of adverb pairs semi-automatically by sampling adjectives that can be derivationally transformed into adverbs (e.g. adding the suffix *-ly*) from the USF, and assessing the correctness of such derivation in WordNet. The resulting pairs include, for instance, *primarily – mainly*, *softly – firmly*, *roughly – reliably*, etc. We include a total of 123 adverb pairs into the final English Multi-SimLex. Note that this is the first time adverbs are included into any semantic similarity dataset.

Fulfillment of Construction Criteria. The final ENG Multi-SimLex dataset spans 1,051 noun pairs, 469 verb pairs, 245 adjective pairs, and 123 adverb pairs.[2] As mentioned above, the criterion C1 has been fulfilled by relying only on word pairs that already underwent meticulous sampling processes in prior work, integrating them into a single resource. As a consequence, Multi-SimLex allows for fine-grained analyses over different POS classes, concreteness levels, similarity spectra, frequency intervals, relation types, morphology, lexical fields, and it also includes some challenging orthographically similar examples (e.g., *infection – inflection*).[3] We ensure that the criteria C2 and C3 are satisfied by using similar annotation guidelines as Simlex-999, SimVerb-3500, and SEMEVAL-500 that explicitly target semantic similarity. In what follows, we outline the carefully tailored process of translating and annotating Multi-SimLex datasets in all target languages.

5. Multi-SimLex: Translation and Annotation

We now detail the development of the final Multi-SimLex resource, describing our language selection process, as well as translation and annotation of the resource, including the steps taken to ensure and measure the quality of this resource. We also provide key data statistics and preliminary cross-lingual comparative analyses.

Language Selection. Multi-SimLex comprises eleven languages in addition to English. The main objective for our inclusion criteria has been to balance language prominence (by number of speakers of the language) for maximum impact of the resource, while simultaneously having a diverse suite of languages based on their typological features (such as morphological type and language family). Table 1 summarizes key information about the languages currently included in Multi-SimLex. We have included a mixture of fusional, agglutinative, isolating, and introflexive languages that come from eight different language families. This includes languages that are very widely used such as Chinese Mandarin and Spanish, and low-resource languages such as Welsh and Kiswahili. We hope to further include additional languages and inspire other researchers to contribute to the effort over the lifetime of this project.

The work on data collection can be divided into two crucial phases: 1) a translation phase where the extended English language dataset with 1,888 pairs (described in §4) is translated into eleven target languages, and 2) an annotation phase where human raters scored each pair in the translated set as well as the English set. Detailed guidelines for both phases are available online at: `https://multisimlex.com`.

[2] There is a very small number of adjective and verb pairs extracted from CARD-660 and SEMEVAL-500 as well. For instance, the total number of verbs is 469 since we augment the original 222 SimLex-999 verb pairs with 244 SimVerb-3500 pairs and 3 SEMEVAL-500 pairs; and similarly for adjectives.

[3] Unlike SEMEVAL-500 and CARD-660, we do not explicitly control for the equal representation of concept pairs across each similarity interval for several reasons: a) Multi-SimLex contains a substantially larger number of concept pairs, so it is possible to extract balanced samples from the full data; b) such balance, even if imposed on the English dataset, would be distorted in all other monolingual and cross-lingual datasets; c) balancing over similarity intervals arguably does not reflect a true distribution “in the wild” where most concepts are only loosely related or completely unrelated.