improvements towards a larger, more representative, and more reliable lexical similarity dataset in English and, consequently, in all other languages.

The Final Output: English Multi-SimLex. In order to ensure that the criterion C1 is satisfied, we consolidate and integrate the data already carefully sampled in prior work into a single, comprehensive, and representative dataset. This way, we can control for diversity, frequency, and other properties while avoiding to perform this time-consuming selection process from scratch. Note that, on the other hand, the word pairs chosen for English are scored from scratch as part of the entire Multi-SimLex annotation process, introduced later in §5. We now describe the external data sources for the final set of word pairs:

1) *Source: SimLex-999*. (Hill, Reichart, and Korhonen 2015). The English Multi-SimLex has been initially conceived as an extension of the original SimLex-999 dataset. Therefore, we include all 999 word pairs from SimLex, which span 666 noun pairs, 222 verb pairs, and 111 adjective pairs. While SimLex-999 already provides examples representing different POS classes, it does not have a sufficient coverage of different linguistic phenomena: for instance, it contains only very frequent concepts, and it does not provide a representative set of verbs (Gerz et al. 2016).

2) *Source: SemEval-17: Task 2* (henceforth SEMEVAL-500; Camacho-Collados et al. 2017). We start from the full dataset of 500 concept pairs to extract a total of 334 concept pairs for English Multi-SimLex a) which contain only single-word concepts, b) which are not named entities, c) where POS tags of the two concepts are the same, d) where both concepts occur in the top 250K most frequent word types in the English Wikipedia, and e) do not already occur in SimLex-999. The original concepts were sampled as to span all the 34 domains available as part of BabelDomains (Camacho-Collados and Navigli 2017), which roughly correspond to the main high-level Wikipedia categories. This ensures topical diversity in our sub-sample.

3) *Source: CARD-660* (Pilehvar et al. 2018). 67 word pairs are taken from this dataset focused on rare word similarity, applying the same selection criteria a) to e) employed for SEMEVAL-500. Words are controlled for frequency based on their occurrence counts from the Google News data and the ukWaC corpus (Baroni et al. 2009). CARD-660 contains some words that are very rare (*logboat*), domain-specific (*erythroleukemia*) and slang (*2mrw*), which might be difficult to translate and annotate across a wide array of languages. Hence, we opt for retaining only the concept pairs above the threshold of top 250K most frequent Wikipedia concepts, as above.

4) *Source: SimVerb-3500* (Gerz et al. 2016) Since both CARD-660 and SEMEVAL-500 are heavily skewed towards noun pairs, and nouns also dominate the original SimLex-999, we also extract additional verb pairs from the verb-specific similarity dataset SimVerb-3500. We randomly sample 244 verb pairs from SimVerb-3500 that represent all similarity spectra. In particular, we add 61 verb pairs for each of the similarity intervals: $[0, 1.5), [1.5, 3), [3, 4.5), [4.5, 6]$. Since verbs in SimVerb-3500 were originally chosen from VerbNet (Kipper, Snyder, and Palmer 2004; Kipper et al. 2008), they cover a wide range of verb classes and their related linguistic phenomena.

5) *Source: University of South Florida* (USF; Nelson, McEvoy, and Schreiber 2004) norms, the largest database of free association for English. In order to improve the representation of different POS classes, we sample additional adjectives and adverbs from the USF norms following the procedure established by Hill, Reichart, and Korhonen (2015); Gerz et al. (2016). This yields additional 122 adjective pairs, but only a limited number of adverb pairs (e.g., *later – never*, *now – here*, *once – twice*). Therefore, we also create a set