additional languages. This can help and create a hugely valuable, large-scale semantic resource for multilingual NLP research.

The core Multi-SimLex we release with this paper already enables researchers to carry out novel linguistic analysis as well as establishes a benchmark for evaluating representation learning models. Based on our preliminary analyses, we found that speakers of closely related languages tend to express equivalent similarity judgments. In particular, geographical proximity seems to play a greater role than family membership in determining the similarity of judgments across languages. Moreover, we tested several state-of-the-art word embedding models, both static and contextualized representations, as well as several (supervised and unsupervised) post-processing techniques, on the newly released Multi-SimLex. This enables future endeavors to improve multilingual representation learning with challenging baselines. In addition, our results provide several important insights for research on both monolingual and cross-lingual word representations:

1) Unsupervised post-processing techniques (mean centering, elimination of top principal components, adjusting similarity orders) are always beneficial independently of the language, although the combination leading to the best scores is language-specific and hence needs to be tuned.

2) Similarity rankings obtained from word embeddings for nouns are better aligned with human judgments than all the other part-of-speech classes considered here (verbs, adjectives, and, for the first time, adverbs). This confirms previous generalizations based on experiments on English.

3) The factor having the greatest impact on the quality of word representations is the availability of raw texts to train them in the first place, rather than language properties (such as family, geographical area, typological features).

4) Massively multilingual pretrained encoders such as M-BERT (Devlin et al. 2019) and XLM-100 (Conneau and Lample 2019) fare quite poorly on our benchmark, whereas pretrained encoders dedicated to a single language are more competitive with static word embeddings such as fastText (Bojanowski et al. 2017). Moreover, for language-specific encoders, parameter reduction techniques reduce performance only marginally.

5) Techniques to inject clean lexical semantic knowledge from external resources into distributional word representations were proven to be effective in emphasizing the relation of semantic similarity. In particular, methods capable of transferring such knowledge from resource-rich to resource-lean languages (Ponti et al. 2019c) increased the correlation with human judgments for most languages, except for those with limited unlabelled data.

Future work can expand our preliminary, yet large-scale study on the ability of pretrained encoders to reason over word-level semantic similarity in different languages. For instance, we have highlighted how sharing the same encoder parameters across multiple languages may harm performance. However, it remains unclear if, and to what extent, the input language embeddings present in XLM-100 but absent in M-BERT help mitigate this issue. In addition, pretrained language embeddings can be obtained both from typological databases (Littell et al. 2017) and from neural architectures (Malaviya, Neubig, and Littell 2017). Plugging these embeddings into the encoders in lieu of embeddings trained end-to-end as suggested by prior work (Tsvetkov et al. 2016; Ammar et al. 2016; Ponti et al. 2019b) might extend the coverage to more resource-lean languages.

Another important follow-up analysis might involve the comparison of the performance of representation learning models on multilingual datasets for both word-level