also be encountered, although more rarely. An example are synonyms where a word is common and the other infrequent, such as *to seize* and *to commandeer*. Hill, Reichart, and Korhonen (2015) revealed that while similarity measures based on the WordNet graph (Wu and Palmer 1994) and human judgments of association in the University of South Florida Free Association Database (Nelson, McEvoy, and Schreiber 2004) do correlate, a number of pairs follow opposite trends. Several studies on human cognition also point in the same direction. For instance, semantic priming can be triggered by similar words without association (Lucas 2000). On the other hand, a connection with cue words is established more quickly for topically related words rather than for similar words in free association tasks (De Deyne and Storms 2008).

A key property of semantic similarity is its *gradience*: pairs of words can be similar to a different degree. On the other hand, the relation of *synonymy* is binary: pairs of words are synonyms if they can be substituted in all contexts (or most contexts, in a looser sense), otherwise they are not. While synonyms can be conceived as lying on one extreme of the semantic similarity continuum, it is crucial to note that their definition is stated in purely relational terms, rather than invoking their referential properties (Lyons 1977; Cruse 1986; Coseriu 1967). This makes behavioral studies on semantic similarity fundamentally different from lexical resources like WordNet (Miller 1995), which include paradigmatic relations (such as synonymy).

2.2 Similarity for NLP: Intrinsic Evaluation and Semantic Specialization

The ramifications of the distinction between similarity and association are profound for distributional semantics. This paradigm of lexical semantics is grounded in the distributional hypothesis, formulated by Firth (1957) and Harris (1951). According to this hypothesis, the meaning of a word can be recovered empirically from the contexts in which it occurs within a collection of texts. Since both pairs of topically related words and pairs of purely similar words tend to appear in the same contexts, their associated meaning confounds the two distinct relations (Hill, Reichart, and Korhonen 2015; Schwartz, Reichart, and Rappoport 2015; Vulić et al. 2017b). As a result, distributional methods obscure a crucial facet of lexical meaning.

This limitation also reflects onto word embeddings (WEs), representations of words as low-dimensional vectors that have become indispensable for a wide range of NLP applications (Collobert et al. 2011; Chen and Manning 2014; Melamud et al. 2016, *inter alia*). In particular, it involves both *static* WEs learned from co-occurrence patterns (Mikolov et al. 2013; Levy and Goldberg 2014; Bojanowski et al. 2017) and *contextualized* WEs learned from modeling word sequences (Peters et al. 2018; Devlin et al. 2019, *inter alia*). As a result, in the induced representations, geometrical closeness (measured e.g. through cosine distance) conflates genuine similarity with broad relatedness. For instance, the vectors for antonyms such as *sober* and *drunk*, by definition dissimilar, might be neighbors in the semantic space under the distributional hypothesis. Turney (2012), Kiela and Clark (2014), and Melamud et al. (2016) demonstrated that different choices of hyper-parameters in WE algorithms (such as context window) emphasize different relations in the resulting representations. Likewise, Agirre et al. (2009) and Levy and Goldberg (2014) discovered that WEs learned from texts annotated with syntactic information mirror similarity better than simple local bag-of-words neighborhoods.

The failure of WEs to capture semantic similarity, in turn, affects model performance in several NLP applications where such knowledge is crucial. In particular, Natural Language Understanding tasks such as statistical dialog modeling, text simplification, or semantic text similarity (Mrkšić et al. 2016; Kim et al. 2016; Ponti et al. 2019c), among