HPLT Analytics report

HPLTAnalytics

General overview

Corpus	Date	Language
hplt-v3-apc_Arab	10/3/2025	Levantine Arabic

Volumes

Docs	Segments	Unique segments	Duplication ratio	Tokens	Characters	Size
253	3,494	3,044 (87.12 %)	12.88%	47K	234,940	413.79 KB

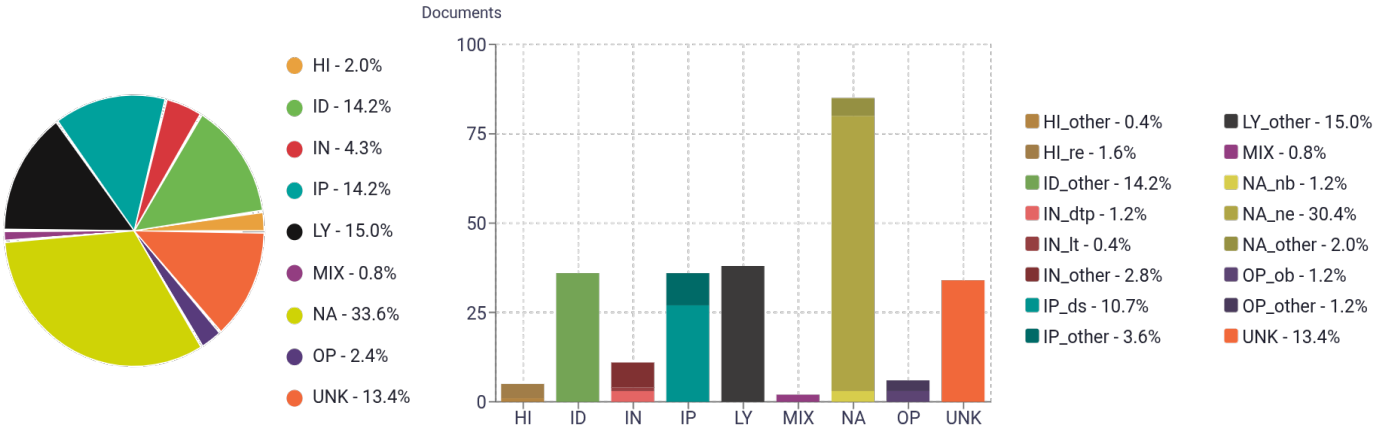
Top 10 domains

Domain	Docs	% of total
palestinerememb...	18	7.11%
art-en.com	11	4.35%
hellooha.com	9	3.56%
babycenter.com	7	2.77%
gidny.com	6	2.37%
fxsolve.com	5	1.98%
bukja.net	5	1.98%
wordpress.com	3	1.19%
nawaret.com	3	1.19%
ilcode.com	3	1.19%

Top 10 TLDs

Domain	Docs	% of total
com	187	73.91%
net	26	10.28%
org	13	5.14%
ps	3	1.19%
me	2	0.79%
co.il	2	0.79%
ws	1	0.40%
tv	1	0.40%
today	1	0.40%
tech	1	0.40%

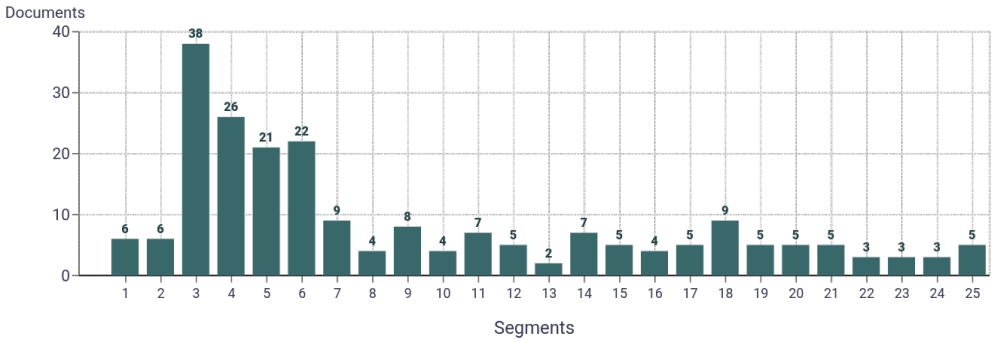
Register labels



MT:1.2% | 3 Documents

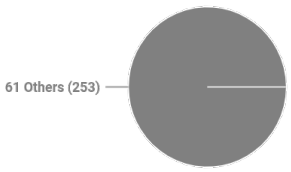
Documents size (in segments)

≤ 25 segments **85.77%** (217 documents)
> 25 segments **14.23%** (36 documents)



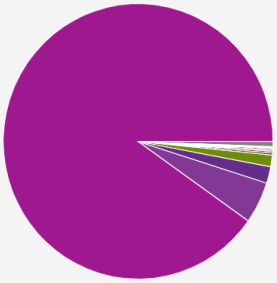
Document collections

CC = 71.94%
IA = 28.06%



Language Distribution

Number of segments in the Levantine Arabic corpus

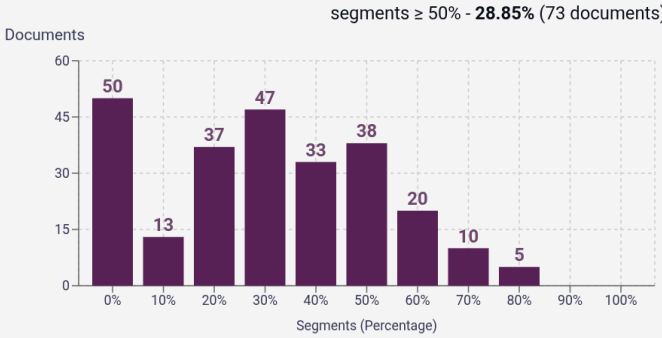


- Arabic - 3.2K **(90.2%)**
- Egyptian Arabic - 171 **(4.9%)**
- English - 68 **(1.9%)**
- Persian - 48 **(1.4%)**
- Italian - 8 **(0.2%)**
- French - 7 **(0.2%)**
- Romanian - 6 **(0.2%)**
- Urdu - 5 **(0.1%)**
- Sindhi - 5 **(0.1%)**
- Greek - 3 **(0.1%)**
- 15 Others - 21 **(0.6%)**

*Levantine Arabic identification might be inaccurate because it is not supported by FastSpell.

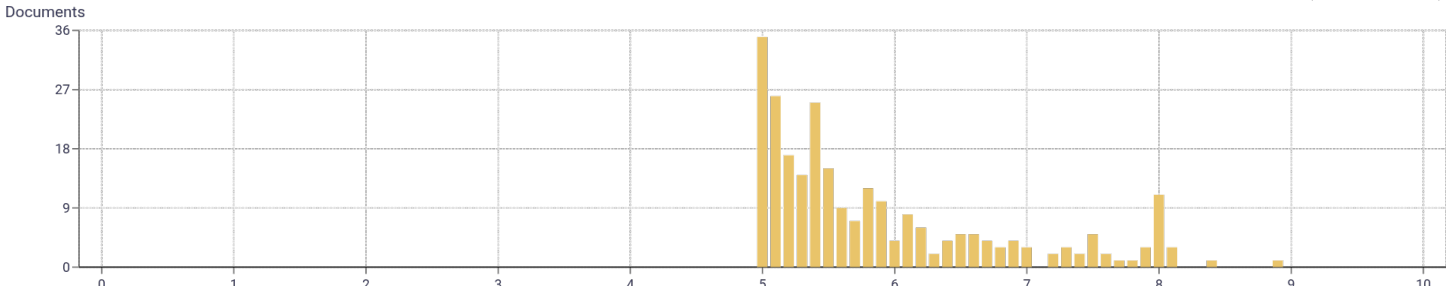
Percentage of segments in Levantine Arabic inside documents

segments < 50% - **71.15%** (180 documents)
segments ≥ 50% - **28.85%** (73 documents)



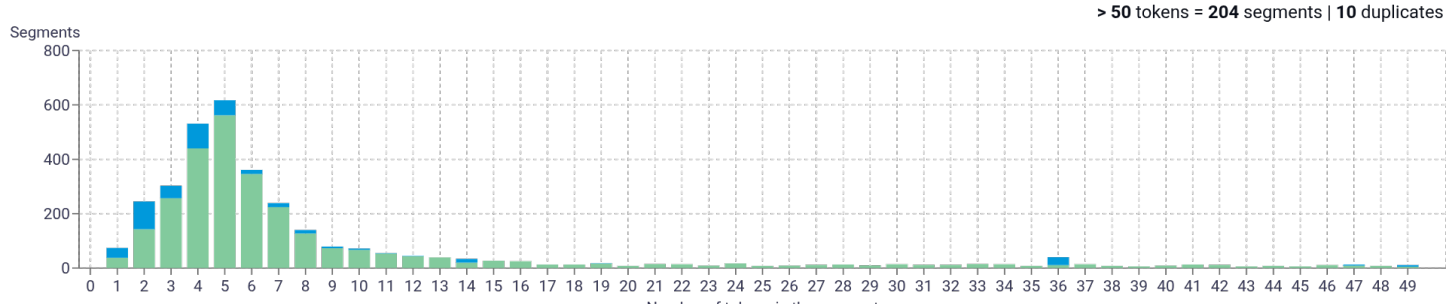
Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (253 documents)

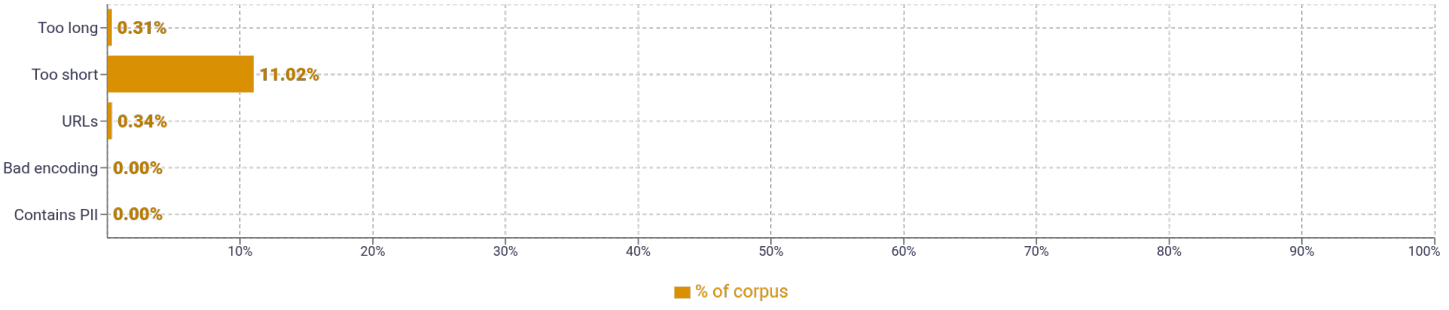


Segment length distribution by token

≤ 49 tokens = **3.3K** segments | **440** duplicates
> 50 tokens = **204** segments | **10** duplicates



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS
1	الله | 125 · عم | 118 · ان | 103 · خبر | 95 · علي | 94
2	خبر عاجل | 84 · تلميع سيارة | 67 · تقرير لجهاز | 51 · لجهاز منزلي | 44 · اجمل تقرير | 40
3	تقرير لجهاز منزلي | 44 · المنتدى لا تستطيع | 44 · اجمل تقرير لجهاز | 30 · بعمولة عبر شركة | 23 · التسويق بعمولة عبر | 23
4	اجمل تقرير لجهاز منزلي | 27 · التسويق بعمولة عبر شركة | 23 · مشاركاتك في هذا المنتدى | 22 · يعبر عن الرأي الشخصي | 18 · مسؤولة عن هذه الآراء | 18
5	يعبر عن الرأي الشخصي لمؤلفها | 18 · وفلسطين في الذاكرة غير مسؤولة | 18 · صحة المعلومات ولكن لا تضمن | 18 · تحاول فلسطين في الذاكرة التدقيق | 18 · المعلومات ولكن لا تضمن صحتها | 18

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				