lingual natural language inference. XNLI was created by translating examples from the English MultiNLI dataset, and projecting its sentence labels (Williams, Nangia, and Bowman 2018). Other recent multilingual datasets target the task of question answering based on reading comprehension: i) MLQA (Lewis et al. 2019) includes 7 languages ii) XQuAD (Artetxe, Ruder, and Yogatama 2019) 10 languages; iii) TyDiQA (Clark et al. 2020) 9 widely spoken typologically diverse languages. While MLQA and XQuAD result from the translation from an English dataset, TyDiQA was built independently in each language. Another multilingual dataset, PAWS-X (Yang et al. 2019), focused on the paraphrase identification task and was created translating the original English PAWS (Zhang, Baldridge, and He 2019) into 6 languages. We believe that Multi-SimLex can substantially contribute to this endeavor by offering a comprehensive multilingual benchmark for the fundamental lexical level relation of semantic similarity. In future work, Multi-SimLex also offers an opportunity to investigate the correlations between word-level semantic similarity and performance in downstream tasks such as QA and NLI across different languages.

4. The Base for Multi-SimLex: Extending English SimLex-999

In this section, we discuss the design principles behind the English (ENG) Multi-SimLex dataset, which is the basis for all the Multi-SimLex datasets in other languages, as detailed in §5. We first argue that a new, more balanced, and more comprehensive evaluation resource for lexical semantic similarity in English is necessary. We then describe how the 1,888 word pairs contained in the ENG Multi-SimLex were selected in such a way as to represent various linguistic phenomena within a single integrated resource.

Construction Criteria. The following criteria have to be satisfied by any high-quality semantic evaluation resource, as argued by previous studies focused on the creation of such resources (Hill, Reichart, and Korhonen 2015; Gerz et al. 2016; Vulić et al. 2017a; Camacho-Collados et al. 2017, *inter alia*):

(C1) Representative and diverse. The resource must cover the full range of diverse concepts occurring in natural language, including different word classes (e.g., nouns, verbs, adjectives, adverbs), concrete and abstract concepts, a variety of lexical fields, and different frequency ranges.

(C2) Clearly defined. The resource must provide a clear understanding of which semantic relation exactly is annotated and measured, possibly contrasting it with other relations. For instance, the original SimLex-999 and SimVerb-3500 explicitly focus on true semantic similarity and distinguish it from broader relatedness captured by datasets such as MEN (Bruni, Tran, and Baroni 2014) or WordSim-353 (Finkelstein et al. 2002).

(C3) Consistent and reliable. The resource must ensure consistent annotations obtained from non-expert native speakers following simple and precise annotation guidelines.

In choosing the word pairs and constructing ENG Multi-SimLex, we adhere to these requirements. Moreover, we follow good practices established by the research on related resources. In particular, since the introduction of the original SimLex-999 dataset (Hill, Reichart, and Korhonen 2015), follow-up works have improved its construction protocol across several aspects, including: 1) coverage of more lexical fields, e.g., by relying on a diverse set of Wikipedia categories (Camacho-Collados et al. 2017), 2) infrequent/rare words (Pilehvar et al. 2018), 3) focus on particular word classes, e.g., verbs (Gerz et al. 2016), 4) annotation quality control (Pilehvar et al. 2018). Our goal is to make use of these