

# SIMSTYL – The Simulated Text Corpus



The simulated text corpus (SIMSTYL) is composed of 2 different text groups, where each group is characterized by a specific different style. In a single group there are two different texts of the same style.

This small corpus is represented in the following table and can be freely downloaded and used for a purpose of reproducibility.

The simulated SIMSTYL corpus

	Simulated Author style 1		Simulated Author style 2	
	Text 1A	Text 1B	Text 2A	Text 2B
Text	aaa red xyz red ggg ggg red hat drk gyy red hat red hat zzz red kkk hat dee red hat fir red hat hhh red hat ppp yyy red	ccc red ryz red gug gpg red hat erk gzy red hat red hat zoz red kok hat dce red hat irr red hat huh red hat upp zyy red	qbc qef xuz zef glg gxg red hat irk ghy red hat red hat azz agc dqf fir dze red hat fri red hat pqp red hat hhh rrr rte	ubc tef iuz olg gig des red hat ipk ghf red hat red hat zkz anc dnf for see red hat fir red hat sqp red hat phh xbc xef
Document size	30 words	30 words	30 words	30 words
Occurence of the word bigram "red hat"	6	6	6	6