HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-ltz_Latn	9/18/2025	Luxembourgish

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
407,481	7,965,654	6,135,723 (77.03 %)	241M	1,331,272,155	1.28 GB

Top 10 domains

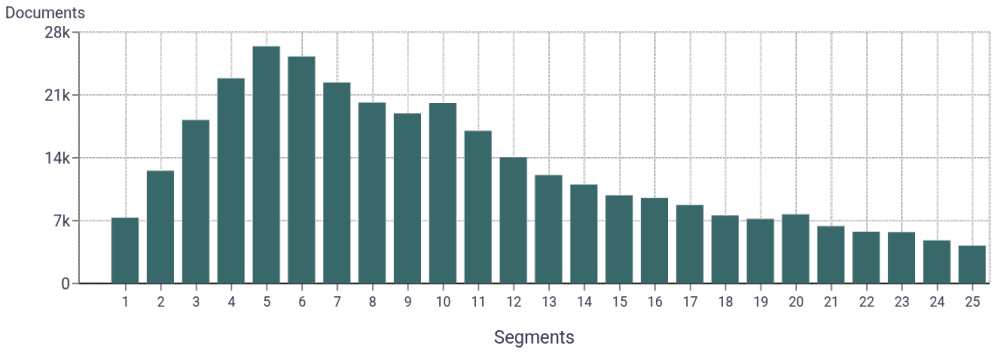
Domain	Docs	% of total
rtl.lu	37K	9.03%
wikipedia.org	28K	6.92%
100komma7.lu	14K	3.45%
moien.lu	12K	3.00%
eldo.lu	9.9K	2.42%
martech.zone	5.1K	1.25%
adr.lu	4.3K	1.05%
estatisticasgra...	2.8K	0.69%
meteoboulaide.com	2.8K	0.68%
piraten.lu	2.6K	0.63%

Top 10 TLDs

Domain	Docs	% of total
com	166K	40.84%
lu	154K	37.67%
org	38K	9.26%
net	6.4K	1.58%
eu	6.2K	1.51%
zone	5.1K	1.25%
de	4K	0.97%
fr	1.6K	0.39%
news	1.5K	0.36%
online	1.4K	0.35%

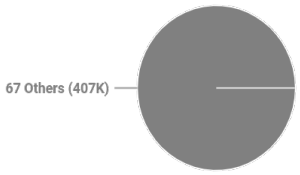
Documents size (in segments)

≤ 25 segments **79.97%** (326K documents)
> 25 segments **20.03%** (82K documents)



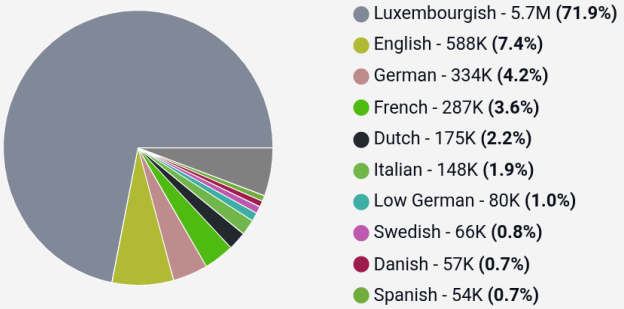
Document collections

CC = 95.19%
IA = 4.81%



Language Distribution

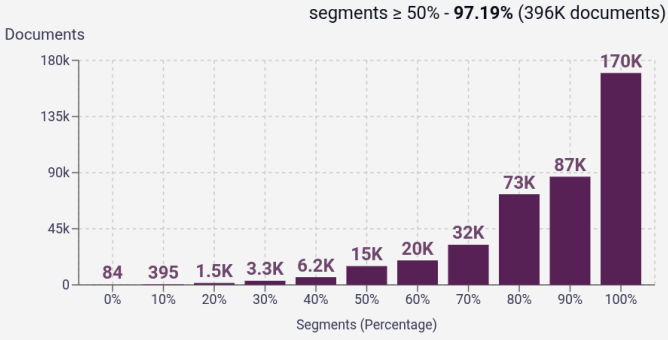
Number of segments in the Luxembourgish corpus



- Luxembourgish - 5.7M **(71.9%)**
- English - 588K **(7.4%)**
- German - 334K **(4.2%)**
- French - 287K **(3.6%)**
- Dutch - 175K **(2.2%)**
- Italian - 148K **(1.9%)**
- Low German - 80K **(1.0%)**
- Swedish - 66K **(0.8%)**
- Danish - 57K **(0.7%)**
- Spanish - 54K **(0.7%)**
- 163 Others - 451K **(5.7%)**

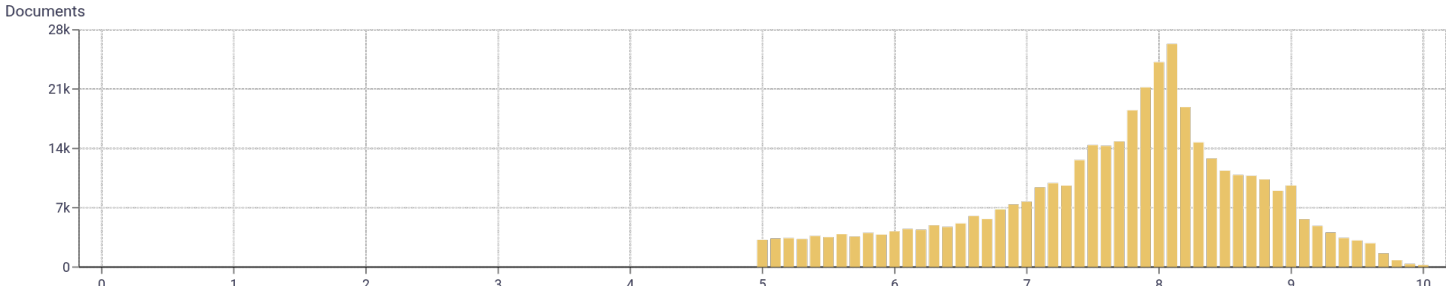
Percentage of segments in Luxembourgish inside documents

segments < 50% - **2.81%** (11K documents)
segments ≥ 50% - **97.19%** (396K documents)



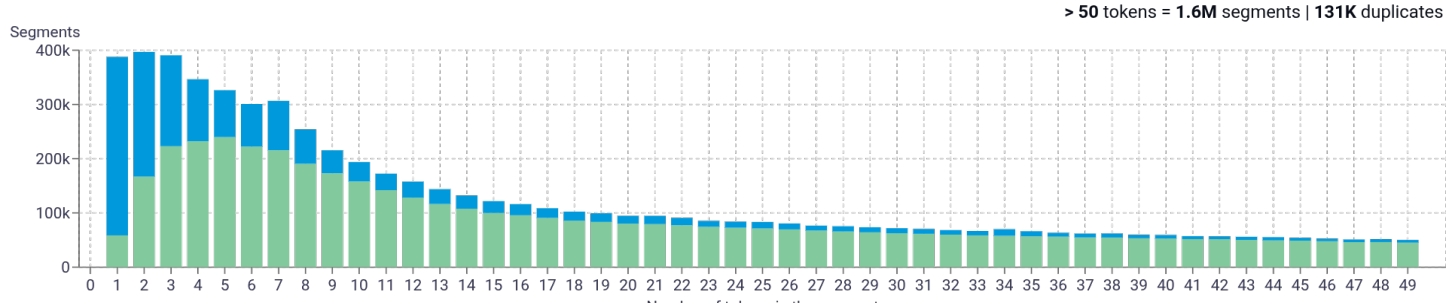
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (407K documents)

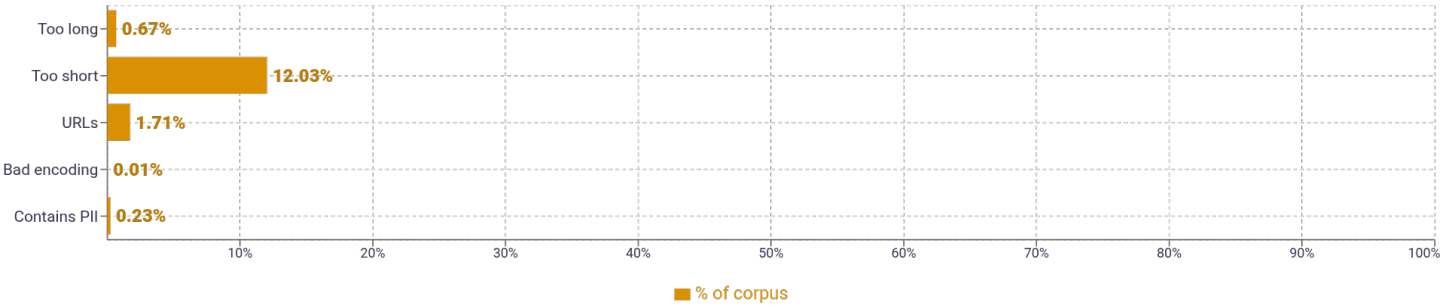


Segment length distribution by token

≤ 49 tokens = **6.4M** segments | **1.7M** duplicates
> 50 tokens = **1.6M** segments | **131K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	d | 4,771,521 · wéi | 1,281,942 · dës | 412,048 · vill | 407,480 · joer | 362,740
2	wéi d | 73,923 · quelltext änneren | 63,871 · well d | 27,688 · iwwert d | 26,716 · egal ob | 26,569
3	wéi och ëmmer | 39,855 · einfach ze benotzen | 12,710 · zur selwechter zäit | 9,788 · lues a lues | 8,838 · aart a weis | 8,575
4	éischt fir ze kommentéieren | 9,885 · iwwerdeems déi lescht noriicht | 5,055 · invité vun der redaktioun | 3,281 · ëffentlech-rechtleche radio zu lëtzebuerg | 3,269 · proposéiert programmer op lëtzebuergesch | 3,252
5	iwwerdeems déi lescht noriicht uginn | 5,055 · eenzegen ëffentlech-rechtleche radio zu lëtzebuerg | 3,253 · déifsttemperaturen an der nächster nuecht | 2,338 · nächster nuecht leie bei ronn | 2,337 · childhood story plus untold biografie | 1,658

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

Register labels

Name	Abbr.	Name	Abbr.	Name	Abbr.
Machine-translated	MT	**How-to or instructions**	**HI**	Description of a thing or person	dtp
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				