

General overview

Corpus	Date	Language
heb_Hebr.jsonl.tsv	6/13/2025	Hebrew (he)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
17,116,813	466,421,580	205,312,172 (44.02 %)	12B	56,374,344,692	92.68 GB

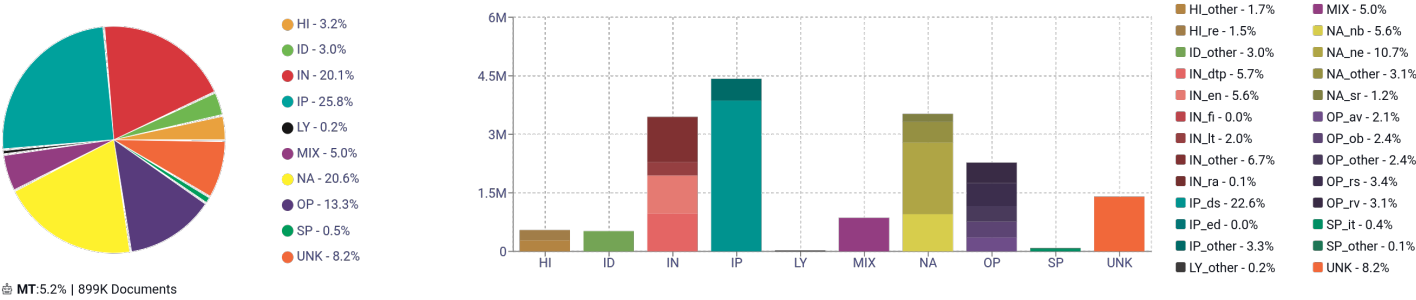
Top 10 domains

Domain	Docs	% of total
wikipedia.org	862K	5.03%
paika.net	318K	1.85%
blogspot.com	244K	1.42%
blogspot.co.il	237K	1.39%
articles.co.il	234K	1.37%
walla.co.il	195K	1.14%
wordpress.com	185K	1.08%
mako.co.il	165K	0.96%
nana10.co.il	146K	0.85%
ynet.co.il	142K	0.83%

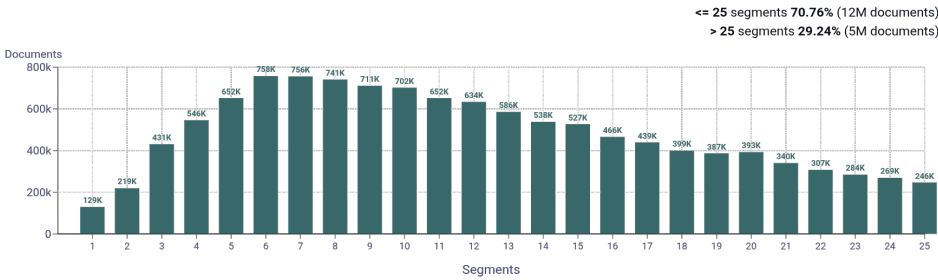
Top 10 TLDs

Domain	Docs	% of total
co.il	9M	52.69%
com	3.8M	22.01%
org	1.5M	9.01%
org.il	1.1M	6.57%
net	721K	4.21%
ac.il	237K	1.39%
info	92K	0.54%
gov.il	56K	0.33%
muni.il	51K	0.30%
me	36K	0.21%

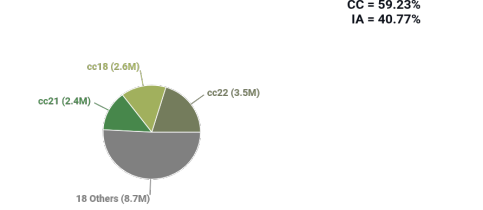
Register labels



Documents size (in segments)

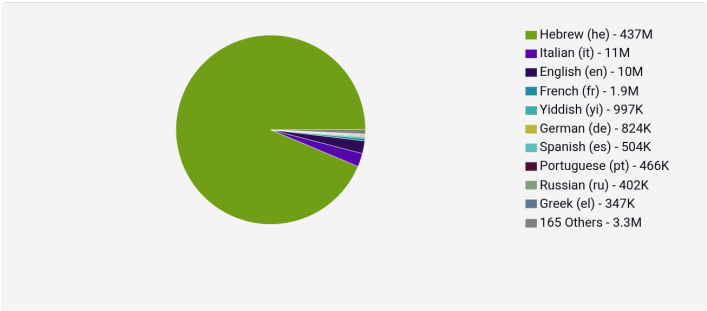


Documents by collection

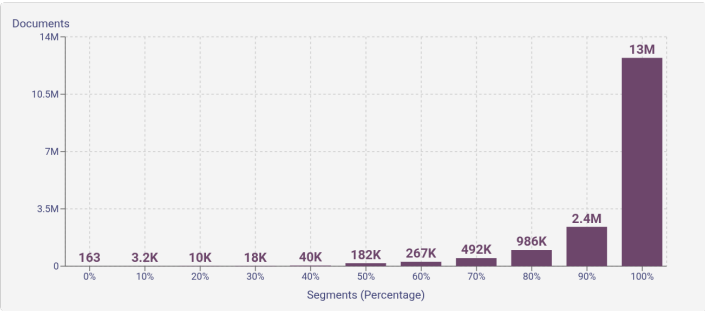


Language Distribution

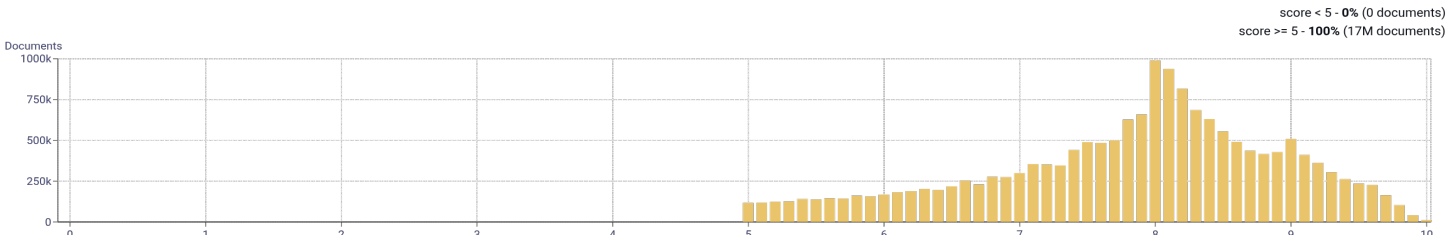
Number of segments in the Hebrew (he) corpus



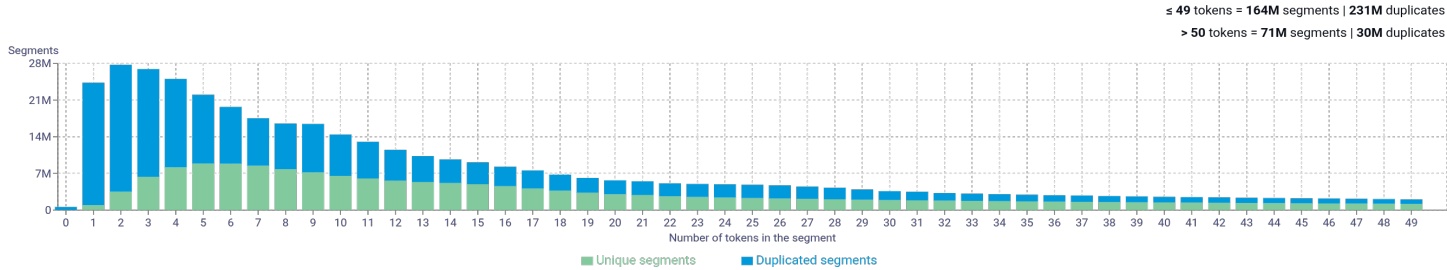
Percentage of segments in Hebrew (he) inside documents



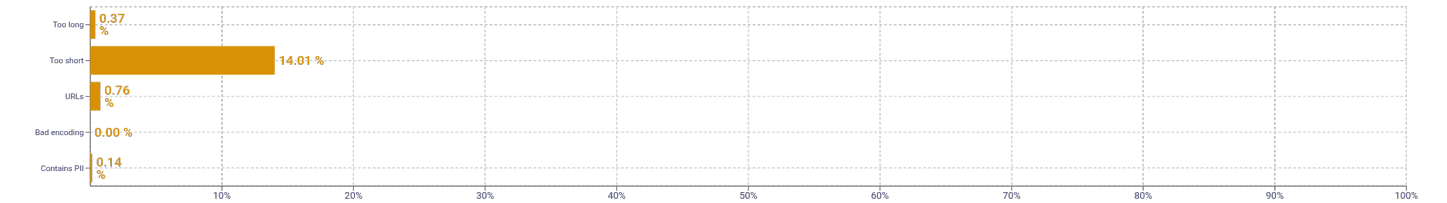
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

Size	n-grams
1	19205192 כ 16477791 אוד 15904810 כד 14865813 ה 13628560 וז
2	3826078 קוד מקור 3820120 עריכת קוד 2177707 בית המשפט 2029513 תל אביב 1455913 לאחר מכן
3	955376 בסופו של דבר 288632 בית המשפט העליון 270592 בית המשפט המחוזי 265904 בדיכום הבאים לאחר 257189 הגב לתענה וז
4	242947 בדיכום הבאים לאחר פסיקה 120056 המצגת להלן הינה אישית 120056 הרצת המצגת להלן הינה 117719 אישית ומבוססת על הניסיון 117718 האישיית של המשתמש שכתב
5	120056 המצגת להלן הינה אישית ומבוססת 120056 הרצת המצגת להלן הינה אישית 117719 הינה אישית ומבוססת על הניסיון 117719 אישית ומבוססת על הניסיון והחוויה 117717 הניסיון והחוויה האישיית של המשתמש

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 *number of 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>, , , 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				