HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
HPLT-v2-fra_Latn.tsv	10/3/2024	French (fr)

Volumes

Docs	Segments	Characters	Size
401,831,660	10,557,090,126	1,447,270,554,770	1.37 TB

Top 10 domains

Domain	Docs	% of total
canalblog.com	5.8M	1.44%
blogspot.com	5.5M	1.36%
wikipedia.org	3.2M	0.81%
wordpress.com	3.1M	0.78%
over-blog.com	3.1M	0.76%
blogspot.fr	2.6M	0.66%
lefigaro.fr	1.2M	0.31%
lexpress.fr	970K	0.24%
herokuapp.com	901K	0.22%
skyrock.com	899K	0.22%

Top 10 TLDs

Domain	Docs	% of total
com	182M	45.32%
fr	120M	29.87%
org	24M	6.00%
net	16M	3.98%
be	11M	2.66%
ca	8.3M	2.06%
ch	7M	1.75%
info	5.4M	1.34%
eu	3.6M	0.90%
co	1.4M	0.35%

Register labels

Pie chart legend:
- HI - 3.3%
- ID - 3.6%
- IN - 15.7%
- IP - 30.2%
- LY - 0.3%
- MIX - 7.7%
- NA - 25.1%
- OP - 7.8%
- SP - 0.8%
- UNK - 5.5%

🤖 **MT**:2.6% | 10M Documents

Bar chart legend:
- HI_other - 1.7%
- HI_re - 1.6%
- ID_other - 3.6%
- IN_dtp - 7.2%
- IN_en - 1.5%
- IN_fi - 0.0%
- IN_lt - 1.0%
- IN_other - 5.7%
- IN_ra - 0.2%
- IP_ds - 25.6%
- IP_ed - 0.0%
- IP_other - 4.6%
- LY_other - 0.3%
- MIX - 7.7%
- NA_nb - 7.7%
- NA_ne - 12.0%
- NA_other - 3.3%
- NA_sr - 2.2%
- OP_av - 0.9%
- OP_ob - 2.0%
- OP_other - 1.3%
- OP_rs - 0.8%
- OP_rv - 2.9%
- SP_it - 0.6%
- SP_other - 0.1%
- UNK - 5.5%
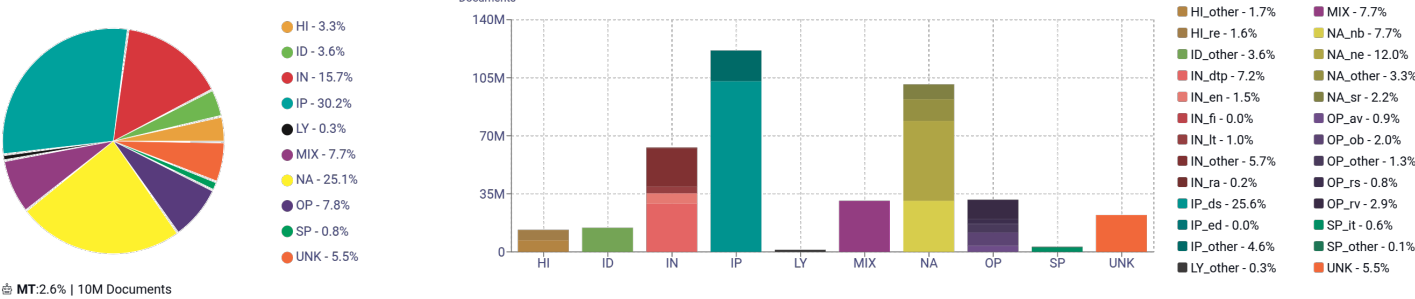
Documents size (in segments)
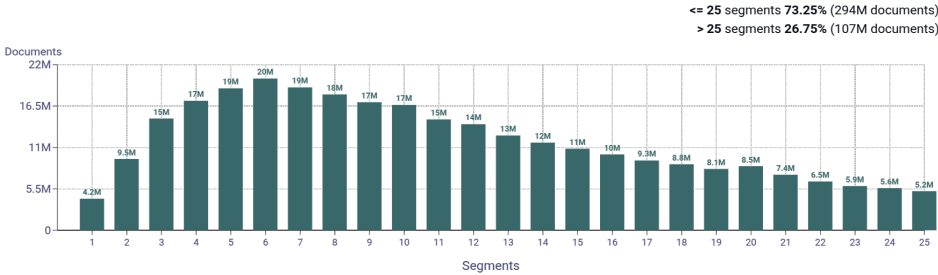
<= 25 segments **73.25%** (294M documents)
> 25 segments **26.75%** (107M documents)

Documents by collection

CC = 65.26%
IA = 34.74%

cc18 (78M), cc22 (108M), 19 Others (215M)
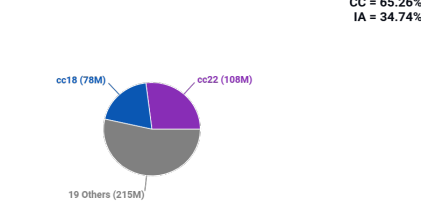
Language Distribution

Number of segments in the French (fr) corpus

- French (fr) - 9.1B
- English (en) - 698M
- Italian (it) - 361M
- Spanish (es) - 73M
- German (de) - 64M
- Portuguese (pt) - 38M
- Dutch (nl) - 35M
- Catalan (ca) - 20M
- Russian (ru) - 16M
- Swedish (sv) - 13M
- 165 Others - 157M
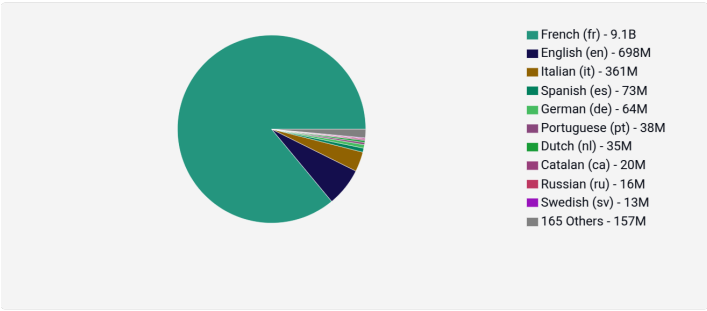
Percentage of segments in French (fr) inside documents
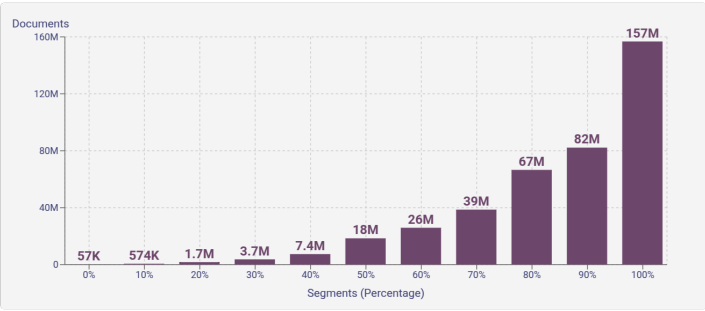
0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
57K	574K	1.7M	3.7M	7.4M	18M	26M	39M	67M	82M	157M

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score >= 5 - **100%** (402M documents)

Segment noise distribution

- Too long: 0.87 %
- Too short: 14.33 %
- URLs: 2.14 %
- Bad encoding: 0.00 %
- Contains PII: 0.46 %
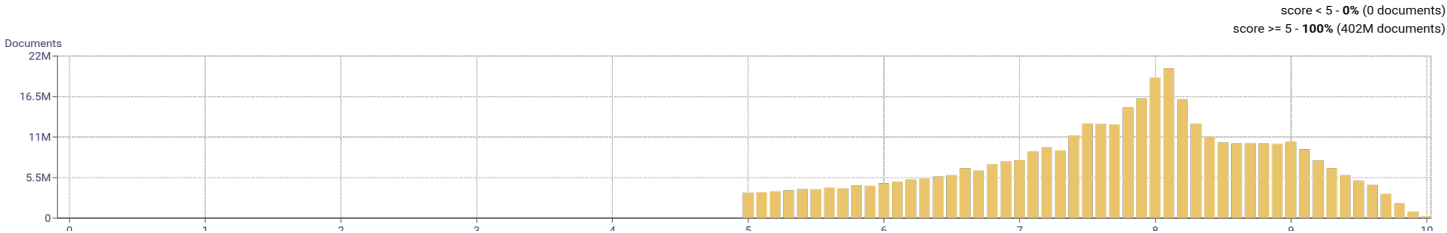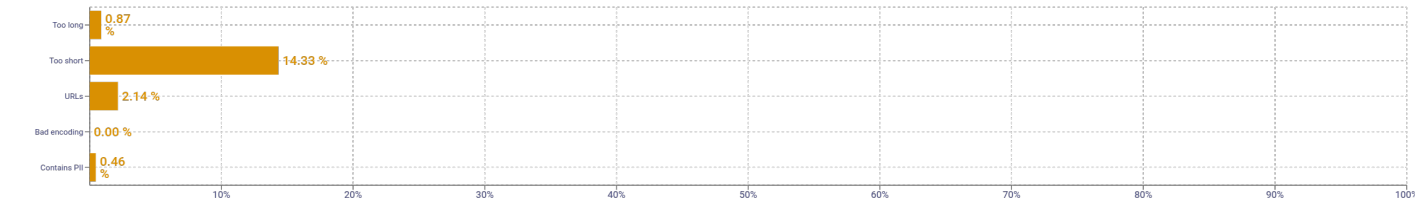
About HPLT Analytics

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.
Machine-translated	MT
Lyrical	LY
Spoken	SP
Interview	it
Interactive discussion	ID
Narrative	NA
News report	ne
Sports report	sr
Narrative blog	nb

Name	Abbr.
How-to or instructions	HI
Recipe	re
Informational persuasion	IP
Description with intent to sell	ds
News & opinion blog or editorial	ed
Informational description	IN
Enciclopedia article	en
Research article	ra

Name	Abbr.
Description of a thing or person	dtp
FAQ	fi
Legal terms & conditions	lt
Opinion	OP
Review	rv
Opinion blog	ob
Denominational religious blog or sermon	rs
Advice	av