HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-ydd_Hebr	9/24/2025	Yiddish (ydd)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
162,585	4,305,559	3,088,225 (71.73 %)	135M	711,061,465	1.18 GB

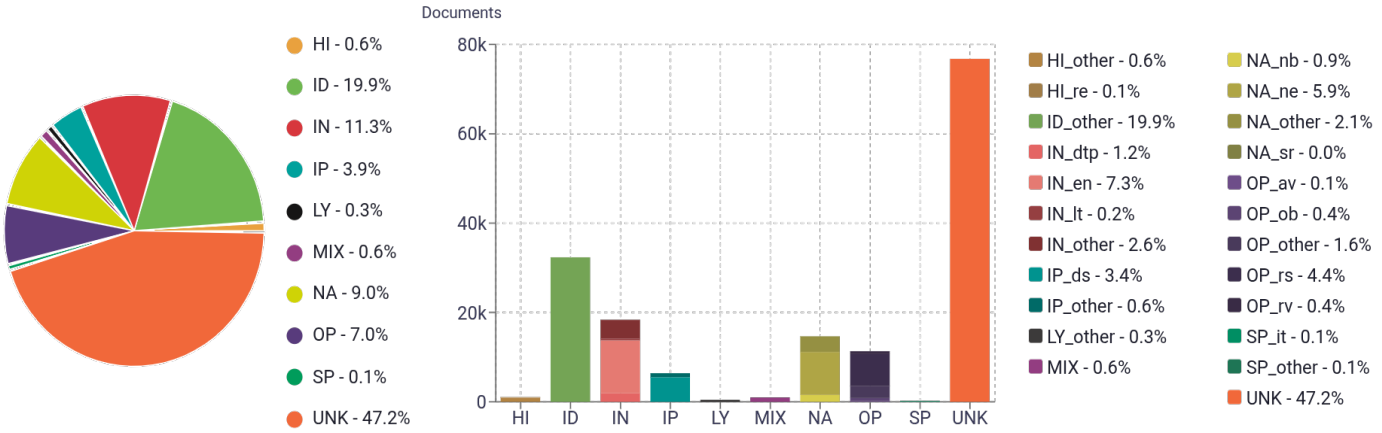
Top 10 domains

Domain	Docs	% of total
ivelt.com	18K	11.19%
kaveshtiebel.com	13K	8.02%
wikipedia.org	10K	6.18%
eureporter.co	6.3K	3.85%
yiddish.news	5.3K	3.24%
yiddishworld.com	5.1K	3.15%
soft-free-downl...	3.4K	2.07%
martech.zone	3.1K	1.88%
breslevcenter.com	2.5K	1.52%
creativosonline...	2.3K	1.44%

Top 10 TLDs

Domain	Docs	% of total
com	104K	63.80%
org	20K	12.44%
co	6.4K	3.91%
news	5.7K	3.48%
net	5.5K	3.41%
zone	3.1K	1.88%
ru	2.2K	1.32%
org.il	1.9K	1.19%
gov	1.7K	1.01%
de	1.2K	0.75%

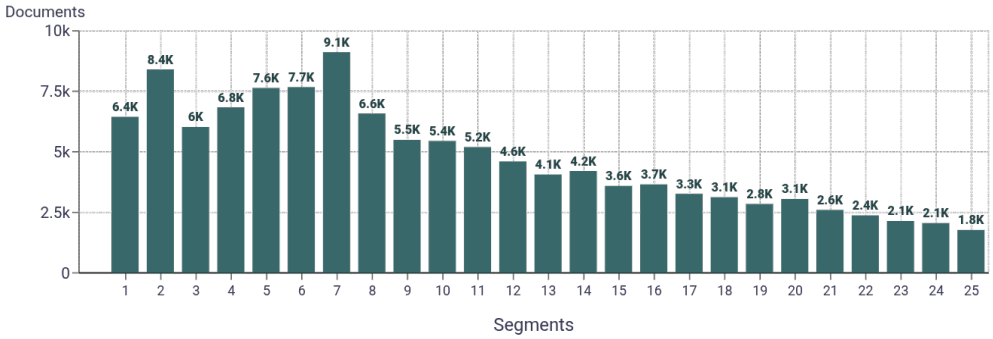
Register labels



MT:44.4% | 72K Documents

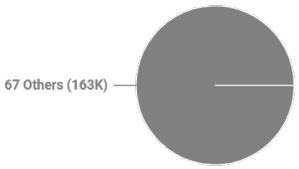
Documents size (in segments)

≤ 25 segments **72.72%** (118K documents)
> 25 segments **27.28%** (44K documents)



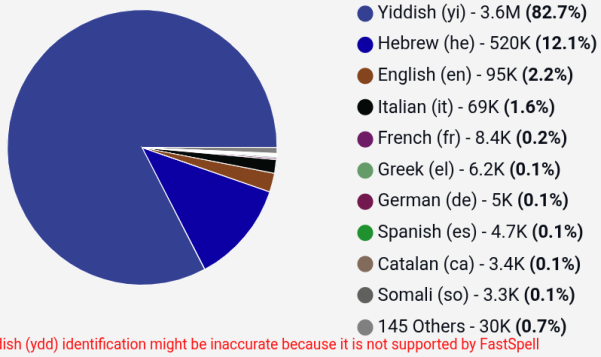
Document collections

CC = 81.51%
IA = 18.49%



Language Distribution

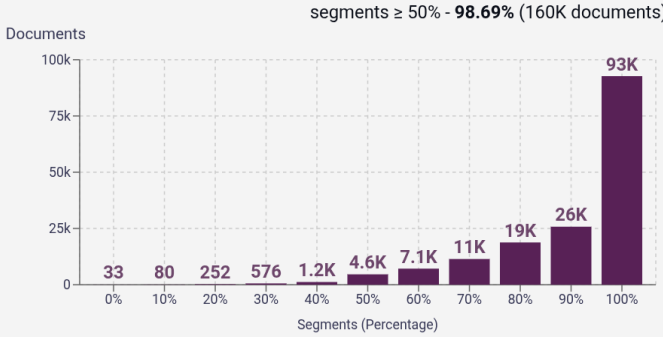
Number of segments in the Yiddish (ydd) corpus



*Yiddish (ydd) identification might be inaccurate because it is not supported by FastSpell

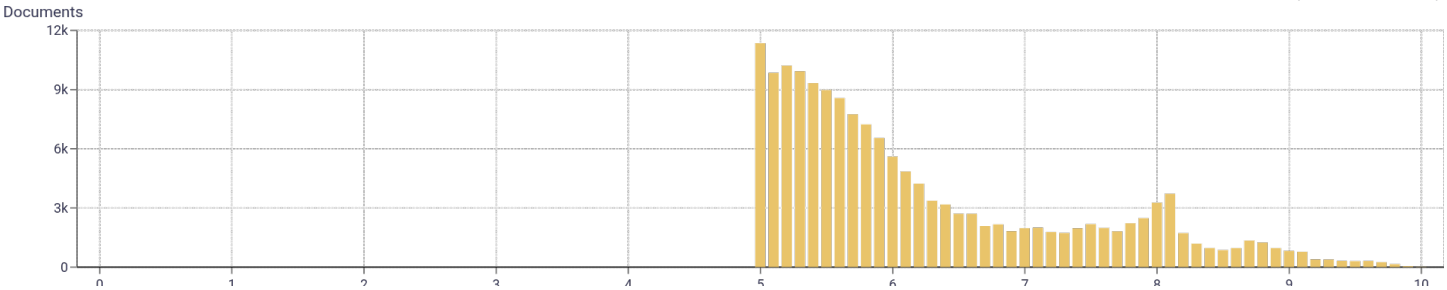
Percentage of segments in Yiddish (ydd) inside documents

segments < 50% - **1.31%** (2.1K documents)
segments ≥ 50% - **98.69%** (160K documents)



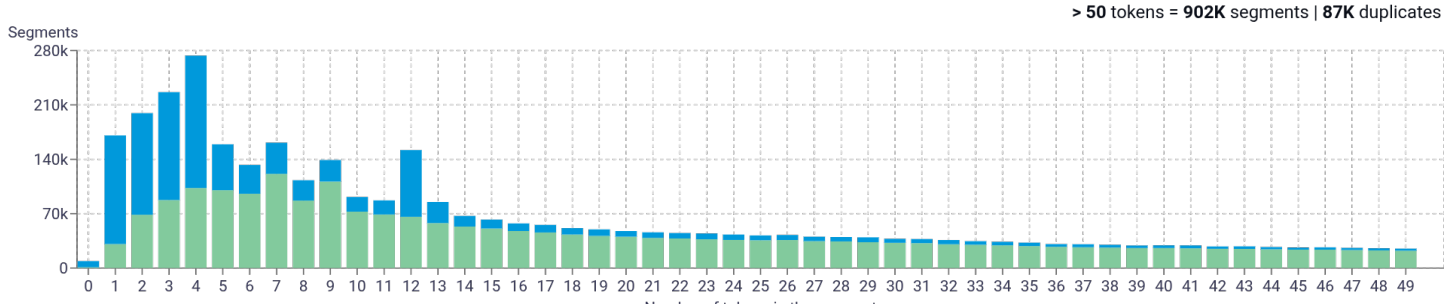
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (163K documents)

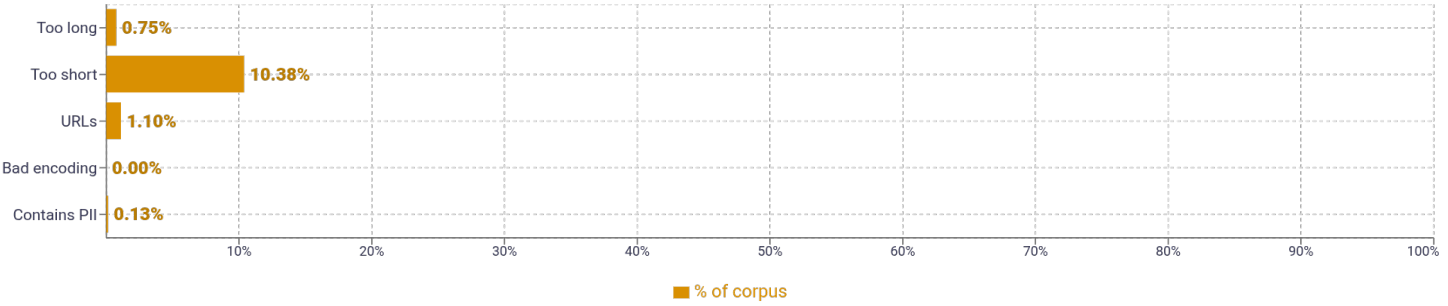


Segment length distribution by token

≤ 49 tokens = **3.4M** segments | **1.1M** duplicates
> 50 tokens = **902K** segments | **87K** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	2,411,400 | איז · 1,530,256 | אַ · 1,188,962 | עס · 994,723 | מיט · 984,320 | האט
2	398,609 | עס איז · 198,243 | איז געווען · 169,515 | האט געשריבן · 142,315 | איז אַ · 111,616 | איר קענען
3	71,497 | זיך איינגעשריבען אום · 37,943 | עס איז אַ · 32,387 | עס איז געווען · 18,610 | אַז עס איז · 17,257 | ער איז געווען
4	13,500 | איינער פון די מערסט · 10,682 | באנוצערס וואס דרייען זיך · 9,589 | וואס דרייען זיך דא · 9,351 | נישטא קיין איינגעשריבענע באנוצערס · 5,493 | עס איז געווען אַ
5	11,483 | זייט דער ערשטער צו באַמערקן · 9,547 | באנוצערס וואס דרייען זיך דא · 4,932 | איז איינער פון די מערסט · 4,732 | באניצער וואס לייענען דעם פארום · 2,694 | סטאַרי פלוס ונטאָלד ביאגראפיע פאַקס

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				