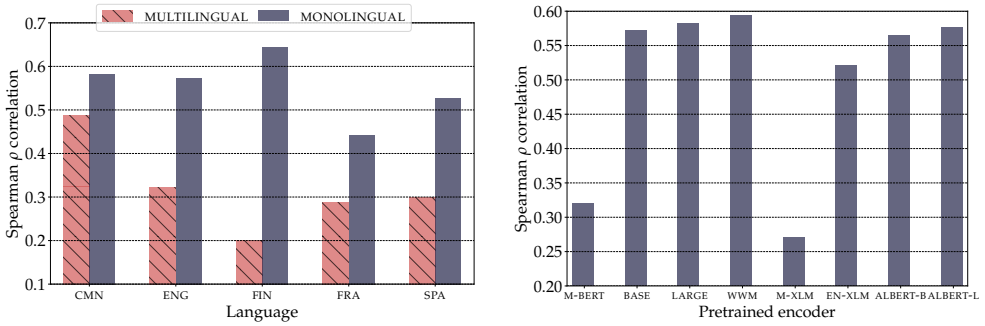

(a) Monolingual vs multilingual (b) Pretrained ENG encoders

Figure 5: **(a)** A performance comparison between monolingual pretrained language encoders and massively multilingual encoders. For four languages (CMN, ENG, FIN, SPA), we report the scores with monolingual uncased BERT-BASE architectures and multilingual uncased M-BERT model, while for FRA we report the results of the multilingual XLM-100 architecture and a monolingual French FlauBERT model (Le et al. 2019), which is based on the same architecture as XLM-100. **(b)** A comparison of various pretrained encoders available for English. All these models are post-processed via ABTT (-3).

From the results in Table 5, it is clear that monolingual pretrained encoders yield much more reliable word-level representations. The gains are visible even for languages such as CMN which showed reasonable performance with M-BERT and are substantial on all test languages. This further confirms the validity of language-specific pretraining in lieu of multilingual training, if sufficient monolingual data are available. Moreover, a comparison of pretrained English encoders in Figure 5b largely follows the intuition: the larger BERT-LARGE model yields slight improvements over BERT-BASE, and we can improve a bit more by relying on word-level (i.e., lexical-level) masking.Finally, light-weight ALBERT model variants are quite competitive with the original BERT models, with only modest drops reported, and ALBERT-L again outperforms ALBERT-B. Overall, it is interesting to note that the scores obtained with monolingual pretrained encoders are on a par with or even outperform static FT word embeddings: this is a very intriguing finding per se as it shows that such subword-level models trained on large corpora can implicitly capture rich lexical semantic knowledge.

Similarity-Specialized Word Embeddings. Conflating distinct lexico-semantic relations is a well-known property of distributional representations (Turney and Pantel 2010; Melamud et al. 2016). Semantic specialization fine-tunes distributional spaces to emphasize a particular lexico-semantic relation in the transformed space by injecting external lexical knowledge (Glavaš, Ponti, and Vulić 2019). Explicitly discerning between true semantic similarity (as captured in Multi-SimLex) and broad conceptual relatedness benefits a number of tasks, as discussed in §2.1.[17] Since most languages lack dedicated lexical resources, however, one viable strategy to steer monolingual word vector spaces to emphasize semantic similarity is through cross-lingual transfer of lexical knowledge, usually through a shared cross-lingual word vector space (Ruder, Vulić, and Søgaard

17 For an overview of specialization methods for semantic similarity, we refer the interested reader to the recent tutorial (Glavaš, Ponti, and Vulić 2019).