low-resource languages typically lack similar evaluation data. Even if some resources do exist, they are limited in their *size* (e.g., 500 pairs in Turkish (Ercan and Yıldız 2018), 500 in Farsi (Camacho-Collados et al. 2017), or 300 in Finnish (Venekoski and Vankka 2017)) and *coverage* (e.g., all datasets which originated from the original English SimLex-999 contain only high-frequent concepts, and are dominated by nouns). This is why, as our departure point, we introduce a **larger and more comprehensive** English word similarity dataset spanning 1,888 concept pairs (see §4).

Most importantly, semantic similarity datasets in different languages have been created using heterogeneous construction procedures with different guidelines for translation and annotation, as well as different rating scales. For instance, some datasets were obtained by directly translating the English SimLex-999 in its entirety (Leviant and Reichart 2015; Mrkšić et al. 2017) or in part (Venekoski and Vankka 2017). Other datasets were created from scratch (Ercan and Yıldız 2018) and yet others sampled English concept pairs differently from SimLex-999 and then translated and reannotated them in target languages (Camacho-Collados et al. 2017). This heterogeneity makes these datasets incomparable and precludes systematic cross-linguistic analyses. In this article, consolidating the lessons learned from previous dataset construction paradigms, we propose a carefully designed **translation and annotation protocol** for developing monolingual Multi-SimLex datasets with aligned concept pairs for typologically diverse languages. We apply this protocol to a set of 12 languages, including a mixture of major languages (e.g., Mandarin, Russian, and French) as well as several low-resource ones (e.g., Kiswahili, Welsh, and Yue Chinese). We demonstrate that our proposed dataset creation procedure yields data with high inter-annotator agreement rates (e.g., the average mean inter-annotator agreement for Welsh is 0.742).

The unified construction protocol and alignment between concept pairs enables a series of quantitative analyses. Preliminary studies on the influence that polysemy and cross-lingual variation in lexical categories (see §2.3) have on similarity judgments are provided in §5. Data created according to Multi-SimLex protocol also allow for probing into whether similarity judgments are universal across languages, or rather depend on linguistic affinity (in terms of linguistic features, phylogeny, and geographical location). We investigate this question in §5.4. Naturally, Multi-SimLex datasets can be used as an intrinsic evaluation benchmark to assess the quality of lexical representations based on monolingual, joint multilingual, and transfer learning paradigms. We conduct a systematic evaluation of several state-of-the-art representation models in §7, showing that there are large gaps between human and system performance in all languages. The proposed construction paradigm also supports the automatic creation of 66 cross-lingual Multi-SimLex datasets by interleaving the monolingual ones. We outline the construction of the cross-lingual datasets in §6, and then present a quantitative evaluation of a series of cutting-edge cross-lingual representation models on this benchmark in §8.

Contributions. We now summarize the main contributions of this work:

1) Building on lessons learned from prior work, we create a more comprehensive lexical semantic similarity dataset for the English language spanning a total of 1,888 concept pairs balanced with respect to similarity, frequency, and concreteness, and covering four word classes: nouns, verbs, adjectives and, for the first time, adverbs. This dataset serves as the main source for the creation of equivalent datasets in several other languages.

2) We present a carefully designed and rigorous language-agnostic translation and annotation protocol. These well-defined guidelines will facilitate the development of future Multi-SimLex datasets for other languages. The proposed protocol eliminates