Data set	Corpus	Sentences	Words
Train	PSC	521251	10684799
Validation	PSC	5265	107511
Evaluation	NCP	3491	71895