resources and evaluation benchmarks. This alternative practice has had a profound impact on distributional semantics and representation learning (Hill, Reichart, and Korhonen 2015). While some prominent English word pair datasets such as WordSim-353 (Finkelstein et al. 2002), MEN (Bruni, Tran, and Baroni 2014), or Stanford Rare Words (Luong, Socher, and Manning 2013) did not discriminate between similarity and relatedness, the importance of this distinction was established by Hill, Reichart, and Korhonen (2015, see again the discussion in §2.1) through the creation of SimLex-999. This inspired other similar datasets which focused on different lexical properties. For instance, SimVerb-3500 (Gerz et al. 2016) provided similarity ratings for 3,500 English verbs, whereas CARD-660 (Pilehvar et al. 2018) aimed at measuring the semantic similarity of infrequent concepts.

Semantic Similarity Datasets in Other Languages. Motivated by the impact of datasets such as SimLex-999 and SimVerb-3500 on representation learning in English, a line of related work focused on creating similar resources in other languages. The dominant approach is translating and reannotating the entire original English SimLex-999 dataset, as done previously for German, Italian, and Russian (Leviant and Reichart 2015), Hebrew and Croatian (Mrkšić et al. 2017), and Polish (Mykowiecka, Marciniak, and Rychlik 2018). Venekoski and Vankka (2017) apply this process only to a subset of 300 concept pairs from the English SimLex-999. On the other hand, Camacho-Collados et al. (2017) sampled a new set of 500 English concept pairs to ensure wider topical coverage and balance across similarity spectra, and then translated those pairs to German, Italian, Spanish, and Farsi (SEMEVAL-500). A similar approach was followed by Ercan and Yıldız (2018) for Turkish, by Huang et al. (2019) for Mandarin Chinese, and by Sakaizawa and Komachi (2018) for Japanese. Netisopakul, Wohlgenannt, and Pulich (2019) translated the concatenation of SimLex-999, WordSim-353, and the English SEMEVAL-500 into Thai and then reannotated it. Finally, Barzegar et al. (2018) translated English SimLex-999 and WordSim-353 to 11 resource-rich target languages (German, French, Russian, Italian, Dutch, Chinese, Portuguese, Swedish, Spanish, Arabic, Farsi), but they did not provide details concerning the translation process and the resolution of translation disagreements. More importantly, they also did not reannotate the translated pairs in the target languages. As we discussed in § 2.3 and reiterate later in §5, semantic differences among languages can have a profound impact on the annotation scores; particulary, we show in §5.4 that these differences even roughly define language clusters based on language affinity.

A core issue with the current datasets concerns a lack of one unified procedure that ensures the comparability of resources in different languages. Further, concept pairs for different languages are sourced from different corpora (e.g., direct translation of the English data versus sampling from scratch in the target language). Moreover, the previous SimLex-based multilingual datasets inherit the main deficiencies of the English original version, such as the focus on nouns and highly frequent concepts. Finally, prior work mostly focused on languages that are widely spoken and do not account for the variety of the world's languages. Our long-term goal is devising a standardized methodology to extend the coverage also to languages that are resource-lean and/or typologically diverse (e.g., Welsh, Kiswahili as in this work).

Multilingual Datasets for Natural Language Understanding. The Multi-SimLex initiative and corresponding datasets are also aligned with the recent efforts on procuring multilingual benchmarks that can help advance computational modeling of natural language understanding across different languages. For instance, pretrained multilingual language models such as multilingual BERT (Devlin et al. 2019) or XLM (Conneau and Lample 2019) are typically probed on XNLI test data (Conneau et al. 2018b) for cross-