HPLT Analytics report

General overview

Corpus	Analytics date	Language
ltz_Latn.jsonl.tsv	9/6/2024	Luxembourgish (lb)

Volumes

Docs	Segments	Unique segments	Tokens	Size	Characters
246,930	5,058,584	2,319,774 (45.86 %)	130M	692.32 MB	705,589,105

Top 10 domains

Domain	Docs	% of total
wikipedia.org	102K	41.48
rtl.lu	17K	6.88
100komma7.lu	4.3K	1.73
adr.lu	3.3K	1.35
eldo.lu	2.6K	1.05
recetin.com	1.9K	0.78
vsaduidoma.com	1.9K	0.75
merchanttrue.news	1.7K	0.68
moien.lu	1.6K	0.65
piraten.lu	1.4K	0.59

Top 10 TLDs

Domain	Docs	% of total
org	106K	42.79
lu	79K	32.02
com	44K	17.97
net	2.5K	1.02
de	2.4K	0.98
eu	1.9K	0.76
news	1.7K	0.68
zone	1.1K	0.44
top	914	0.37
info	868	0.35

Documents size (in segments)

<= 25 segments **79.05%** (195K documents)
> 25 segments **20.95%** (52K documents)
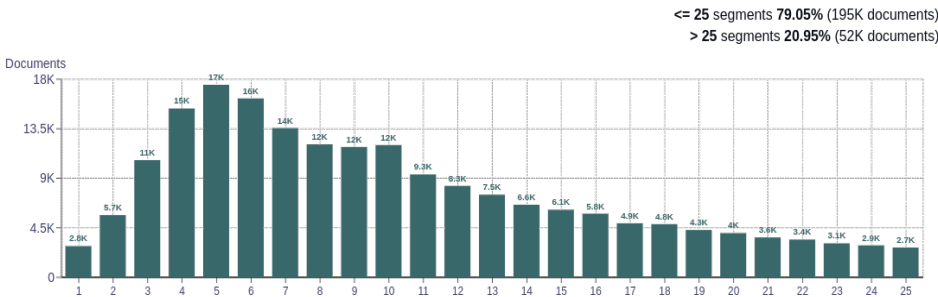
Documents by collection

cc22 (78K)
cc21 (32K)
cc18 (28K)
18 Others (109K)
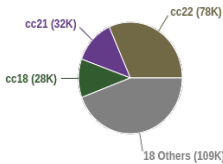
Language Distribution

Number of segments

- Luxembourgish (lb) - 2.8M
- German (de) - 955K
- English (en) - 332K
- French (fr) - 259K
- Dutch (nl) - 101K
- Italian (it) - 93K
- Low German (nds) - 89K
- Danish (da) - 40K
- Spanish (es) - 35K
- Swedish (sv) - 34K
- 160 Others - 272K
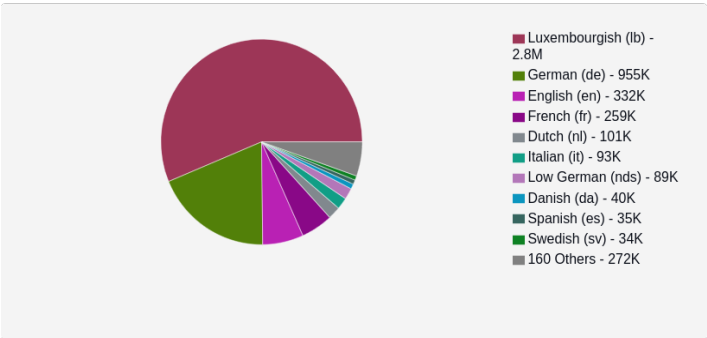
Percentage of segments in Luxembourgish (lb) inside documents
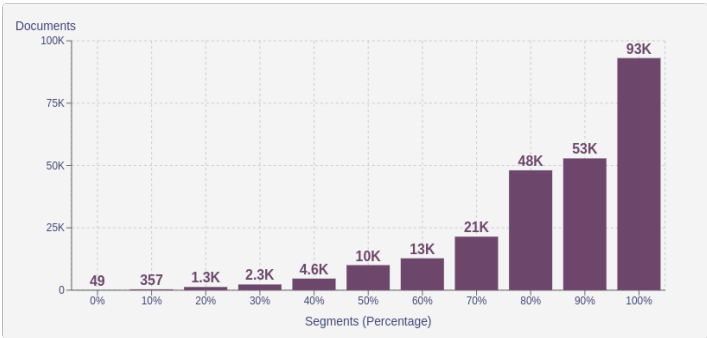
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (247K documents)

Segment length distribution by token

<= 49 tokens = **1.8M** segments | **2.4M** duplicates
> 50 tokens = **803K** segments | **296K** duplicates

Number of tokens in the segment

- Unique segments
- Duplicated segments

Segment noise distribution

Too long	0.00 %
Too short	9.88 %
URLs	1.61 %
Bad encoding	0.01 %
Contains PII	0.18 %
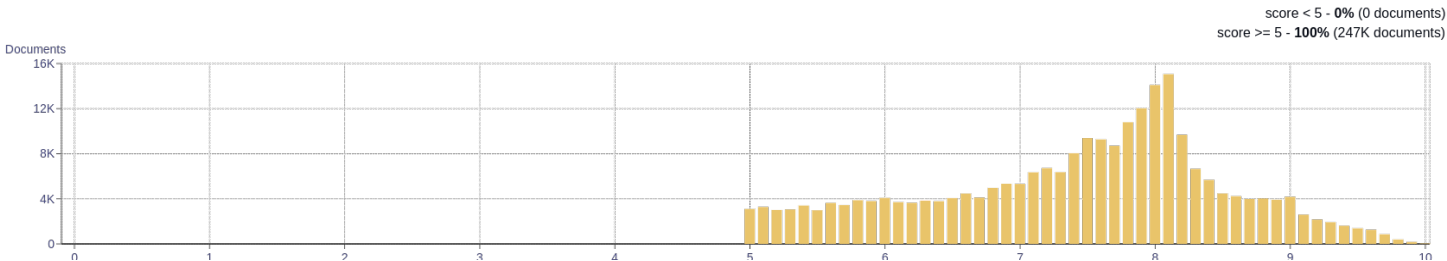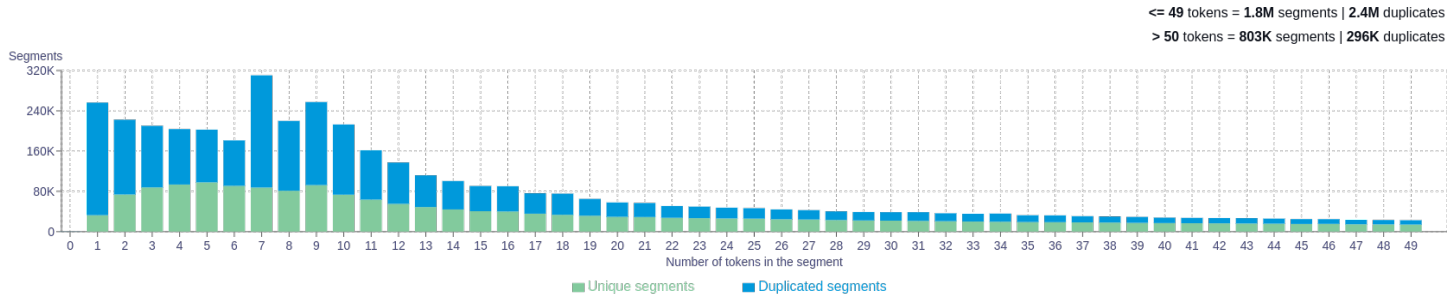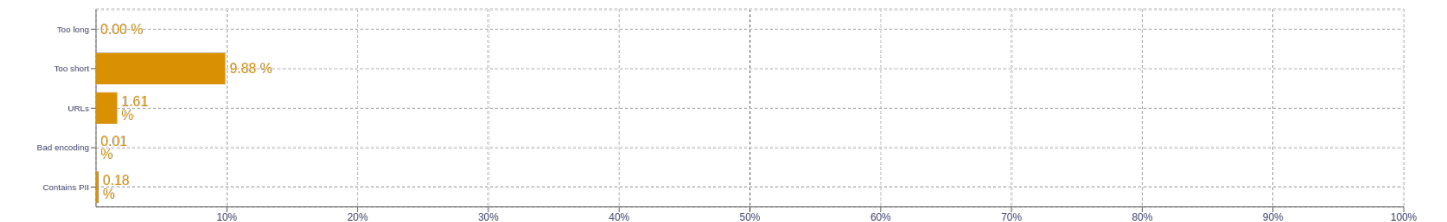
Frequent n-grams

Size	n-grams
1	d | 2444318 änneren | 598020 wéi | 495945 quelltext | 278694 joer | 213882
2	quelltext änneren | 278184 wéi d | 32317 stad lëtzebuerg | 17204 z. b. | 16733 well d | 15165
3	detailer vum foussballlännermatch | 10296 wéi och ëmmer | 8743 lëtzebuergesch foussballnationalequipe verléiert | 8111 websäit vun european | 6363 qualifikatioun fir d | 5607
4	brandenburg an der havel | 9905 kader vun der qualifikatioun | 5469 websäit vun european football | 5058 informatioun doriwwer am artikel | 4813 lëscht vun de lëtzebuerger | 4145
5	verléiert an der stad lëtzebuerg | 3773 foussballnationalequipe verléiert an der stad | 3158 säit befaasst sech mam joer | 2926 déifsttemperaturen an der nächster nuecht | 1750 nächster nuecht leie bei ronn | 1749

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