others, suffer the most. As a consequence, resources providing clean information on semantic similarity are key in mitigating the side effects of the distributional signal. In particular, such databases can be employed for the *intrinsic evaluations* of specific WE models as a proxy of their reliability for downstream applications (Collobert and Weston 2008; Baroni and Lenci 2010; Hill, Reichart, and Korhonen 2015); intuitively, the more WEs are misaligned with human judgments of similarity, the more their performance on actual tasks is expected to be degraded. Moreover, word representations can be *specialized* (a.k.a. retrofitted) by disentangling word relations of similarity and association. In particular, linguistic constraints sourced from external databases (such as synonyms from WordNet) can be injected into WEs (Faruqui et al. 2015; Wieting et al. 2015; Mrkšić et al. 2017; Lauscher et al. 2019; Kamath et al. 2019, *inter alia*) in order to enforce a particular relation in a distributional semantic space while preserving the original adjacency properties.

2.3 Similarity and Language Variation: Semantic Typology

In this work, we tackle the concept of (true) semantic similarity from a multilingual perspective. While the same meaning representations may be shared by all human speakers at a deep cognitive level, there is no one-to-one mapping between the words in the lexicons of different languages. This makes the comparison of similarity judgments across languages difficult, since the meaning overlap of translationally equivalent words is sometimes far less than exact. This results from the fact that the way languages 'partition' semantic fields is partially arbitrary (Trier 1931), although constrained cross-lingually by common cognitive biases (Majid et al. 2007). For instance, consider the field of colors: English distinguishes between *green* and *blue*, whereas Murle (South Sudan) has a single word for both (Kay and Maffi 2013).

In general, *semantic typology* studies the variation in lexical semantics across the world's languages. According to (Evans 2011), the ways languages categorize concepts into the lexicon follow three main axes: 1) *granularity*: what is the number of categories in a specific domain?; 2) *boundary location*: where do the lines marking different categories lie?; 3) *grouping and dissection*: what are the membership criteria of a category; which instances are considered to be more prototypical? Different choices with respect to these axes lead to different lexicalization patterns.[1] For instance, distinct senses in a polysemous word in English, such as *skin* (referring to both the body and fruit), may be assigned separate words in other languages such as Italian *pelle* and *buccia*, respectively (Rzymski et al. 2020). We later analyze whether similarity scores obtained from native speakers also loosely follow the patterns described by semantic typology.

3. Previous Work and Evaluation Data

Word Pair Datasets. Rich expert-created resources such as WordNet (Miller 1995; Fellbaum 1998), VerbNet (Kipper Schuler 2005; Kipper et al. 2008), or FrameNet (Baker, Fillmore, and Lowe 1998) encode a wealth of semantic and syntactic information, but are expensive and time-consuming to create. The scale of this problem gets multiplied by the number of languages in consideration. Therefore, crowd-sourcing with non-expert annotators has been adopted as a quicker alternative to produce smaller and more focused semantic

1 More formally, *colexification* is a phenomenon when different meanings can be expressed by the same word in a language (François 2008). For instance, the two senses which are distinguished in English as *time* and *weather* are co-lexified in Croatian: the word *vrijeme* is used in both cases.