cross-lingual vector space starting from two monolingual ones, the final correlation scores seem substantially higher than the ones obtained by the single massively multilingual M-BERT model.

Finally, the results in Figure 6a again verify the usefulness of unsupervised post-processing also in cross-lingual settings. We observe improved performance with both M-BERT and XLM-100 when mean centering (+MC) is applied, and further gains can be achieved by using ABTT on the mean-centered vector spaces. A similar finding also holds for static cross-lingual word embeddings[23], where applying ABBT (-10) yields higher scores on 61/66 language pairs.

Fully Unsupervised vs. Weakly Supervised Cross-Lingual Embeddings. The results in Table 15 indicate that fully unsupervised cross-lingual learning fails for a large number of language pairs. However, recent work (Vulić et al. 2019) has noted that these sub-optimal non-alignment solutions with the UNSUPER model can be avoided by relying on (weak) cross-lingual supervision spanning only several thousands or even hundreds of word translation pairs. Therefore, we examine 1) if we can further improve the results on cross-lingual Multi-SimLex resorting to (at least some) cross-lingual supervision for resource-rich language pairs; and 2) if such available word-level supervision can also be useful for a range of languages which displayed near-zero performance in Table 15. In other words, we test if recent "tricks of the trade" used in the rich literature on CLWE learning reflect in gains on cross-lingual Multi-SimLex datasets.

First, we reassess the findings established on the bilingual lexicon induction task (Søgaard, Ruder, and Vulić 2018; Vulić et al. 2019): using at least some cross-lingual supervision is always beneficial compared to using no supervision at all. We report improvements over the UNSUPER model for all 10 language pairs in Table 16, even though the UNSUPER method initially produced strong correlation scores. The importance of self-learning increases with decreasing available seed dictionary size, and the +SL model always outperforms UNSUPER with 1k seed pairs; we observe the same patterns also with even smaller dictionary sizes than reported in Table 16 (250 and 500 seed pairs). Along the same line, the results in Table 17 indicate that at least some supervision is crucial for the success of static CLWEs on resource-leaner language pairs. We note substantial improvements on all language pairs; in fact, the VECMAP model is able to learn a more reliable mapping starting from clean supervision. We again note large gains with self-learning.

Multilingual vs. Bilingual Contextualized Embeddings. Similar to the monolingual settings, we also inspect if massively multilingual training in fact dilutes the knowledge necessary for cross-lingual reasoning on a particular language pair. Therefore, we compare the 100-language XLM-100 model with i) a variant of the same model trained on a smaller set of 17 languages (XLM-17); ii) a variant of the same model trained specifically for the particular language pair (XLM-2); and iii) a variant of the bilingual XLM-2 model that also leverages bilingual knowledge from parallel data during joint training (XLM-2++). We again use the pretrained models made available by Conneau and Lample (2019), and we refer to the original work for further technical details.

The results are summarized in Figure 6b, and they confirm the intuition that massively multilingual pretraining can damage performance even on resource-rich languages and language pairs. We observe a steep rise in performance when the

23 Note that VECMAP does mean centering by default as one of its preprocessing steps prior to learning the mapping function (Artetxe, Labaka, and Agirre 2018b; Vulić et al. 2019).