HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-ron_Latn	9/18/2025	Romanian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
95,911,906	2,165,142,698	1,156,254,328 (53.40 %)	63B	337,220,739,476	325.83 GB

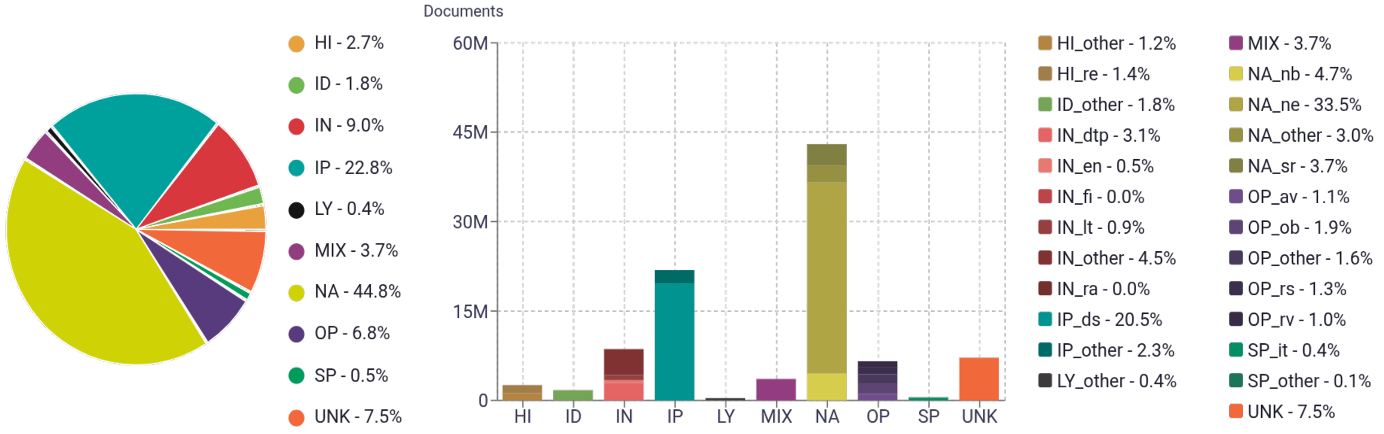
Top 10 domains

Domain	Docs	% of total
wordpress.com	1.8M	1.83%
blogspot.com	1.7M	1.78%
ziare.com	754K	0.79%
adevarul.ro	686K	0.72%
hotnews.ro	630K	0.66%
9am.ro	614K	0.64%
blogspot.ro	579K	0.60%
wall-street.ro	491K	0.51%
mediafax.ro	440K	0.46%
bzi.ro	416K	0.43%

Top 10 TLDs

Domain	Docs	% of total
ro	71M	73.80%
com	13M	13.33%
md	3.2M	3.38%
net	2.3M	2.38%
org	1.3M	1.36%
eu	1.1M	1.12%
info	1M	1.06%
tv	174K	0.18%
ru	169K	0.18%
news	152K	0.16%

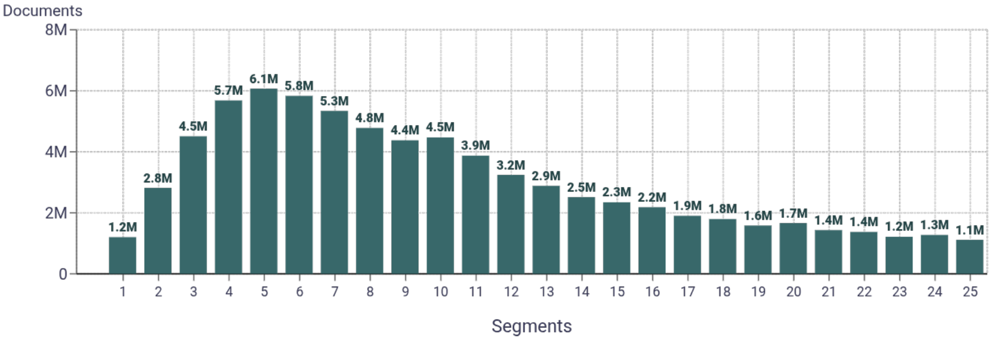
Register labels



MT:4.1% | 3.9M Documents

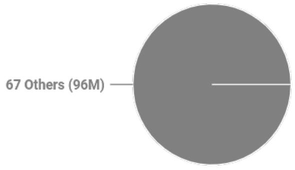
Documents size (in segments) ⓘ

≤ 25 segments **78.69%** (75M documents)
> 25 segments **21.31%** (20M documents)



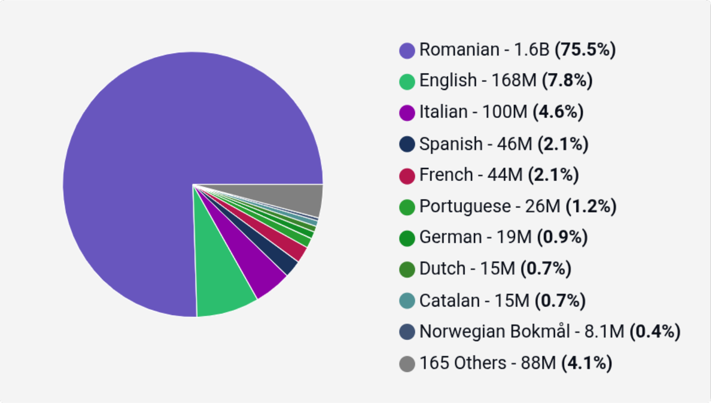
Document collections

CC = 86.96%
IA = 13.04%



Language Distribution

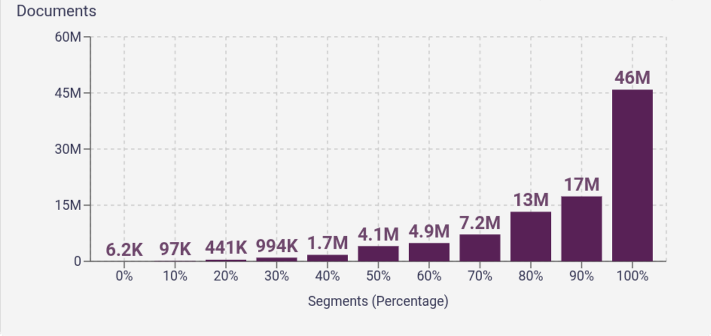
Number of segments in the Romanian corpus



Percentage of segments in Romanian inside documents

segments < 50% - **3.42%** (3.3M documents)

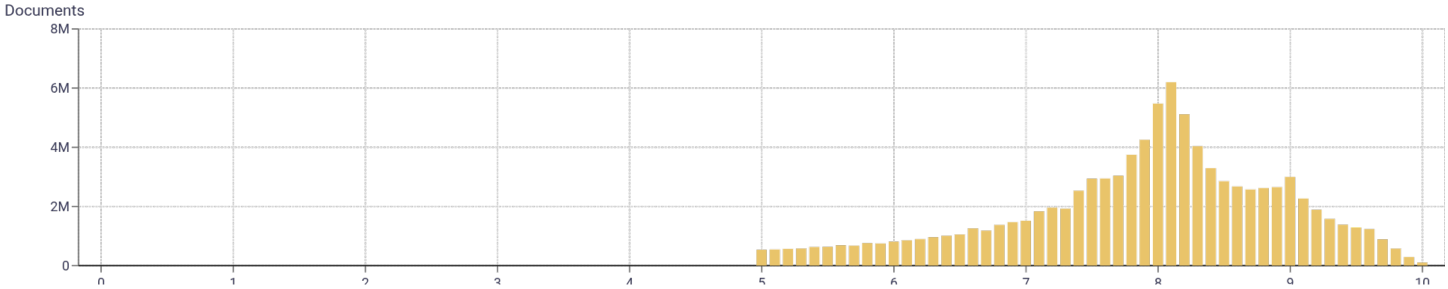
segments ≥ 50% - **96.58%** (93M documents)



Distribution of documents by document score

score < 5 - **0%** (0 documents)

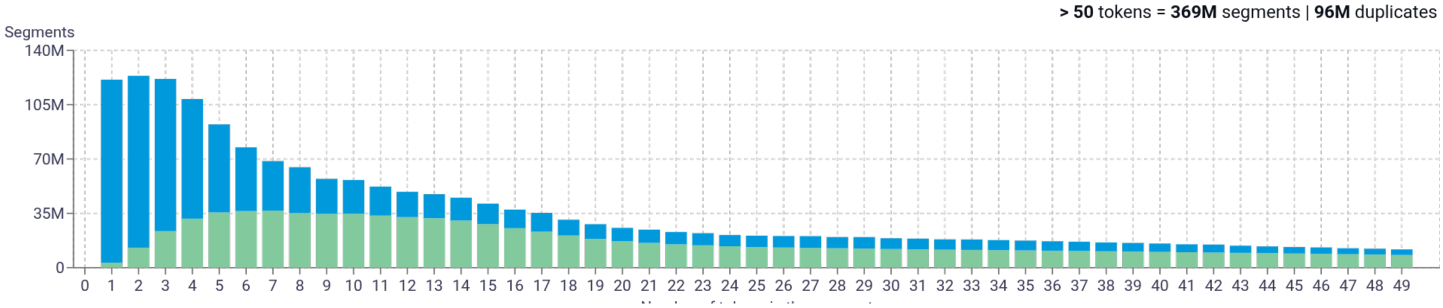
score ≥ 5 - **100%** (96M documents)



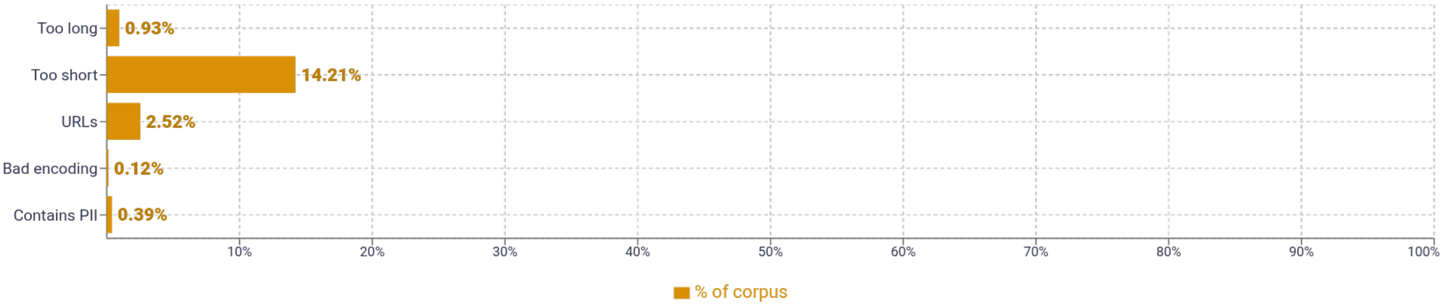
Segment length distribution by token

≤ 49 tokens = **1.8B** segments | **916M** duplicates

> 50 tokens = **369M** segments | **96M** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	și | 612,555,527 ani | 77,789,635 trebuie | 69,011,428 s-a | 67,620,221 mare | 63,748,967
2	precum și | 6,540,787 anul trecut | 4,708,244 de-a lungul | 4,506,220 s-ar putea | 3,641,559 of the | 3,375,296
3	punct de vedere | 5,641,858 milioane de euro | 3,995,069 pierderea în greutate | 2,842,600 având în vedere | 2,657,687 pierdere în greutate | 2,604,107
4	datelor cu caracter personal | 765,940 statele unite ale americii | 748,783 rugăm să ne contactați | 739,052 lipsa unui acord scris | 574,463 acord scris din partea | 572,480
5	sursa si daca inserati vizibil | 563,378 precizati sursa si daca inserati | 563,351 facebook pentru a primi periodic | 528,645 9am sau conecteaza-te prin facebook | 475,200 aboneaza-te la 9am sau conecteaza-te | 475,047

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				