HPLT Analytics report

General overview

Corpus	Date	Language
hplt-v3-mkd_Cyrl	9/24/2025	Macedonian

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
6,791,451	97,567,523	58,291,826 (59.75 %)	3B	16,333,470,805	27.26 GB

Top 10 domains

Domain	Docs	% of total
daily.mk	185K	2.72%
makfax.com.mk	110K	1.62%
slobodnaevropa.mk	93K	1.37%
netpress.com.mk	88K	1.29%
novamakedonija....	86K	1.26%
airbnb.com	72K	1.06%
republika.mk	71K	1.04%
kurir.mk	70K	1.03%
wikipedia.org	69K	1.02%
mkd.mk	59K	0.86%

Top 10 TLDs

Domain	Docs	% of total
mk	3.9M	57.16%
com	1.1M	16.11%
com.mk	778K	11.46%
org	414K	6.09%
org.mk	108K	1.60%
gov.mk	90K	1.33%
net	49K	0.71%
edu.mk	47K	0.69%
co.uk	39K	0.57%
net.tr	34K	0.50%

Register labels

- HI - 2.3%
- ID - 0.3%
- IN - 6.2%
- IP - 12.3%
- LY - 0.1%
- MIX - 2.5%
- NA - 60.3%
- OP - 5.2%
- SP - 0.7%
- UNK - 10.1%

- HI_other - 1.1%
- HI_re - 1.2%
- ID_other - 0.3%
- IN_dtp - 2.0%
- IN_en - 1.2%
- IN_fi - 0.0%
- IN_lt - 0.4%
- IN_other - 2.6%
- IN_ra - 0.0%
- IP_ds - 10.8%
- IP_other - 1.5%
- LY_other - 0.1%
- MIX - 2.5%
- NA_nb - 1.1%
- NA_ne - 46.8%
- NA_other - 3.6%
- NA_sr - 8.8%
- OP_av - 1.7%
- OP_ob - 1.2%
- OP_other - 1.2%
- OP_rs - 0.7%
- OP_rv - 0.4%
- SP_it - 0.5%
- SP_other - 0.2%
- UNK - 10.1%
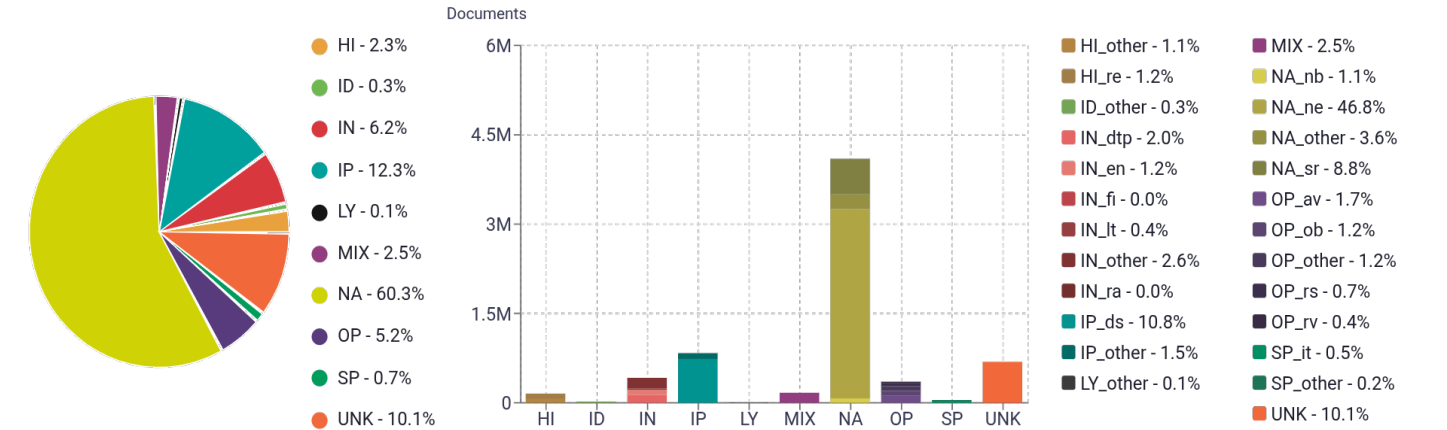
**MT**:8.9% | 605K Documents

Documents size (in segments) ⓘ

≤ 25 segments **86.66%** (5.9M documents)
> 25 segments **13.34%** (906K documents)

Document collections

CC = 88.64%
IA = 11.36%

67 Others (6.8M)
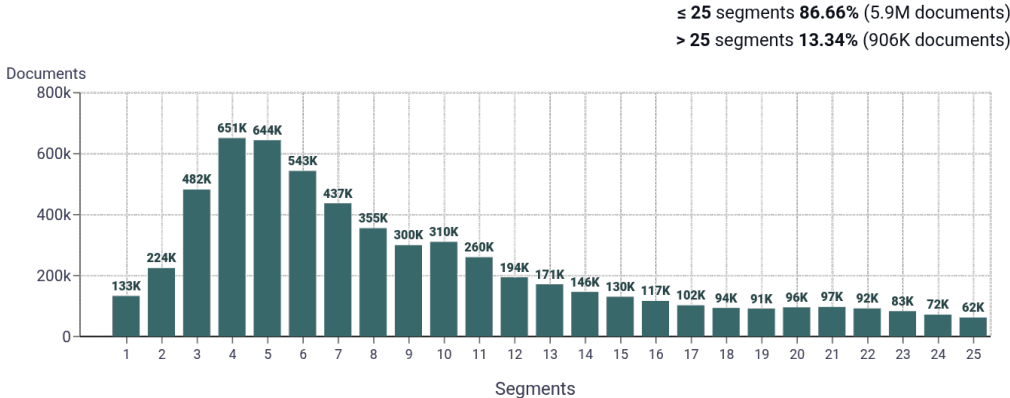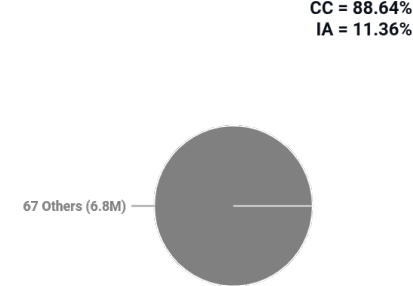
Language Distribution

Number of segments in the Macedonian corpus

- Macedonian - 87M **(88.8%)**
- English - 2.5M **(2.5%)**
- Serbian - 1.8M **(1.8%)**
- Italian - 1.6M **(1.6%)**
- Russian - 1.3M **(1.3%)**
- French - 771K **(0.8%)**
- Bulgarian - 747K **(0.8%)**
- Ukrainian - 512K **(0.5%)**
- German - 372K **(0.4%)**
- Portuguese - 182K **(0.2%)**
- 164 Others - 1.3M **(1.3%)**
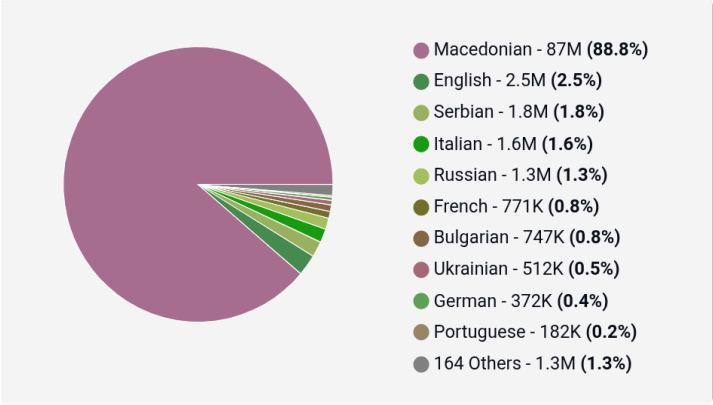
Percentage of segments in Macedonian inside documents
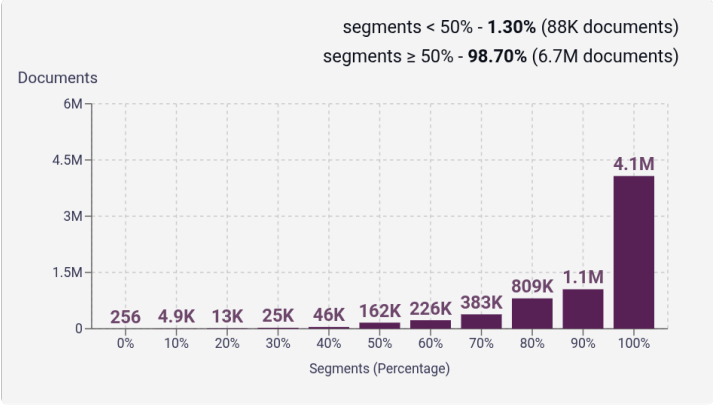
segments < 50% - **1.30%** (88K documents)
segments ≥ 50% - **98.70%** (6.7M documents)

Documents

0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
256	4.9K	13K	25K	46K	162K	226K	383K	809K	1.1M	4.1M

Segments (Percentage)

Distribution of documents by document score

score < 5 - **0%** (0 documents)
score ≥ 5 - **100%** (6.8M documents)

Segment length distribution by token

≤ **49** tokens = **81M** segments | **37M** duplicates
> **50** tokens = **17M** segments | **2.8M** duplicates

Segment noise distribution

- Too long — **0.84%**
- Too short — **11.71%**
- URLs — **0.92%**
- Bad encoding — **0.00%**
- Contains PII — **0.20%**

■ % of corpus
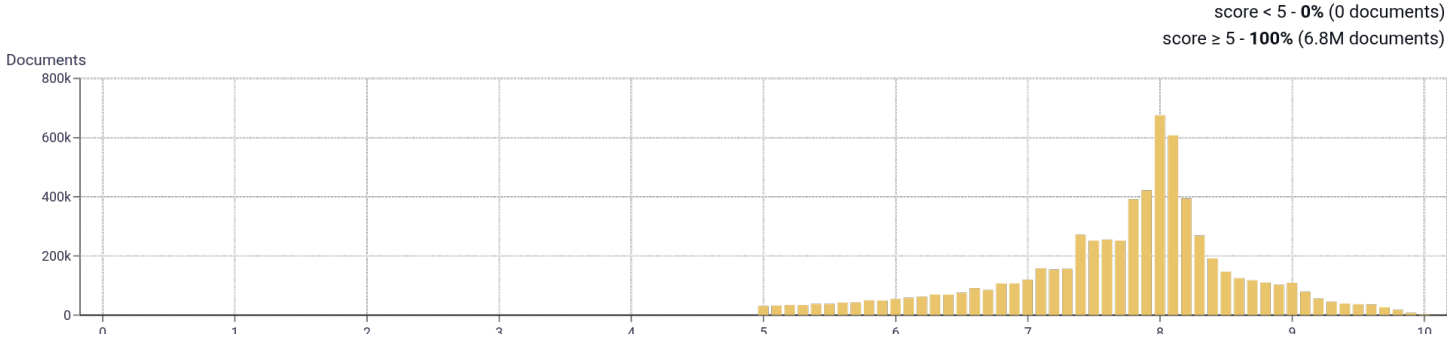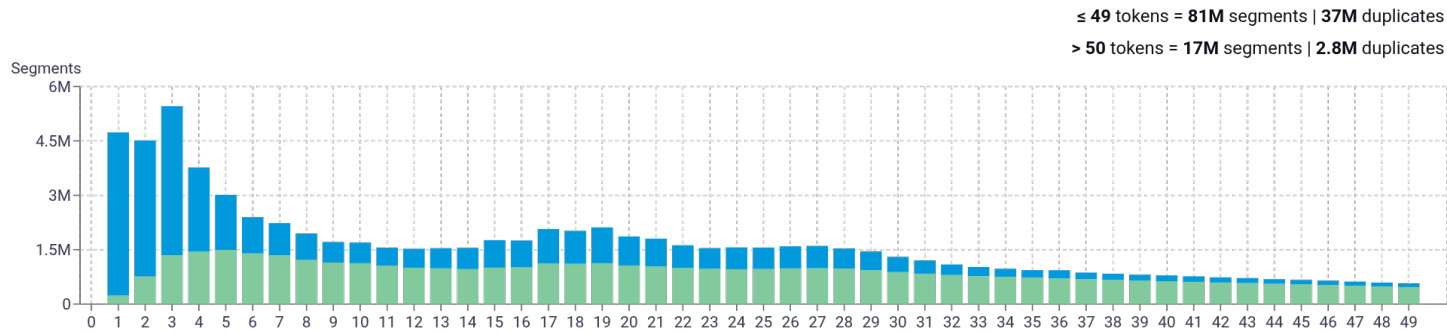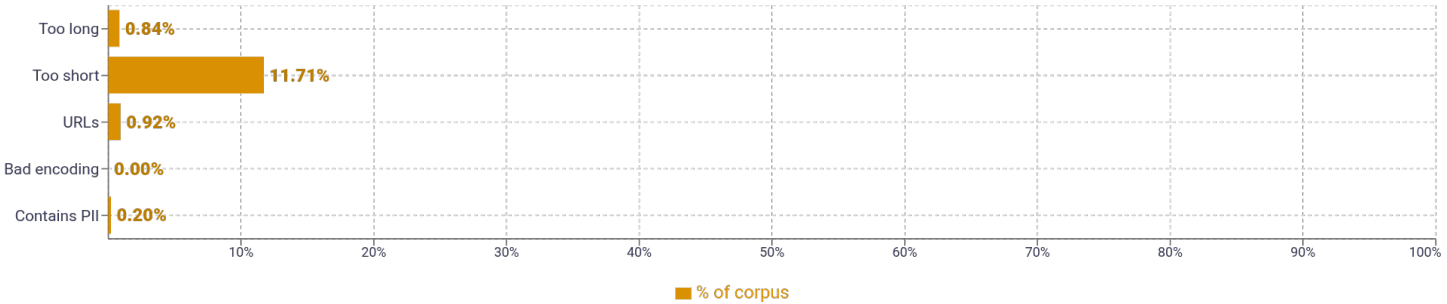
Frequent n-grams

SIZE	N-GRAMS	
1	година | 6,089,057 македонија | 4,121,460 време | 3,603,079 години | 3,008,923 дел | 2,789,299	⧉
2	ве молиме | 815,878 република македонија | 644,988 станува збор | 524,223 северна македонија | 499,523 милиони евра | 437,952	⧉
3	република северна македонија | 221,843 лигата на шампионите | 136,497 можат да бидат | 134,569 ве молиме контактирајте | 131,144 министерството за здравство | 115,663	⧉
4	министерството за внатрешни работи | 88,892 министерот за надворешни работи | 71,335 министер за надворешни работи | 66,752 are closed for this | 66,507 comments are closed for | 66,504	⧉
5	comments are closed for this | 66,504 are closed for this post | 66,486 изнајмување на сместувања за одмор | 50,024 view this post on instagram | 40,348 сноси никаква одговорност за коментарите | 37,881	⧉

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				