

Length of the data with BPE tokenization vs BPE vocabulary size
Results: number of types=16051, training data length=124695

