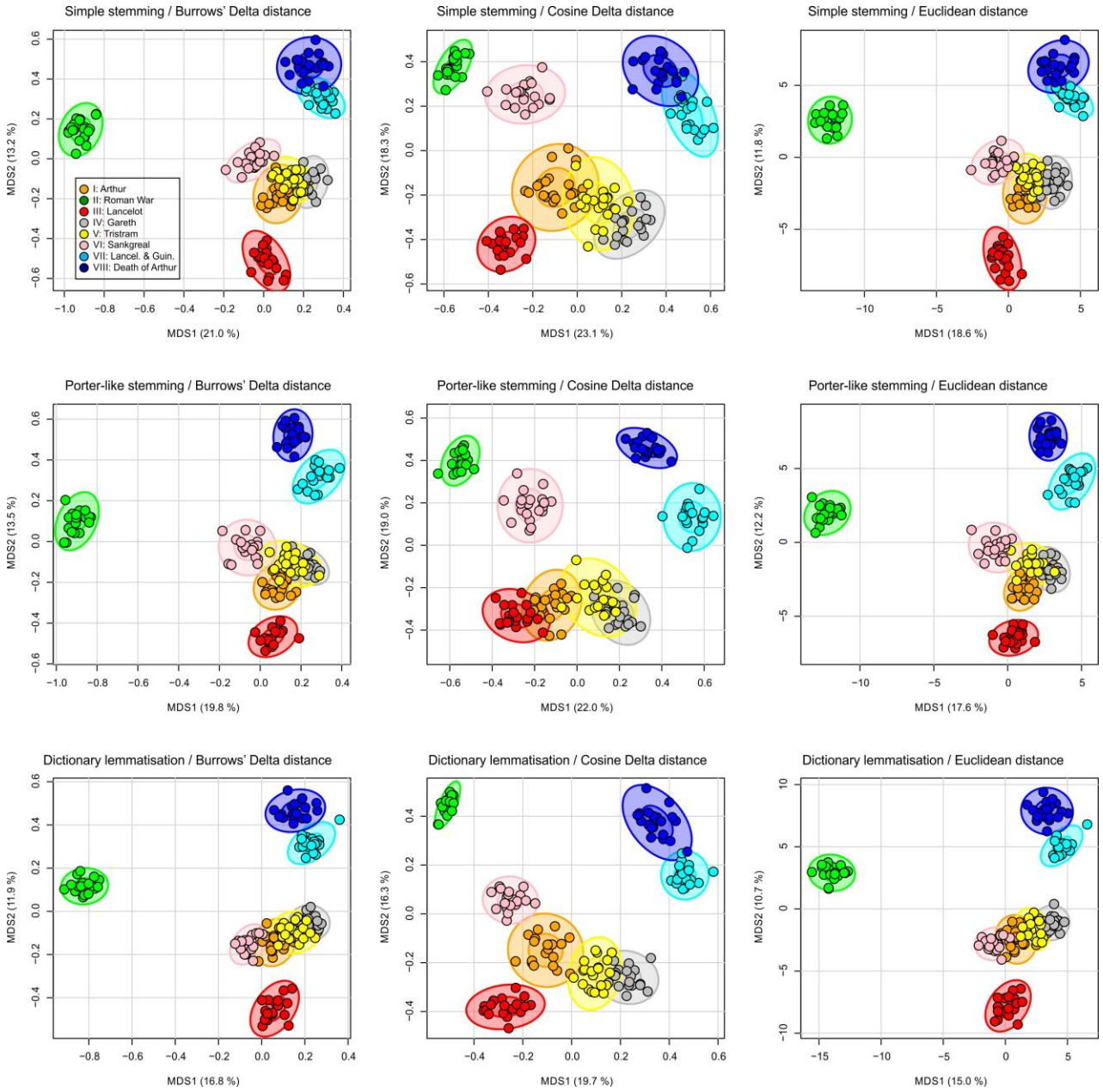

Supplementary Figure S1: The first two MDS dimensions of three different types of tokenization ("simple stemming", "Porter-like stemming" and "dictionary lemmatization") that were combined with three different types of distance functions. Colours and symbols have the same meaning as in Fig. 1.