HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-ukr_Cyrl	9/19/2025	Ukrainian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
80,026,949	1,605,316,878	980,871,409 (61.10 %)	43B	243,132,195,763	409.41 GB

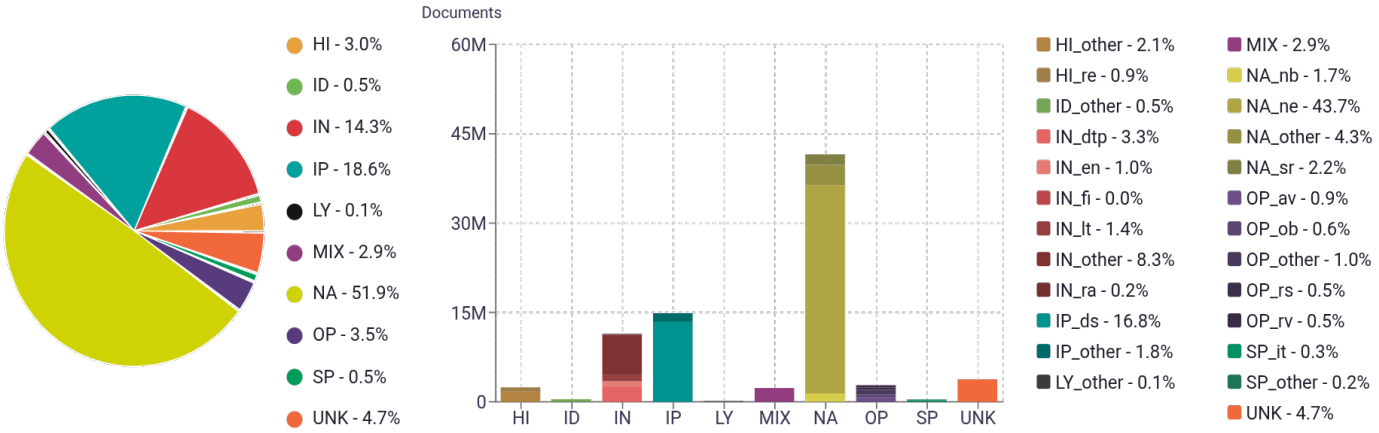
Top 10 domains

Domain	Docs	% of total
24tv.ua	700K	0.87%
pp.ua	655K	0.82%
obozrevatel.com	587K	0.73%
korrespondent.net	538K	0.67%
unian.ua	532K	0.66%
co.ua	495K	0.62%
wikipedia.org	468K	0.59%
segodnya.ua	450K	0.56%
ukrinform.ua	446K	0.56%
tsn.ua	414K	0.52%

Top 10 TLDs

Domain	Docs	% of total
ua	15M	18.73%
com.ua	15M	18.64%
com	13M	16.49%
gov.ua	3.5M	4.41%
net	3.4M	4.28%
org.ua	3.4M	4.26%
in.ua	3M	3.71%
org	3M	3.70%
info	2.6M	3.27%
ru	1.3M	1.62%

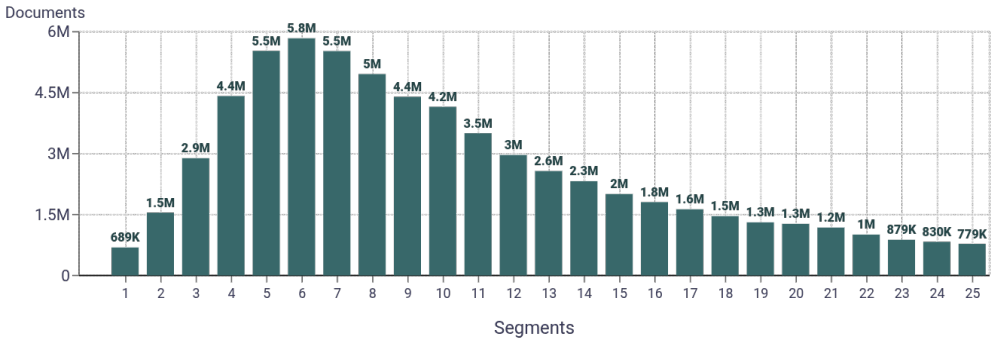
Register labels



MT:1.5% | 1.2M Documents

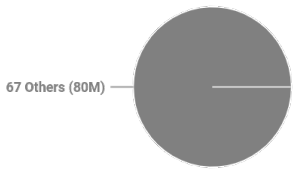
Documents size (in segments)

≤ 25 segments **81.77%** (65M documents)
> 25 segments **18.23%** (15M documents)



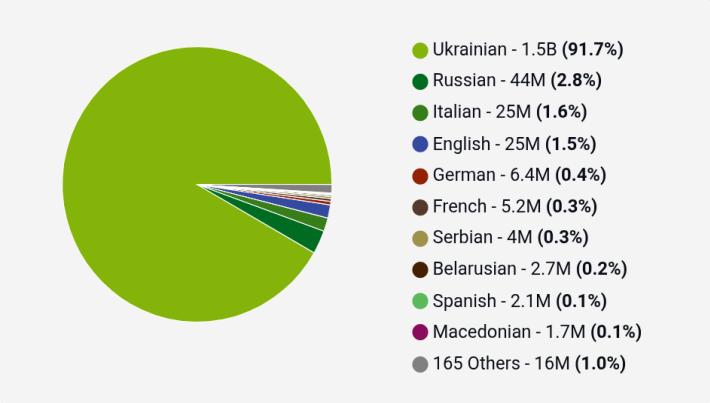
Document collections

CC = 91.44%
IA = 8.56%

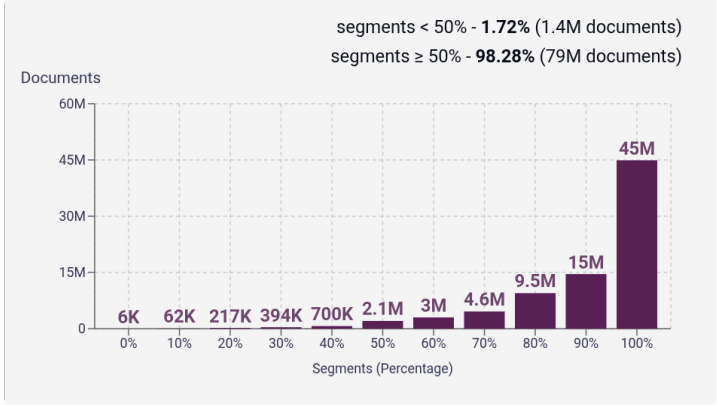


Language Distribution

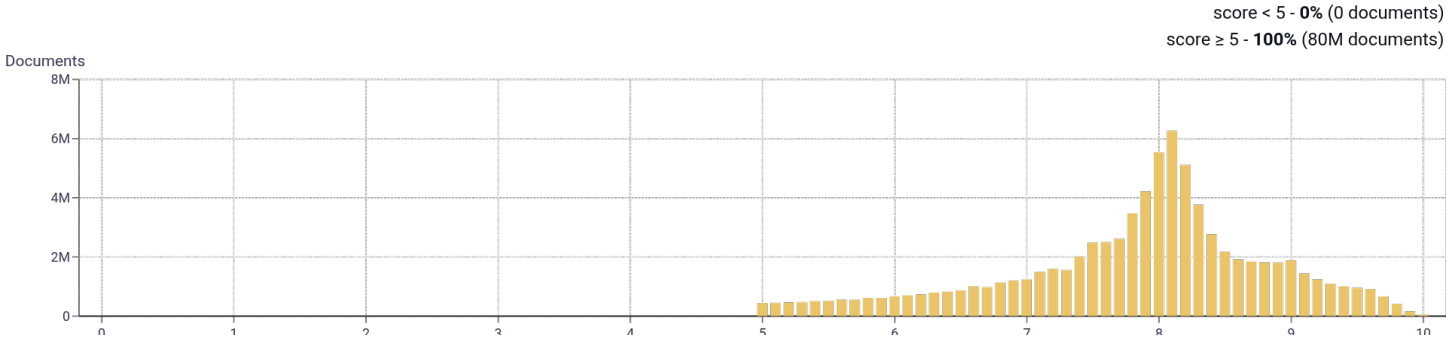
Number of segments in the Ukrainian corpus



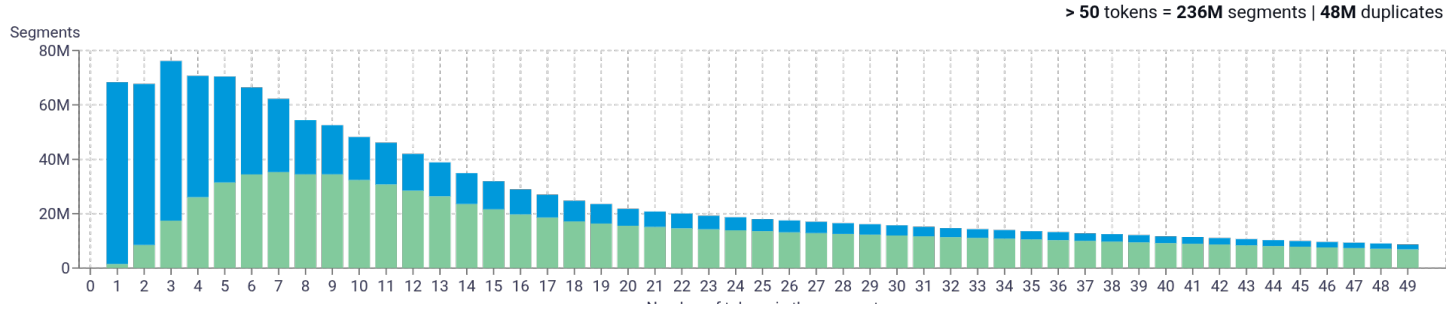
Percentage of segments in Ukrainian inside documents



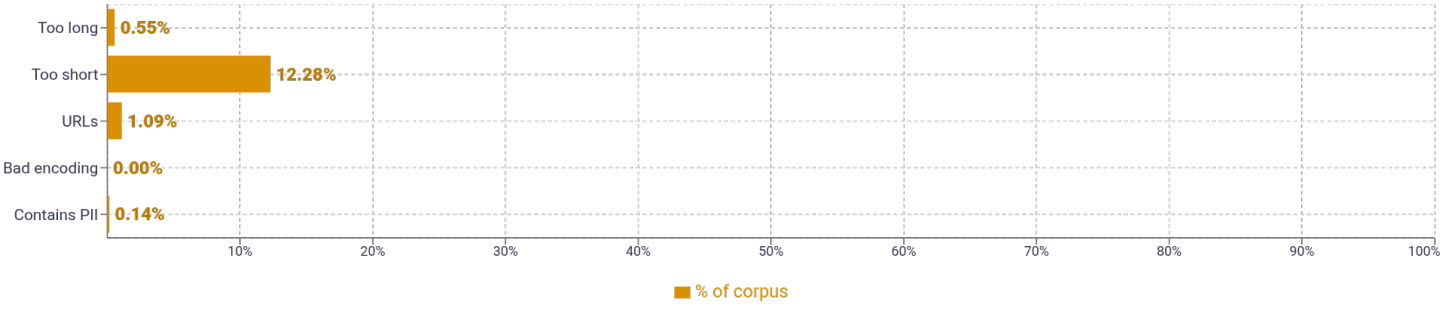
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	україни | 87,651,142 · його | 77,759,999 · із | 64,398,939 · він | 62,009,551 · при | 60,438,583
2	під час | 26,690,085 · може бути | 11,192,604 · при цьому | 10,398,687 · крім того | 8,252,612 · таким чином | 6,413,679
3	який відрізняється тим | 1,699,295 · кабінету міністрів україни | 1,540,524 · якщо у вас | 1,458,388 · полягає в тому | 1,455,864 · освіти і науки | 1,437,145
4	освіти і науки україни | 808,824 · якщо ви помітили помилку | 541,459 · президент україни володимир зеленський | 505,227 · товариство з обмеженою відповідальністю | 474,358 · міністерства освіти і науки | 428,191
5	виділіть необхідний текст і натисніть | 373,872 · необхідний текст і натисніть ctrl | 373,266 · міністерства освіти і науки україни | 342,863 · вітаю тебе з днем народження | 279,138 · змін до деяких законодавчих актів | 246,330

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				