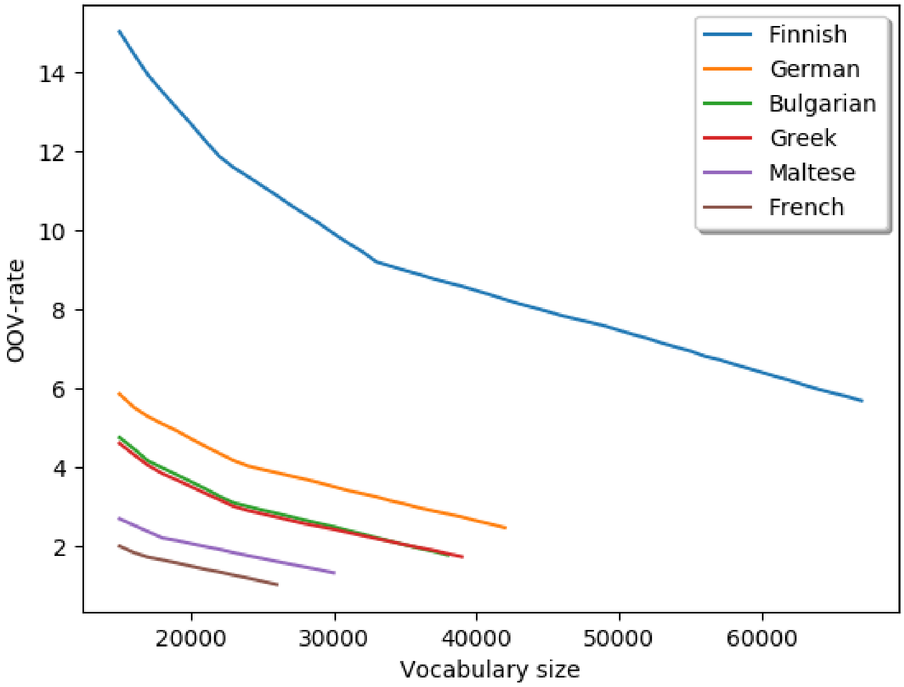
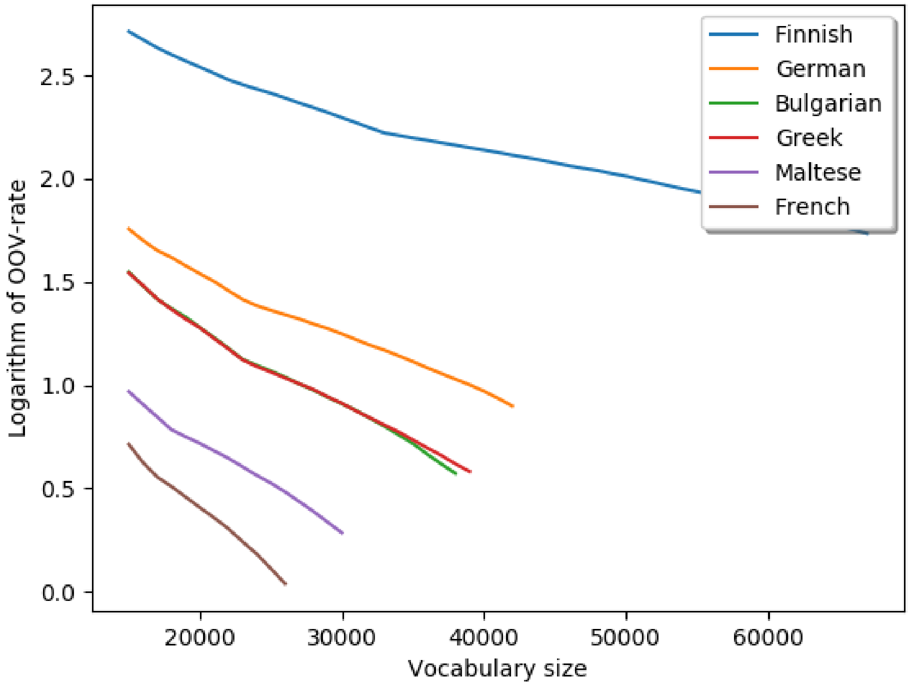
2.2 Out-of-vocabulary (OOV) Words

d) We handled six language corpora and calculated OOV-rates for Finnish, German, Bulgarian, Greek, Maltese and French on different partitions of the vocabulary.





Vocabulary size of these language corpora are different: Finnish – 67462, German – 42696, Bulgarian – 38319, Greek – 39397, Maltese – 30508 and French – 26651.

Morphologically rich languages have higher OOV rates for the same vocabulary size. We can conclude that Finnish is highly diversified language with the richest morphology. Other languages are less diversified in the following order: German, Bulgarian, Greek, Maltese and French. Moreover, Bulgarian and Greek have the graph lines which are very close to each other.

e) Morphologically rich languages such as Finnish and also German, Bulgarian and Greek are more challenging for statistical modelling than others. The reasons are:

   - Larger corpora are necessary for development of the LM.
   - The LM itself is heavier. More complex model can’t be constructed because the process will be time- and resources-consuming.
   - Morphological diversity introduces some kinds of ambiguities.
   - Large amount of out-of-vocabulary words.