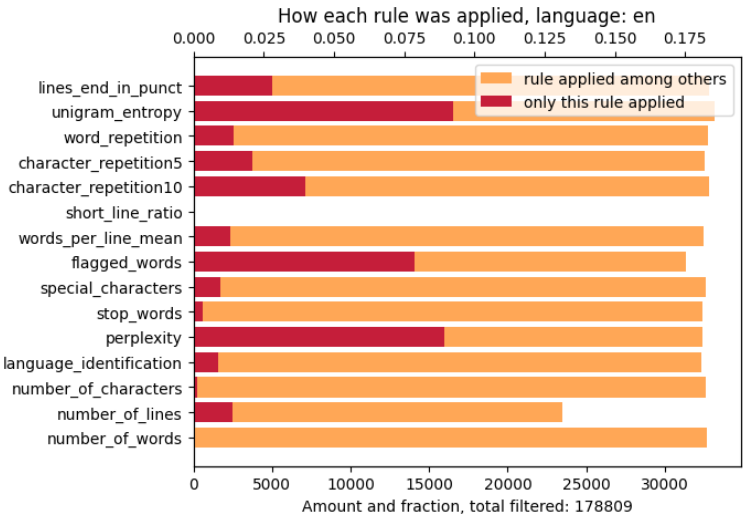
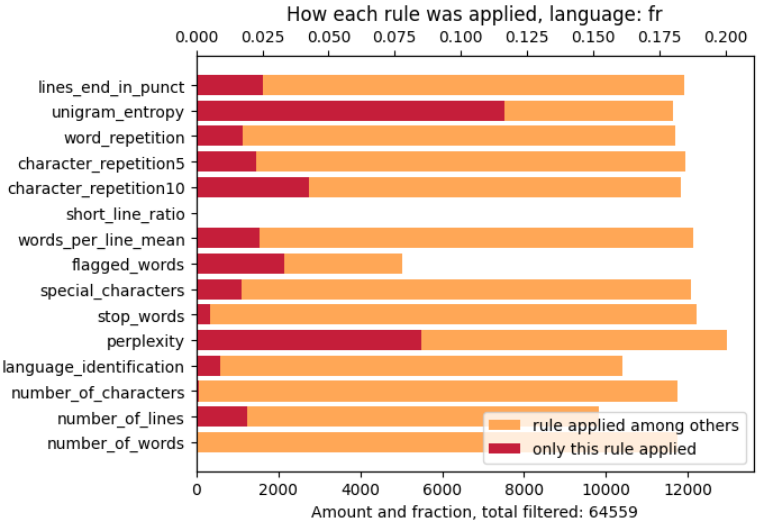
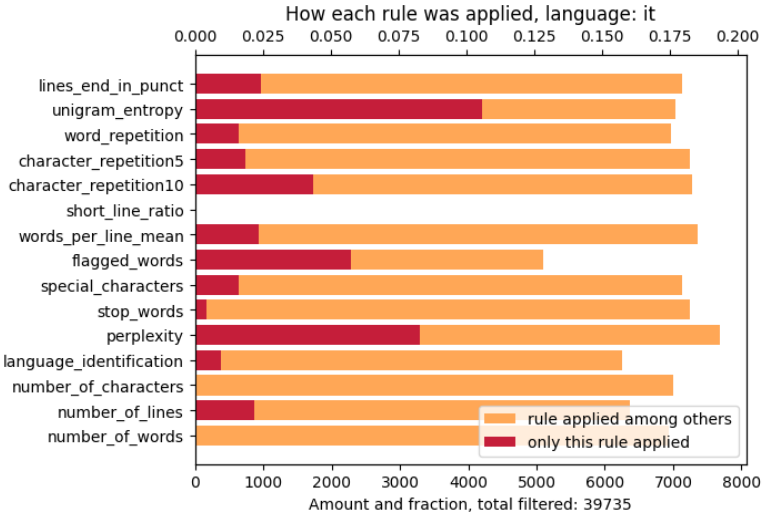
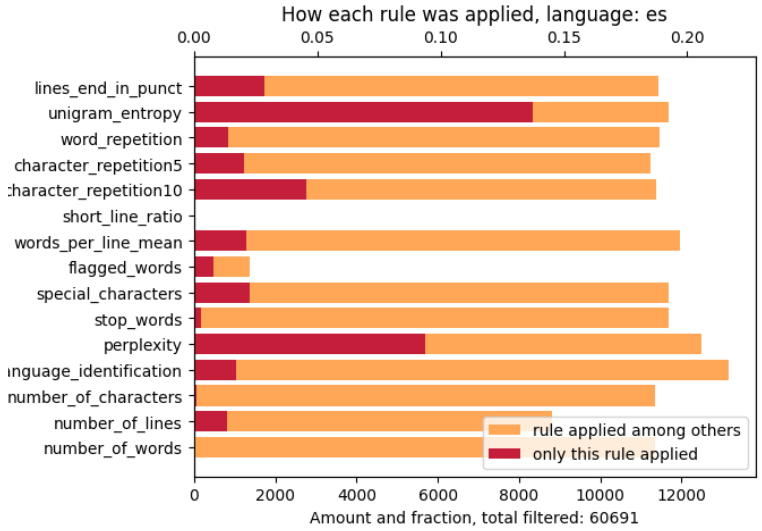
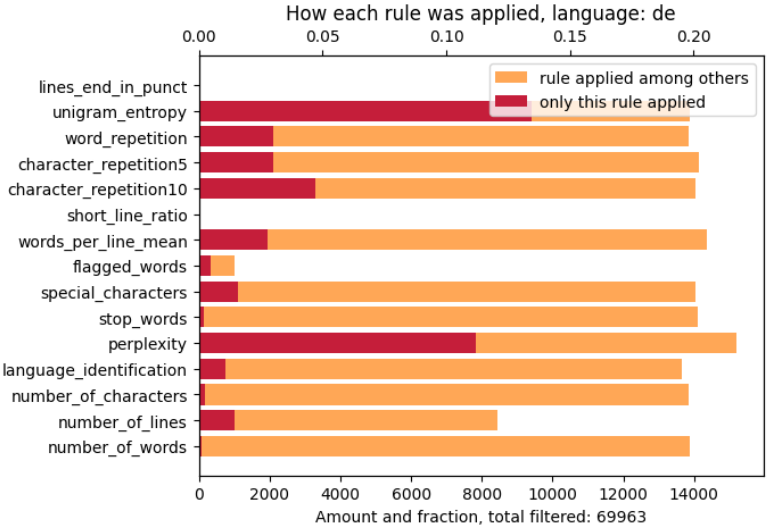
Filter rule analysis, N = 25 (0.5%), English N = 10 (0.2%) 30.11.2023







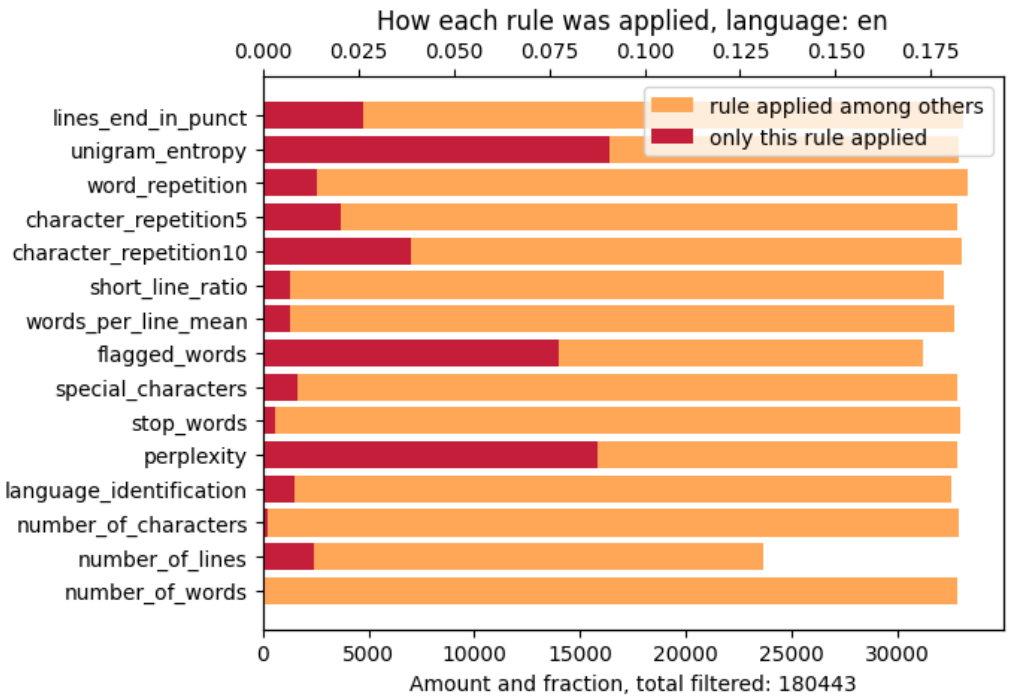




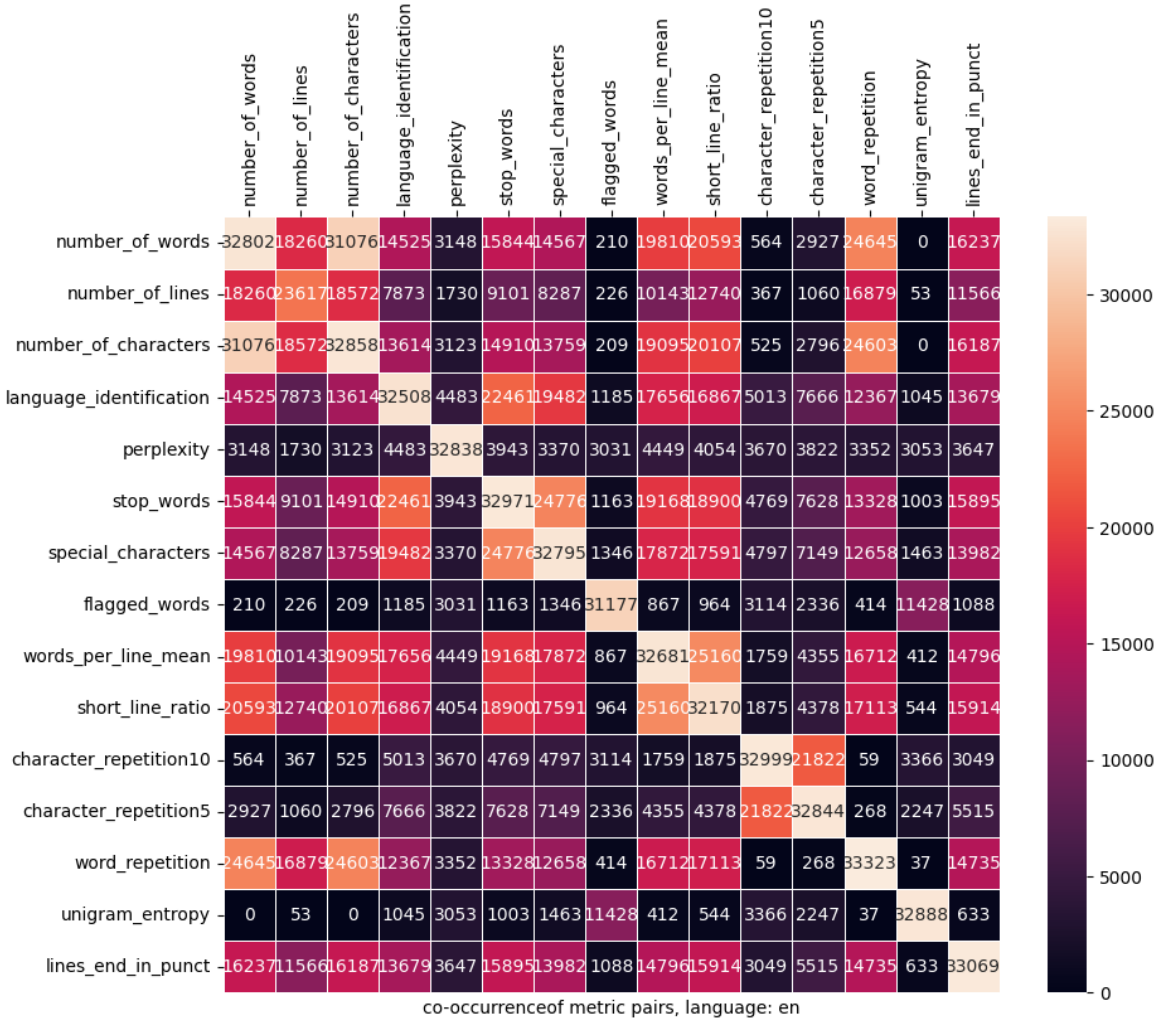
Fraction remaining after filtering:

- English: 0.45714285714285713
- French: 0.4544478903470597
- Italian: 0.44316764528650904
- Spanish: 0.4710931780946073
- German: 0.5000178659482173

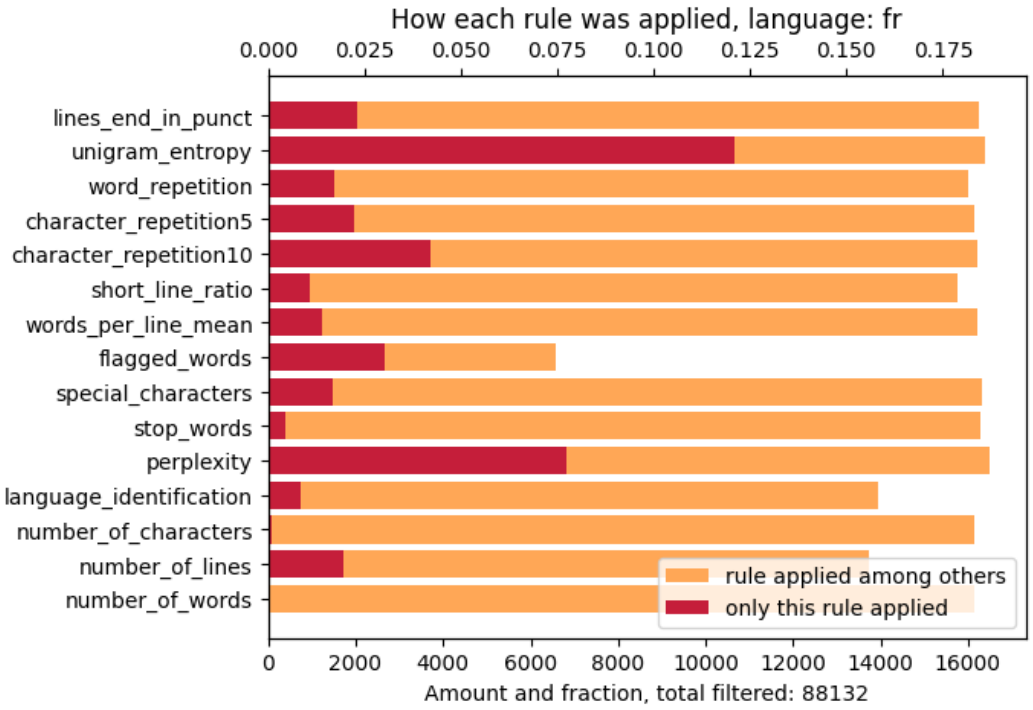
Short line ratio as approximated 100 chars + UT1 11.12.2023

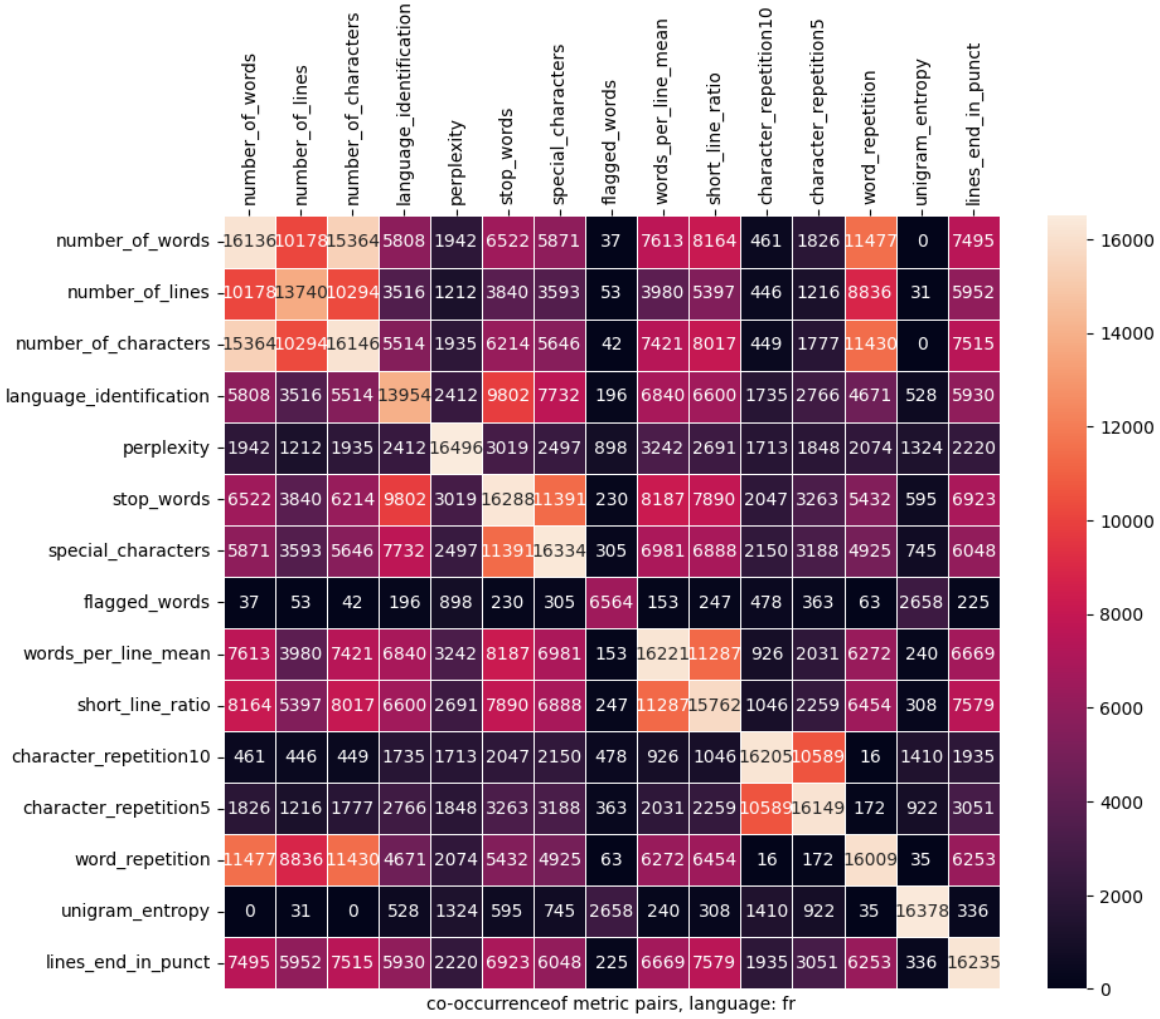


En	metrics keep	metrics discard
ut1 not flagged	148479	180279
ut1 flagged	68	164
total	148547	180443
%	45.15	54.85



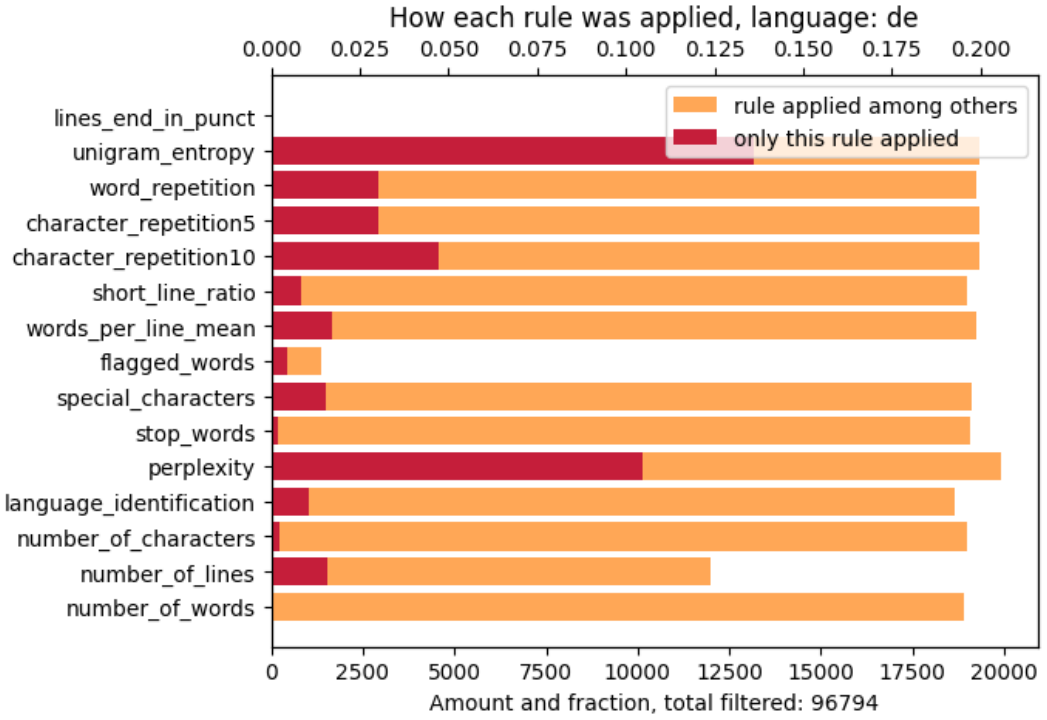
Short line ratio as approximated 100 chars + UT1 11.12.2023



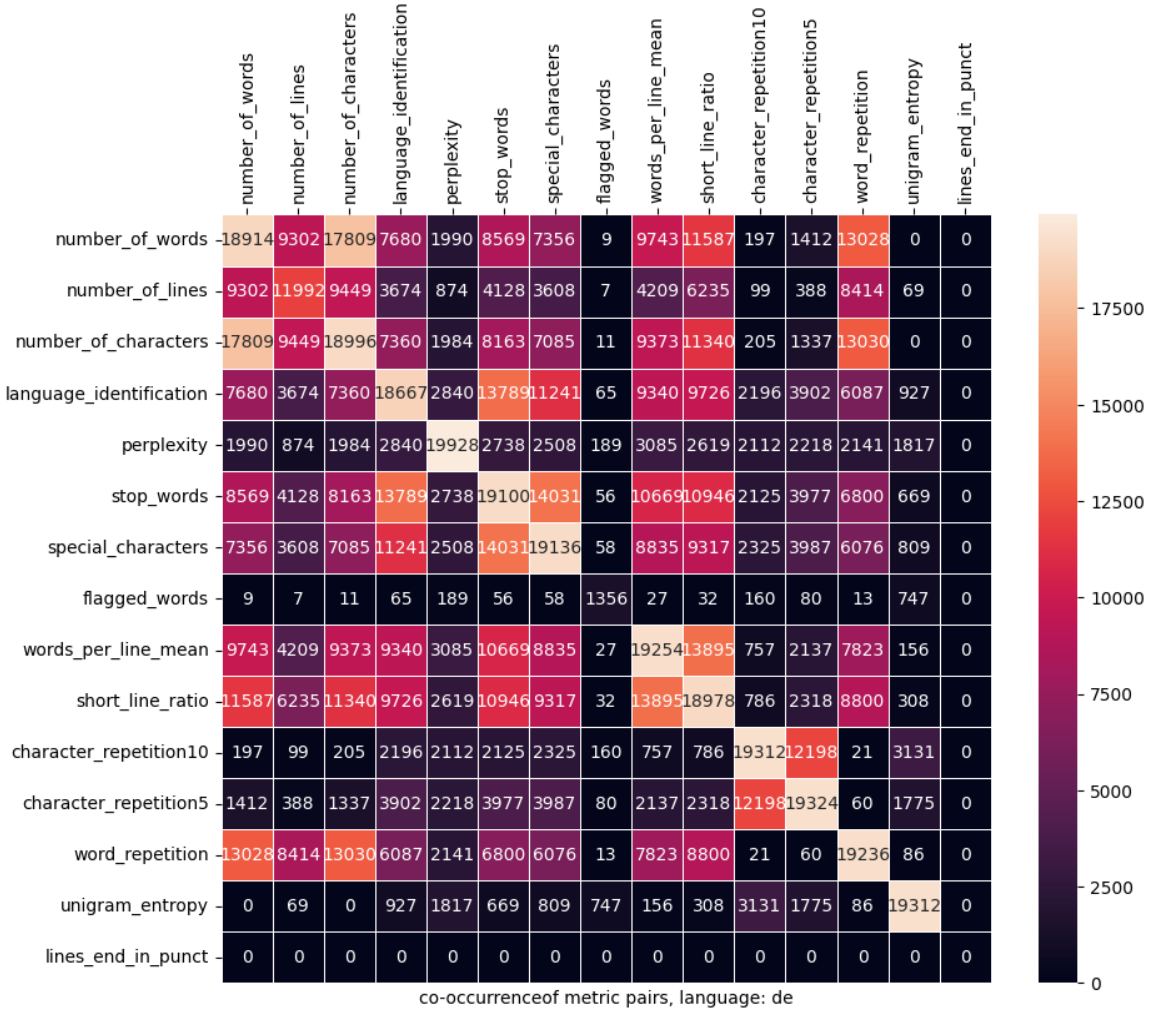


Fr	metrics keep	metrics discard
ut1 not flagged	73374	88076
ut1 flagged	37	56
total	73411	88132
%	45.44	54.66

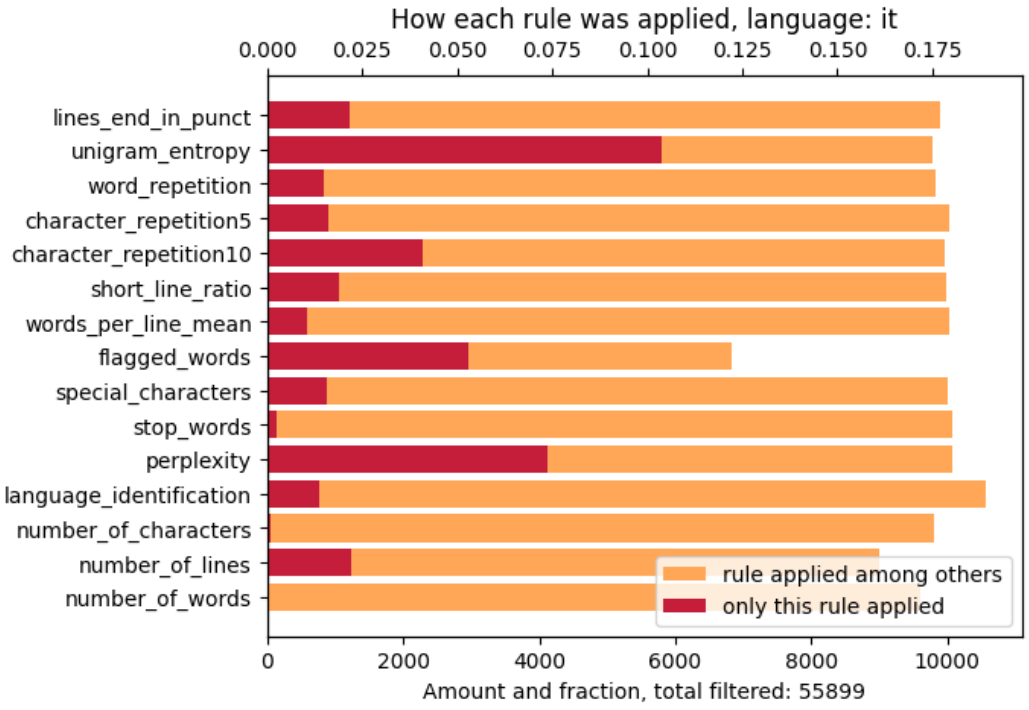
Short line ratio as approximated 100 chars + UT1 11.12.2023

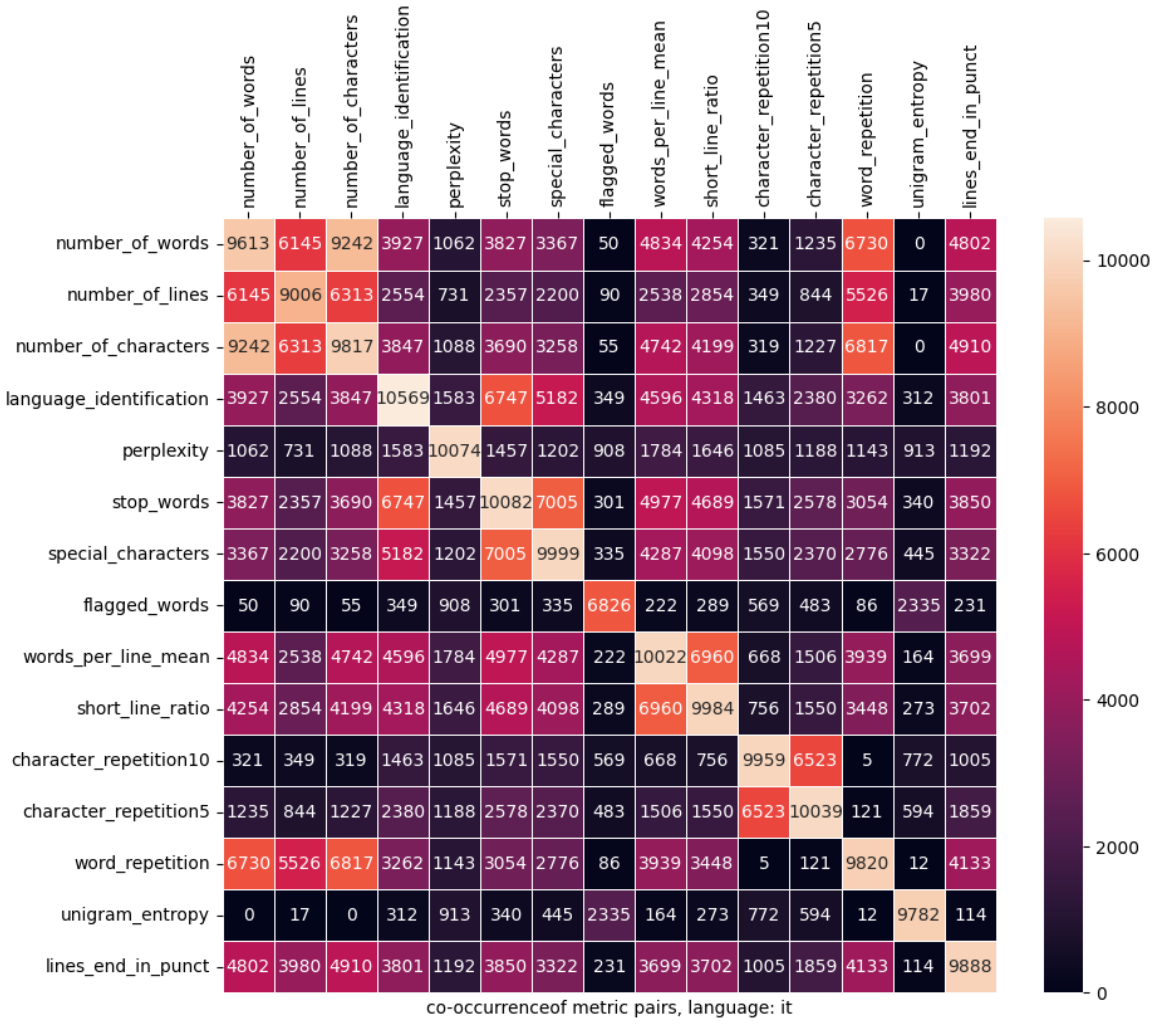


De	metrics keep	metrics discard
ut1 not flagged	96498	96766
ut1 flagged	18	28
total	96516	96794
%	49.93	50.07



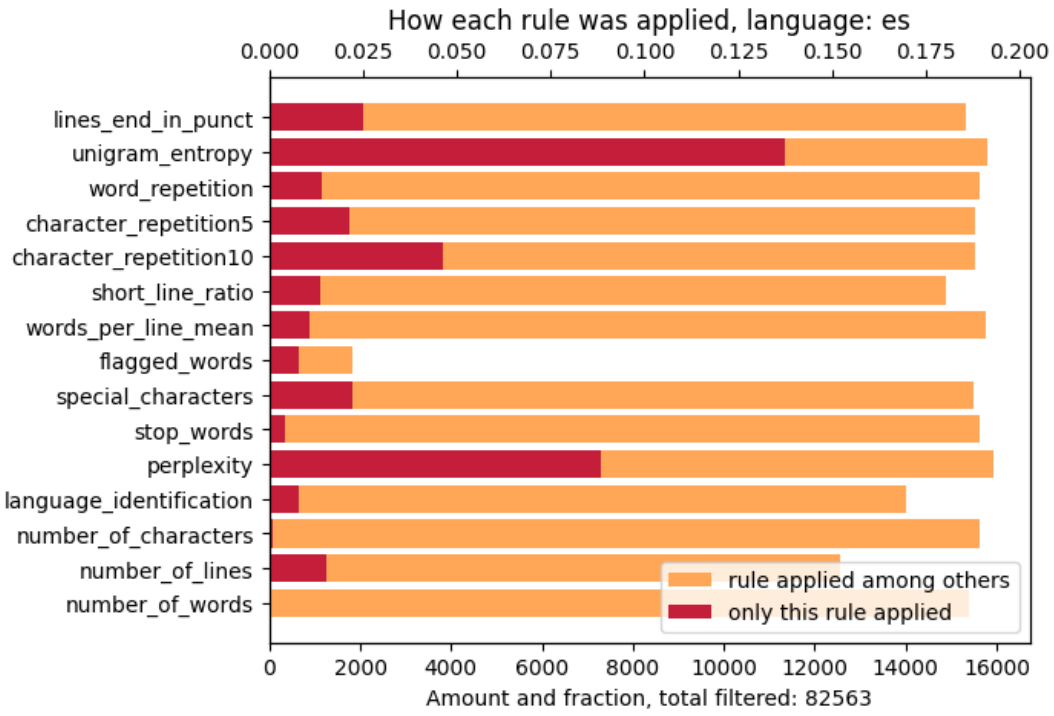
Short line ratio as approximated 100 chars + UT1 11.12.2023

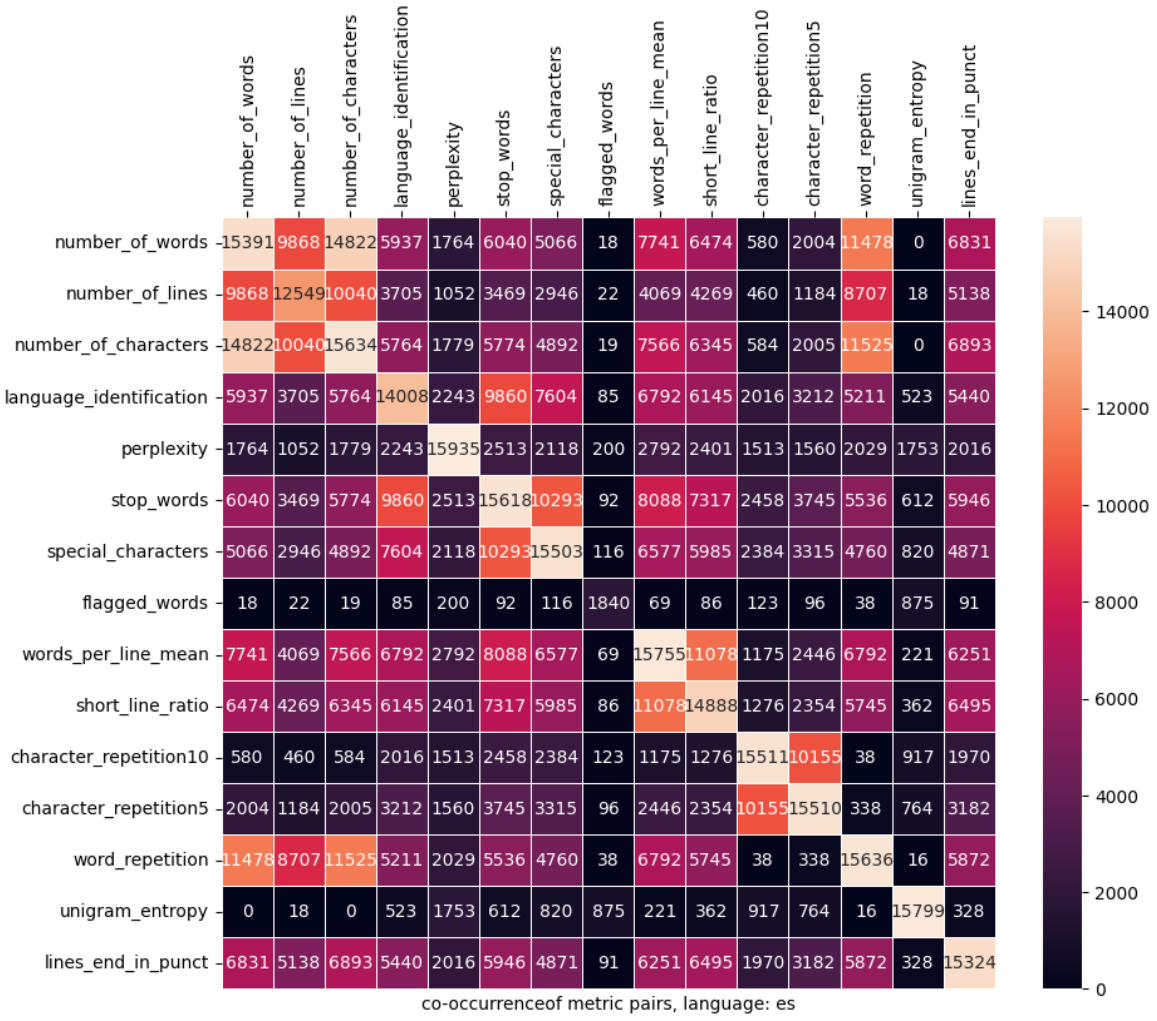




lt	metrics keep	metrics discard
ut1 not flagged	42973	55888
ut1 flagged	2	11
total	42975	55899
%	43.46	56.54

Short line ratio as approximated 100 chars + UT1 11.12.2023





Es	metrics keep	metrics discard
ut1 not flagged	74241	82532
ut1 flagged	19	31
total	74260	82563
%	47.35	52.65

FR	#word	#line	#char	lang. ident	perpl exity	stop words	spec char	flagged words	WPL mean	SLR (<100)	char rep 10	char rep 5	word rep	entro py	punct	total =56
adult porn	4	9	4	1	11	2	1	23	3	3	7	4	3	13	5	43
chat														1		1
adult phishing porn											2	1				2
dating	4	4	4						5	4			4			5
gambling	1	1	1	1		2	2			1		1	1		1	3
mixed adult										1				1		2

*one error with loading the value of ut1 (?)

EN	#word	#line	#char	lang. ident	perpl exity	stop words	spec char	flagged words	WPL mean	SLR (<100)	char rep 10	char rep 5	word rep	entro py	punct	total =163*
adult porn	20	9	23	18	23	24	26	84	32	34	14	17	23	23	29	141
filehosting				1		1	1		1	1			1		1	1
phishing	3		3	2		3	1		1	1		1	2		2	4
violence								1			1		1			2
gambling														1		1
mixed adult	3		3	3	6	2	2	3	3			1	4	2	3	14

DE	#words	#lines	#char	lang. ident	perplexity	stop words	spec char	flagged word	WPL mean	SLR (<100)	char rep 10	char rep 5	word rep	entropy	punct	total =28
adult porn	6	5	7	5	4	4	4	9	5	5			5	4		18
violence					4			2			5	2		9		10

It and Es were too tiny, they had adult/porn and violence.