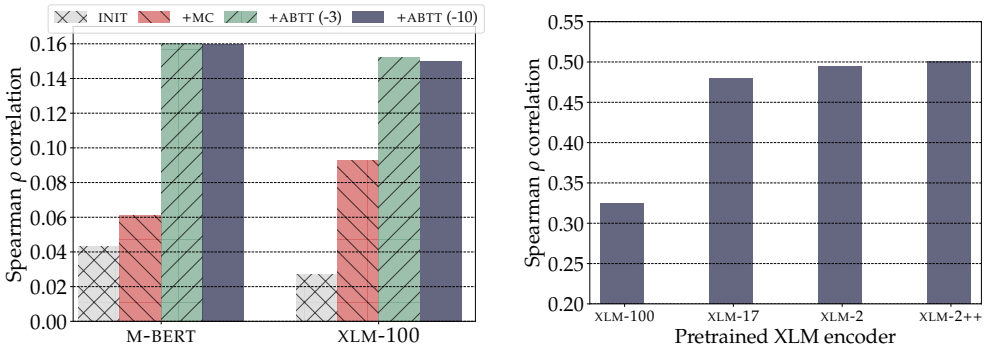

(a) Average scores (b) Scores on ENG-FRA

Figure 6: Further performance analyses of cross-lingual Multi-SimLex datasets. (a) Spearman's ρ correlation scores averaged over all 66 cross-lingual Multi-SimLex datasets for two pretrained multilingual encoders (M-BERT and XLM). The scores are obtained with different configurations that exclude (INIT) or enable unsupervised post-processing. (b) A comparison of various pretrained encoders available for the English-French language pair, see the main text for a short description of each benchmarked pretrained encoder.

of Table 15): many correlation scores are in fact no-correlation results, accentuating the problem of fully unsupervised cross-lingual learning for typologically diverse languages and with fewer amounts of monolingual data (Vulić et al. 2019). The scores are particularly low across the board for lower-resource languages such as Welsh and Kiswahili. It also seems that the lack of monolingual data is a larger problem than typological dissimilarity between language pairs, as we do observe reasonably high correlation scores with VECMAP for language pairs such as CMN-SPA, HEB-EST, and RUS-FIN. However, typological differences (e.g., morphological richness) still play an important role as we observe very low scores when pairing CMN with morphologically rich languages such FIN, EST, POL, and RUS. Similar to prior work of Vulić et al. (2019) and Doval et al. (2019), given the fact that unsupervised VECMAP is the most robust unsupervised CLWE method at present (Glavaš et al. 2019), our results again question the usefulness of fully unsupervised approaches for a large number of languages, and call for further developments in the area of unsupervised and weakly supervised cross-lingual representation learning.

The scores of M-BERT and XLM-100[22] lead to similar conclusions as in the monolingual settings. Reasonable correlation scores are achieved only for a small subset of resource-rich language pairs (e.g., ENG, FRA, SPA, CMN) which dominate the multilingual M-BERT training. Interestingly, the scores indicate a much higher performance of language pairs where YUE is one of the languages when we use M-BERT instead of VECMAP. This boils down again to the fact that YUE, due to its specific language script, has a good representation of its words and subwords in the shared M-BERT vocabulary. At the same time, a reliable VECMAP mapping between YUE and other languages cannot be found due to a small monolingual YUE corpus. In cases when VECMAP does not yield a degenerate

22 The XLM-100 scores are not reported for brevity; they largely follow the patterns observed with M-BERT. The aggregated scores between the two encoders are also very similar as indicated by Figure 6a.