

General overview

Corpus	Date	Language
hplt-v3-deu_Latn	9/19/2025	German (de)

Volumes

Docs	Segments	Unique segments	Tokens	Characters	Size
645,363,459	14,374,232,816	7,542,997,070 (52.48 %)	396B	2,411,334,286,965	2.23 TB

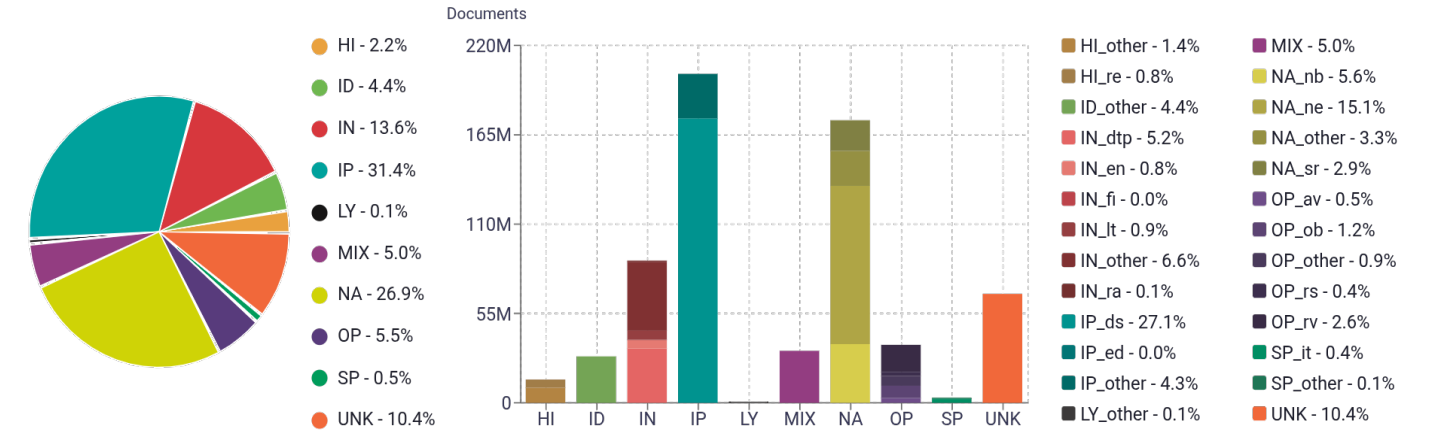
Top 10 domains

Domain	Docs	% of total
blogspot.com	5.5M	0.85%
wordpress.com	4.7M	0.73%
gutefrage.net	1.9M	0.29%
wikipedia.org	1.6M	0.25%
blogspot.de	1.2M	0.19%
docplayer.org	1.2M	0.19%
webwiki.de	1.1M	0.17%
t-online.de	1.1M	0.17%
derstandard.at	1M	0.16%
rp-online.de	939K	0.15%

Top 10 TLDs

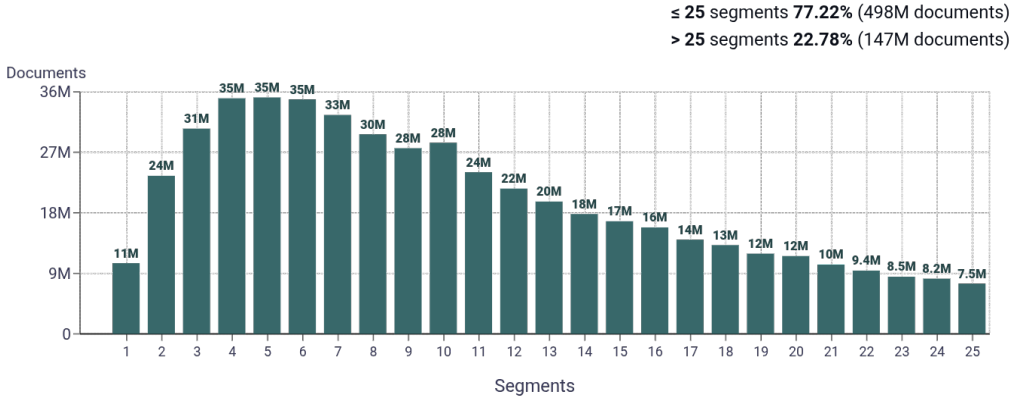
Domain	Docs	% of total
de	369M	57.24%
com	117M	18.07%
at	36M	5.52%
ch	31M	4.73%
net	17M	2.62%
org	15M	2.30%
eu	11M	1.65%
info	7.4M	1.15%
nl	2.5M	0.38%
biz	2M	0.31%

Register labels

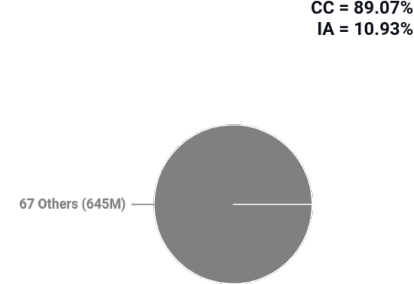


MT:7.4% | 48M Documents

Documents size (in segments) ⓘ

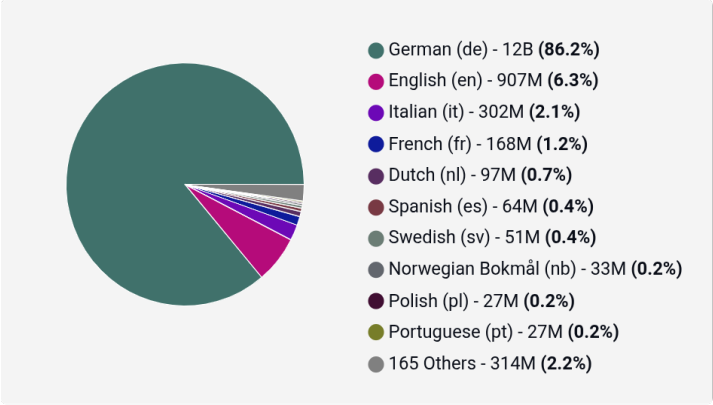


Document collections

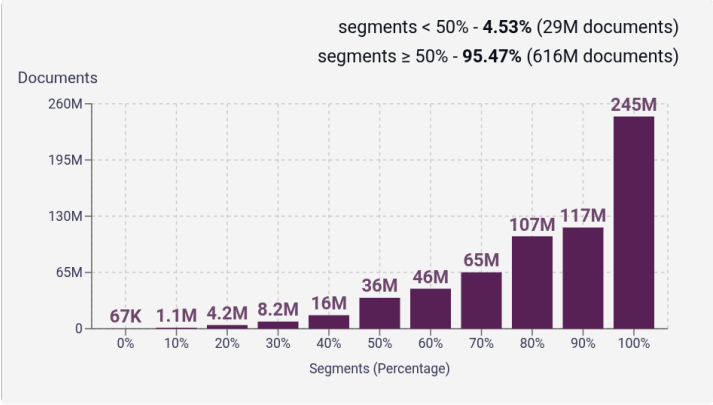


Language Distribution

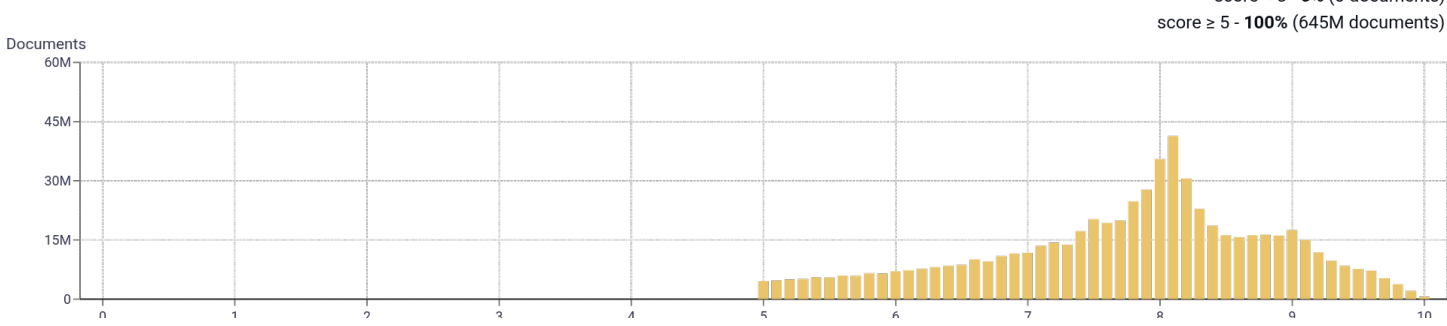
Number of segments in the German (de) corpus



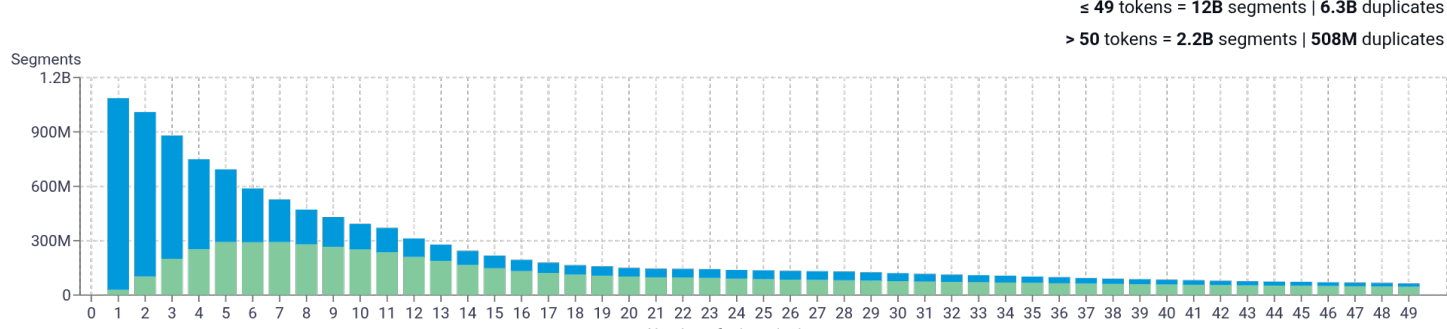
Percentage of segments in German (de) inside documents



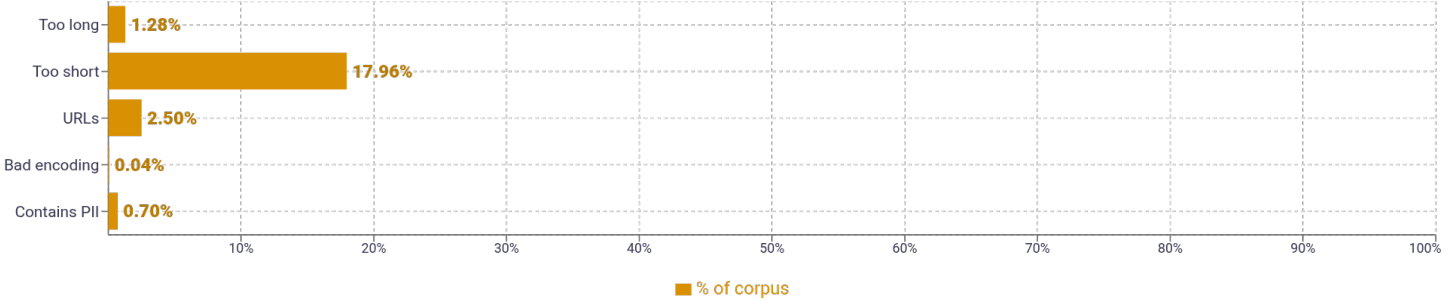
Distribution of documents by document score



Segment length distribution by token



Segment noise distribution



Frequent n-grams

SIZE	N-GRAMS	
1	<div>gibt   406,174,153</div> <div>immer   392,574,787</div> <div>schon   380,213,158</div> <div>the   328,035,679</div> <div>ab   313,033,903</div>	
2	<div>online casino   48,725,907</div> <div>darüber hinaus   40,240,657</div> <div>of the   39,058,335</div> <div>z. b.   32,044,604</div> <div>weitere informationen   28,875,123</div>	
3	<div>book of ra   24,774,949</div> <div>lieb und wert   17,190,365</div> <div>aufs hohe ross   14,636,566</div> <div>hohe ross setzen   14,633,826</div> <div>fahrenden zug aufspringen   10,019,862</div>	
4	<div>aufs hohe ross setzen   14,438,933</div> <div>auge auf etwas werfen   8,533,029</div> <div>vertreterin des schönen geschlechts   8,353,049</div> <div>neue sau durchs dorf   7,829,654</div> <div>sau durchs dorf treiben   7,256,122</div>	
5	<div>neue sau durchs dorf treiben   7,157,571</div> <div>for the website to function   1,967,260</div> <div>schreiben sie mir in pm   1,715,307</div> <div>e-mail-adresse ist vor spambots geschützt   1,606,494</div> <div>verstehen sich inklusive der gesetzlichen   1,325,593</div>	

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 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>, <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				