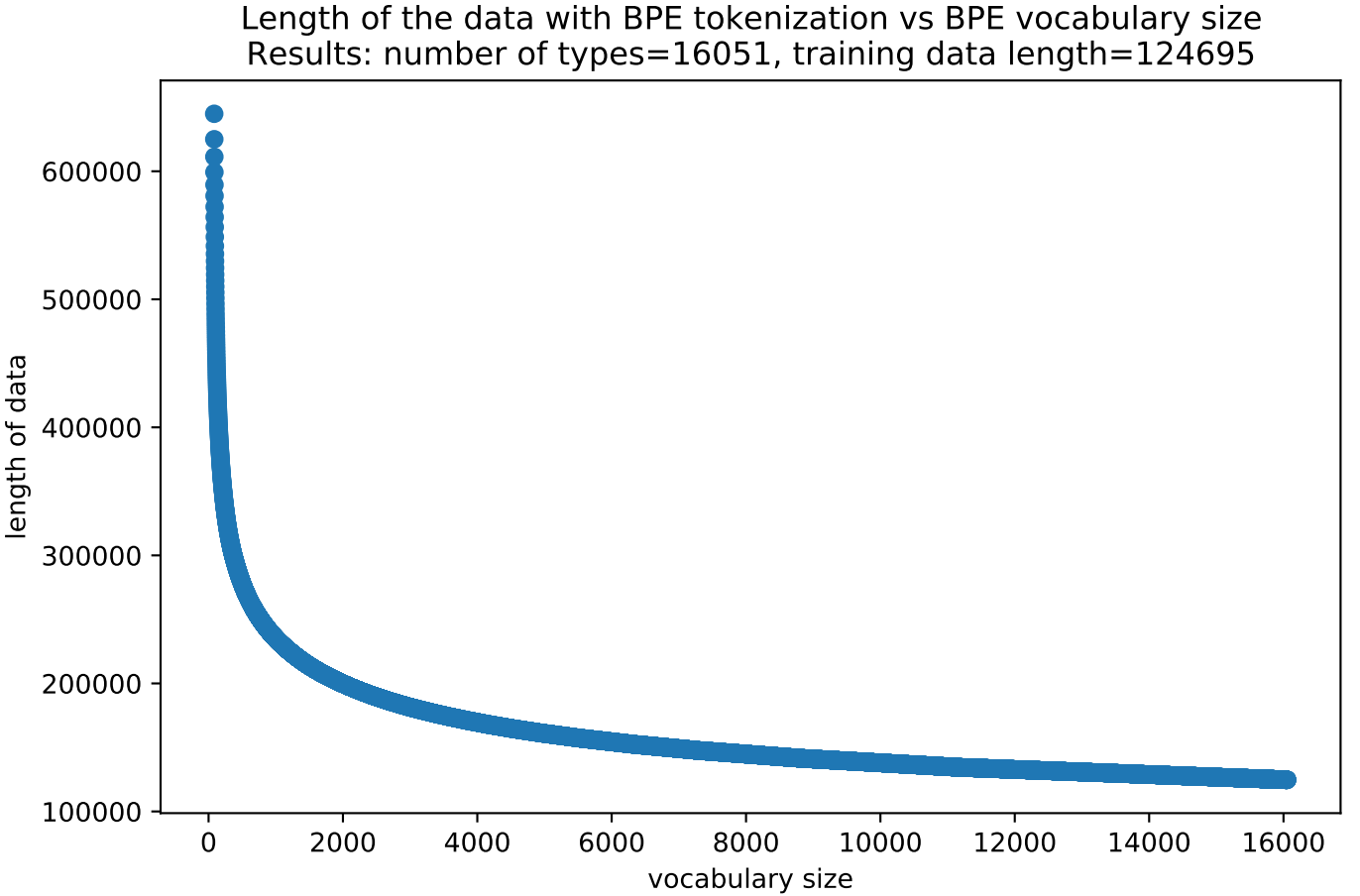
Length of the data with BPE tokenization vs BPE vocabulary size
Results: number of types=16051, training data length=124695
length of data
600000
500000
400000
300000
200000
100000
0
2000
4000
6000
8000
10000
12000
14000
16000
vocabulary size