

General overview

Corpus	Date	Language
hplt-v3-fra_Latn	9/25/2025	French

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
603,878,673	15,629,875,449	7,572,233,924 (48.45 %)	440B	2,259,182,542,118	2.13 TB

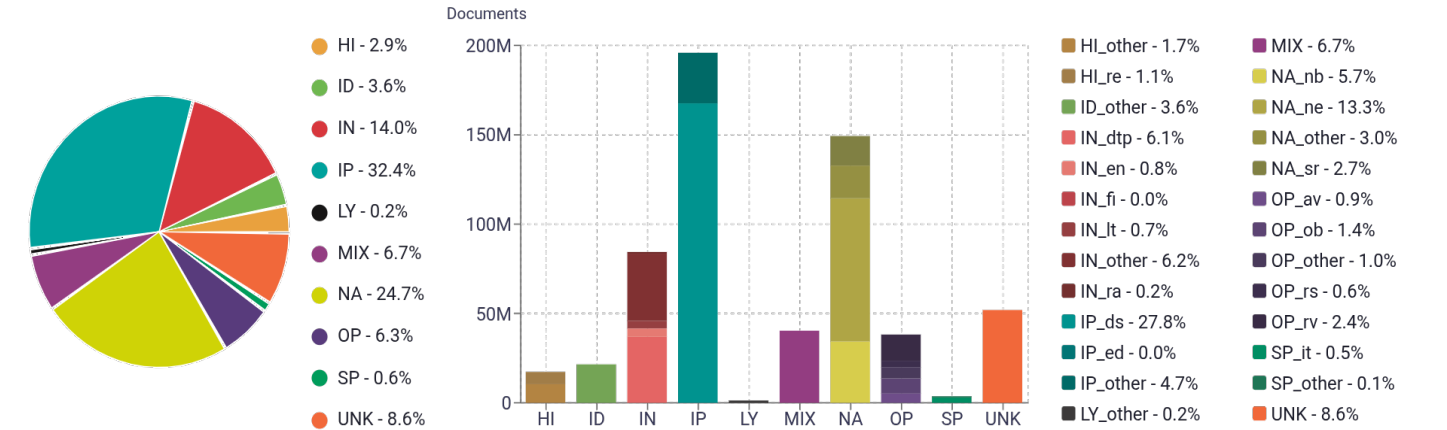
Top 10 domains

Domain	Docs	% of total
herokuapp.com	6.4M	1.06%
canalblog.com	6.4M	1.05%
blogspot.com	5.7M	0.95%
wordpress.com	4.2M	0.69%
over-blog.com	4M	0.66%
lefigaro.fr	3.1M	0.51%
francetvinfo.fr	1.7M	0.29%
wikipedia.org	1.4M	0.23%
blogspot.fr	1.4M	0.22%
web.app	1.3M	0.22%

Top 10 TLDs

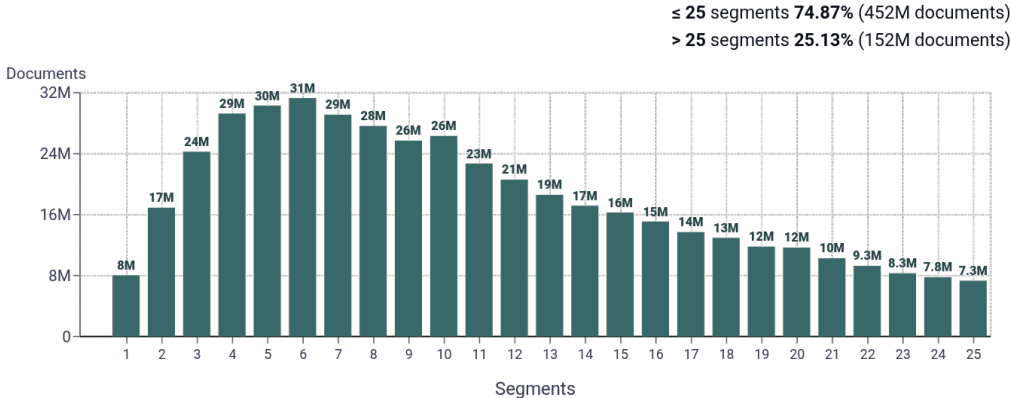
Domain	Docs	% of total
com	268M	44.34%
fr	181M	30.02%
org	29M	4.78%
net	23M	3.79%
be	16M	2.67%
ca	12M	1.92%
ch	11M	1.81%
info	8.7M	1.43%
eu	6.5M	1.08%
ma	2.3M	0.38%

Register labels

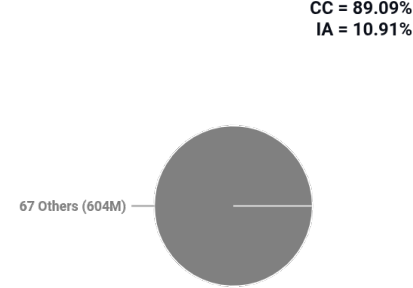


MT:5.7% | 34M Documents

Documents size (in segments) ⓘ

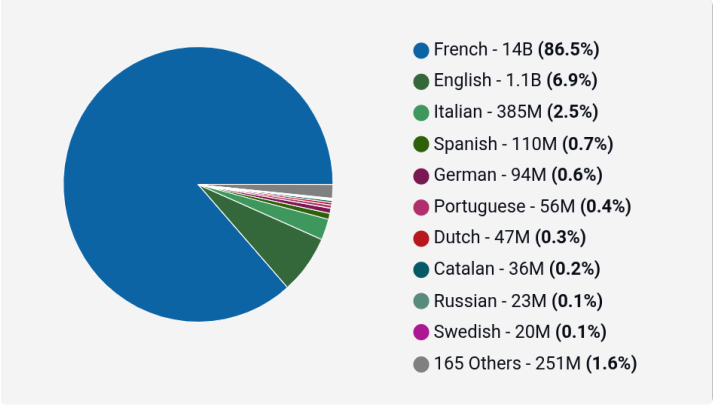


Document collections

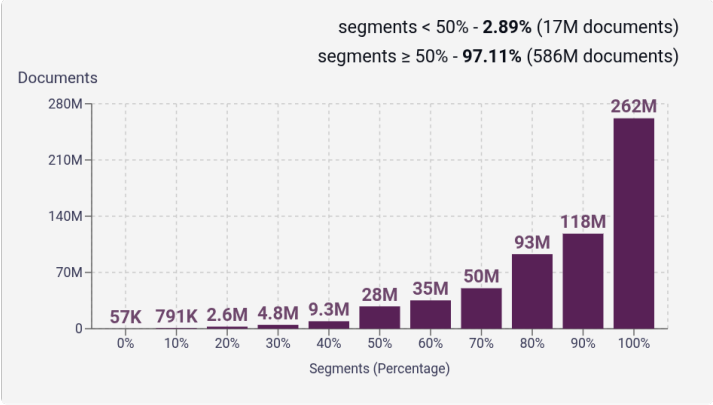


Language Distribution

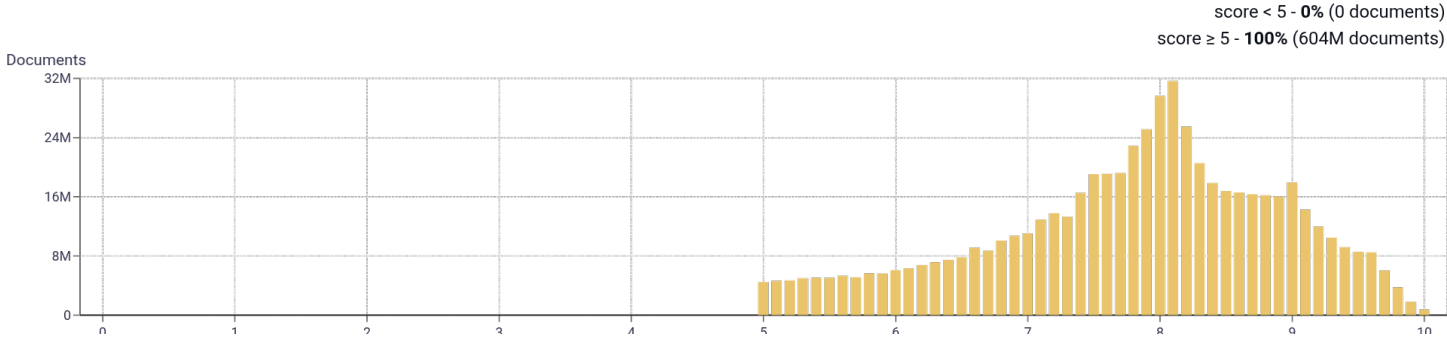
Number of segments in the French corpus



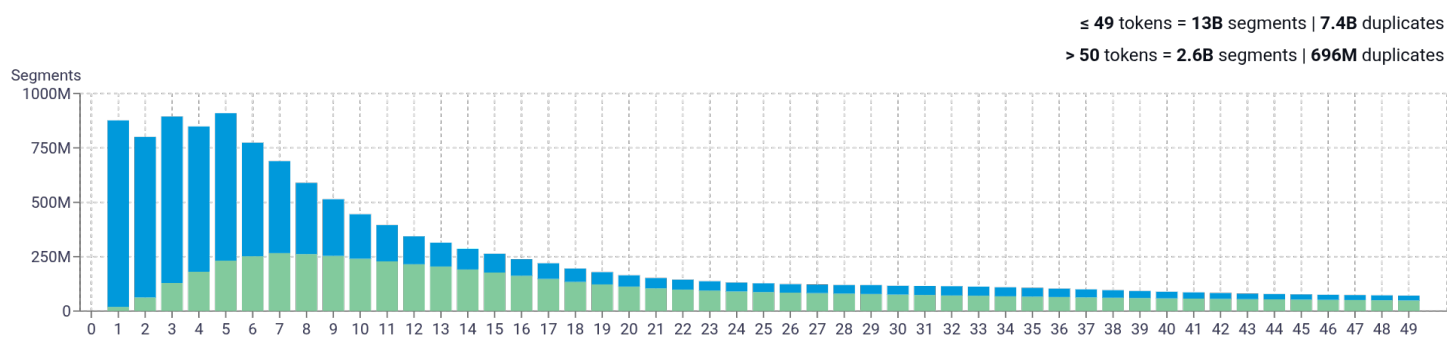
Percentage of segments in French inside documents



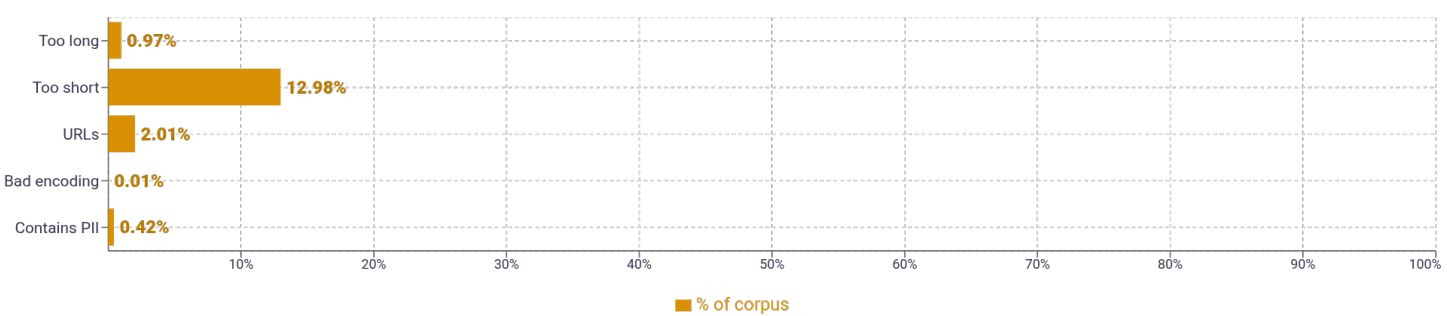
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>a 2,373,516,525</div> <div>plus 1,702,629,010</div> <div>cv 1,056,698,125</div> <div>cette 807,024,016</div> <div>tout 711,983,173</div>	
2	<div>peut être 99,423,476</div> <div>modele cv 62,442,591</div> <div>plan cul 56,974,393</div> <div>aujourd' hui 54,786,937</div> <div>exemple cv 52,251,995</div>	
3	<div>site de rencontre 100,103,939</div> <div>lire la suite 68,286,703</div> <div>exemple de cv 58,188,534</div> <div>plus en détail 56,297,997</div> <div>lettre de motivation 42,490,532</div>	
4	<div>site de rencontre gratuit 17,008,291</div> <div>comment faire un cv 9,999,503</div> <div>président de la république 7,256,419</div> <div>fur et à mesure 5,905,125</div> <div>société à responsabilité limitée 5,654,697</div>	
5	<div>cv et lettre de motivation 6,586,996</div> <div>hésitez pas à nous contacter 4,616,407</div> <div>gaz à effet de serre 3,410,115</div> <div>si vous avez des questions 2,821,706</div> <div>a mettre dans un cv 2,415,677</div>	

About HPLT Analytics

Volumes - Segments

Segments correspond to paragraph and list boundaries as defined by HTML elements (<p>, , , 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 $\frac{\text{number of types (uniques)}}{\text{number of tokens}}$, 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>, , , 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				