HPLT Analytics report

General overview

Corpus	Analytics date	Language
fon_Latn.jsonl.tsv	12/5/2024	Fon (fon)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
1,226	14,764	9,689 (65.63 %)	1.6M	6.21 MB	5,321,478

Top 10 domains

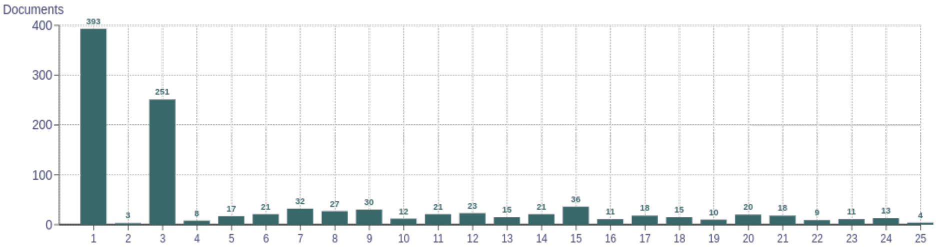
Domain	Docs	% of total
jw.org	602	49.10
bible.is	573	46.74
saxwe.net	19	1.55
spip.net	10	0.82
unicode.org	6	0.49
wikimedia.org	4	0.33
ohchr.org	2	0.16
mp3songspk.info	2	0.16
3songspk.com	1	0.08
songspkking.com	1	0.08

Top 10 TLDs

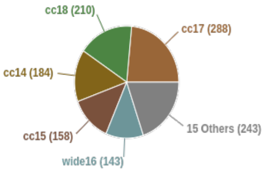
Domain	Docs	% of total
org	618	50.41
is	573	46.74
net	29	2.37
com	4	0.33
info	2	0.16

Documents size (in segments)

<= 25 segments **84.75%** (1K documents)
> 25 segments **15.25%** (187 documents)

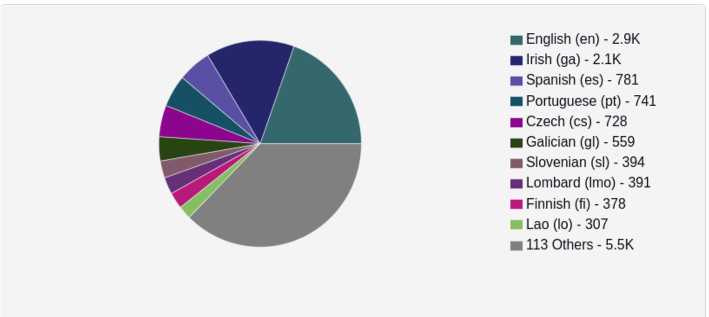


Documents by collection



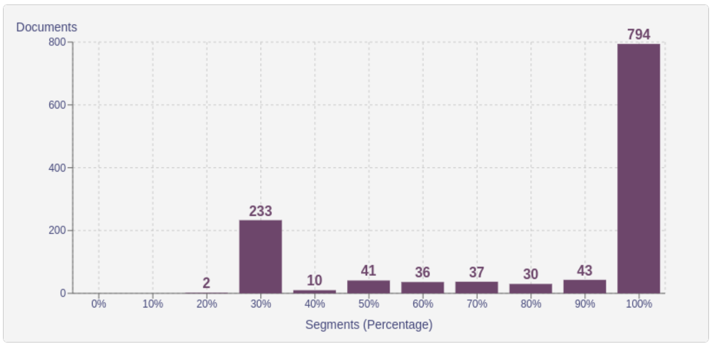
Language Distribution

Number of segments



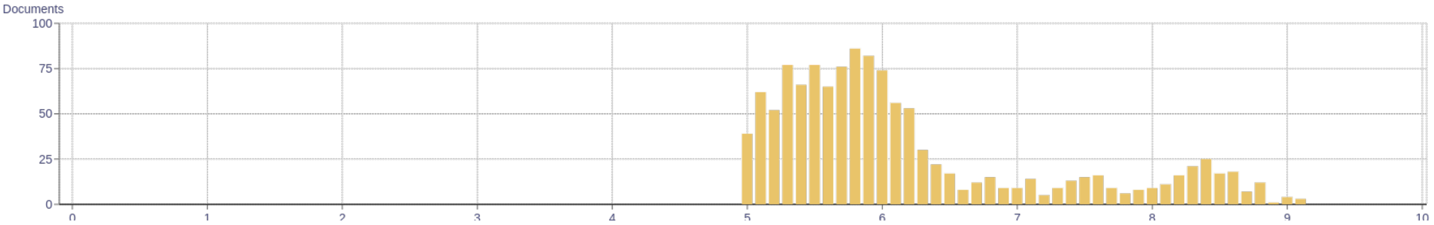
*Fon (fon) identification might be inaccurate because language is not supported by Fasttext

Percentage of segments in Fon (fon) inside documents



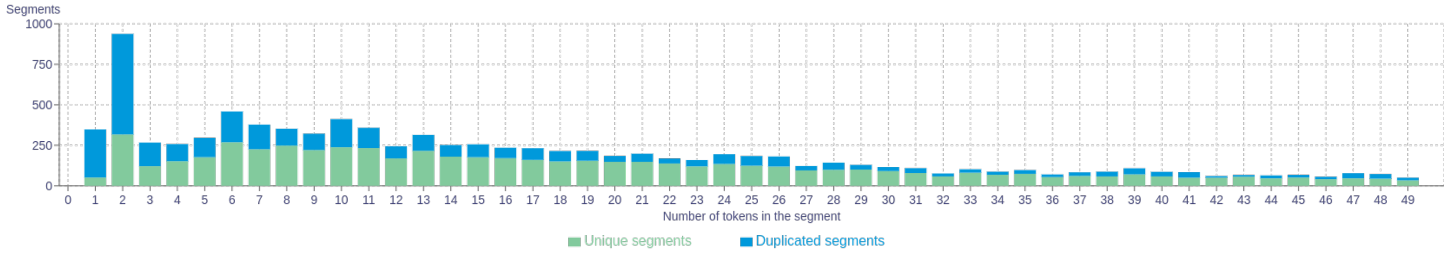
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (1.2K documents)

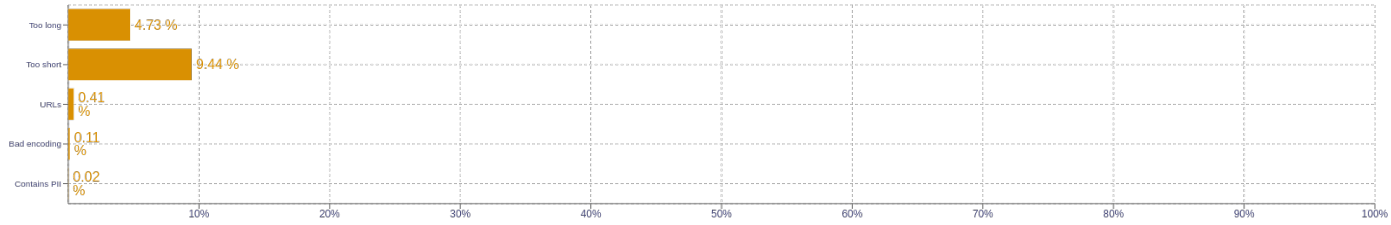


Segment length distribution by token

<= 49 tokens = **6K** segments | **3.6K** duplicates
> 50 tokens = **5.1K** segments | **1.5K** duplicates



Segment noise distribution



Frequent n-grams

Size	n-grams
1	lɛ | 32423 ɛ | 17979 na | 17871 mɛ | 15998 tɔn | 12563
2	tɔn lɛ | 2367 kɔn tè | 1830 ɖó na | 1489 u je | 1243 mawu tɔn | 1243
3	je ɔ je | 889 di ɛɛ numa | 870 nì kɔn tè | 731 kě ɔ dɛ | 700 tè wlò popoɛ | 538
4	u je ɔ je | 308 tèǒ nì kɔn tè | 308 n tèǒ joo nɛ | 242 ngmà je ne sùsu | 219 tǐi doo tǐ jleě | 187
5	bǐì tentǎn jǔu wǎn ane | 131 ɛ mǐ ɛ kpa kpa | 121 sà ɔɔ wlò kǔ mǒ | 110 alɔ mǐ bɔ mǐ na | 101 ke ɛ toon jleě mǒ | 99

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