summarized in Table 12 are the first evidence that also confirms its importance for semantic similarity in a wide array of languages. In sum, as a general rule of thumb, we suggest to always mean-center representations for semantic tasks.

The results further indicate that additional post-processing methods such as ABTT and UNCOVEC on top of mean-centered vector spaces can lead to further gains in most languages. The gains are even visible for languages which start from high correlation scores: for instance., CMN with CC+Wiki FT increases from 0.534 to 0.583, from 0.315 to 0.526 with Wiki FT, and from 0.408 to 0.487 with M-BERT. Similarly, for RUS with CC+Wiki FT we can improve from 0.422 to 0.500, and for FRA the scores improve from 0.578 to 0.613. There are additional similar cases reported in Table 12.

Overall, the unsupervised post-processing techniques seem universally useful across languages, but their efficacy and relative performance does vary across different languages. Note that we have not carefully fine-tuned the hyper-parameters of the evaluated post-processing methods, so additional small improvements can be expected for some languages. The main finding, however, is that these post-processing techniques are robust to semantic similarity computations beyond English, and are truly language independent. For instance, removing dominant latent (PCA-based) components from word vectors emphasizes semantic differences between different concepts, as only shared non-informative latent semantic knowledge is removed from the representations.

In summary, pretrained word embeddings do contain more information pertaining to semantic similarity than revealed in the initial vectors. This way, we have corroborated the hypotheses from prior work (Mu, Bhat, and Viswanath 2018; Artetxe et al. 2018) which were not previously empirically verified on other languages due to a shortage of evaluation data; this gap has now been filled with the introduction of the Multi-SimLex datasets. In all follow-up experiments, we always explicitly denote which post-processing configuration is used in evaluation.

POS-Specific Subsets. We present the results for subsets of word pairs grouped by POS class in Table 13. Prior work based on English data showed that representations for nouns are typically of higher quality than those for the other POS classes (Schwartz, Reichart, and Rappoport 2015, 2016; Vulić et al. 2017b). We observe a similar trend in other languages as well. This pattern is consistent across different representation models and can be attributed to several reasons. First, verb representations need to express a rich range of syntactic and semantic behaviors rather than purely referential features (Gruber 1976; Levin 1993; Kipper et al. 2008). Second, low correlation scores on the adjective and adverb subsets in some languages (e.g., POL, CYM, SWA) might be due to their low frequency in monolingual texts, which yields unreliable representations. In general, the variance in performance across different word classes warrants further research in class-specific representation learning (Baker, Reichart, and Korhonen 2014; Vulić et al. 2017b). The scores further attest the usefulness of unsupervised post-processing as almost all class-specific correlation scores are improved by applying mean-centering and ABTT. Finally, the results for M-BERT and XLM-100 in Table 13 further confirm that massively multilingual pretraining cannot yield reasonable semantic representations for many languages: in fact, for some classes they display no correlation with human ratings at all.

Differences across Languages. Naturally, the results from Tables 12 and 13 also reveal that there is variation in performance of both static word embeddings and pretrained encoders across different languages. Among other causes, the lowest absolute scores with FT are reported for languages with least resources available to train monolingual word embeddings, such as Kiswahili, Welsh, and Estonian. The low performance on Welsh is especially indicative: Figure 1 shows that the ratings in the Welsh dataset match up very