inside documents



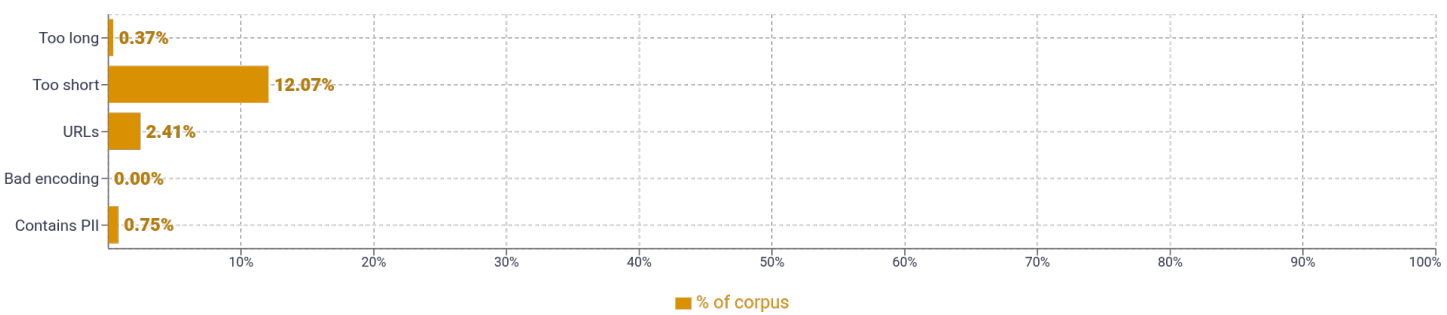
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	og 4,056,720av 1,101,525vit 698,649ein 682,917eru 668,208	
2	tá ið 106,344vit hava 60,435partur av 47,699og so 42,137eins og 35,195	
3	siggja hetta innihald 13,716vinarlíga broyt tínar 13,715broyt tínar kennifíla 13,715vp ikki veit 13,341so til vp 13,340	
4	vinarlíga broyt tínar kennifíla 13,715privatlívsstillingar fyrri at siggja 13,715skriva so til vp 13,340varð fyrstu ferð lögð 10,955greinin varð fyrstu ferð 10,954	
5	og privatlívsstillingar fyrri at siggja 13,715varð fyrstu ferð lögð út 10,953greinin varð fyrstu ferð lögð 10,953mest lisið í farnu viku 6,995løgtingslóg um broyting í løgtingslóg 2,742	

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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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>, , , 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				