HyperLex: A Large-Scale Evaluation of Graded Lexical Entailment

Ivan Vulić*
LTL, University of Cambridge

Daniela Gerz*
LTL, University of Cambridge

Douwe Kiela**
Facebook AI Research

Felix Hill†
Google DeepMind

Anna Korhonen*
LTL, University of Cambridge

We introduce HyperLex—a data set and evaluation resource that quantifies the extent of the semantic category membership, that is, type-of *relation, also known as hyponymy–hypernymy or lexical entailment (LE) relation between 2,616 concept pairs. Cognitive psychology research has established that typicality and category/class membership are computed in human semantic memory as a gradual rather than binary relation. Nevertheless, most NLP research and existing large-scale inventories of concept category membership (WordNet, DBPedia, etc.) treat category membership and LE as binary. To address this, we asked hundreds of native English speakers to indicate typicality and strength of category membership between a diverse range of concept pairs on a crowdsourcing platform. Our results confirm that category membership and LE are indeed more gradual than binary. We then compare these human judgments with the predictions of automatic systems, which reveals a huge gap between human performance and state-of-the-art LE, distributional and representation learning models, and substantial differences between the models themselves. We discuss a pathway for improving semantic models to overcome this discrepancy, and indicate future application areas for improved graded LE systems.*

* Language Technology Lab (LTL), Department of Theoretical and Applied Linguistics, University of Cambridge, 9 West Road, CB3 9DP Cambridge, UK. E-mail: {iv250|dsg40|alk23}@cam.ac.uk.
** Facebook AI Research, 770 Broadway, NY 10003, New York City, NY, USA. E-mail: dkiela@fb.com.
† Google DeepMind, 7 Pancras Square, London NC14AG, UK. E-mail: felixhill@google.com.

Submission received: 30 August 2016; revised version received: 11 May 2017; accepted for publication: 7 July 2017.

doi:10.1162/COLI_a_00301

1. Introduction

Most native speakers of English, in almost all contexts and situations, would agree that *dogs*, *cows*, or *cats* are *animals*, and that *tables* or *pencils* are not. However, for certain concepts, membership of the animal category is less clear-cut. Whether lexical concepts such as *dinosaur*, *human being*, or *amoeba* are considered animals seems to depend on the context in which such concepts are described, the perspective of the speaker or listener, and even the formal scientific knowledge of the interlocutors. Despite this indeterminacy, when communicating, humans intuitively reason about such relations between concepts and categories (Quillian 1967; Collins and Quillian 1969). Indeed, the ability to quickly perform inference over such networks and arrive at coherent knowledge representations is crucial for human language understanding.

The Princeton WordNet lexical database (Miller 1995; Fellbaum 1998) is perhaps the best-known attempt to formally represent such a semantic network. In WordNet, concepts are organized in a hierarchical fashion, in an attempt to replicate observed aspects of human semantic memory (Collins and Quillian 1972; Beckwith et al. 1991). One of the fundamental relations between concepts in WordNet is the so-called TYPE-OF or **hyponymy–hypernymy** relation that exists between category concepts such as *animal* and their constituent members such as *cat* or *dog*. The type-of relation is particularly important in language understanding because it underlines the **lexical entailment** (LE) relation. Simply put, an instantiation of a member concept such as a cat *entails* the existence of an *animal*. This lexical entailment in turns governs many cases of phrasal and sentential entailment: If we know that *a cat is in the garden*, we can quickly and intuitively conclude that *an animal is in the garden,* too.[1]

Because of this fundamental connection to language understanding, the automatic detection and modeling of lexical entailment has been an area of much focus in natural language processing (Bos and Markert 2005; Dagan, Glickman, and Magnini 2006; Baroni et al. 2012; Beltagy et al. 2013, inter alia). The ability to effectively detect and model both lexical and phrasal entailment in a human-like way may be critical for numerous related applications, such as question answering, information retrieval, information extraction, and text summarization and generation (Androutsopoulos and Malakasiotis 2010). For instance, in order to answer a question such as *"Which mammal has a strong bite?"*, a question-answering system has to know that a jaguar or a grizzly bear are types of mammals, whereas a crocodile or a piranha are not.

Although inspired to some extent by theories of human semantic memory, large-scale inventories of semantic concepts, such as WordNet, typically make many simplifying assumptions, particularly regarding the nature of the *type-of* relation, and consequently the effect of LE. In WordNet, for instance, all semantic relations are represented in a binary way (i.e., concept X entails Y) rather than gradual (e.g., X entails Y to a certain degree). However, since at least the pioneering experiments on prototypes (Rosch 1973, 1975), it has been known that, for a given semantic category, certain member concepts are consistently understood as more central to the category than others (even when controlling for clearly confounding factors such as frequency) (Coleman and Kay 1981; Medin, Altom, and Murphy 1984; Lakoff 1990; Hampton 2007). In other words, WordNet and similar resources fail to capture the fact that category membership is a gradual

[1] Because of dual and inconsistent use in prior work, in this work we use the term *lexical entailment (LE)* in its stricter definition: it refers precisely to the taxonomical *hyponymy–hypernymy* relation, also known as TYPE-OF, or IS-A relation. More details on the distinction between taxonomical and substitutable LE are provided in Section 2.

semantic phenomenon. This limitation of WordNet also characterizes much of the LE research in NLP, as we discuss later in Section 3.

To address these limitations, the present work is concerned with **graded lexical entailment**: the degree of the LE relation between two concepts on a continuous scale. Thanks to the availability of crowdsourcing technology, we conduct a variant of the seminal behavioral data collection by Rosch (1973), but on a massive scale. To do so, we introduce the idea of graded or *soft* LE, and design a human rating task for (X, Y) concept pairs based on the following question: *To what degree is X a type of Y?*. We arrive at a data set with 2,616 concept pairs, each rated by at least 10 human raters, scored by the degree to which they exhibit typicality and semantic category membership and, equivalently, LE. Using this data set, **HyperLex**,[2] we investigate two questions:

- **(Q1)** Do we observe the same effects of typicality, graded membership, and graded lexical entailment in human judgments as observed by Rosch? Do humans intuitively distinguish between central and non-central members of a category/class? Do humans distinguish between full and partial membership in a class as discussed by Kamp and Partee (1995)?
- **(Q2)** Is the current LE modeling and representation methodology as applied in NLP research and technology sufficient to accurately capture graded lexical entailment automatically? What is the gap between current automatic systems and human performance in the graded LE task?

The article is structured as follows. We define and discuss graded LE in Section 2. In Section 3, we survey benchmarking resources from the literature that pertain to semantic category membership, LE identification, or evaluation, and motivate the need for a new, more expressive resource. In Section 4, we describe the design and development of HyperLex, and outline the various semantic dimensions (such as POS usage, hypernymy levels, and concreteness levels) along which these concept pairs are designed to vary.

This allows us to address Q1 in Section 5, where we present a series of qualitative analyses of the data gathered and collated into HyperLex. High inter-annotator agreement scores (pairwise and mean Spearman's ρ correlations around 0.85 on the entire data set; similar correlations on noun and verb subsets) indicate that participants found it unproblematic to rate consistently the graded LE relation for the full range of concepts. These analyses reveal that the data in HyperLex enhance, rather than contradict or undermine, the information in WordNet, in the sense that hyponymy–hypernymy pairs receive highest average ratings in HyperLex compared with all other WordNet relations. We also show that participants are able to capture the implicit asymmetry of the graded LE relation by examining ratings of (X, Y) and reversed (Y, X) pairs. Most importantly, our analysis shows that the effects of typicality, vagueness, and gradual nature of LE are indeed captured in human judgments. For instance, graded LE scores indicate that humans rate concepts such as *to talk* or *to speak* as more typical instances of the class *to communicate* than concepts such as *to touch*, or *to pray*.

In Section 6 we then turn our attention to Q2: We evaluate the performance of a wide range of LE detection or measurement approaches. This review covers: (i) distributional models relying on the distributional inclusion hypothesis (Geffet and Dagan

[2] HyperLex is available online at: `http://people.ds.cam.ac.uk/iv250/hyperlex.html`.

2005; Lenci and Benotto 2012) and semantic generality computations (Santus et al. 2014); (ii) multi-modal approaches (Kiela et al. 2015); (iii) WordNet-based approaches (Pedersen, Patwardhan, and Michelizzi 2004); and (iv) a selection of state-of-the-art recent word embeddings, some optimized for similarity on semantic similarity data sets (Mikolov et al. 2013b; Levy and Goldberg 2014; Wieting et al. 2015, inter alia), others developed to better capture the asymmetric LE relation (Vilnis and McCallum 2015; Vendrov et al. 2016). Because of its size, and unlike other word pair scoring data sets such as SimLex-999 or WordSim-353, in HyperLex we provide standard train/dev/test splits (both *random* and *lexical* [Levy et al. 2015; Shwartz, Goldberg, and Dagan 2016]) so that HyperLex can be used for supervised learning. We therefore evaluate several prominent supervised LE architectures (Baroni et al. 2012; Roller, Erk, and Boleda 2014; Weeds et al. 2014, inter alia). Although we observe interesting differences in the models, our findings indicate clearly that none of the currently available models or approaches accurately model the relation of graded LE reflected in human subjects. This study therefore calls for new paradigms and solutions capable of capturing the gradual nature of semantic relations such as hypernymy in hierarchical semantic networks.

In Section 8, we turn to the future and discuss potential applications of the graded LE concept and HyperLex. We conclude in Section 9 by summarizing the key aspects of our contribution. HyperLex offers robust, data-driven insight into how humans perceive the concepts of typicality and graded membership within the graded LE relation. We hope that this will in turn incentivize research into language technology that both reflects human semantic memory more faithfully and interprets and models linguistic entailment more effectively.

2. Graded Lexical Entailment

Note on Terminology. Because of dual and inconsistent use in prior work, in this work we use the term *lexical entailment (LE)* in its stricter definition. It refers precisely to the taxonomical **asymmetric** *hyponymy–hypernymy* relation, also known as IS-A, or TYPE-OF relation (Hearst 1992; Snow, Jurafsky, and Ng 2004; Weeds, Weir, and McCarthy 2004; Pantel and Pennacchiotti 2006; Do and Roth 2010, inter alia), e.g., *snake* is a TYPE-OF *animal*, *computer* is a TYPE-OF *machine*.

This is different from an alternative definition (Zhitomirsky-Geffet and Dagan 2009; Kotlerman et al. 2010; Turney and Mohammad 2015) as **substitutable** lexical entailment: This relation holds for a pair of words (X, Y) if a possible meaning of one word (i.e., X) entails a meaning of the other, and the entailing word can substitute the entailed one in some typical contexts. This definition is looser and more general than the TYPE-OF definition, as it also encompasses other lexical relations such as synonymy, metonymy, meronymy, and so forth.[3]

Definitions. The classical definition of **ungraded lexical entailment** is as follows: Given a concept word pair (X, Y), Y is a hypernym of X if and only if X is a type of Y, or equivalently every X is a Y.[4] On the other hand, **graded lexical entailment** defines the strength of the lexical entailment relation between the two concepts. Given the concept

[3] For instance, Turney and Mohammad (2015) argue that in the sentences *Jane dropped the glass* and *Jane dropped something fragile*, the concept *glass* should entail *fragile*.

[4] Other variants of the same definition replace TYPE-OF with KIND-OF or INSTANCE-OF.

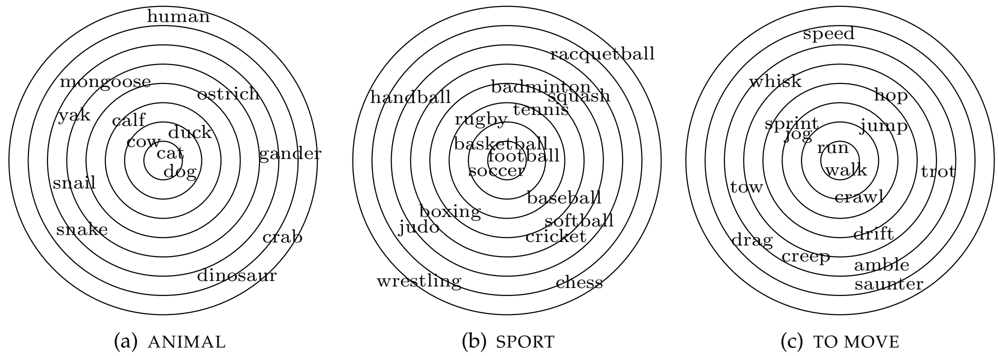
pair (X, Y) and the entailment strength s, the triplet (X, Y, s) defines to what degree Y is a hypernym of X (i.e., *to what degree X is a type of Y*), where the degree is quantified by s (e.g., to what degree *snake* is a TYPE-OF *animal*).

It may be observed as approximate or **soft entailment**, a weaker form of the classical entailment variant (Esteva et al. 2012; Bankova et al. 2016). By imposing a threshold *thr* on s, all graded relations may be easily converted to discrete ungraded decisions.

(Proto)typicality, Graded Membership, and Graded LE. The graded LE relation as described by the intuitive question "to what degree is X a type of Y?" encompasses two distinct phenomena described in cognitive science research (cf. Hampton 2007). First, it can be seen as the measure of **typicality** in graded cognitive categorization (Rosch 1973, 1975; Medin, Altom, and Murphy 1984; Lakoff 1990), where some instances of a category are more central than others, as illustrated in Fig. 1(a)–Fig. 1(c). It measures to what degree some class instance X is a prototypical example of class/concept Y. For instance, when humans are asked to give an example instance of the concept *sport*, it turns out that *football* and *basketball* are more frequently cited than *wrestling*, *chess*, *softball*, or *racquetball*. Another viewpoint stresses that "prototypes serve as reference points for the categorization of not-so-clear instances" (Taylor 2003). Osherson and Smith (1997) make further developments to the theory of (proto)typicality by recognizing that there exist concepts "that lack prototypes while possessing degrees of exemplification." They list the famous example of the concept *building* without a clear prototype; however, people tend to agree that most banks are more typical buildings than, say, barns or pile dwellings.

Second, the graded LE relation also arises when one asks about the applicability of concepts to objects: The boundaries between a category and its instances are much more often fuzzy and vague than unambiguous and clear-cut (Kamp and Partee 1995). In other words, the **graded membership** (often termed **vagueness**) measures the graded applicability of a concept to different instances. For instance, it is not clear to what extent different objects in our surroundings (e.g., *tables*, *pavements*, *washing machines*, *stairs*, *benches*) could be considered members of the category *chair* despite the fact that such objects can be used as "objects on which one can sit."

The notions of typicality and graded membership are not limited to concrete or nominal concepts, as similar gradience effects are detected for more complex and abstract concepts (e.g., *"To what degree is THESIS an instance/type of STATEMENT?"*) (Coleman and Kay 1981), or action verbs (Pulman 1983) and adjectives (Dirven and Taylor 1986).



(a) ANIMAL (b) SPORT (c) TO MOVE

Figure 1
Toy examples illustrating the "typicality" or centrality of various class instances X of the Y classes (a) *animal*, (b) *sport*, (c) *(to) move*.

In short, graded membership or vagueness quantifies "whether or not and to what degree an instance falls within a conceptual category," whereas typicality reflects "how representative an exemplar is of a category" (Hampton 2007). The subtle distinction between the two is discussed and debated at length from the philosophical and psychological perspective (Osherson and Smith 1981; Kamp and Partee 1995; Osherson and Smith 1997; Hampton 2006, 2007; Blutner, Pothos, and Bruza 2013; Decock and Douven 2014). In our crowdsourcing study with non-expert workers, we have deliberately avoided any explicit differentiation between the two phenomena captured by the same intuitive "to-what-degree" question, reducing the complexity of the study design and allowing for their free variance in the collected data in terms of their quantity and representative concept pairs. In addition, the distinction is often not evident for verb concepts. We leave further developments with respect to the two related phenomena of typicality and vagueness for future work, and refer the interested reader to the aforementioned literature. However, we already provide preliminary qualitative analyses in this article (see Section 5) suggesting that *both* phenomena are captured in graded LE ratings.

Relation to Relational Similarity. A strand of related research on **relational similarity** (Turney 2006; Jurgens et al. 2012) also assigns the score s to a pair of concepts (X, Y). Note that there exists a fundamental difference between relational similarity and graded LE. In the latter, s refers to the degree of the LE relation in the (X, Y) pair, that is, to the levels of typicality and graded membership of the instance X for the class Y, whereas the former quantifies the typicality of the pair (X, Y) for some fixed lexical relation class R (Bejar, Chaffin, and Embretson 1991; Vylomova et al. 2016), for example, to what degree the pair *(snake, animal)* reflects a typical LE relation or a typical synonymy relation.[5]

Graded LE vs. Semantic Similarity. A plethora of current evaluations in NLP and representation learning almost exclusively focus on semantic similarity and relatedness. Semantic similarity as quantified by, for example, SimLex-999 or SimVerb-3500 (Gerz et al. 2016) may be redefined as **graded synonymy relation**. The graded scores there, in fact, refer to the strength of the synonymy relation between any pair of concepts (X, Y). One could say that semantic similarity aims to answer the question *to what degree X and Y are similar*.[6] Therefore, an analogy between previously annotated semantic similarity data sets and our objective to construct a graded LE data set may be utilized to introduce the graded LE task and facilitate the construction of HyperLex.

3. Design Motivation

3.1 Lexical Entailment Evaluations in NLP

Because the work in NLP and human language understanding focuses on the ungraded version of the LE relation, we briefly survey main ungraded LE evaluation protocols in Section 3.1.1, followed by an overview of benchmarking LE evaluation sets in

[5] For instance, given the lexical relation classification scheme of Bejar, Chaffin, and Embretson (1991), LE or CLASS–INCLUSION is only one of the 10 high-level relation classes.

[6] From the SimLex-999 guidelines: "Two words are synonymys if they have very similar meanings. Synonyms represent the same type or category (...) you are asked to compare word pairs and to rate how *similar* they are..." Synonymy and LE capture different aspects of meaning regarding semantic hierarchies/taxonomies: For example, whereas the pair *(mouse, rat)* receives a score of 7.78 in SimLex-999 (on the scale 0–10), the same pair has a graded LE score of 2.22 in HyperLex.

Section 3.1.2. We show that none of the existing evaluation protocols coupled with existing evaluation sets enables a satisfactory evaluation of the capability of statistical models to capture graded LE. As opposed to existing evaluation sets, by collecting human judgments through a crowdsourcing study our new HyperLex evaluation set also enables qualitative linguistic analysis on how humans perceive and rate graded lexical entailment.

3.1.1 Evaluation Protocols. Evaluation protocols for the lexical entailment or type-of relation in NLP, based on the classical definition of ungraded LE, may be roughly clustered as follows:

(i) Entailment Directionality. Given two words (X, Y) that are known to stand in a lexical entailment relation, the system has to predict the relation **directionality**, that is, which word is the hypernym and which word is the hyponym. More formally, the following mapping is defined by the directionality function f_{dir}:

$$f_{dir} : (X, Y) \rightarrow \{-1, 1\} \tag{1}$$

f_{dir} simply maps to 1 when Y is the hypernym, and to -1 otherwise.

(ii) Entailment Detection. The system has to predict whether there exists a lexical entailment relation between two words, or the words stand in some other relation (synonymy, meronymy–holonymy, causality, no relation, etc.). A more detailed overview of lexical relations is available in related work (Hendrickx et al. 2010; Jurgens et al. 2012; Vylomova et al. 2016). The following mapping is defined by the **detection** function f_{det}:

$$f_{det} : (X, Y) \rightarrow \{0, 1\} \tag{2}$$

f_{det} simply maps to 1 when (X, Y) stand in a lexical entailment relation, irrespective to the actual directionality of the relation, and to 0 otherwise.

(iii) Entailment Detection and Directionality. This recently proposed evaluation protocol (Weeds et al. 2014; Kiela et al. 2015) combines (i) and (ii). The system first has to detect whether there exists a lexical entailment relation between two words (X, Y), and then, if the relation holds, it has to predict its directionality (i.e., the correct hypernym). The following mapping is defined by the joint **detection and directionality** function $f_{det+dir}$:

$$f_{det+dir} : (X, Y) \rightarrow \{-1, 0, 1\} \tag{3}$$

$f_{det+dir}$ maps to 1 when (X, Y) stand in a lexical entailment relation and Y is the hypernym, to -1 if X is the hypernym, and to 0 if X and Y stand in some other lexical relation or no relation.

Standard Modeling Approaches. These decisions are typically based on the **distributional inclusion hypothesis** (Geffet and Dagan 2005) or a **lexical generality** measure (Herbelot and Ganesalingam 2013). The intuition supporting the former is that the class (i.e., *extension*) denoted by a hyponym is included in the class denoted by the hypernym, and therefore hyponyms are expected to occur in a subset of the contexts of their hypernyms. The intuition supporting the latter hints that typical characteristics constituting the *intension* (i.e., concept) expressed by a hypernym (e.g., *move* or *eat* for the concept *animal*) are semantically more general than the characteristics forming the intension[7] of its hyponyms (e.g., *bark* or *has tail* for the concept *dog*). In other words, superordinate concepts such as *animal* or *appliance* are semantically less informative than their hyponyms (Murphy 2003), which is also reflected in less specific contexts for hypernyms.

Unsupervised (distributional) models of lexical entailment were instigated by the early work of Hearst (1992) on prototypicality patterns (e.g., the pattern "X such as Y" indicates that Y is a hyponym of X). The current unsupervised models typically replace the symmetric cosine similarity measure that works well for semantic similarity computations (Bullinaria and Levy 2007; Mikolov et al. 2013a) with an asymmetric similarity measure optimized for entailment (Weeds, Weir, and McCarthy 2004; Clarke 2009; Kotlerman et al. 2010; Lenci and Benotto 2012; Herbelot and Ganesalingam 2013; Santus et al. 2014).

Supervised models, on the other hand, learn the asymmetric operator from a training set, differing mostly in the feature selection to represent each candidate pair of words (Baroni et al. 2012; Fu et al. 2014; Rimell 2014; Roller, Erk, and Boleda 2014; Weeds et al. 2014; Fu et al. 2015; Roller and Erk 2016; Shwartz, Goldberg, and Dagan 2016).[8] An overview of the supervised techniques also discussing their main shortcomings is provided by Levy et al. (2015); a thorough discussion of differences between unsupervised and supervised entailment models is provided by Turney and Mohammad (2015).

Why is HyperLex Different? In short, regardless of the chosen methodology, the evaluation protocols (directionality or detection) may be straightforwardly translated into binary decision problems: (1) distinguishing between hypernyms and hyponyms; (2) distinguishing between lexical entailment and other relations.

HyperLex, on the other hand, targets a different type of evaluation. The graded entailment function f_{graded} defines the following mapping:

$$f_{graded} : (X, Y) \rightarrow \mathbb{R}_0^+ \tag{4}$$

f_{graded} outputs the strength of the lexical entailment relation $s \in \mathbb{R}_0^+$.

By adopting the graded LE paradigm, HyperLex thus measures the degree of lexical entailment between words X and Y constituting the order-sensitive pair (X, Y). From another perspective, it measures the typicality and graded membership of the instance X for the class/category Y. From the relational similarity viewpoint (Jurgens et al. 2012; Zhila et al. 2013), it also measures the prototypicality of the pair (X, Y) for the LE relation.

3.1.2 Evaluation Sets

BLESS. Introduced by Baroni and Lenci (2011), the original **BLESS** evaluation set includes 200 concrete English nouns as target concepts (i.e., X-s from the pairs (X, Y)), equally divided between animate and inanimate entities. A total of 175 concepts were extracted from the McRae feature norms data set (McRae et al. 2005), and the remaining 25 were selected manually by the authors. These concepts were then paired to 8,625

[7] The terms *intension* and *extension* assume classical intensional and extensional definitions of a concept (van Benthem and ter Meulen 1996; Baronett 2012).

[8] Typical choices are feature vector concatenation ($\vec{X} \oplus \vec{Y}$), difference ($\vec{Y} - \vec{X}$), or element-wise multiplication ($\vec{X} \odot \vec{Y}$), where $\vec{X}$ and $\vec{Y}$ are feature vectors of concepts X and Y.

different relatums (i.e., Y-s), yielding a total of 26,554 (X, Y) pairs, where 14,440 contain a meaningful lexical relation and 12,154 are paired randomly. The lexical relations represented in BLESS are lexical entailment, co-hyponymy, meronymy, attribute, event, and random/no relation.

The use of its hyponymy–hypernymy/LE subset of 1,337 (X, Y) pairs is then 2-fold. First, for directionality evaluations (Santus et al. 2014; Kiela et al. 2015), only the LE subset is used. Note that the original BLESS data are always presented with the hyponym first, so gold annotations are implicitly provided here. Second, for detection evaluations (Roller, Erk, and Boleda 2014; Santus et al. 2014; Levy et al. 2015), the pairs from the LE subset are taken as positive pairs, and all the remaining pairs are considered negative pairs. That way, the evaluation data effectively measure a model's ability to predict the positive LE relation. Another evaluation data set based on BLESS was introduced by Santus et al. (2015). Following the standard annotation scheme, it comprises 7,429 noun pairs in total, and 1,880 LE pairs in particular, covering a wider range of relations than BLESS (i.e., the data set now includes synonymy and antonymy pairs).

Adaptations of the original BLESS evaluation set were proposed recently. First, relying on its LE subset, Weeds et al. (2014) created another data set called **WBLESS** (Kiela et al. 2015) consisting of 1,976 concept pairs in total. Only (X, Y) pairs where Y is the hypernym are annotated as positive examples. It also contains reversed LE pairs (i.e., X is the hypernym), cohyponymy pairs, meronymy–holonymy pairs, and randomly matched nouns balanced across different lexical relations; all are annotated as negative examples. Because of its construction, WBLESS is used solely for experiments on LE detection. Weeds et al. (2014) created another data set in a similar fashion, consisting of 5,835 noun pairs, targeting co-hyponymy detection.

For the combined detection and directionality evaluation, a variant evaluation set called **BiBLESS** was proposed (Kiela et al. 2015). It is built on WBLESS, but now explicitly distinguishes direction in LE pairs. Examples of concept pairs in all BLESS variants can be found in Table 1. A majority of alternative ungraded LE evaluation sets briefly discussed here have a structure very similar to BLESS and its variants.

Kotlerman et al. (2010). Based on the original data set of Zhitomirsky-Geffet and Dagan (2009), this evaluation set (Kotlerman et al. 2010) contains 3,772 word pairs in total. The structure is similar to BLESS: 1,068 pairs are labeled as positive examples (i.e., 1 or *entails* iff X entails Y), and 2,704 labeled as negative examples, including the reversed positive pairs. The assignment of binary labels is described in detail by Zhitomirsky-Geffet and Dagan (2009). The class sizes are not balanced, and because of its design, although each pair is unique, 30 high-frequent nouns occur in each pair in the data set. Note that

Table 1
Example pairs from BLESS data set variants.

Variant	Pair	Annotation
BLESS	(cat, animal)	1
WBLESS	(cat, animal)	1
	(cat, monkey)	0
	(animal, cat)	0
BiBLESS	(cat, animal)	1
	(cat, monkey)	0
	(animal, cat)	-1

this data set has been annotated according to the broader definition of substitutable LE (see Section 2).

Baroni et al. (2012). The $N_1 \models N_2$ evaluation set contains 2,770 nominal concept pairs, with 1,385 pairs labeled as positive examples (i.e., 1 or *entails*) (Baroni et al. 2012). The remaining 1,385 pairs labeled as negatives were created by inverting the positive pairs and randomly matching concepts from the positive pairs. The pairs and annotations were extracted automatically from WordNet and then validated manually by the authors (e.g., the abstract concepts with a large number of hyponyms such as *entity* or *object* were removed from the pool of concepts).

Levy et al. (2014). A similar data set for the standard LE evaluation may be extracted from manually annotated entailment graphs of subject–verb–object tuples (i.e., propositions) (Levy, Dagan, and Goldberger 2014): Noun LEs were extracted from entailing tuples that were identical except for one of the arguments, thus propagating the proposition-level entailment to the word level. This data set was built for the medical domain and adopts the looser definition of substitutable LE.

Custom Evaluation Sets. A plethora of relevant work on ungraded LE do not rely on established evaluation resources, but simply extract ad hoc LE evaluation data using distant supervision from readily available semantic resources and knowledge bases such as WordNet (Miller 1995), DBPedia (Auer et al. 2007), Freebase (Tanon et al. 2016), Yago (Suchanek, Kasneci, and Weikum 2007), or dictionaries (Gheorghita and Pierrel 2012). Although plenty of the custom evaluation sets are available online, there is a clear tendency to construct a new custom data set in every subsequent paper that uses the same evaluation protocol for ungraded LE.

A standard practice (Snow, Jurafsky, and Ng 2004, 2006; Bordes et al. 2011; Riedel et al. 2013; Socher et al. 2013; Weeds et al. 2014; Shwartz, Goldberg, and Dagan 2016; Vendrov et al. 2016, inter alia) is to extract positive and negative pairs by coupling concepts that are directly related in at least one of the resources. Only pairs standing in an unambiguous hypernymy/LE relation, according to the set of indicators from Table 2, are annotated as positive examples (i.e., again 1 or *entailing*, Table 1) (Shwartz et al. 2015). All other pairs standing in other relations are taken as negative instances. Using related rather than random concept pairs as negative instances enables detection experiments. We adopt a similar construction principle regarding wide coverage of different lexical relations in HyperLex. This decision will support a variety of interesting analyses related to graded LE and other relations.

Table 2
Indicators of LE/hypernymy relation in structured semantic resources.

Resource	Relation
WordNet	instance hypernym, hypernym
Wikidata	subclass of, instance of
DBPedia	type
Yago	subclass of

Jurgens et al. (2012). Finally, the evaluation resource most similar in spirit to HyperLex is the data set of Jurgens et al. (2012) (`https://sites.google.com/site/semeval2012task2/`) created for measuring degrees of relational similarity. It contains 3,218 word pairs labeled with 79 types of lexical relations from the relation classification scheme of Bejar et al. (1991).

The data set was constructed using two phases of crowdsourcing. First, for each of the 79 subcategories, human subjects were shown paradigmatic examples of word pairs in the given subcategory. They were then asked to generate more pairs of the same semantic relation type. Second, for each of the 79 subcategories, other subjects were shown word pairs that were generated in the first phase, and they were asked to rate the pairs according to their degree of prototypicality for the given semantic relation type. This is different from HyperLex, where all word pairs, regardless of their actual relation, were scored according to the degree of lexical entailment between them.

Bejar et al.'s hierarchical classification system contains ten high-level categories, with five to ten subcategories each. Only one high-level category, CLASS–INCLUSION, refers to the true relation of ungraded LE or hyponymy–hypernymy, and the scores in the data set do not reflect graded LE. The data set aims at a wide coverage of different fine-grained relations: It comprises a small sample of manually generated instances (e.g., the number of distinct pairs for the CLASS–INCLUSION class is 200) for each relation scored according to their prototypicality only for that particular relation. For more details concerning the construction of the evaluation set, we refer the reader to the original work. Also, for details on how to convert the data set to an evaluation resource for substitutable LE, we refer the reader to the work of Turney and Mohammad (2015).

HyperLex: A Short Summary of Motivation. The usefulness of these evaluation sets is evident from their wide usage in the LE literature over recent years: They guided the development of semantic research focused on taxonomic relations. However, none of the evaluation sets contains graded LE ratings. Therefore, HyperLex may be considered as a more informative data collection: It enables a new evaluation protocol focused on gradience of the TYPE-OF relation rooted in cognitive science (Hampton 2007). As discussed in Section 2, graded annotations from HyperLex may be easily converted to ungraded annotations: HyperLex may also be used in the standard format of previous LE evaluation sets (see Table 1) for detection and directionality evaluation protocols (see later in Section 7.2).

Second, a typical way to evaluate word representation quality at present is by judging the similarity of representations assigned to similar words. The most popular semantic similarity evaluation sets such as SimLex-999 or SimVerb-3500 consist of word pairs with similarity ratings produced by human annotators. HyperLex is the first resource that can be used for the intrinsic evaluation (Schnabel et al. 2015; Faruqui et al. 2016) of LE-based vector space models (Vendrov et al. 2016); see later in Section 6.6. Encouraged by high inter-annotator agreement scores and evident large gaps between the human and system performance (see Section 7), we believe that HyperLex will guide the development of a new generation of representation-learning architectures that induce hypernymy/LE-specialized word representations, as opposed to the current ubiquitous word representations targeting exclusively semantic similarity and/or relatedness (see later the discussion in Sections 7.4 and 8).

Finally, HyperLex provides a wide coverage of different semantic phenomena related to LE: graded membership vs. typicality (see Section 2), entailment depths, concreteness levels, word classes (nouns and verbs), word pairs standing in other lexical relations, etc. Besides its primary purpose as an evaluation set, such a large-scale and diverse crowdsourced semantic resource (2,616 pairs in total!) enables novel linguistic and cognitive science analyses regarding human typicality and vagueness judgments, as well as taxonomic relationships (discussed in Section 5).

4. The HyperLex Data Set

Construction Criteria. Hill, Reichart, and Korhonen 2015 argue that comprehensive high-quality evaluation resources have to satisfy the following three criteria:

- *(C1) Representative*: The resource covers the full range of concepts occurring in natural language.
- *(C2) Clearly defined*: A clear understanding is needed of what exactly the gold standard measures; that is, the data set has to precisely define the annotated relation (e.g., relatedness as with WordSim-353, similarity as with SimLex-999, or in this case *graded lexical entailment*).
- *(C3) Consistent and reliable*: Untrained native speakers must be able to quantify the target relation consistently relying on simple instructions.

The choice of word pairs and construction of the evaluation set were steered by these requirements. Criterion C1 was satisfied by sampling a sufficient number of pairs from the University of Southern Florida (USF) Norms data set (Nelson, McEvoy, and Schreiber 2004). As shown in prior work (Hill, Reichart, and Korhonen 2015), the USF data set provides an excellent range of different semantic relations (e.g., synonyms vs. hypernyms vs. meronyms vs. cohyponyms) and semantic phenomena (e.g., it contains concrete vs. abstract word pairs, noun pairs vs. verb pairs). This, in turn, guarantees a wide coverage of distinct semantic phenomena in HyperLex. We discuss USF and the choice of concept words in more detail in Section 4.1.

Criteria C2 and C3 were satisfied by providing clear and precise annotation guidelines that accurately outline the lexical entailment relation and its graded variant in terms of the synonymous definition based on the TYPE-OF relationship (Fromkin, Rodman, and Hyams 2013) for average native speakers of English without any linguistic background. We discuss the guidelines and questionnaire structure in Sections 4.3 and 4.4.

Final Output. The HyperLex evaluation set contains **noun pairs** (2,163 pairs) and **verb pairs** (453 pairs) annotated for the strength of the lexical entailment relation between the words in each pair. Because the LE relation is asymmetric and the score always quantifies to what degree X is a type of Y, pairs (X, Y) and (Y, X) are considered distinct pairs. Each concept pair is rated by at least 10 human raters. The rating scale goes from 0 (no type-of relationship at all) to 10 (perfect type-of relationship). Several examples from HyperLex are provided in Table 3.

In its 2,616 word pairs, HyperLex contains 1,843 distinct noun types and 392 distinct verb types. In comparison, SimLex-999 as the standard crowdsourced evaluation benchmark for representation learning architectures focused on the synonymy relation contains 751 distinct nouns and 170 verbs in its 999 word pairs. In another comparison, the LE benchmark BLESS (see Section 3.1.2) contains relations where one of the words in each pair comes from the set of 200 distinct concrete noun types.

Table 3
Example word pairs from HyperLex. The order of words in each pair is fixed, for example, the pair *chemistry / science* should be read as *"Is CHEMISTRY a type of SCIENCE?"*

Pair	HyperLex LE Rating
chemistry / science	10.0
motorcycle / vehicle	9.85
pistol / weapon	9.62
to ponder / to think	9.40
to scribble / to write	8.18
gate / door	6.53
thesis / statement	6.17
to overwhelm / to defeat	4.75
shore / beach	3.33
vehicle / motorcycle	1.09
enemy / crocodile	0.33
ear / head	0.00

4.1 Choice of Concepts

Sources: USF and WordNet. To ensure a wide coverage of a variety of semantic phenomena (C1), the choice of candidate pairs is steered by two standard semantic resources available online: (1) the USF norms data set[9] (Nelson, McEvoy, and Schreiber 2004), and (2) WordNet[10] (Miller 1995).

USF was used as the primary source of concept pairs. It is a large database of free association data collected for English, generated by presenting human subjects with one of 5,000 cue concepts and asking them to write the first word coming to mind that is associated with that concept. Each cue c was normed in this way by over 10 participants, resulting in a set of associates a for each cue, for a total of over 72,000 (c, a) pairs. For each such pair, the proportion of participants who produced associate a when presented with cue c can be used as a proxy for the strength of association between the two words.

The norming process guarantees that two words in a pair have a degree of semantic association that correlates well with semantic relatedness, reflected in different lexical relations between words in the pairs. Inspecting the pairs manually revealed a good range of semantic relationship values represented—for example, there were examples of ungraded LE pairs (*car / vehicle*, *biology / science*), cohyponym pairs (*peach / pear*), synonyms or near-synonyms (*foe / enemy*), meronym–holonym pairs (*heel / boot*), and antonym pairs (*peace / war*). USF also covers different POS categories: nouns (*winter / summer*), verbs (*to elect / to select*), and adjectives (*white / gray*), at the same time spanning word pairs at different levels of concreteness (*panther / cat* vs. *wave / motion* vs. *hobby / interest*). The rich annotations of the USF data (e.g., concreteness scores, association strength) can be combined with graded LE scores to yield additional analyses and insight.

WordNet was used to automatically assign a fine-grained lexical relation to each pair in the pool of candidates; this guided the sampling process to ensure a wide coverage of word pairs standing in different lexical relations (Shwartz, Goldberg, and Dagan 2016).

[9] http://w3.usf.edu/FreeAssociation/.
[10] https://wordnet.princeton.edu/.

Lexical Relations. To guarantee the coverage of a wide range of semantic phenomena, we have conditioned the cohort/pool used for sampling on the lexical relation between the words in each pair. As mentioned earlier, the information was extracted from WordNet. We consider the following lexical relations in HyperLex:

1. `hyp-N`: (X, Y) pairs where X is a hyponym of Y according to WordNet. N is the path length between the two concepts in the WordNet hierarchy, for example, the pair *cathedral / building* is assigned the `hyp-3` relation. Because of unavailability of a sufficient number of pairs for longer paths, we have grouped all pairs with the path length ≥ 4 into a single class `hyp` ≥ 4. It was shown that pairs that are separated by fewer levels in the WordNet hierarchy are both more strongly associated and rated as more similar (Hill, Reichart, and Korhonen 2015). This fine-grained division over LE levels enables analyses based on the semantic distance in a concept hierarchy.
2. `rhyp-N`: The same as `hyp-N`, now with the order reversed: X is now a hypernym of Y. Such pairs were included to investigate the inherent asymmetry of the type-of relation and how human subjects perceive it.
3. `cohyp`: X and Y are two instances of the same category, that is, they share a hypernym (e.g., *dog* and *elephant* are instances of the category *animal*). For simplicity, we retain only (X, Y) pairs that share a direct hypernym.
4. `mero`: It denotes the PART–WHOLE relation, where X always refers to the meronym (i.e., PART), and Y to the holonym (i.e., WHOLE): *finger / hand*, *letter / alphabet*. By its definition, this relation is observed only between nominal concepts.
5. `syn`: X and Y are synonyms and near-synonyms, for example, *movement / motion*, *attorney / lawyer*. In case of polysemous concepts, at least one sense has to be synonymous with a meaning of the other concept, for example, *author / writer*.
6. `ant`: X and Y are antonyms, for example, *beginning / end*, *to unite / to divide*.
7. `no-rel`: X and Y do not stand in any lexical relation, including the ones not present in HyperLex (e.g., causal relations, space–time relations), and are also not semantically related. This relation specifies that there is no apparent semantic connection between the two concepts at all, for example, *chimney / swan*, *nun / softball*.

POS Category. HyperLex includes subsets of pairs from two principle meaning-bearing POS categories: nouns and verbs.[11] This decision will enable finer-grained analyses based on the two main POS categories. It is further supported by recent research in distributional semantics showcasing that different word classes (e.g., nouns vs. verbs) require different modeling approaches and distributional information to reach per-class peak performances (Schwartz, Reichart, and Rappoport 2015). In addition, we expect verbs to have fuzzier category borders because of their high variability and polysemy,

[11] We have decided to leave out adjectives: They are represented in USF to a lesser extent than nouns and verbs, and it is thus not possible to sample large enough subsets of adjective pairs across different lexical relations and lexical entailment levels, namely, only `syn` and `ant` adjective pairs are available in USF.

increased abstractness, and a wide range of syntactic–semantic behavior (Jackendoff 1972; Levin 1993; Gerz et al. 2016).

Pools of Candidate Concept Pairs. The initial pools for sampling were selected as follows. First, we extracted all possible noun pairs (N / N) and verb pairs (V / V) from USF based on the associated POS tags available as part of USF annotations. Concept pairs of other and mixed POS (e.g., *puzzle / solve*, *meet / acquaintance*) were excluded from the pool of candidate pairs.[12] To ensure that semantic association between concepts in a pair is not accidental, we then discarded all such USF pairs that had been generated by two or fewer participants in the original USF experiments.[13] We also excluded all concept pairs containing a multi-word expression (e.g., *put down / insult*, *stress / heart attack*), pairs containing a named entity (e.g., *Europe / continent*), and pairs containing a potentially offensive concept (e.g., *weed / pot*, *heroin / drug*).[14]

All remaining pairs were then assigned a lexical relation according to WordNet. In case of duplicate (X, Y) and (Y, X) pairs, only one variant (i.e., (X, Y)) was retained. In addition, all `rhyp-N` pairs at this stage were reversed into `hyp-N` pairs. All `no-rel` pairs from USF were also discarded at this stage to prevent the inclusion of semantically related pairs with a prominent degree of association in the `no-rel` subset of HyperLex.

In the final step, all remaining pairs were divided into per-relation pools of candidate noun and verb pairs for each represented relation: `hyp-N`, `cohyp`, `mero`, `syn`, `ant`. Two additional pools were created for `rhyp-N` and `no-rel` after the sampling process.

4.2 Sampling Procedure

The candidate pairs were then sampled from the respective per-relation pools. The final number of pairs per each relation and POS category was influenced by: (1) the number of candidates in each pool (therefore, HyperLex contains significantly more noun pairs); (2) the focus on LE (therefore, HyperLex contains more `hyp-N` pairs at different LE levels); (3) the wide coverage of most prominent lexical relations (therefore, each lexical relation is represented by a sufficient number of pairs); and (4) logistic reasons (we were unable to rate all candidates in a crowdsourcing study and had to sample a representative subset of candidates for each relation and POS category in the first place).

Step 1: Initial Sampling. First, pairs for lexical relations `hyp-N`, `cohyp`, `mero`, `syn`, and `ant` were sampled from their respective pools. WordNet, although arguably the best choice for our purpose, is not entirely reliable as a gold standard resource, with occasional inconsistencies and debatable precision regarding the way lexical relations have been encoded: for example, *silly* is a hyponym of *child* according to WordNet. Therefore, all sampled pairs were manually checked by the authors plus two native English speakers in several iterations. Only such sampled pairs where the majority of human checkers agreed on the lexical relation were retained. If a pair was discarded, another substitute pair was randomly sampled if available, and again verified against human judgments.

[12] POS categories are generally considered to reflect very broad ontological classes (Fellbaum 1998). We thus felt it would be very difficult, or even counter-intuitive, for annotators to rate mixed POS pairs.

[13] The numbers are available as part of USF annotations.

[14] Note that the pairs with the same concept without any offensive connotation were included in the pools, for example, *weed / grass*, *weed / plant*, or *ecstasy / feeling*.

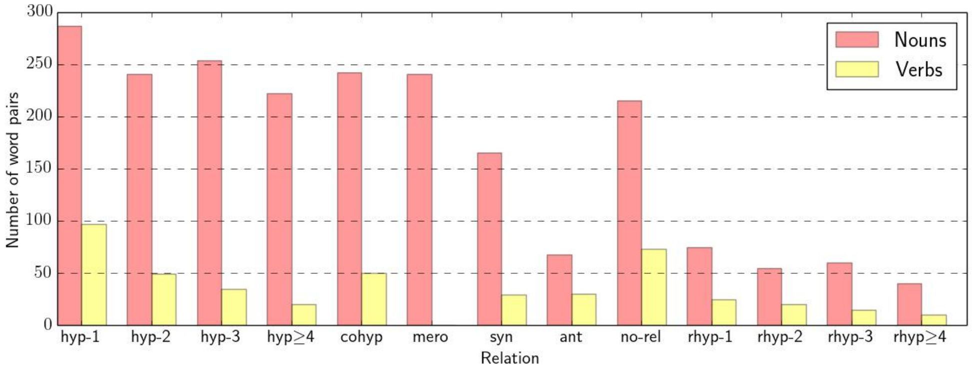


Figure 2
A total number of noun and verb pairs in HyperLex representing different fine-grained semantic relations extracted from WordNet.

Step 2: Reverse and No-Rel Pairs. Before the next step, the pool for `rhyp-N` was generated by simply reversing the order of concepts in all previously sampled (X, Y) `hyp-N` pairs. The pool for `no-rel` was generated by pairing up the concepts from the pairs extracted in Step 1 at random using the Cartesian product. From these random parings, we excluded those that coincidentally occurred elsewhere in USF (and therefore had a degree of association), as well as those that were assigned any lexical relation according to WordNet. From the remaining pairs, we accepted only those in which both concepts had been subject to the USF norming procedure, ensuring that these non-USF pairs were indeed unassociated rather than simply not normed. `rhyp-N` and `no-rel` were then sampled from these two pools, followed by another manual check. The `rhyp-N` pairs will be used to test the asymmetry of human and system judgments (see later in Tables 7 and 8), which is immanent to the LE relation.

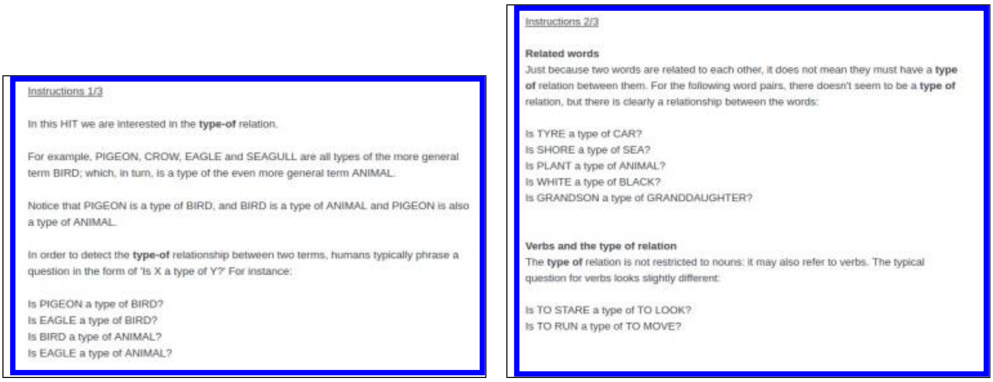
Figure 2 shows the exact numbers of noun and verb pairs across different lexical relations represented in HyperLex. The final set of 2,616 distinct word pairs[15] was then annotated in a crowdsourcing study (Sections 4.3 and 4.4).

4.3 Question Design and Guidelines

Here, we detail the exact annotation guidelines followed by the participants in the crowdsourcing study. In order to accurately outline the LE relation to average native speakers of English without any linguistic background, we have deliberately eschewed the usage of any expert linguistic terminology in the annotation guidelines, and have also avoided addressing the subtle differences between typicality and vagueness (Section 2). For instance, terms such as *hypernymy*, *lexical entailment*, *prototypicality*, or *taxonomy* were never explicitly defined using any precise linguistic formalism.

(Page 1). We have adopted a simpler and more intuitive definition of lexical entailment instead, based on the *type-of* relationship between words in question (Fromkin, Rodman, and Hyams 2013), illustrated by a set of typical examples in the guidelines (see Figure 3(a)).

[15] The final number was obtained after randomly discarding a small number of pairs for each relation in order to distribute the pairs of both POS categories into tranches of equal size in the crowdsourcing study (see Section 4.4).



Instructions 1/3

In this HIT we are interested in the **type-of** relation.

For example, PIGEON, CROW, EAGLE and SEAGULL are all types of the more general term BIRD; which, in turn, is a type of the even more general term ANIMAL.

Notice that PIGEON is a type of BIRD, and BIRD is a type of ANIMAL and PIGEON is also a type of ANIMAL.

In order to detect the **type-of** relationship between two terms, humans typically phrase a question in the form of 'Is X a type of Y?' For instance:

Is PIGEON a type of BIRD?
Is EAGLE a type of BIRD?
Is BIRD a type of ANIMAL?
Is EAGLE a type of ANIMAL?

(a) Page 1

Instructions 2/3

Related words
Just because two words are related to each other, it does not mean they must have a **type of** relation between them. For the following word pairs, there doesn't seem to be a **type of** relation, but there is clearly a relationship between the words:

Is TYRE a type of CAR?
Is SHORE a type of SEA?
Is PLANT a type of ANIMAL?
Is WHITE a type of BLACK?
Is GRANDSON a type of GRANDDAUGHTER?

Verbs and the type of relation
The **type of** relation is not restricted to nouns: it may also refer to verbs. The typical question for verbs looks slightly different:

Is TO STARE a type of TO LOOK?
Is TO RUN a type of TO MOVE?

(b) Page 2(a)

Figure 3
Page 1 and Page 2(a) of HyperLex annotation guidelines.

(Page 2). Following that, a clear distinction was made between words standing in a broader relationship of *semantic relatedness* and words standing in an actual type-of relation (see Fig. 3(b)). We have included typical examples of related words without any entailment relation, including meronymy pairs (*tyre / car*), cohyponymy pairs (*plant / animal*), and antonymy pairs (*white / black*), and pairs in other lexical relations (e.g., *shore / sea*). Because HyperLex also contains verbs, we have also provided several examples for a type-of relation between verbs (see Fig. 3(b)).

Potential polysemy issues have been addressed by stating, using intuitive examples, that *two words stand in a type-of relation if any of their senses stand in a type-of relation*. However, we acknowledge that this definition is vague, and the actual disambiguation process was left to the annotators and their intuition as native speakers. A similar context-free rating was used in the construction of other word pair scoring data sets such as SimLex-999 or WordSim-353.[16] In the next step, we have explicitly stressed that the type-of relation is *asymmetric* (see Fig. 4(a)).

(Page 3). The final page explains the main idea behind graded lexical entailment, graded membership, and prototypical class instances, according to the theories from cognitive science (Rosch 1973, 1975; Lakoff 1990; Hampton 2007; Divjak and Arppe 2013) by providing another illustrative set of examples (see Fig. 4(b)). The main goals of the study were then quickly summarized in the final paragraph, and the annotators were reminded to think in terms of the type-of relationship throughout the study.

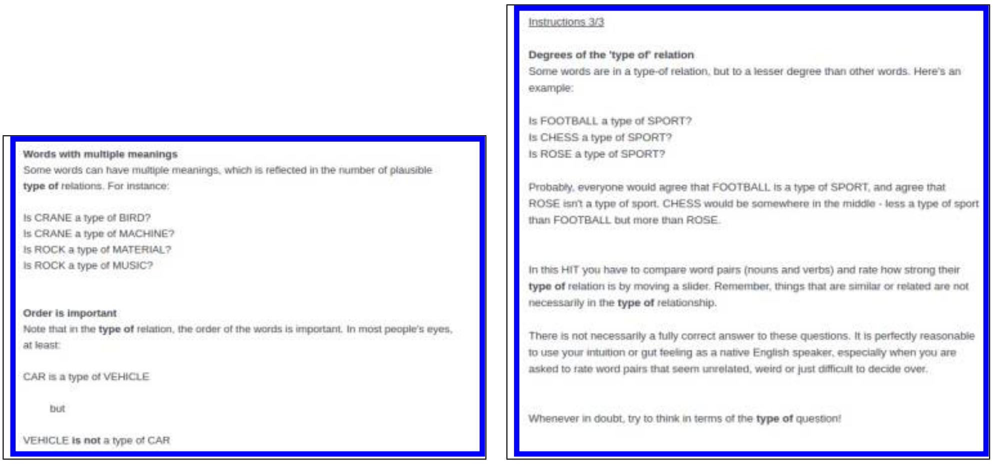
4.4 Questionnaire Structure and Participants

We use the Prolific Academic (PA) crowdsourcing platform,[17] an online marketplace very similar to Amazon Mechanical Turk and to CrowdFlower. Although PA was used to recruit participants, the actual questionnaire was hosted on Qualtrics.[18] Unlike other

[16] Determining the set of exact senses for a given concept, and then the set of contexts that represent those senses, introduces a high degree of subjectivity into the design process. Furthermore, in the infrequent case that some concept X in a pair (X, Y) is genuinely (etymologically) polysemous, Y can provide sufficient context to disambiguate X (Hill, Reichart, and Korhonen 2015; Leviant and Reichart 2015).

[17] `https://prolific.ac/`. (We chose PA for logistic reasons.)

[18] `https://www.qualtrics.com/`.



Words with multiple meanings
Some words can have multiple meanings, which is reflected in the number of plausible **type of** relations. For instance:

Is CRANE a type of BIRD?
Is CRANE a type of MACHINE?
Is ROCK a type of MATERIAL?
Is ROCK a type of MUSIC?

Order is important
Note that in the **type of** relation, the order of the words is important. In most people's eyes, at least:

CAR is a type of VEHICLE

but

VEHICLE **is not** a type of CAR

(a) Page 2(b)

Instructions 3/3

Degrees of the 'type of' relation
Some words are in a type-of relation, but to a lesser degree than other words. Here's an example:

Is FOOTBALL a type of SPORT?
Is CHESS a type of SPORT?
Is ROSE a type of SPORT?

Probably, everyone would agree that FOOTBALL is a type of SPORT, and agree that ROSE isn't a type of sport. CHESS would be somewhere in the middle - less a type of sport than FOOTBALL but more than ROSE.

In this HIT you have to compare word pairs (nouns and verbs) and rate how strong their **type of** relation is by moving a slider. Remember, things that are similar or related are not necessarily in the **type of** relationship.

There is not necessarily a fully correct answer to these questions. It is perfectly reasonable to use your intuition or gut feeling as a native English speaker, especially when you are asked to rate word pairs that seem unrelated, weird or just difficult to decide over.

Whenever in doubt, try to think in terms of the **type of** question!

(b) Page 3

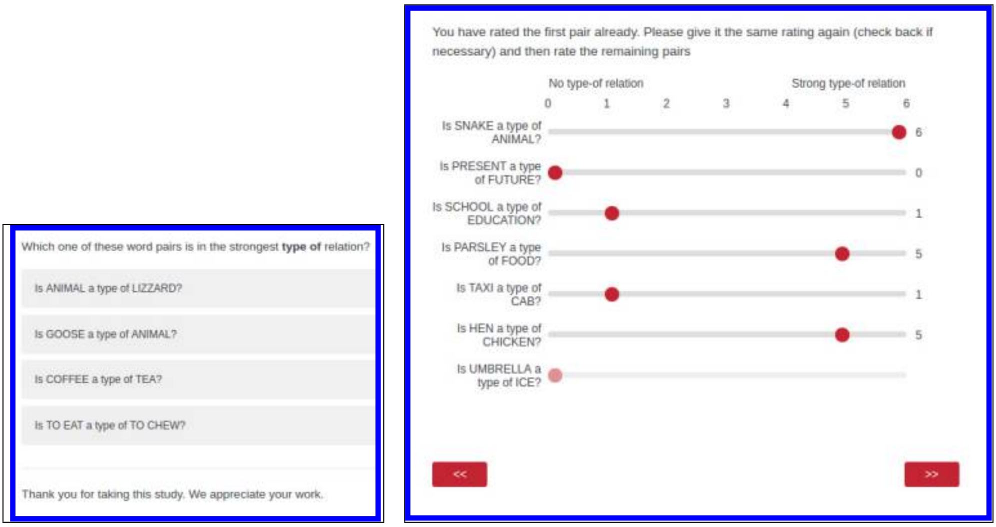
Figure 4
Page 2(b) and Page 3 of HyperLex annotation guidelines.

crowdsourcing platforms, PA collects and stores detailed demographic information from the participants upfront. This information was used to carefully select the pool of eligible participants. We restricted the pool to native English speakers with a 90% approval rate (maximum rate on PA), of age, 18–50 years, born and currently residing in the United States or the UK.

Immediately after the guidelines, similar to the SimLex-999 questionnaire, a *qualification question* is posed to the participant to test whether she/he understood the guidelines and is allowed to proceed with the questionnaire. The question is shown in Figure 5(a). In the case of an incorrect answer, the study terminates for the participant without collecting any ratings. In case of a correct answer, the participant begins rating pairs by moving a slider, as shown in Figure 5(b). Having a slider attached to the question *"Is X a type of Y?"* implicitly translates the posed question to the question *"To what degree is X a type of Y?"* (Section 2). The pairs are presented to the participant in groups of six or seven. As with SimLex-999, this group size was chosen because the (relative) rating of a set of pairs implicitly requires pairwise comparisons between all pairs in that set. Therefore, larger groups would have significantly increased the cognitive load on the annotators. Because concept pairs were presented to raters in batches defined according to POS, another advantage of grouping was the clear break (submitting a set of ratings and moving to the next page) between the tasks of rating noun and verb pairs. For better inter-group calibration, from the second group onward the last pair of the previous group became the first pair of the present group. The participants were then asked to re-assign the rating previously attributed to the first pair before rating the remaining new items (Figure 5(b)).

It is also worth stressing that we have decided to retain the type-of structure of each question explicitly for all word pairs so that raters are constantly reminded of the targeted lexical relation, that is, all (X, Y) word pairs are rated according to the question *"Is X a type of Y?"*, as shown in Figure 5(b). For verbs, we have decided to use the infinitive form in each question, for example, *"Is TO RUN a type of TO MOVE?"*

Following a standard practice in crowdsourced word pair scoring studies (Finkelstein et al. 2002; Luong, Socher, and Manning 2013; Hill, Reichart, and Korhonen 2015), each of the 2,616 concept pairs has to be assigned at least 10 ratings from 10 different accepted



Which one of these word pairs is in the strongest **type of** relation?

- Is ANIMAL a type of LIZZARD?
- Is GOOSE a type of ANIMAL?
- Is COFFEE a type of TEA?
- Is TO EAT a type of TO CHEW?

Thank you for taking this study. We appreciate your work.

You have rated the first pair already. Please give it the same rating again (check back if necessary) and then rate the remaining pairs

	No type-of relation 0 … 6 Strong type-of relation
Is SNAKE a type of ANIMAL?	6
Is PRESENT a type of FUTURE?	0
Is SCHOOL a type of EDUCATION?	1
Is PARSLEY a type of FOOD?	5
Is TAXI a type of CAB?	1
Is HEN a type of CHICKEN?	5
Is UMBRELLA a type of ICE?	

(a) Qualification question (b) Survey structure

Figure 5
(a) Qualification question. The correct answer is *"Is GOOSE a type of ANIMAL?"* (b) A group of noun pairs to be rated by moving the sliders. The rating slider was initially at position 0, and it was possible to attribute a rating of 0, although it was necessary to have actively moved the slider to that position to proceed to the next page. The first pair is repeated from the previous page, and the last pair will be repeated on the next page.

annotators. We collected ratings from more than 600 annotators in total. To distribute the workload, we divided the 2,616 pairs into 45 tranches, with 79 pairs each: 50 are unique to one tranche, and 20 manually chosen pairs are in all tranches to ensure consistency. The use of such consistency pairs enabled control for possible systematic differences between annotators and tranches, which could be detected by variation on this set of 20 pairs shared across all tranches. The remaining 9 are duplicate pairs displayed to the same participant multiple times to detect inconsistent annotations. The number of noun and verb pairs is the same across all tranches (64 of 79 and 15 of 79, respectively). Each annotator was asked to rate the pairs in a single tranche only. Participants took 10 minutes on average to complete one tranche, including the time spent reading the guidelines and answering the qualification question.

4.5 Post-Processing

Eighty-five percent of total exclusions occurred because of crowdsourcers answering the qualification question incorrectly: We did not collect any ratings from such workers. In the post-processing stage, we additionally excluded ratings of annotators who (a) did not give equal ratings to duplicate pairs; (b) showed suspicious rating patterns (e.g., randomly alternating between two ratings, using one single rating throughout the study, or assigning random ratings to pairs from the consistency set). The final acceptance rate was 85.7% (if we also count the workers who answered the qualification question incorrectly for the total number of assignments) and 97.5% (with such workers excluded from the counts). We then calculated the average of all ratings from the accepted raters (≥ 10) for each word pair. The score was finally scaled linearly from the 0–6 to the 0–10 interval as also done by Hill, Reichart, and Korhonen (2015).

Table 4
A comparison of HyperLex IAA with several prominent crowdsourced semantic similarity/relatedness evaluation benchmarks that also provide scores for word pairs. Numbers in parentheses refer to the total number of word pairs in each evaluation set.

Benchmark	IAA-1	IAA-2
WordSim (353) (Finkelstein et al. 2002)	0.611	0.756
WS-Sim (203) (Agirre et al. 2009)	0.667	0.651
SimLex (999) (Hill, Reichart, and Korhonen 2015)	0.673	0.778
HyperLex (2616)	**0.854**	**0.864**
HyperLex: Nouns (2163)	0.854	0.864
HyperLex: Verbs (453)	0.855	0.862

5. Analysis

Inter-Annotator Agreement. We report two different inter-annotator agreement (IAA) measures. **IAA-1 (pairwise)** computes the average pairwise Spearman's ρ correlation between any two raters. This is a common choice in previous data collection in distributional semantics (Padó, Padó, and Erk 2007; Reisinger and Mooney 2010a; Silberer and Lapata 2014; Hill, Reichart, and Korhonen 2015).

A complementary measure would smooth individual annotator effects. For this aim, our **IAA-2 (mean)** measure compares the average correlation of a human rater with the average of all the other raters. It arguably serves as a better "upper bound" than IAA-1 for the performance of automatic systems. HyperLex obtains $\rho = 0.854$ (IAA-1) and $\rho = 0.864$ (IAA-2), a very good agreement compared to other prominent crowdsourced benchmarks for semantic evaluation which also used word pair scoring (see Table 4).[19] We also report IAAs over different groups of pairs according to the relation extracted from WordNet in Table 5.

We acknowledge the fact that the grading process at places requires specific world-knowledge (e.g., *to what degree is SNAKE a type of REPTILE?*, *to what degree is TOMATO a type of FRUIT?*), or is simply subjective and demographically biased (e.g., *to what degree is TO PRAY a type of TO COMMUNICATE?*), which needs principled qualitative analyses. However, the HyperLex inter-rater agreement scores suggest that participants were able to understand the characterization of graded lexical entailment presented in the instructions and to apply it to concepts of various types (e.g. nouns vs. verbs, concrete vs. abstract concepts, different lexical relations from WordNet) consistently.

Typicality in Human Judgments. In the first analysis, we investigate the straightforward question: Are some concepts really more (proto)typical of semantically broader higher-level classes? Several examples of prominent high-level taxonomic categories, along with

[19] Note that the IAAs are not computed on the entire data set, but are in fact computed per tranche, as one worker annotated only one tranche. Exactly the same IAA computation was used previously by Hill, Reichart, and Korhonen (2015).

Table 5
Inter-annotator agreements, measured by average pairwise Spearman's ρ correlation over different fine-grained lexical relations extracted from WordNet.

	hyp−1	hyp−2	hyp−3	hyp≥4	cohyp	mero	syn	ant	no−rel	rhyp−1	rhyp−2	rhyp−3	rhyp≥4
IAA-1	0.850	0.844	0.859	0.848	0.857	0.856	0.860	0.858	0.854	0.855	0.842	0.868	0.856
IAA-2	0.866	0.847	0.872	0.851	0.875	0.876	0.883	0.858	0.859	0.845	0.850	0.846	0.859

LE scores, are shown in Table 6. We might draw several preliminary insights based on the presented lists. There is an evident prototyping effect present in human judgments: Concepts such as *cat*, *monkey*, or *cow* are more typical instances of the class *animal* than the more peculiar instances, such as *mongoose* or *snail*, according to HyperLex annotators. Instances of the class *sport* also seem to be sorted accordingly, as higher scores are assigned to arguably more prototypical sports such as *basketball*, *volleyball*, or *soccer*, and less prototypical sports such as *racquetball* or *wrestling* are assigned lower scores.

Nonetheless, the majority of `hyp-N` pairs (X, *animal*) or (X, *sport*), where X is a hyponym of *animal/sport* according to WordNet (WN), are indeed assigned reasonably high graded LE scores. It suggests that humans are able to: (1) judge the LE relation consistently and decide that a concept indeed stands in a type-of relation with another concept, and (2) grade the LE relation by assigning more strength to more prototypical class instances. Similar patterns are visible with other class instances from Table 6, as well as with other prominent nominal classes (e.g., *bird*, *appliance*, *science*). We also observe the same effect with verbs, for example, *(drift, move, 8.58), (hustle, move, 7.67), (tow, move, 7.37), (wag, move, 6.80), (unload, move, 6.22)*.

We also analyze if the effects of graded membership/vagueness (see the discussion in Section 2) are also captured in the ratings, and our preliminary qualitative analysis suggests so. For instance, an interesting example quantifies the graded membership in the class *group*: *(gang, group, 9.25), (legion, group, 7.67), (conference, group, 6.80), (squad, group, 8.33), (caravan, group, 5.00), (grove, group, 3.58), (herd, group, 9.23), (fraternity, group, 8.72), (staff, group, 6.28)*. Although we have not explicitly distinguished between typicality and graded membership in our annotation guidelines, with both subsumed under the TYPE-OF formulation of graded lexical entailment, the listed examples suggest that human subjects are able to quantify both in a satisfying manner.

Table 6
Graded LE scores for instances of several prominent taxonomical categories/classes represented in HyperLex (i.e., the categories are the word Y in each (X, Y, s) graded LE triplet).

TYPE-OF	**animal**		**food**		**sport**		**person**		**vehicle**
cat	10.0	sandwich	10.0	basketball	10.0	girl	9.85	car	10.0
monkey	10.0	pizza	10.0	hockey	10.0	customer	9.08	limousine	10.0
cow	10.0	rice	10.0	volleyball	10.0	clerk	8.97	motorcycle	9.85
bat	9.52	hamburger	9.75	soccer	9.87	citizen	8.63	van	9.75
mink	9.17	mushroom	9.07	baseball	9.75	nomad	8.63	automobile	9.58
snake	8.75	pastry	8.83	softball	9.55	poet	7.78	tractor	9.37
snail	8.62	clam	8.20	cricket	9.37	guest	7.22	truck	9.23
mongoose	8.33	snack	7.78	racquetball	9.03	mayor	6.67	caravan	8.33
dinosaur	8.20	oregano	5.97	wrestling	8.85	publisher	6.03	buggy	8.20
crab	7.27	rabbit	5.83	recreation	2.46	climber	5.00	bicycle	8.00
plant	0.13	dinner	4.85	–	–	idol	4.28	vessel	6.38

Hypernymy/LE Levels. Graded LE scores in HyperLex averaged for each WordNet relation are provided in Table 7. Note that the LE level is extracted as the shortest direct path between two concept words in the WordNet taxonomy, where X-s in each (X, Y) pair always refer to the less general concept (i.e., hyponym).

Graded LE scores for nouns increase with the increase of the LE level (i.e., WN path length) between the concepts. A longer WN path implies a clear difference in semantic generality between nominal concepts, which seems to be positively correlated with the degree of the LE relation and ease of human judgment. A similar finding in directionality and detection experiments on BLESS and its variants was reported by Kiela et al. (2015). They demonstrate that their model is less accurate on concepts with short paths (i.e., the lowest results are reported for WN `hyp-1` pairs from BLESS), and the performance increases with the increase of the WN path length. The tendency is explained by the lower difference in generality between concepts with short paths, which may be difficult to discern for a statistical model. The results from Table 7 show that human raters also display a similar tendency when rating nominal pairs.

Another factor underlying the observed scores might be the link between HyperLex and the source USF norms. Because USF contains free association norms, one might assume that more prototypical instances are generated more frequently as responses to cue words in the original USF experiments. This, in turn, reflects in their greater presence in HyperLex, especially for concept pairs with longer WN distances.

Further, nominal concepts higher in the WN hierarchy typically refer to semantically very broad but well-defined categories such as *animal*, *food*, *vehicle*, or *appliance* (see again Table 6). Semantically more specific instances of such concepts are easier to judge as *true* hyponyms (using the ungraded LE terminology), which also reflects in higher LE ratings for such instances. However, gradience effects are clearly visible even for pairs with longer WN distances (Table 6).

The behavior with respect to the LE level is reversed for verbs: the average scores decrease over increasing LE levels. We attribute this effect to a higher level of abstractness and ambiguity present in verb concepts higher in the WN hierarchy stemming from a fundamental cognitive difference: Gentner (2006) showed that children find verbs harder to learn than nouns, and Markman and Wisniewski (1997) present evidence that different

Table 7
Average HyperLex scores across all pairs, and noun and verb pairs representing finer-grained semantic relations extracted from WordNet.

	All	Nouns	Verbs
hyp-1	7.86	7.99	7.49
hyp-2	8.10	8.31	7.08
hyp-3	8.16	8.39	6.55
hyp≥4	8.33	8.62	5.12
cohyp	3.54	3.29	4.76
mero	3.14	3.14	-
syn	6.83	6.69	7.66
ant	1.47	1.57	1.25
no-rel	0.85	0.64	1.48
rhyp-1	4.75	4.17	6.45
rhyp-2	4.19	3.44	6.15
rhyp-3	3.07	2.72	4.47
rhyp≥4	2.85	2.54	4.11

Table 8
A selection of scored (X, Y) word pairs from HyperLex holding the `hyp-1`, `hyp-2`, and `hyp-3` relation according to WordNet along with the HyperLex score for the actual (X, Y) pair (*scr*) and the HyperLex score for the reversed (Y, X) pair (i.e., `rhyp-N` relations): *rscr*. The reported percentages on top refer to the ratio of (X, Y) pairs for each relation where *scr* > *rscr*.

hyp-1 vs rhyp-1 (89%)			hyp-2 vs rhyp-2 (95%)			hyp-3 vs rhyp-3 (96%)		
Pair	`scr`	`rscr`	Pair	`scr`	`rscr`	Pair	`scr`	`rscr`
(computer, machine)	**9.83**	2.43	(gravity, force)	**9.50**	3.58	(flask, container)	**9.37**	1.83
(road, highway)	**9.67**	4.30	(professional, expert)	**6.37**	6.03	(elbow, joint)	**7.18**	1.07
(dictator, ruler)	**9.87**	6.22	(therapy, treatment)	**9.17**	4.10	(nylon, material)	**9.75**	1.42
(truce, peace)	**8.00**	6.38	(encyclopedia, book)	**8.93**	2.22	(choir, group)	**8.72**	2.43
(remorse, repentance)	**7.63**	3.50	(empathy, feeling)	**8.85**	2.42	(beer, beverage)	**9.25**	0.67
(disagreement, conflict)	**8.78**	8.67	(shovel, tool)	**9.70**	2.57	(reptile, animal)	**9.87**	1.17
(navigator, explorer)	6.80	**7.63**	(fraud, deception)	**9.52**	8.17	(parent, ancestor)	**7.00**	6.17
(ring, jewelry)	**10.0**	2.78	(bed, furniture)	**9.75**	2.63	(note, message)	**9.00**	6.07
(solution, mixture)	6.52	**7.37**	(verdict, judgment)	**9.67**	7.57	(oven, appliance)	**9.83**	1.33
(spinach, vegetable)	**10.0**	0.55	(reader, person)	**7.43**	3.33	(king, leader)	**8.67**	4.55
(surgeon, doctor)	**8.63**	4.05	(vision, perception)	3.82	**6.25**	(hobby, activity)	**7.12**	6.83
(hint, suggestion)	**8.75**	7.03	(daughter, child)	**9.37**	2.78	(prism, shape)	**7.50**	2.70

cognitive operations are used when comparing two nouns or two verbs. For instance, it is intuitive to assume that human subjects find it easier to grade instances of the class *animal* than instances of verb classes such as *to get*, *to set*, or *to think*.

LE Directionality. Another immediate analysis investigates whether the inherent asymmetry of the type-of relation is captured by the human annotations in HyperLex. Several illustrative example pairs and their reverse pairs split across different LE levels are shown in Table 8. Two important conclusions may be drawn from the analysis.

First, human raters are able to capture the asymmetry; the strong majority of `hyp-N` pairs is rated higher than their `rhyp-N` counterparts: 94% of all `hyp-N` pairs for which exists the `rhyp-N` counterpart are assigned a higher rating. Second, the ability to clearly detect the correct LE direction seems to rise with the increase of semantic distance in WordNet: (1) We notice decreasing average scores for the `rhyp-N` relation as N increases (see Table 7); (2) We notice a higher proportion of `hyp-N` concept pairs scoring higher than their `rhyp-N` counterparts as N increases (see Table 8). There are evident difficulties to decide on the entailment direction with several pairs (e.g., *navigator / explorer*, *solution / mixture*, *disagreement / conflict*), especially for the taxonomically closer `hyp-1` pairs, a finding aligned with prior work on LE directionality (Rimell 2014; Kiela et al. 2015).

Other Lexical Relations. Another look into Table 7, where graded LE scores are averaged across lexical relations, indicates the expected order of all other lexical relations sorted by the average per-relation scores (i.e., `syn` > `cohyp` > `mero` > `ant` > `no-rel`). `no-rel` and `ant` pairs have the lowest graded LE scores by a large margin. `no-rel` pairs are expected to have completely non-overlapping semantic fields, which facilitates human judgment. With antonyms, the graded LE question may be implicitly reformulated as *To what degree is X a type of* $\neg X$*?* (e.g., *winner / loser, to depart / to arrive*), which intuitively should result in low graded LE scores: The HyperLex ratings confirm the intuition.

Low scores for `cohyp` pairs in comparison to `hyp-N` pairs indicate that the annotators are able to effectively distinguish between the two related but different well-defined taxonomic relations (i.e., `hyp-N` vs `cohyp`). High scores for `syn` pairs are also aligned with our expectations and agree with intuitions from prior work on ungraded LE (Rei and

Table 9
A total number of concept pairs in each of the four coarse-grained groups based on the concepts' concreteness ratings: Both concepts are concrete (USF concreteness rating ≥ 4) $\rightarrow G_1$; both abstract $\rightarrow G_2$; one concrete and one abstract concept with a difference in concreteness $\leq 1 \rightarrow G_3$ or $> 1 \rightarrow G_4$. `rhyp-N` pairs are not counted.

	G1	G2	G3	G4
# Pairs	979	259	883	344

Briscoe 2014). In a slightly simplified view, given that two synonyms may be observed as two different utterances of the same semantic concept X, the graded LE question may be rephrased as *To what degree is X a type of X?* One might say that `syn` could be seen as a special case: the degenerate taxonomic `hyp-0` relation. Such an implicit reformulation of the posed question naturally results in higher scores for `syn` pairs on average.

Concreteness. Differences in human and computational concept learning and representation have been attributed to the effects of **concreteness**, the extent to which a concept has a directly perceptible physical referent (Paivio 1991; Hill, Korhonen, and Bentz 2014). Because the main focus of this work is not on the distinction between abstract and concrete concepts, we have not explicitly controlled for the balanced amount of concrete/abstract pairs in HyperLex. However, because the source USF data set provides concreteness scores, we believe that HyperLex will also enable various additional analyses regarding this dimension in future work.

Here, we report the number of pairs in four different groups based on concreteness ratings of two concepts in each pair. The four groups are as follows: (G_1) both concepts are concrete (USF concreteness rating ≥ 4); (G_2) both concepts are abstract (USF rating $<$ 4); (G_3) one concept is considered concrete and another abstract, with their difference in ratings ≤ 1; (G_4) one concept is considered concrete and another abstract, with their difference in ratings > 1.

The statistics regarding HyperLex pairs divided into groups $G_1 - G_4$ is presented in Table 9. `rhyp-N` pairs are not counted as they are simply reversed `hyp-N` pairs present in HyperLex. Concept pairs where at least one concreteness rating is missing in the USF data are also not taken into account. Although HyperLex contains more concrete pairs overall, there is also a large sample of highly abstract pairs and mixed pairs. For instance, HyperLex contains 125 highly abstract concept pairs, with both concepts scoring ≤ 3 in concreteness (e.g., *misery / sorrow, hypothesis / idea, competence / ability*, or *religion / belief*). This preliminary coarse-grained analysis already hints that HyperLex provides a good representation of concepts across the entire concreteness scale. This will also facilitate further analyses related to concept concreteness and its influence on the automatic construction of semantic taxonomies.

Data Splits: Random and Lexical. A common problem in scored/graded word pair data sets is the lack of a standard split to development and test sets (Faruqui et al. 2016). Custom splits (e.g., 10-fold cross-validation) make results incomparable with others. Further, because of their limited size, they also do not support supervised learning, and do not provide splits into training, development, and test data. The lack of standard splits in such word pair data sets stems mostly from small size and poor coverage—issues that we have solved with HyperLex.

Table 10
HyperLex data splits: Basic statistics. The number of pairs is always provided in the following format #Overall (#N + #V), where #N and #V denote the number of noun and verb pairs respectively. The columns represent groups/buckets of word pairs according to their LE scores.

	Rating Intervals					
Split	All	$[0,2\rangle$	$[2,4\rangle$	$[4,6\rangle$	$[6,8\rangle$	$[8,10]$
HyperLex–All	2616 (2163+453)	604 (504+100)	350 (304+46)	307 (243+64)	515 (364+151)	840 (748+92)
Random Split						
Train	1831 (1514+317)	423 (353+70)	245 (213+32)	215 (170+45)	361 (255+106)	587 (523+64)
Dev	130 (108+22)	30 (25+5)	17 (15+2)	15 (13+2)	26 (18+8)	42 (37+5)
Test	655 (541+114)	151 (126+25)	88 (76+12)	77 (60+17)	128 (91+37)	211 (188+23)
Lexical Split						
Train	1133 (982+151)	253 (220+33)	140 (122+18)	129 (109+20)	195 (148+47)	416 (383+33)
Dev	85 (71+14)	20 (18+2)	13 (11+2)	11 (8+3)	17 (10+7)	24 (24+0)
Test	269 (198+71)	65 (52+13)	37 (29+8)	41 (31+10)	63 (37+26)	63 (49+14)

We provide two standard data splits into train, dev, and test data: **random** and **lexical**. In the random split, 70% of all pairs were reserved for training, 5% for development, and 25% for testing. The subsets were selected by random sampling, but controlling for a broad coverage in terms of similarity ranges (i.e., non-similar and highly similar pairs, as well as pairs of medium similarity are represented). Some statistics are available in Table 10. A manual inspection of the subsets revealed that a good range of lexical relations is represented in the subsets.

The lexical split, advocated by Levy et al. (2015) and Shwartz, Goldberg, and Dagan (2016), prevents the effect of **lexical memorization**: supervised distributional lexical inference models tend to learn an independent property of a single concept in the pair instead of learning a relation between the two concepts.[20] To prevent such behavior, we split HyperLex into a train and test set with zero lexical overlap. We tried to retain roughly the same 70%/25%/5% ratio in the lexical split. Note that the lexical split discards all "cross-set" training–test concept pairs. Therefore, the number of instances in each subset is lower than with the random split. Statistics are again given in Table 10.

We believe that the provided standardized HyperLex data splits will enable easy and direct comparisons of various LE modeling architectures in unsupervised and supervised settings. Following arguments from prior work, we hold that it is important to provide both data set splits, as they can provide additional possibility to assess differences between models. It is true that training a model on a lexically split data set may result in a more general model (Levy et al. 2015), which is able to better reason over pairs consisting of two unseen concepts during inference. However, Shwartz, Goldberg, and Dagan (2016) argue that a random split emulates a more typical "real-life" reasoning scenario, where inference involves an unseen concept pair (X, Y), in which X and/or Y have already been observed separately. Models trained on a random split may introduce the model with a concept's "prior belief" of being a frequent hypernym or a hyponym. This information can be effectively exploited during inference.

[20] For instance, if the training set contains concept pairs *(dog / animal)*, *(cow, animal)*, and *(cat, animal)*, all assigned very high LE scores or annotated as positive examples in case of ungraded LE evaluation, the algorithm may learn that *animal* is a prototypical hypernym, assigning any new $(X, animal)$ pair a very high score, regardless of the actual relation between X and *animal*; additional analyses provided in Section 7.3.

6. Evaluation Set-up and Models

Evaluation Set-up. We compare the performance of prominent models and frameworks focused on modeling lexical entailment on our new HyperLex evaluation set now measuring the strength of the lexical entailment relation. Because of the evident similarity of the graded evaluation with standard protocols in the semantic similarity (i.e., synonymy detection) literature (Finkelstein et al. 2002; Agirre et al. 2009; Hill, Reichart, and Korhonen 2015; Schwartz, Reichart, and Rappoport 2015, inter alia), we adopt the same evaluation set-up. Each evaluated model assigns a score to each pair of words measuring the strength of lexical entailment relation between them.[21]

As in prior work on intrinsic evaluations with word pair scoring evaluation sets, all reported scores are Spearman's ρ correlations between the ranks derived from the scores of the evaluated models and the human scores provided in HyperLex. In this work, we evaluate off-the-shelf **unsupervised models** and insightful baselines on the entire HyperLex. We also report on preliminary experiments exploiting provided data splits for **supervised learning**.

6.1 Directional Entailment Measures

Note that all **directional entailment measures** (DEMs) available in the literature have "pre-embedding" origins and assume traditional count-based vector spaces (Turney and Pantel 2010; Baroni, Dinu, and Kruszewski 2014) based on counting word-to-word corpus co-occurrence. Distributional features are typically words co-occurring with the target word in a chosen context (e.g., a window of neighboring words, a sentence, a document, a dependency-based context).

This collection of models is grounded on variations of the distributional inclusion hypothesis (Geffet and Dagan 2005): If X is a semantically narrower term than Y, then a significant number of salient distributional features of X are included in the feature vector of Y as well. We closely follow the work from Lenci and Benotto (2012) in the presentation. Let $Feat_X$ denote the set of distributional features ft for a concept word X, and let $w_X(ft)$ refer to the weight of the feature ft for X. The most common choices for the weighting function in traditional count-based distributional models are positive variants of pointwise mutual information (PMI) (Bullinaria and Levy 2007) and local mutual information (LMI) (Evert 2008).

WeedsPrec (DEM_1). This DEM quantifies the weighted inclusion of the features of a concept word X within the features of a concept word Y (Weeds and Weir 2003; Weeds, Weir, and McCarthy 2004; Kotlerman et al. 2010):

$$DEM_1(X,Y) = \frac{\sum_{ft \in Feat_X \cap Feat_Y} w_X(ft)}{\sum_{ft \in Feat_X} w_X(ft)} \tag{5}$$

WeedsSim (DEM_2). It computes the geometrical average of WeedsPrec (DEM_1) or any other asymmetric measure (e.g., APinc from Kotlerman et al. [2010]) and the symmetric

[21] Note that, unlike with similarity scores, the score now refers to an asymmetric relation stemming from the question *"Is X a type of Y"* for the word pair (X,Y). Therefore, the scores for two reverse pairs (X,Y) and (Y,X) should be different; see also Table 8.

similarity $sim(X,Y)$ between X and Y, measured by cosine (Weeds, Weir, and McCarthy 2004), or the Lin measure (Lin 1998) as in the balAPinc measure of Kotlerman et al. (2010):

$$DEM_2(X,Y) = DEM_1(X,Y) \cdot sim(X,Y) \tag{6}$$

ClarkeDE (DEM_3). A close variation of DEM_1 was proposed by Clarke (2009):

$$DEM_3(X,Y) = \frac{\sum_{ft \in Feat_X \cap Feat_Y} min(w_X(ft), w_Y(ft))}{\sum_{ft \in Feat_X} w_X(ft)} \tag{7}$$

InvCL (DEM_4). A variation of DEM_3 was introduced by Lenci and Benotto (2012). It takes into account both the inclusion of context features of X in context features of Y and non-inclusion of features of Y in features of X.[22]

$$DEM_4(X,Y) = \sqrt{DEM_3(X,Y) \cdot (1 - DEM_3(Y,X))} \tag{8}$$

6.2 Generality Measures

Another related view towards the TYPE-OF relation is as follows. Given two semantically related words, a key aspect of detecting lexical entailment is the generality of the hypernym compared to the hyponym. For example, *bird* is more general than *eagle*, having a broader intension and a larger extension. This property has led to the introduction of lexical entailment measures that compare the entropy/semantic content of distributional word representations, under the assumption that a more general term has a higher-entropy distribution (Herbelot and Ganesalingam 2013; Rimell 2014; Santus et al. 2014). From this group, we show the results with the SLQS model (Santus et al. 2014) demonstrating the best performance in prior work.

SLQS. It is an entropy-based measure which quantifies the specificity/generality level of related terms. First, the top n most associated context features (i.e., typically context words as in the original work of Santus et al. [2014]) are identified (e.g., using positive PMI or LMI); for each identified context feature cn, its entropy $H(cn)$ is defined as:

$$H(cn) = -\sum_{i=1}^{n} P(ft_i|cn) \log_2 P(ft_i|c) \tag{9}$$

where ft_i, $i = 1, \ldots, n$ is the i-th context feature, and $P(ft_i|cn)$ is computed as the ratio of the co-occurrence frequency (cn, ft_i) and the total frequency of cn. For each concept word X, it is possible to compute its median entropy E_X over the N most associated context features. A higher value E_X implies a higher semantic generality of the concept word X. The initial SLQS measure, called SLQS–BASIC, is then defined as:

$$SLQS(X,Y) = 1 - \frac{E_X}{E_Y} \tag{10}$$

[22] For example, if *animal* is a hypernym of *crocodile*, one expects that (i) a number of context features of *crocodile* are also features of *animal*, and (ii) that a number of context features of *animal* are not context features of *crocodile*. As a broader concept, *animal* is also found in contexts in which also occur *animals* other than *crocodiles*.

This measure may be directly used in standard ungraded LE directionality experiments because $SLQS(X, Y) > 0$ implies that X is a type of Y (see Table 1). Another variant of SLQS, called SLQS–SIM, is tailored to LE detection experiments: It resembles the DEM_2 measure from Equation (6); the only difference is that, because SLQS can now produce negative scores, all such scores are set to 0.

6.3 Visual Generality Measures

Kiela et al. (2015) showed that such generality-based measures need not be linguistic in nature, and proposed a series of visual and multi-modal models for LE directionality and detection. We briefly outline the two best performing ones in their experiments.

Deselaers and Ferrari (2011) previously showed that sets of images corresponding to terms at higher levels in the WordNet hierarchy have greater visual variability than those at lower levels. They exploit this tendency, using sets of images associated with each concept word as returned by Google's image search. The intuition is that the set of images returned for the broader concept *animal* will consist of pictures of different kinds of animals, that is, exhibiting greater visual variability and lesser concept specificity; on the other hand, the set of images for *bird* will consist of pictures of different birds, and the set for *owl* will mostly consist only of images of owls.

The generality of a set of n images for each concept X is then computed. The first model relies on the **image dispersion** measure (Kiela et al. 2014). It is the average pairwise cosine distance between all image representations[23] $\{\overrightarrow{i_{X,1}}, \ldots, \overrightarrow{i_{X,n}}\}$ for X:

$$id(X) = \frac{2}{n(n-1)} \sum_{j<k\leq n} 1 - cos(\overrightarrow{i_{X,j}}, \overrightarrow{i_{X,k}}) \tag{11}$$

Another similar measure instead of calculating the pairwise distance calculates the distance to the centroid $\overrightarrow{\mu_X}$ of $\{\overrightarrow{i_{X,1}}, \ldots, \overrightarrow{i_{X,n}}\}$:

$$cent(X) = \frac{1}{n} \sum_{1\leq j\leq n} 1 - cos(\overrightarrow{i_{X,j}}, \overrightarrow{\mu_X}) \tag{12}$$

Final Model. The following formula summarizes the visual model for ungraded LE directionality and detection that we also test in graded evaluations:

$$s_\theta(X, Y) = \begin{cases} 1 - \frac{f(X)+\alpha}{f(Y)} & \text{if } cos(\vec{X}, \vec{Y}) \geq \theta \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

f is one of the functions for image generality given by Equations (11) and (12). The model relying on Equation (11) is called VIS–ID, and the other is called VIS–CENT. α is a tunable threshold which sets a minimum difference in generality for LE identification, driven by the idea that non-LE pairs also have non-identical generality scores. To avoid false positives where one word is more general but the pair is not semantically related,

[23] As is common practice in multi-modal semantics, each image representation is obtained by extracting the 4096-dimensional pre-softmax layer from a forward pass in a convolutional neural network (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2015) that has been trained on the ImageNet classification task using Caffe (Jia et al. 2014; Russakovsky et al. 2015).

a second threshold θ is used, which sets f to zero if the two concepts have low cosine similarity. Finally, $\vec{X}$ and $\vec{Y}$ are representations of concept words used to compute their semantic similarity (Turney and Pantel 2010; Kiela and Bottou 2014).

6.4 Concept Frequency Ratio

Concept word frequency ratio (FR) is used as a proxy for lexical generality: It is a surprisingly competitive baseline in the standard (binary) LE evaluation protocols (see Section 3.1.1 and later Section 7.2) (Weeds, Weir, and McCarthy 2004; Santus et al. 2014; Kiela et al. 2015, inter alia). FR also relies on Equation (13); the only difference is that $f(X) = freq(X)$, where $freq(X)$ is a simple word frequency count obtained from a large corpus.

6.5 WordNet-Based Similarity Measures

A variant of Equation (13) may also be used with any standard WordNet-based similarity measure to quantify the degree of TYPE-OF relation:

$$s(X, Y) = f_{WN}(X, Y) \tag{14}$$

where $f_{WN}(X, Y)$ returns a similarity score based on the WordNet path between two concepts. We use three different standard measures for f_{WN}, resulting in three variant WN-based models:

(1) WN–BASIC: f_{WN} returns a score denoting how similar two concepts are, based on the shortest path that connects the concepts in the WN taxonomy.

(2) WN–LCH: Leacock–Chodorow similarity function (Leacock and Chodorow 1998) returns a score denoting how similar two concepts are, based on their shortest connecting path (as above) and the maximum depth of the taxonomy in which the concepts occur. The score is then $-\log(path/2 \cdot depth)$, where *path* is the shortest connecting path length and *depth* the taxonomy depth.

(3) WN–WUP: Wu–Palmer similarity function (Wu and Palmer 1994; Pedersen, Patwardhan, and Michelizzi 2004) returns a score denoting how similar two concepts are, based on their depth in the taxonomy and that of their most specific ancestor node.

Note that all three WN-based similarity measures are not well-suited for graded LE experiments by their design: For example, they will rank direct co-hyponyms as more similar than distant hyponymy–hypernymy pairs.

6.6 Order Embeddings

Following trends in semantic similarity (or graded synonymy computations, see Section 2 again), Vendrov et al. (2016) have recently demonstrated that it is possible to construct a *vector space* or a *word embedding* model that specializes in the lexical entailment relation, rather than in the more popular similarity/synonymy relation. The model is then applied to a variety of tasks including ungraded LE detection and directionality.

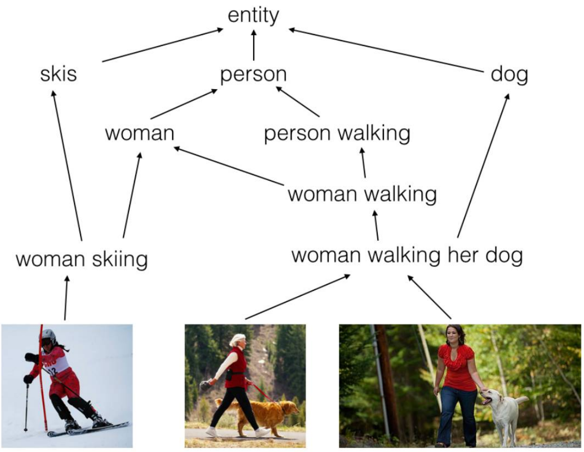


Figure 6
A slice of the visual–semantic hierarchy. The toy example is taken from Vendrov et al. (2016), inspired by the resource of Young et al. (2014).

The order embedding model exploits the partial order structure of a visual–semantic hierarchy (see Figure 6) by learning a mapping which is not distance-preserving but order-preserving between the visual–semantic hierarchy and a partial order over the embedding space. It learns a mapping from a partially ordered set $(U, \preceq_U)$ into a partially ordered embedding space $(V, \preceq_V)$: the ordering of a pair in U is then based on the ordering in the embedding space. The chosen embedding space is the reversed product order on $\mathbb{R}_+^N$, defined by the conjunction of total orders on each coordinate:

$$\vec{X} \preceq \vec{Y} \quad \text{iff} \quad \bigwedge_{i=1}^{N} X_i \geq Y_i \tag{15}$$

for all vectors $\vec{X}$ and $\vec{Y}$ with nonnegative coordinates, where the vectors $\vec{X}$ and $\vec{Y}$ are order embeddings of concept words X and Y.[24] With a slight abuse of notation, X_i refers to the i-th coordinate of vector $\vec{X}$, the same for Y_i. The ordering criterion, however, is too restrictive to impose as a hard constraint. Therefore, an approximate order-embedding is sought: A mapping that violates the order-embedding condition, imposed as a soft constraint, as little as possible. In particular, the penalty L for an ordered pair $(\vec{X}, \vec{Y})$ of points/vectors in $\mathbb{R}_+^N$ is defined as:

$$L(\vec{X}, \vec{Y}) = ||\max(0, \vec{Y} - \vec{X})||^2 \tag{16}$$

$L(\vec{X}, \vec{Y}) = 0$ implies that $X \preceq Y$, according to the reversed product order. If the order is not satisfied, the penalty is positive. The model requires a set of positive pairs PP (i.e., true LE pairs) and a set of negative pairs NP for training. To learn an approximate

[24] Smaller coordinates imply higher position in the partial order. The origin is then the top element of the order, representing the most general concept.

mapping to an order embedding space, a max-margin loss is used. It encourages positive examples to have zero penalty, and negatives to have penalty greater than a margin γ:

$$\sum_{(X,Y)\in PP} L(\vec{X},\vec{Y}) + \sum_{(X',Y')\in NP} \max(0, \gamma - L(\vec{X'},\vec{Y'})) \tag{17}$$

Positive and negative examples are task-dependent. For the standard ungraded LE evaluations, positive pairs for the training set *PP* are extracted from the WordNet hierarchy. The set *NP* is obtained by artificially constructing "corrupted" pairs (Socher et al. 2013), that is, by replacing one of the two concepts from positive examples with a randomly selected concept. This model is called ORDEREMB.

Graded LE with Order Embeddings. Order embeddings are trained for the binary LE detection task, but not explicitly for the graded LE task. To measure how much one such off-the-shelf order embedding model captures LE on the continuous scale, we test three different distance measures:

(1) ORDEREMB–COS: A standard cosine similarity is used on vector representations.

(2) ORDEREMB–DISTALL: The sum of the absolute distance between all coordinates of the vectors $\vec{X}$ and $\vec{Y}$ is used as a distance function:

$$DistAll(\vec{X},\vec{Y}) = \sum_i |Y_i - X_i| \tag{18}$$

This measure is based on the training penalty defined by Equation (16). The idea is that for order embeddings the space is sorted based on the degree of hypernymy/hyponymy violation in each dimension: The absolute coordinate distance may be used as an indicator of the LE strength.

(3) ORDEREMB–DISTPOS: This variant extends the *DistAll* distance by only adding up those coordinates fulfilling the criterion defined in the reversed product order in Equation (15):

$$DistPos(\vec{X},\vec{Y}) = \sum_i \begin{cases} |Y_i - X_i|, & \text{if } X_i \geq Y_i \\ 0, & \text{otherwise} \end{cases} \tag{19}$$

6.7 Standard ("Similarity") Embeddings

A majority of other word embedding models available in the literature target the symmetric relation of semantic relatedness and similarity, and the strength of the similarity relation is modeled by a symmetric similarity measure such as cosine.

It was shown that human subjects often consider "closer" LE pairs quite semantically similar (Geffet and Dagan 2005; Agirre et al. 2009; Hill, Reichart, and Korhonen 2015).[25] For instance, pairs *(assignment, task)* or *(author, creator)* are judged as strong LE pairs (with average scores 9.33 and 9.30 in HyperLex, respectively); they are assigned the

[25] 'Closeness' or hypernymy level for (X, Y) may be measured by the shortest WN path connecting X and Y.

labels hyp-1 and hyp-2, respectively, according to WordNet and are also considered semantically very similar (their SimLex-999 scores are 8.70 and 8.02). In another example, the WordNet syn pairs *(foe, enemy)* and *(summit, peak)* have graded LE scores of 9.72 and 9.58 in HyperLex. The rationale behind these experiments is then to test to what extent these symmetric models are capable of quantifying the degree of lexical entailment, and to what degree these two relations are interlinked.

We test the following benchmarking semantic similarity models: (1) Unsupervised models that learn from distributional information in text, including the skip-gram negative-sampling model (*SGNS*) (Mikolov et al. 2013b) with various contexts (BOW = bag of words; DEPS = dependency contexts) as described by Levy and Goldberg (2014); and (2) Models that rely on linguistic hand-crafted resources or curated knowledge bases. We evaluate models that currently hold the peak scores in word similarity tasks: sparse binary vectors built from linguistic resources (NON-DISTRIBUTIONAL [Faruqui and Dyer 2015]), vectors fine-tuned to a paraphrase database (PARAGRAM [Wieting et al. 2015]), and further refined using linguistic constraints (PARAGRAM+CF [Mrkšić et al. 2016]). Because these models are not the main focus of this work, the reader is referred to the relevant literature for detailed descriptions.

6.8 Gaussian Embeddings

An alternative approach to learning word embeddings was proposed by Vilnis and McCallum (2015). They represent words as Gaussian densities rather than points in the embedding space. Each concept X is represented as a multivariate K-dimensional Gaussian parameterized as $\mathcal{N}(\boldsymbol{\mu}_X, \boldsymbol{\sigma}_X)$, where $\boldsymbol{\mu}_X$ is a K-dimensional vector of means, and $\boldsymbol{\sigma}_X$ in the most general case is a $K \times K$ covariance matrix.[26]

Word types are embedded into soft regions in space: The intersection of these regions could be straightforwardly used to compute the degree of lexical entailment. This allows a natural representation of hierarchies using, for example, the asymmetric Kullback–Leibler (KL) divergence. KL divergence between Gaussian probability distributions is straightforward to calculate, naturally asymmetric, and has a geometric interpretation as an inclusion between families of ellipses.

To train the model, they define an energy function that returns a similarity-like measure of the two probabilities. It is possible to train the model to better capture "standard semantic similarity" (see Section 6.7) by using expected likelihood (EL) as the energy function. On the other hand, KL divergence is a natural energy function for representing entailment between concepts—a low KL divergence from X to Y indicates that we can encode Y easily as X, implying that Y entails X. This can be interpreted as a soft form of inclusion between the level sets of ellipsoids generated by the two Gaussians—if there is a relatively high expected log-likelihood ratio (negative KL), then most of the mass of Y lies inside X.

We refer the reader to the original work (He et al. 2015; Vilnis and McCallum 2015) for a detailed description and low-level modeling steps. We evaluate two variants of the model on the graded LE task following Vilnis and McCallum (2015): (i) WORD2GAUSS–EL–COS and WORD2GAUSS–EL–KL use EL in training, but the former uses cosine between vectors of means as a measure of similarity between concepts, and the latter

[26] Vilnis and McCallum (2015) use a simplification where $\boldsymbol{\sigma}_X$ is represented as a K-dimensional vector (so-called *diagonal* Gaussian embeddings) or a scalar (*spherical* embeddings).

relies on the (asymmetric) KL divergence between full Gaussians; (ii) WORD2GAUSS–KL–COS and WORD2GAUSS–KL–KL use KL divergence as the energy function.

7. Results and Discussion

7.1 Training Data and Parameters

Because we evaluate a plethora of heterogeneous models and architectures on the graded LE task, we first provide a quick overview of their training set-up regarding training data, their parameter settings, and other modeling choices.

DEMs and SLQS. Directional entailment measures DEM_1–DEM_4 and both SLQS variants (i.e., SLQS–BASIC and SLQS–SIM) are based on the cleaned, tokenized, and lowercased Polyglot Wikipedia (Al-Rfou, Perozzi, and Skiena 2013). We have used two set-ups for the induction of word representations, the only difference being that in *Set-up 1* context/feature vectors are extracted from the Polyglot Wiki directly based on bigram co-occurrence counts, whereas in *Set-up 2*, these vectors are extracted from the TYPEDM tensor (Baroni and Lenci 2010) as in the original work of Lenci and Benotto (2012).[27] Both set-ups use the positive LMI weighting calculated on syntactic co-occurrence links between each word and its context word (Gulordava and Baroni 2011): $LMI(w_1, w_2) = C(w_1, w_2) * \log_2 \frac{C(w_1,w_2)*Total}{C(w_1)C(w_2)}$, where $C(w)$ is the unigram count in the Polyglot Wiki for the word w, $C(w_1, w_2)$ is the dependency based co-occurrence count of the two tokens w_1 and w_2, namely $(w_1, (dep\_rel, w_2))$, and *Total* is the number of all such tuples. The Polyglot Wiki was parsed with Universal Dependencies (Nivre et al. 2015) as in the work of Vulić and Korhonen (2016).[28] The context vocabulary (i.e., words w_2) is restricted to the 10K most frequent words in the Polyglot Wiki. The same two set-ups were used for the SLQS model. We also use frequency counts collected from the Polyglot Wiki for the frequency ratio model. WordNet-based similarity measures rely on the latest WordNet 3.1 release.

Word Embeddings. We use 300-dimensional pre-trained order embeddings of Vendrov et al. (2016), available online.[29] For the detailed description of the training procedure, we refer the reader to the original paper. Gaussian embeddings are trained on the Polyglot Wiki with the vocabulary of the top 200K most frequent single words. We train 300-dimensional representations using the online tool and default settings suggested by Vilnis and McCallum (2015):[30] spherical embeddings trained for 200 epochs on a max-margin objective with margin set to 2.

We also use pre-trained standard "semantic similarity" word embeddings available online from various sources. 300-dimensional SGNS–BOW/DEPS vectors are also trained on the Polyglot Wiki: these are the same vectors from Levy and Goldberg (2014).[31] The

[27] TypeDM is a variant of the Distributional Memory (DM) framework, where distributional info is represented as a set of weighted word–link–word tuples $\langle\langle w_1, l, w_2\rangle, \delta\rangle$ where w_1 and w_2 are word tokens, l is a syntactic co-occurrence link between the words (e.g., a typed dependency link), and δ is a weight assigned to the tuple (e.g., LMI or PMI).

[28] We have also experimented with the TypeDM scores directly and negative LMI values. We do not report these results as they are significantly lower than the reported results obtained by the other two set-ups.

[29] `https://github.com/ivendrov/order-embedding`.

[30] `https://github.com/seomoz/word2gauss`.

[31] `https://levyomer.wordpress.com/2014/04/25/dependency-based-word-embeddings/`.

300-dimensional PARAGRAM vectors are the same as those published by Wieting et al. (2015),[32] and their extension using a retrofitting procedure (PARAGRAM+CF) has been made available by Mrkšić et al. (2016).[33] Sparse NON-DISTRIBUTIONAL vectors of Faruqui and Dyer (2015) are also available online.[34]

7.2 Results

Because of a wide variety of models and a large space of results, it is not feasible to present all results at once or provide detailed analyses across all potential dimensions of comparison. Therefore, we have decided to make a gradual selection of the most interesting experiments and results, and stress (what we consider to be) the most important aspects of the HyperLex evaluation set and models in our comparisons.

Experiment I: Ungraded LE Approaches. First, we evaluate a series of state-of-the-art traditional LE modeling approaches in the graded LE task on the entire HyperLex evaluation set. The models are described in Sections 6.1–6.5. A summary of the results is provided in Table 11. Comparing model scores with the inter-annotator agreements suggests that the graded LE task, although well-defined and understandable by average native speakers, poses a challenge for current ungraded LE models. The absolute difference in scores between human and system performance indicates that there is vast room for improvement in future work. The gap also illustrates the increased difficulty of the graded LE task compared with previous ungraded LE evaluations (see also Experiment IV). For instance, the best unsupervised LE directionality and detection models from Table 11 reach over 70% and up to 90% in precision scores (Santus et al. 2014; Kiela et al. 2015, inter alia) on BLESS and other data sets discussed in Section 3.1.2.

Previous work on ungraded LE evaluation also detected that frequency is a surprisingly competitive baseline in LE detection/directionality experiments (Herbelot and Ganesalingam 2013; Weeds et al. 2014; Kiela et al. 2015). This finding stems from an assumption that the informativeness of a concept decreases and generality increases as frequency of the concept increases (Resnik 1995). Although the assumption is a rather large simplification (Herbelot and Ganesalingam 2013), the results based on simple frequency scores in this work further suggest that the FR model may be used as a very competitive baseline in the graded LE task.

The results also reveal that visual approaches are competitive to purely textual distributional ones. In Table 11, we have set the parameters according to Kiela et al. (2015). Varying the α parameter leads to even better results—for example, the VIS–ID model scores $\rho = 0.229$ and VIS–CENT scores $\rho = 0.228$ with $\alpha = 1$. This finding supports recent trends in multi-modal semantics and calls for more expressive multi-modal LE models, as discussed previously by Kiela et al. (2015).

To our own surprise, the FR model was the strongest model in this first comparison, whereas directional measures fall short of all other approaches, although prior work suggested that they are tailored to capture the LE relation in particular. As we do not observe any major difference between two set-ups for DEMs and SLQS, all subsequent experiments use Set-up 1. The observed strong correlation between frequency and graded

[32] `http://ttic.uchicago.edu/~wieting/`.

[33] `https://github.com/nmrksic/counter-fitting`.

[34] `https://github.com/mfaruqui/non-distributional`.

LE supports the intuition that prototypical class instances will be more often cited in text, and therefore are simply more frequent.

Even WN-based measures do not lead to huge improvements over DEMs and fall short of FR. Because WordNet lacks annotations pertinent to the idea of graded LE, such simple WN-based measures cannot quantify the actual LE degree. The inclusion of the basic "semantic relatedness detector" (as controlled by the parameter θ) does not lead to any significant improvements (e.g., as evident from the comparison of SLQS–SIM vs. SLQS–BASIC, or DEM_2 vs. DEM_1). In summary, the large gap between human and system performances along with the FR superiority over more sophisticated LE approaches from prior work unambiguously calls for the next generation of distributional models tailored for graded lexical entailment in particular.

Experiment II: Word Embeddings. In the next experiment, we evaluate a series of state-of-the-art word embedding architectures, covering order embeddings (Section 6.6), standard semantic similarity embeddings optimized on SimLex-999 and related word similarity tasks (Section 6.7), and Gaussian embeddings (Section 6.8). A summary of the results is provided in Table 12. The scores again reveal the large gap between the system performance and human ability to consistently judge the graded LE relation. The scores on average are similar to or even lower than scores obtained in Experiment I. One trivial reason behind the failure is as follows: Word embeddings typically apply the cosine similarity in the Euclidean space to measure the distance between X and Y. In practice, this leads to the symmetry: $dist(X, Y) = dist(Y, X)$ for each pair (X, Y), which is an undesired model behavior for graded LE in practice, as corroborated by our analysis

Table 11
Results in the graded LE task over all HyperLex concept pairs obtained by the sets of most prominent LE models available in the literature (see Section 6.1–Section 6.5). SET-UP 1 and SET-UP 2 refer to different training set-ups for DEMs and SLQS. All results are Spearman's ρ correlation scores. IAA ρ scores are provided to quantify the upper bound for the graded LE task.

Model	**Set-up 1**	**Set-up 2**
FR ($\alpha = 0.02, \theta = 0.25$)	0.279	0.240
FR ($\alpha = 0, \theta = 0$)	0.268	0.265
DEM_1	0.162	0.162
DEM_2	0.171	0.180
DEM_3	0.150	0.150
DEM_4	0.153	0.153
SLQS–BASIC	0.225	0.221
SLQS–SIM	0.228	0.226
WN–BASIC	0.207	0.207
WN–LCH	0.214	0.214
WN–WUP	0.234	0.234
VIS–ID ($\alpha = 0.02, \theta = 0$)	0.203	0.203
VIS–CENT ($\alpha = 0.02, \theta = 0$)	0.209	0.209
IAA—1	0.854	0.854
IAA—2	0.864	0.864

Table 12
Results (Spearman's ρ correlation scores) in the graded LE task on HyperLex using a selection of state-of-the-art pre-trained word embedding models (see Section 6.6-Section 6.8). All word embeddings, excluding sparse NON-DISTRIBUTIONAL vectors, are 300-dimensional.

Model	**All**	**Nouns**	**Verbs**
FR ($\alpha = 0.02, \theta = 0.25$)	0.279	0.283	0.239
FR ($\alpha = 0, \theta = 0$)	0.268	0.283	0.091
SGNS–BOW (`win=2`)	0.167	0.148	0.289
SGNS–DEPS	0.205	0.182	0.352
NON-DISTRIBUTIONAL	0.158	0.115	0.543
PARAGRAM	0.243	0.200	0.492
PARAGRAM+CF	0.320	0.267	0.629
ORDEREMB–COS	0.156	0.162	0.005
ORDEREMB–DISTALL	0.180	0.180	0.130
ORDEREMB–DISTPOS	0.191	0.195	0.120
WORD2GAUSS–EL–COS	0.192	0.171	0.207
WORD2GAUSS–EL–KL	0.206	0.192	0.209
WORD2GAUSS–KL–COS	0.190	0.179	0.160
WORD2GAUSS–KL–KL	0.201	0.189	0.172
IAA–1	0.854	0.854	0.855
IAA–2	0.864	0.864	0.862

of asymmetry in human judgments (see Tables 7 and 8). This finding again calls for a new methodology capable of tackling the asymmetry of the graded LE problem in future work.

Dependency-based contexts (SGNS–DEPS) seem to have a slight edge over ordinary bag-of-words contexts (SGNS–BOW), which agrees with findings from prior work on ungraded LE (Roller and Erk 2016; Shwartz, Santus, and Schlechtweg 2017). We observe no clear advantage with ORDEREMB and WORD2GAUSS, two word embedding models tailored for capturing the hierarchical LE relation naturally in their training objective. We notice slight but encouraging improvements with ORDEREMB when resorting to more sophisticated distance metrics, for example, moving from the symmetric straightforward COS measure to DISTPOS with ORDEREMB, or using KL instead of COS with WORD2GAUSS.

As discussed in Section 6.6, the off-the-shelf ORDEREMB model was trained for the binary ungraded LE detection task: Its expressiveness for graded LE thus remains limited. One line of future work might utilize the ORDEREMB framework with a true graded LE objective, and investigate new ORDEREMB-style representation models fully adapted to the graded LE setting.

Lexical Entailment and Similarity. Hill, Reichart, and Korhonen (2015) report that there is strong correlation between `hyp-N` word pairs and semantic similarity as judged by human raters. For instance, given the same $[0, 10]$ continuous rating scale in SimLex-999, the average similarity score in SimLex-999 for SimLex-999 `hyp-1` pairs is 6.62; it is 6.19 for `hyp-2` pairs, and 5.70 for `hyp-3` and `hyp-4`. In fact, the only group scoring higher than `hyp-N` pairs in SimLex-999 are `syn` pairs with the average score of 7.70. Therefore, we also evaluate state-of-the-art word embedding models obtaining peak scores on SimLex-999, some of them even obtaining scores above the SimLex-999 IAA–1. The rationale is to test whether HyperLex really captures the fine-grained and subtle notion of graded lexical entailment, or whether the HyperLex annotations were largely driven by decisions at the broader level of semantic similarity.

Another look into Table 12 indicates an evident link between the LE relation and semantic similarity. Positive correlation scores for all models reveal that pairs with high graded LE scores naturally imply some degree of semantic similarity (e.g., *author / creator*). However, the scores with similarity-specialized models are much lower than the human performance in the graded LE task, which suggests that they cannot capture intricacies of the task accurately. More importantly, there is a dramatic drop in performance when evaluating exactly the same models in the semantic similarity task (i.e., graded synonymy) on SimLex-999 vs. the graded LE task on HyperLex. For instance, two best performing word embedding models on SimLex-999 are PARAGRAM and PARAGRAM+CF, reaching Spearman's ρ correlation of 0.685 and 0.742, respectively, with SimLex-999 IAA–1 = 0.673, IAA–2 = 0.778. At the same time, the two models score 0.243 and 0.320 on HyperLex, respectively, where the increase in scores for PARAGRAM+CF may be attributed to its explicit control of antonyms through dictionary-based constraints.

A similar decrease in scores is observed with other models in our comparisons, for example, SGNS–BOW falls from 0.415 on SimLex-999 to 0.167 on HyperLex. To further examine this effect, we have performed a simple experiment using only the intersection of the two evaluation sets comprising 111 word pairs in total (91 nouns and 20 verbs) for evaluation. The results of selected embedding models on the 111 pairs are shown in Figure 7. It is evident that all state-of-the-art word embedding models are significantly better at capturing semantic similarity.

In summary, the analysis of results with distributed representation models on SimLex-999 and HyperLex suggests that the human understanding of the graded LE relation is not conflated with semantic similarity. Human scores assigned to word pairs in both SimLex-999 and HyperLex reflect truly the nature of the annotated relation: semantic similarity in case of SimLex-999 and graded lexical entailment in case of HyperLex.

Experiment III: Nouns vs. Verbs. Next, given the theoretical likelihood of variation in model performance across POS categories mentioned in Section 4.1, we assess the differences

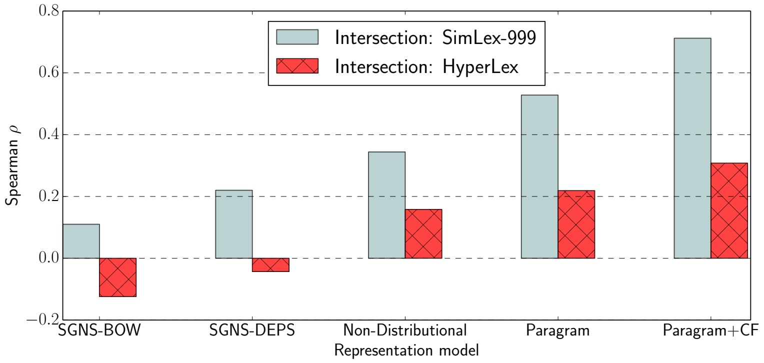


Figure 7
Results on the intersection subset of 111 concept pairs annotated both in SimLex-999 (for similarity) and in HyperLex (for graded LE).

Table 13
Results in the graded LE task over all HyperLex noun and verb pairs separately. All DEM and SLQS model variants are using Set-up 1.

Model	**Nouns**	**Verbs**
FR ($\alpha = 0.02, \theta = 0.25$)	0.283	0.239
FR ($\alpha = 0, \theta = 0$)	0.283	0.091
DEM_1	0.180	0.018
DEM_2	0.170	0.047
DEM_3	0.164	0.108
DEM_4	0.167	0.109
SLQS–BASIC	0.224	0.247
SLQS–SIM	0.229	0.232
WN–BASIC	0.240	0.263
WN–LCH	0.214	0.260
WN–WUP	0.214	0.269
VIS–ID ($\alpha = 1, \theta = 0$)	0.253	0.137
VIS–CENT ($\alpha = 1, \theta = 0$)	0.252	0.132
IAA–1	0.854	0.855
IAA–2	0.864	0.862

in results on noun (N) and verb (V) subsets of HyperLex. The results of "traditional" LE models (Experiment I) are provided in Table 13. Table 12 shows results of word embedding models. IAA scores on both POS subsets are very similar and reasonably high, implying that human raters did not find it more difficult to rate verb pairs. However, we observe differences in performance over the two POS-based HyperLex subsets. First, DEMs obtain much lower scores on the verb subset. It may be attributed to a larger variability of context features for verbs, which also affects the pure distributional models relying on the distributional inclusion hypothesis. WN-based approaches, relying on an external curated knowledge base, do not show the same pattern, with comparable results over pairs of both word classes. Visual models also score better on nouns, which may again be explained by the increased level of abstractness when dealing with verbs. This, in turn, leads to a greater visual variability and incoherence in visual concept representations.

For word embedding models, we notice that scores for the V subset are significantly higher than for the N subset. To isolate the influence of test set size, we have also repeated experiments with random subsets of the N subset, equal to the V subset in size (453 pairs). We observe the same trend even with such smaller N test sets, leading to the conclusion that difference in results stems from the fundamental difference in how humans perceive nouns and verbs. Human raters seem to associate the LE relation with similarity more frequently in case of verbs, and they do it consistently (based on the IAA scores). We speculate that it is indeed easier for humans to think in terms of semantic taxonomies when dealing with real-world entities (e.g., concrete nouns), than with more abstract events and actions, as expressed by verbs. Another reason could be that, when humans make judgments over verb semantics, syntactic features become more important and implicitly influence the judgments. This effect is supported by the

Table 14
Results in the ungraded LE directionality task (precision) using a subset of 940 HyperLex pairs converted to the ungraded directionality data set. Graded LE results (Spearman's ρ correlation) on the same subset are also provided for comparison purposes, using the best model configurations from Tables 11 and 12.

Model	Directionality			Graded LE		
	All	**Nouns**	**Verbs**	**All**	**Nouns**	**Verbs**
FR ($\alpha = 0, \theta = 0$)	0.760	0.778	0.636	0.089	0.104	0.032
DEM$_1$	0.700	0.696	0.726	– 0.072	– 0.102	– 0.071
DEM$_2$	0.700	0.696	0.726	– 0.070	– 0.050	– 0.042
DEM$_3$	0.696	0.684	0.777	0.036	0.063	0.115
DEM$_4$	0.696	0.684	0.777	0.036	0.064	0.110
SLQS–BASIC	0.747	0.734	0.835	0.088	0.121	– 0.036
SLQS–SIM	0.749	0.734	0.851	0.163	0.126	– 0.012
ORDEREMB	0.578	0.578	0.571	0.048	0.068	0.029

research on automatic acquisition of verb semantics, in which syntactic features have proven particularly important (Kipper et al. 2008; Korhonen 2010, inter alia).

We leave the underlying causes at the level of speculation. A deeper exploration here is beyond the scope of this work, but this preliminary analysis already highlights how the principal word classes integrated in HyperLex are pertinent to a range of questions concerning distributional, lexical, and cognitive semantics.

Experiment IV: Ungraded vs. Graded LE. We also analyze the usefulness of HyperLex as a data set for ungraded LE evaluations and study the differences between graded LE and one ungraded LE task: hypernymy/LE directionality (see Section 3.1.1). First, we have converted a subset of HyperLex into a data set for LE directionality experiments similar to BLESS by retaining only `hyp-N` pairs from HyperLex (as indicated by WordNet) with the graded LE score ≥ 7.0. The subset contains 940 (X, Y) pairs in total (of which 121 pairs are verb pairs), where Y in each pair may be seen as the hypernym. Following that, we run a selection of ungraded LE models from Section 6 tailored to capture directionality, and compare the scores of the same models in the graded LE task on this HyperLex subset containing "true hyponymy–hypernymy" pairs.

The frequency baseline considers the more frequent concept as the hypernym in the pair. For DEM_1-DEM_4 models (Section 6.1), the prediction of directionality is based on the asymmetry of the measure: If $DEM_i(X, Y) > DEM_i(Y, X)$, it means that the inclusion of the features of X within the features of Y is higher than the reverse, which in turn implies that Y is the hypernym in the pair. Further, $SLQS(X, Y) > 0$ implies that Y is a semantically more general concept and is therefore the hypernym (see Section 6.2).[35] With ORDEREMB, smaller coordinates mean higher position in the partial order: We

[35] Following the same idea, also discussed by Lazaridou, Pham, and Baroni (2015) and Kiela et al. (2015), a concept with a higher word embedding standard deviation or embedding entropy could be considered semantically more general and therefore the hypernym. However, we do not report the scores with word embeddings as they were only slightly better than the random baseline with the precision of 0.5.

compute and compare $DistPos(\vec{X}, \vec{Y})$ and $DistPos(\vec{Y}, \vec{X})$ scores to find the hypernym. The results provided as binary precision scores are summarized in Table 14. They reveal that frequency is a strong indicator of directionality, but further improvements, especially for verbs, may be achieved by resorting to asymmetric and generality measures. The reasonably high scores observed in our ungraded directionality experiments are also reported for the detection task in prior work (Shwartz, Santus, and Schlechtweg 2017). The graded LE results on the HyperLex subset are prominently lower than the results with the same models on the entire HyperLex: This shows that fine-grained differences in human ratings in the high end of the graded LE spectrum are even more difficult to capture with current statistical models.

The main message conveyed by the results from Table 14 is that the output from the models built for ungraded LE indeed cannot be used as an estimate of graded LE. In other words, the relative entropy or the measure of distributional inclusion between two concepts can be used to reliably detect which concept is the hypernym in the directionality task, or to distinguish between LE and other relations in the detection task, but it leads to a poor global estimate of the LE strength for graded LE experiments.

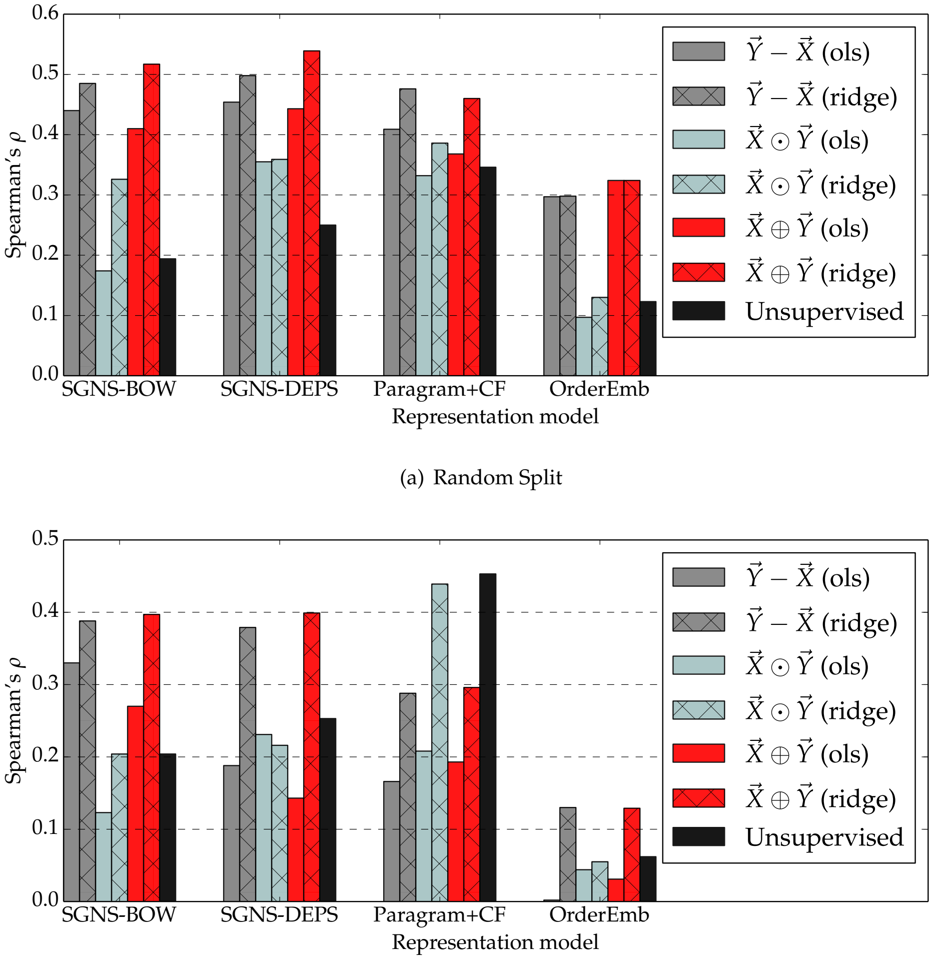
7.3 Supervised Settings: Regression Models

We also conduct preliminary experiments in supervised settings, relying on the random and lexical splits of HyperLex introduced in Section 5 (see Table 10). We experiment with several well-known supervised models from the literature. They typically represent concept pairs as a combination of each concept's embedding vector: concatenation $\vec{X} \oplus \vec{Y}$ (Baroni et al. 2012), difference $\vec{Y} - \vec{X}$ (Roller, Erk, and Boleda 2014; Weeds et al. 2014; Fu et al. 2014), or element-wise multiplication $\vec{X} \odot \vec{Y}$ (Levy et al. 2015). Based on state-of-the-art word embeddings such as SGNS–BOW or PARAGRAM, these methods are easy to apply, and show very good results in ungraded LE tasks (Baroni et al. 2012; Weeds et al. 2014; Roller, Erk, and Boleda 2014). Using two standardized HyperLex splits, the experimental set-up is as follows: We learn a regression model on the TRAINING set, optimize parameters (if any) on DEV, and test the model's prediction power on TEST. We experiment with two linear regression models: (1) standard ordinary least squares (OLS), and (2) ridge regression or Tikhonov regularization (RIDGE) (Myers 1990).

Ridge regression is a variant of least squares regression in which a regularization term is added to the training objective to favor solutions with certain properties. The regularization term is the Euclidean L2-norm of the inferred vector of regression coefficients. This term ensures that the regression favors lower coefficients and a smoother solution function, which should provide better generalization performance than simple OLS linear regression. The RIDGE objective is to minimize the following:

$$||\vec{a}\mathbf{Q} - \vec{s}||_2^2 + ||\mathbf{\Gamma}\vec{a}||_2^2 \tag{20}$$

where $\vec{a}$ is the vector of regression coefficients, and $\mathbf{Q}$ is a matrix of feature representations for each concept pair (X, Y) obtained using concatenation, difference, or element-wise multiplication. $\vec{s}$ is the vector of graded LE strengths for each concept pair, and $\mathbf{\Gamma}$ is some suitably chosen Tikhonov matrix. We rely on the most common choice: It is a multiple of the identity matrix $\mathbf{\Gamma} = \beta\mathbf{I}$. The effect of regularization is thus varied via the β hyperparameter, which is optimized on the DEV set. Setting $\beta = 0$ reduces the model to the unregularized OLS solution.



(a) Random Split

(b) Lexical Split

Figure 8
Spearman's ρ correlation scores using two different HyperLex data splits: (a) random and (b) lexical (see Section 5). Two linear regression models are used: ordinary least squares (OLS) and ridge regression (RIDGE), both trained on the training subset of each split, and tested on the test subset. Three typical feature transformations from prior work on LE detection/directionality in supervised settings have been tested: feature vector difference ($\vec{Y}-\vec{X}$), element-wise multiplication ($\vec{X}\odot\vec{Y}$), and concatenation ($\vec{X}\oplus\vec{Y}$). We also report baseline ρ correlation scores obtained by simply computing $cos(\vec{X},\vec{Y})$ on each test subset directly, without any model learning (UNSUPERVISED). Performance ceilings are 0.849 (IAA–1), 0.862 (IAA–2) for the random split, and 0.846 (IAA–1), 0.857 (IAA–2) for the lexical split.

Results. Following related work, we rely on the selection of state-of-the-art word embedding models to provide feature vectors $\vec{X}$ and $\vec{Y}$. The results of a variety of tested regression models are summarized in Figure 8(a) (random split) and Figure 8(b) (lexical split). As another reference point, we also report results with several unsupervised models on the two smaller test sets in Table 15.

Table 15
Results on the two TEST sets of HyperLex splits with a selection of unsupervised LE models. Lower scoring model variants are not shown.

Model	**Random**	**Lexical**
FR ($\alpha = 0.02, \theta = 0.25$)	0.299	0.199
DEM$_1$	0.212	0.188
DEM$_2$	0.220	0.142
DEM$_3$	0.142	0.177
DEM$_4$	0.145	0.178
SLQS–SIM	0.223	0.179
WN–BASIC	0.189	0.255
WN–WUP	0.212	0.261
VIS–ID ($\alpha = 1, \theta = 0$)	0.203	0.201
VIS–CENT ($\alpha = 1, \theta = 0$)	0.207	0.209
IAA–1	0.849	0.846
IAA–2	0.862	0.857

The IAA scores from Table 15 again indicate that there is firm agreement between annotators for the two test sets, and that automatic systems still display a large gap to the human performance. The scores on the smaller test sets follow patterns similar to those on the entire HyperLex. We see a slight increase of performance for similarity-specialized models (e.g., WN-based models or PARAGRAM+CF) on the lexical split. We attribute this increase to the larger percentage of verb pairs in the lexical test set, shown to be better modeled with similarity-oriented embeddings in the graded LE task. Verb pairs constitute 17.3% of the entire random test set, the same percentage as in the entire HyperLex, whereas the number is 26.4% for the lexical test set.

We reassess that supervised distributional methods indeed perform worse on a lexical split (Levy et al. 2015; Shwartz, Goldberg, and Dagan 2016). Besides operating with a smaller training set in a lexical split, the finding is also explained by the effect of lexical memorization with a random split: if high scores are systematically assigned to training pairs *(X_1, animal)* or *(X_2, appliance)*, the model will simply memorize that each pair *(Y_1, animal)* or *(Y_2, appliance)* should be assigned a high score during inference. The impact of lexical memorization is illustrated by Table 16, using a sample of concept pairs containing "prototypical hypernyms" (Roller and Erk 2016) such as *animal*: The regression models assign high scores even to clear negatives such as *(plant, animal)*. However, the effect of lexical memorization also partially explains the improved performance of all regression models over UNSUPERVISED baselines for a random split, as many *(X_t, animal)* pairs are indeed assigned high scores in the test set.

On the other hand, we also notice that almost all OLS regression models and a large number of RIDGE models in a lexical split cannot beat unsupervised model variants without any model learning. This suggests that the current state-of-the-art methodology in supervised settings is indeed limited in such scenarios and cannot learn satisfying generalizations regarding the type-of relation between words in training pairs. We suspect that another reason behind strong results with the semantically specialized

Table 16
The effects of lexical memorization on the output of regression models when dealing with typical hypernymy concepts higher in the taxonomy (e.g., *animal*, *plant*): HYPERLEX denotes the score assigned to the pair by humans in HyperLex, and OLS and RIDGE refer to the predicted output of the two tested regression models. We use SGNS–DEPS with concatenation ($\vec{X} \oplus \vec{Y}$, see Figure 8); similar trends are observed with other sets of vectors and feature transformations.

Concept pair	HYPERLEX	OLS	RIDGE
(plant, animal)	0.13	6.95	7.39
(mammal, animal)	10.0	7.14	7.43
(animal, mammal)	1.25	6.61	4.99
(rib, animal)	0.35	6.94	7.08
(reader, person)	7.43	7.47	6.97
(foot, plant)	0.42	7.86	6.05
(fungus, plant)	4.75	7.94	7.51
(dismiss, go)	3.97	4.22	4.29
(dinner, food)	4.85	9.36	8.63

PARAGRAM+CF model in the unsupervised setting for the lexical split is the larger percentage of verbs in the lexical test set as well as explicit handling of antonymy, as mentioned earlier. The model explicitly penalizes antonyms through dictionary-based constraints (i.e., pushes them away from each other in the vector space), a property that is desired both for semantic similarity *and* graded LE (see the low scores for the ant relation in Table 7).

The variation in results across the tested supervised model variants also indicates that the performance of a regression model is strongly dependent on the actual choice of the underlying representation model, feature transformation, as well as the chosen regression algorithm. First, the results on a random split reveal that the best unsupervised representation model does not necessarily yield the best supervised model (e.g., higher results are observed with SGNS–DEPS than with PARAGRAM in that setting). ORDEREMB is by far the weakest model in our comparison. Second, there is no clear winner in the comparison of three different feature representations. Whereas vector difference ($\vec{Y} - \vec{X}$) and concatenation ($\vec{Y} \oplus \vec{X}$) seem to yield higher scores overall for a majority of models, element-wise multiplication obtains highest scores overall in a lexical split with PARAGRAM and PARAGRAM+CF. The variation clearly suggests that supervised models have to be carefully tuned in order to perform effectively on the graded LE task.

Finally, consistent improvements of RIDGE over OLS across all splits, models, and feature transformations reveal that the choice of a regression model matters. This preliminary analysis advocates the use of more sophisticated learning algorithms in future work. Another path of research work could investigate how to exploit more training data from resources other than HyperLex to yield improved graded LE models.

7.4 Further Discussion: Specializing Semantic Spaces

Following the growing interest in word representation learning, this work also touches upon the ideas of vector/semantic space specialization: A desirable property of representation models is their ability to steer their output vector spaces according to explicit linguistic and dictionary knowledge (Yu and Dredze 2014; Astudillo et al. 2015; Faruqui et al. 2015; Liu et al. 2015; Wieting et al. 2015; Mrkšić et al. 2016; Vulić et al. 2017, inter alia). Previous work showed that it is possible to build vector spaces specialized for capturing different lexical relations (e.g., antonymy [Yih, Zweig, and Platt 2012; Ono, Miwa, and Sasaki 2015]) or distinguishing between similarity and relatedness (Kiela, Hill, and Clark 2015). Yet, it is to be seen how to build a representation model specialized for the graded LE relation. An analogy with (graded) semantic similarity is appropriate here: It was recently demonstrated that vector space models specializing for similarity and scoring high on SimLex-999 and SimVerb-3500 are able to boost performance of statistical systems in language understanding tasks such as *dialogue state tracking* (Mrkšić et al. 2016, 2017; Vulić et al. 2017). We assume that the specification of what the degree of LE means for each individual pair may also boost performance of statistical end-to-end systems in another language understanding task in future work: natural language inference (Bowman et al. 2015; Parikh et al. 2016; Agić and Schluter 2017).

Owing to their adaptability and versatility, representation architectures inspired by neural networks (e.g., Mrkšić et al. 2016; Vendrov et al. 2016) seem to be a promising avenue for future modeling work on graded lexical entailment in both unsupervised and supervised settings, despite their low performance on the graded LE task at present.

8. Application Areas: A Quick Overview

The proposed data set should have an immediate impact in the cognitive science research, providing means to analyze the effects of typicality and gradience in concept representations (Hampton 2007; Decock and Douven 2014). Besides this, a variety of other research domains share interest in taxonomic relations, automatic methods for their extraction from text, completion of rich knowledge bases, and so on. Here, we provide a quick overview of such application areas for the graded LE framework and HyperLex.

Natural Language Processing. As discussed in depth in Section 3, lexical entailment is an important linguistic task in its own right (Rimell 2014). Graded LE introduces a new challenge and a new evaluation protocol for data-driven distributional LE models. In current binary evaluation protocols targeting ungraded LE detection and directionality, even simple methods modeling lexical generality are able to yield very accurate predictions. However, our preliminary analysis in Section 7.2 demonstrates their fundamental limitations for graded lexical entailment.

In addition to the use of HyperLex as a new evaluation set, we believe that the introduction of graded LE will have implications on how the distributional hypothesis (Harris 1954) is exploited in distributional models targeting taxonomic relations in particular (Rubinstein et al. 2015; Shwartz, Goldberg, and Dagan 2016; Roller and Erk 2016, inter alia). Further, a tight connection of LE with the broader phrase-/sentence-level task of recognizing lexical entailment (Dagan, Glickman, and Magnini 2006; Dagan et al. 2013) should lead to further implications for text generation (Biran and McKeown 2013), metaphor detection (Mohler et al. 2013), question answering (Sacaleanu et al. 2008), paraphrasing (Androutsopoulos and Malakasiotis 2010), and so forth.

Representation Learning. Prior work in representation learning has mostly focused on the relations of semantic similarity and relatedness, as evidenced by the surge in interest in evaluation of word embeddings on data sets such as SimLex-999, WordSim-353, MEN (Bruni, Tran, and Baroni 2014), Rare Words (Luong, Socher, and Manning 2013), and so on. This strong focus towards similarity and relatedness means that other fundamental semantic relations such as lexical entailment have been largely overlooked in the representation learning literature. Notable exceptions building word embeddings for LE have appeared only recently (see the work of Vendrov et al. [2016] and a short overview in Section 7.4), but a comprehensive evaluation resource for intrinsic evaluation of such *LE embeddings* is still missing. There is a pressing need to improve, broaden, and introduce new evaluation protocols and data sets for representation learning architectures (Schnabel et al. 2015; Tsvetkov et al. 2015; Batchkarov et al. 2016; Faruqui et al. 2016; Yaghoobzadeh and Schütze 2016, inter alia).[36] We believe that one immediate application of HyperLex is its use as a comprehensive, wide-coverage large evaluation set for representation-learning architectures focused on the fundamental TYPE-OF taxonomic relation.

Data Mining: Extending Knowledge Bases. Ontologies and knowledge bases such as WordNet, Yago, or DBPedia are useful resources in a variety of applications such as text generation, question answering, information retrieval, or for simply providing structured knowledge to users. Because they typically suffer from incompleteness and a lack of reasoning capability, a strand of research (Snow, Jurafsky, and Ng 2004; Suchanek, Kasneci, and Weikum 2007; Bordes et al. 2011; Socher et al. 2013; Lin et al. 2015) aims to extend existing knowledge bases using patterns or classifiers applied to large text corpora. One of the fundamental relations in all knowledge bases is the TYPE-OF/INSTANCE-OF/IS-A LE relation (see Table 2 in Section 3.1.2). HyperLex may be again used straightforwardly as a wide-coverage evaluation set for such knowledge base extension models: It provides an opportunity to evaluate statistical models that tackle the problem of graded LE.

Cognitive Science. Inspired by theories of prototypicality and graded membership, HyperLex is a repository of human-graded LE scores that could be exploited in cognitive linguistics research (Taylor 2003) and other applications in cognitive science (Gärdenfors 2004; Hampton 2007). For instance, reasoning over lexical entailment is related to analogical transfer: Transferring information from the past experience (the source domain) to the new situation (the target domain) (Gentner 1983; Holyoak 2012)—for example, seeing an unknown animate object called *wampimunk* or *huhblub* which resembles a *dog*—one is likely to conclude that such *huhblubs* are to a large extent types of *animals*, although definitely not prototypical instances such as *dogs*.

Information Search. Graded LE may find application in relational Web search (Cafarella, Banko, and Etzioni 2006; Kato et al. 2009; Kopliku, Pinel-Sauvagnat, and Boughanem 2011). A user of a relational search engine might pose the query: *"List all animals with four legs"* or *"List manners of slow movement."* A system aware of the degree of LE would be better suited to relational search than a simple discrete classifier; the relational engine could rank the output list so that more prototypical instances are cited first (e.g., *dogs*, *cats*, or *elephants* before *huhblubs* or *wampimunks*). This has a direct analogy with how standard search engines rank documents or Web pages in descending order of relevance to the user's query. Further, taxonomy keyword search (Song et al. 2011; Liu et al. 2012; Wu et al. 2012) is another prominent problem in information search and retrieval where such knowledge of lexical entailment relations may be particularly useful.

[36] The need for finding better evaluation protocols for representation learning models is further exemplified by the initiative focused on designing better evaluations for semantic representation models (RepEval): `https://sites.google.com/site/repevalacl16/`
`https://repeval2017.github.io/`.

Beyond the Horizon: Multi-Modal Modeling. From a high-level perspective, autonomous artificial agents will need to jointly model vision and language in order to parse the visual world and communicate with people. Lexical entailment, textual entailment, and image captioning can be seen as special cases of a partial order over unified visual–semantic hierarchies (Deselaers and Ferrari 2011; Vendrov et al. 2016), see also Figure 6. For instance, image captions may be seen as abstractions of images, and they can be expressed at various levels in the hierarchy. The same image may be abstracted as, for example, *A boy and a girl walking their dog*, *People walking their dog*, *People walking*, *A boy, a girl, and a dog*, *Children with a dog*, *Children with an animal*. LE might prove helpful in research on image captioning (Hodosh, Young, and Hockenmaier 2013; Socher et al. 2014; Bernardi et al. 2016) or cross-modal information retrieval (Pereira et al. 2014) based on such visual–semantic hierarchies, but it is yet to be seen whether the knowledge of gradience and prototypicality may contribute to image captioning systems.

Image generality is closely linked to semantic generality, as is evident from recent work (Deselaers and Ferrari 2011; Kiela et al. 2015). The data set could also be very useful in evaluating models that ground language in the physical world (Silberer and Lapata 2012, 2014; Bruni, Tran, and Baroni 2014, inter alia). Future work might also investigate attaching graded LE scores to large hierarchical image databases such as ImageNet (Deng et al. 2009; Russakovsky et al. 2015).

9. Conclusions

Although the ultimate test of semantic models is their usefulness in downstream applications, the research community is still in need of wide-coverage comprehensive gold standard resources for intrinsic evaluation (Camacho-Collados, Pilehvar, and Navigli 2015; Schnabel et al. 2015; Tsvetkov et al. 2015; Gladkova and Drozd 2016; Hashimoto, Alvarez-Melis, and Jaakkola 2016, inter alia). Such resources can measure the general quality of the representations learned by semantic models, prior to their integration in end-to-end systems. We have presented HyperLex, a large wide-coverage gold standard resource for the evaluation of semantic representations targeting the lexical relation of *graded* lexical entailment (LE), also known as hypernymy-hyponymy or TYPE-OF relation, a relation which is fundamental in construction and understanding of concept hierarchies, that is, semantic taxonomies. Given that the problem of concept category membership is central to many cognitive science problems focused on semantic representation, we believe that HyperLex will also find its use in this domain.

The development of HyperLex was principally inspired and motivated by several factors. First, unlike prior work on lexical entailment in NLP, it focuses on the relation of graded or soft lexical entailment at a continuous scale: The relation quantifies the strength of the TYPE-OF relation between concepts rather than simply making a binary decision as with the ungraded LE variant surveyed in Section 3. Graded LE is firmly grounded in cognitive linguistic theory of class prototypes (Rosch 1973, 1975) and graded membership (Hampton 2007), stating that some concepts are more central to a broader category/class than others (prototypicality) or that some concepts are only within the category to some extent (graded membership). For instance, *basketball* is more frequently cited as a prototypical *sport* than is *chess* or *wrestling*. One purpose of HyperLex is to examine the effects of prototypicality and graded membership in human judgments, as well as to provide a large repository (i.e., HyperLex contains 2,616 word pairs in total) of concept pairs annotated for graded lexical entailment. A variety of analyses in Section 5 show that the effects are indeed prominent.

Second, whereas existing gold standards measure the ability of models to capture similarity or relatedness, HyperLex is the first crowdsourced data set with the relation of (graded) lexical entailment as its primary target. As such, it will serve as an invaluable evaluation resource for representation learning architectures tailored for the principal lexical relation, which has plenty of potential applications, as indicated in Section 8. Analysis of the HyperLex ratings from more than 600 annotators, all native English speakers, showed that subjects can consistently quantify graded LE, and distinguish it from a broader notion of similarity/relatedness and other prominent lexical relations (e.g., cohyponymy, meronymy, antonymy) based on simple non-expert intuitive instructions. This is supported by high inter-annotator agreement scores on the entire data set, as well as on different subsets of HyperLex (e.g., POS categories, WordNet relations).

Third, because we wanted HyperLex to be wide-coverage and representative, the construction process guaranteed that the data set covers concept pairs of different POS categories (nouns and verbs), at different levels of concreteness, and concept pairs standing in different relations according to WordNet. The size and coverage of HyperLex makes it possible to compare the strengths and weaknesses of various representation models via statistically robust analyses on specific word classes, and investigate human judgments in relation to such different properties. The size of HyperLex also enables supervised learning, for which we provide two standard data set splits (Levy et al. 2015; Shwartz, Goldberg, and Dagan 2016) into training, test, and development subsets.

To dissect the key properties of HyperLex, we conducted a spectrum of experiments and evaluations with most prominent state-of-the-art classes of lexical entailment and embedding models available in the literature. One clear conclusion is that current lexical entailment models optimized for the ungraded LE variant perform very poorly in general. There is clear room under the inter-rating ceiling to guide the development of the next generation of distributional models: The low performance can be partially mitigated by focusing models on the graded LE variant, and developing new and more expressive architectures for LE in future work. Even analyses with a selection of prominent supervised LE models reveal the huge gap between the human and system performance in the graded LE task. Future work thus needs to find a way to conceptualize and encode the graded LE idea into distributional models to tackle the task effectively. Despite their poor performance at present, we believe that a promising step in that direction are neural net–inspired approaches to LE proposed recently (Vilnis and McCallum 2015; Vendrov et al. 2016), mostly because of their conceptual distinction from other distributional modeling approaches complemented with their modeling adaptability and flexibility. In addition, in order to model hierarchical semantic knowledge more accurately, in future work we may require algorithms that are better suited to fast learning from few examples (Lake et al. 2011), and have some flexibility with respect to sense-level distinctions (Reisinger and Mooney 2010b; Neelakantan et al. 2014; Jauhar, Dyer, and Hovy 2015; Šuster, Titov, and van Noord 2016).

Despite the abundance of reported experiments and analyses in this work, we have only scratched the surface in terms of the possible analyses with HyperLex and use of such models as components of broader phrase-level and sentence-level textual entailment systems, as well as in other applications, as quickly surveyed in Section 8. Beyond the preliminary conclusions from these initial analyses, we believe that the benefits of HyperLex will become evident as researchers use it to probe the relationship between architectures, algorithms, and representation quality for a wide range of concepts. A better understanding of how to represent the full diversity of concepts (with LE grades attached) in hierarchical semantic networks should in turn yield improved methods for encoding and interpreting the hierarchical semantic knowledge that constitutes much of the important information in language.

Acknowledgments

This work is supported by the ERC Consolidator grant (no 648909). DK and FH performed their work while they were still at the University of Cambridge.

References

Agić, Željko and Natalie Schluter. 2017. Baselines and test data for cross-lingual inference. *CoRR*, abs/1704.05347.

Agirre, Eneko, Enrique Alfonseca, Keith B. Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. 2009. A study on similarity and relatedness using distributional and WordNet-based approaches. In *Proceedings of NAACL-HLT*, pages 19–27, Boulder, CO.

Al-Rfou, Rami, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proceedings of CoNLL*, pages 183–192, Sofia.

Androutsopoulos, Ion and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Astudillo, Ramón, Silvio Amir, Wang Ling, Mario Silva, and Isabel Trancoso. 2015. Learning word representations from scarce and noisy data with embedding subspaces. In *Proceedings of ACL*, pages 1074–1084, Beijing.

Auer, Sören, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. 2007. DBPedia: A nucleus for a Web of open data. In *Proceedings of the Semantic Web Conference (ISWC)*, pages 722–735, Busan.

Bankova, Desislava, Bob Coecke, Martha Lewis, and Daniel Marsden. 2016. Graded entailment for compositional distributional semantics. *CoRR*, abs/1601.04908.

Baronett, Stan. 2012. *Logic*, 3rd edition.

Baroni, Marco, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. 2012. Entailment above the word level in distributional semantics. In *Proceedings of EACL*, pages 23–32, Avignon.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of ACL*, pages 238–247, Baltimore, MD.

Baroni, Marco and Alessandro Lenci. 2010. Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4):673–721.

Baroni, Marco and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics (GEMS)*, pages 1–10, Edinburgh.

Batchkarov, Miroslav, Thomas Kober, Jeremy Reffin, Julie Weeds, and David Weir. 2016. A critique of word similarity as a method for evaluating distributional semantic models. In *Proceedings of REPEVAL*, pages 7–12, Berlin.

Beckwith, Richard, Christiane Fellbaum, Derek Gross, and George A. Miller. 1991. WordNet: A lexical database organized on psycholinguistic principles. *Lexical acquisition: Exploiting on-line resources to build a lexicon*, pages 211–231.

Bejar, Isaac I., Roger Chaffin, and Susan Embretson. 1991. *Cognitive and Psychometric Analysis of Analogical Problem Solving*. Springer.

Beltagy, Islam, Cuong Chau, Gemma Boleda, Dan Garrette, Katrin Erk, and Raymond Mooney. 2013. Montague meets Markov: Deep semantics with probabilistic logical form. In *Proceedings of *SEM*, pages 11–21, Atlanta, GA.

van Benthem, Johan and Alice ter Meulen. 1996. *Handbook of Logic and Language*. Elsevier.

Bernardi, Raffaella, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Biran, Or and Kathleen McKeown. 2013. Classifying taxonomic relations between pairs of Wikipedia articles. In *Proceedings of IJCNLP*, pages 788–794, Nagoya.

Blutner, Reinhard, Emmanuel M. Pothos, and Peter Bruza. 2013. A quantum probability perspective on borderline vagueness. *Topics in Cognitive Science*, 5(4):711–736.

Bordes, Antoine, Jason Weston, Ronan Collobert, and Yoshua Bengio. 2011. Learning structured embeddings of knowledge bases. In *Proceedings of AAAI*, pages 301–306, San Francisco, CA.

Bos, Johan and Katja Markert. 2005. Recognising textual entailment with logical inference. In *Proceedings of EMNLP*, pages 628–635, Vancouver.

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*, pages 632–642, Lisbon.

Bruni, Elia, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Bullinaria, John A. and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3):510–526.

Cafarella, M. J., Banko, M., & Etzioni, O. 2006. Relational Web search. Technical Report 2006-04-02, University of Washington, Department of Computer Science and Engineering, Edinburgh.

Camacho-Collados, José, Mohammad Taher Pilehvar, and Roberto Navigli. 2015. A framework for the construction of monolingual and cross-lingual word similarity datasets. In *Proceedings of ACL*, pages 1–7, Beijing.

Clarke, Daoud. 2009. Context-theoretic semantics for natural language: An Overview. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics (GEMS)*, pages 112–119, Athens.

Coleman, Linda and Paul Kay. 1981. Prototype semantics: The English word lie. *Language*, 57(1):26–44.

Collins, Allan M. and Ross M. Quillian. 1969. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8(2):240–247.

Collins, Allan M. and Ross M. Quillian. 1972. Experiments on semantic memory and language comprehension. *Cognition in Learning and Memory*.

Dagan, Ido, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges*, pages 177–190.

Dagan, Ido, Dan Roth, Mark Sammons, and Fabio Massimo Zanzotto. 2013. Recognizing textual entailment: Models and applications. *Synthesis Lectures on Human Language Technologies*, 6(4):1–220.

Decock, Lieven and Igor Douven. 2014. What is graded membership? *Noûs*, 48(4):653–682.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. ImageNet: A large-scale hierarchical image database. In *Proceedings of CVPR*, pages 248–255, Miami, FL.

Deselaers, Thomas and Vittorio Ferrari. 2011. Visual and semantic similarity in ImageNet. In *Proceedings of CVPR*, pages 1777–1784, Colorado Springs, CO.

Dirven, René and John Taylor. 1986. *The conceptualisation of vertical space in English: The case of* tall. Linguistic Agency, University of Duisburg.

Divjak, Dagmar and Antti Arppe. 2013. Extracting prototypes from exemplars: What can corpus data tell us about concept representation? *Cognitive Linguistics*, 24(2):221–274.

Do, Quang and Dan Roth. 2010. Constraints based taxonomic relation classification. In *Proceedings of EMNLP*, pages 1099–1109, Cambridge, MA.

Esteva, Francesc, Lluís Godo, Ricardo O. Rodríguez, and Thomas Vetterlein. 2012. Logics for approximate and strong entailments. *Fuzzy Sets and Systems*, 197:59–70.

Evert, Stefan. 2008. Corpora and collocations. *Corpus Linguistics*, 2:223–233.

Faruqui, Manaal, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of NAACL-HLT*, pages 1606–1615, Denver, CO.

Faruqui, Manaal and Chris Dyer. 2015. Non-distributional word vector representations. In *Proceedings of ACL*, pages 464–469, Beijing.

Faruqui, Manaal, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. In *Proceedings of REPEVAL*, pages 30–35, Berlin.

Fellbaum, Christiane. 1998. *WordNet*. Cambridge, MA, MIT Press.

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

Fromkin, Victoria, Robert Rodman, and Nina Hyams. 2013. *An Introduction to Language*, 10th edition. Wadsworth Cengage Learning.

Fu, Ruiji, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2014. Learning semantic hierarchies via word embeddings. In *Proceedings of ACL*, pages 1199–1209, Baltimore, MD.

Fu, Ruiji, Jiang Guo, Bing Qin, Wanxiang Che, Haifeng Wang, and Ting Liu. 2015. Learning semantic hierarchies: A continuous vector space approach. *IEEE/ACM Transactions on Audio, Speech & Language Processing*, 23(3): 461–471.

Gärdenfors, Peter. 2004. *Conceptual Spaces: The Geometry of Thought*. MIT Press.

Geffet, Maayan and Ido Dagan. 2005. The distributional inclusion hypotheses and lexical entailment. In *Proceedings of ACL*, pages 107–114, Ann Arbor, MI.

Gentner, Dedre. 1983. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170.

Gentner, Dedre. 2006. Why verbs are hard to learn. *Action meets word: How children learn verbs*, pages 544–564. Oxford University Press.

Gerz, Daniela, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. SimVerb-3500: A large-scale evaluation set of verb similarity. In *Proceedings of EMNLP*, pages 2173–2182, Austin, TX.

Gheorghita, Inga and Jean-Marie Pierrel. 2012. Towards a methodology for automatic identification of hypernyms in the definitions of large-scale dictionary. In *Proceedings of LREC*, pages 2614–2618, Istanbul.

Gladkova, Anna and Aleksandr Drozd. 2016. Intrinsic evaluations of word embeddings: What can we do better? In *Proceedings of REPEVAL*, pages 36–42, Berlin.

Gulordava, Kristina and Marco Baroni. 2011. A distributional similarity approach to the detection of semantic change in the Google Books Ngram corpus. In *Proceedings of the GEMS Workshop on GEometrical Models of Natural Language Semantics*, pages 67–71, Edinburgh.

Hampton, James A. 2006. Concepts as prototypes. *Psychology of Learning and Motivation*, 46:79–113.

Hampton, James A. 2007. Typicality, graded membership, and vagueness. *Cognitive Science*, 31(3):355–384.

Harris, Zellig S. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Hashimoto, Tatsunori B., David Alvarez-Melis, and Tommi S. Jaakkola. 2016. Word embeddings as metric recovery in semantic spaces. *Transactions of the ACL*, 4:273–286.

He, Shizhu, Kang Liu, Guoliang Ji, and Jun Zhao. 2015. Learning to represent knowledge graphs with Gaussian embedding. In *Proceedings of CIKM*, pages 623–632, Melbourne.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of COLING*, pages 539–545, Nantes.

Hendrickx, Iris, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of SEMEVAL*, pages 33–38, Uppsala.

Herbelot, Aurélie and Mohan Ganesalingam. 2013. Measuring semantic content in distributional vectors. In *Proceedings of ACL*, pages 440–445, Sofia.

Hill, Felix, Anna Korhonen, and Christian Bentz. 2014. A quantitative empirical analysis of the abstract/concrete distinction. *Cognitive Science*, 38(1):162–177.

Hill, Felix, Roi Reichart, and Anna Korhonen. 2015. SimLex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.

Hodosh, Micah, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Holyoak, Keith J. 2012. Analogy and relational reasoning. In K. J. Holyoak and R. G. Morrison, editors, *The Oxford Handbook of Thinking and Reasoning*. Oxford University Press, pages 234–259.

Jackendoff, Ray S. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press .

Jauhar, Sujay Kumar, Chris Dyer, and Eduard Hovy. 2015. Ontologically grounded multi-sense representation learning for semantic vector space models. In *Proceedings of NAACL-HLT*, pages 683–693, Denver, CO.

Jia, Yangqing, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross B. Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of ACM Multimedia*, pages 675–678, Orlando, FL.

Jurgens, David, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. SemEval-2012 task 2: Measuring degrees of relational similarity. In *Proceedings of SEMEVAL*, pages 356–364, Montreal.

Kamp, Hans and Barbara Partee. 1995. Prototype theory and compositionality. *Cognition*, 57(2):129–191.

Kato, Makoto P., Hiroaki Ohshima, Satoshi Oyama, and Katsumi Tanaka. 2009. Query by analogical example: Relational search using Web search engine indices. In *Proceedings of CIKM*, pages 27–36, Hong Kong.

Kiela, Douwe and Léon Bottou. 2014. Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of EMNLP*, pages 36–45, Doha.

Kiela, Douwe, Felix Hill, and Stephen Clark. 2015. Specializing word embeddings for similarity or relatedness. In *Proceedings of EMNLP*, pages 2044–2048, Lisbon.

Kiela, Douwe, Felix Hill, Anna Korhonen, and Stephen Clark. 2014. Improving multi-modal representations using image dispersion: Why less is sometimes more. In *Proceedings of ACL*, pages 835–841, Baltimore, MD.

Kiela, Douwe, Laura Rimell, Ivan Vulić, and Stephen Clark. 2015. Exploiting image generality for lexical entailment detection. In *Proceedings of ACL*, pages 119–124, Beijing.

Kipper, Karin, Anna Korhonen, Neville Ryant, and Martha Palmer. 2008. A large-scale classification of English verbs. *Language Resources and Evaluation*, 42(1):21–40.

Kopliku, Arlind, Karen Pinel-Sauvagnat, and Mohand Boughanem. 2011. Retrieving attributes using Web tables. In *Proceedings of JCDL*, pages 397–398, Ottawa.

Korhonen, Anna. 2010. Automatic lexical classification: bridging research and practice. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 368(1924):3621–3632.

Kotlerman, Lili, Ido Dagan, Idan Szpektor, and Maayan Zhitomirsky-Geffet. 2010. Directional distributional similarity for lexical inference. *Natural Language Engineering*, 16(4):359–389.

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet classification with deep convolutional neural networks. In *Proceedings of NIPS*, pages 1106–1114, South Lake Tahoe, NV.

Lake, Brenden M., Ruslan Salakhutdinov, Jason Gross, and Joshua B. Tenenbaum. 2011. One shot learning of simple visual concepts. In *Proceedings of CogSci*, pages 2568–2573, Boston, MA.

Lakoff, George. 1990. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. University of Chicago Press.

Lazaridou, Angeliki, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of NAACL-HLT*, pages 153–163, Denver, CO.

Leacock, Claudia and Martin Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

Lenci, Alessandro and Giulia Benotto. 2012. Identifying hypernyms in distributional semantic spaces. In *Proceedings of *SEM*, pages 75–79, Montreal.

Leviant, Ira and Roi Reichart. 2015. Separated by an un-common language: Towards judgment language informed vector space modeling. *CoRR*, abs/1508.00106.

Levin, Beth. 1993. *English Verb Classes and Alternation: A Preliminary Investigation*. University of Chicago Press.

Levy, Omer, Ido Dagan, and Jacob Goldberger. 2014. Focused entailment graphs for open IE propositions. In *Proceedings of CoNLL*, pages 87–97, Baltimore, MD.

Levy, Omer and Yoav Goldberg. 2014. Dependency-based word embeddings. In *Proceedings of ACL*, pages 302–308, Baltimore, MD.

Levy, Omer, Steffen Remus, Chris Biemann, and Ido Dagan. 2015. Do supervised distributional methods really learn lexical inference relations? In *Proceedings of NAACL-HLT*, pages 970–976, Denver, CO.

Lin, Dekang. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of ACL*, pages 768–774, Montreal.

Lin, Yankai, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. 2015. Learning entity and relation embeddings for knowledge graph completion. In *Proceedings of AAAI*, pages 2181–2187, Austin, TX.

Liu, Quan, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of ACL*, pages 1501–1511, Beijing.

Liu, Xueqing, Yangqiu Song, Shixia Liu, and Haixun Wang. 2012. Automatic taxonomy construction from keywords. In *Proceedings of KDD*, pages 1433–1441, Beijing.

Luong, Thang, Richard Socher, and Christopher Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of CoNLL*, pages 104–113, Sofia.

Markman, Arthur B. and Edward J. Wisniewski. 1997. Similar and different: The differentiation of basic-level categories.

Journal of Experimental Psychology: Learning, Memory, and Cognition, 23(1).

McRae, Ken, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

Medin, Douglas L., Mark W. Altom, and Timothy D. Murphy. 1984. Given versus induced category representations: Use of prototype and exemplar information in classification. *Journal of Experimental Psychology*, 10(3):333–352.

Mikolov, Tomas, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of ICLR: Workshop Papers*, Scottsdale, AZ. http://arxiv.org/abs/1301.3781

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119, South Lake Tahoe, NV.

Miller, George A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Mohler, Michael, David Bracewell, Marc Tomlinson, and David Hinote. 2013. Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 27–35, Atlanta, GA.

Mrkšić, Nikola, Diarmuid Ó Séaghdha, Blaise Thomson, Milica Gašić, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, Tsung-Hsien Wen, and Steve J. Young. 2016. Counter-fitting word vectors to linguistic constraints. In *Proceedings of NAACL-HLT*, pages 142–148, San Diego, CA.

Mrkšić, Nikola, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve J. Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of ACL*, pages 1777–1788, Vancouver.

Murphy, Lynne M. 2003. *Semantic Relations and the Lexicon: Antonymy, Synonymy and Other Paradigms*. Cambridge University Press.

Myers, Raymond H. 1990. *Classical and Modern Regression with Applications (Volume 2)*. Duxbury Press.

Neelakantan, Arvind, Jeevan Shankar, Alexandre Passos, and Andrew McCallum. 2014. Efficient non-parametric estimation of multiple embeddings per word in vector space. In *Proceedings of EMNLP*, pages 1059–1069, Doha.

Nelson, Douglas L., Cathy L. McEvoy, and Thomas A. Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods*, 36(3):402–407.

Nivre, Joakim, Željko Agić, Maria Jesus Aranzabe, et al. 2015. Universal Dependencies 1.2. LINDAT/CLARIN digital library at Institute of Formal and Applied Linguistics, Charles University in Prague.

Ono, Masataka, Makoto Miwa, and Yutaka Sasaki. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of NAACL-HLT*, pages 984–989, Denver, CO.

Osherson, Daniel N. and Edward E. Smith. 1981. On the adequacy of prototype theory as a theory of concepts. *Cognition*, 9(1):35–58.

Osherson, Daniel N. and Edward E. Smith. 1997. On typicality and vagueness. *Cognition*, 64(2):189–206.

Padó, Sebastian, Ulrike Padó, and Katrin Erk. 2007. Flexible, corpus-based modelling of human plausibility judgements. In *Proceedings of EMNLP-CoNLL*, pages 400–409, Prague.

Paivio, Allan. 1991. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology*, 45(3):255.

Pantel, Patrick and Marco Pennacchiotti. 2006. Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proceedings of ACL*, pages 113–120, Sydney.

Parikh, Ankur, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of EMNLP*, pages 2249–2255, Austin, TX.

Pedersen, Ted, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity - Measuring the relatedness of concepts. In *Proceedings of AAAI*, pages 1024–1025, San Jose, CA.

Pereira, Jose Costa, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2014. On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):521–535.

Pulman, Stephan Guy. 1983. *Word Meaning and Belief*. Routledge.

Quillian, Ross M. 1967. Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science*, 12(5):410–430.

Rei, Marek and Ted Briscoe. 2014. Looking for hyponyms in vector space. In *Proceedings of CoNLL*, pages 68–77, Baltimore, MD.

Reisinger, Joseph and Raymond J. Mooney. 2010a. A mixture model with sharing for lexical semantics. In *Proceedings of EMNLP*, pages 1173–1182, Cambridge, MA.

Reisinger, Joseph and Raymond J. Mooney. 2010b. Multi-prototype vector-space models of word meaning. In *Proceedings of NAACL-HLT*, pages 109–117, Los Angeles, CA.

Resnik, Philip. 1995. Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of IJCAI*, pages 448–453, Montreal.

Riedel, Sebastian, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. 2013. Relation extraction with matrix factorization and universal schemas. In *Proceedings of NAACL-HLT*, pages 74–84, Atlanta, GA.

Rimell, Laura. 2014. Distributional lexical entailment by topic coherence. In *Proceedings of EACL*, pages 511–519, Gothenburg.

Roller, Stephen and Katrin Erk. 2016. Relations such as hypernymy: Identifying and exploiting Hearst patterns in distributional vectors for lexical entailment. In *Proceedings of EMNLP*, pages 2163–2172, Austin, TX.

Roller, Stephen, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of COLING*, pages 1025–1036, Dublin.

Rosch, Eleanor H. 1973. Natural categories. *Cognitive Psychology*, 4(3):328–350.

Rosch, Eleanor H. 1975. Cognitive representations of semantic categories. *Journal of Experimental Psychology*, 104(3):192–233.

Rubinstein, Dana, Effi Levi, Roy Schwartz, and Ari Rappoport. 2015. How well do distributional models capture different types of semantic knowledge? In *Proceedings of ACL*, pages 726–730, Beijing.

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. 2015. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.

Sacaleanu, Bogdan, Constantin Orasan, Christian Spurk, Shiyan Ou, Oscar Ferrandez, Milen Kouylekov, and Matteo Negri. 2008. Entailment-based question answering for structured data. In *Proceedings of COLING*, pages 173–176, Manchester.

Santus, Enrico, Alessandro Lenci, Qin Lu, and Sabine Schulte im Walde. 2014. Chasing hypernyms in vector spaces with entropy. In *Proceedings of EACL*, pages 38–42, Gothenburg.

Santus, Enrico, Frances Yung, Alessandro Lenci, and Chu-Ren Huang. 2015. EVALution 1.0: An evolving semantic dataset for training and evaluation of distributional semantic models. In *Proceedings of the 4th Workshop on Linked Data in Linguistics: Resources and Applications*, pages 64–69, Beijing.

Schnabel, Tobias, Igor Labutov, David M. Mimno, and Thorsten Joachims. 2015. Evaluation methods for unsupervised word embeddings. In *Proceedings of EMNLP*, pages 298–307, Lisbon.

Schwartz, Roy, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. In *Proceedings of CoNLL*, pages 258–267, Beijing.

Shwartz, Vered, Yoav Goldberg, and Ido Dagan. 2016. Improving hypernymy detection with an integrated path-based and distributional method. In *Proceedings of ACL*, pages 2389–2398, Berlin.

Shwartz, Vered, Omer Levy, Ido Dagan, and Jacob Goldberger. 2015. Learning to exploit structured resources for lexical inference. In *Proceedings of CoNLL*, pages 175–184, Beijing.

Shwartz, Vered, Enrico Santus, and Dominik Schlechtweg. 2017. Hypernyms under siege: Linguistically-motivated artillery for hypernymy detection. In *Proceedings of EACL*, pages 65–75, Valencia.

Silberer, Carina and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of EMNLP*, pages 1423–1433, Jeju.

Silberer, Carina and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of ACL*, pages 721–732, Baltimore, MD.

Simonyan, Karen and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR*, San Diego, CA. http://arxiv.org/abs/1409.1556

Snow, Rion, Daniel Jurafsky, and Andrew Y. Ng. 2004. Learning syntactic patterns for automatic hypernym discovery. In *Proceedings of NIPS*, pages 1297–1304, Vancouver.

Snow, Rion, Daniel Jurafsky, and Andrew Y. Ng. 2006. Semantic taxonomy induction from heterogenous evidence. In *Proceedings of ACL*, pages 801–808, Sydney.

Socher, Richard, Danqi Chen, Christopher D. Manning, and Andrew Y. Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Proceedings of NIPS*, pages 926–934, South Lake Tahoe, NV.

Socher, Richard, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the ACL*, 2:207–218, Baltimore, MD.

Song, Yangqiu, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. 2011. Short text conceptualization using a probabilistic knowledgebase. In *Proceedings of AAAI*, pages 2330–2336, San Francisco, CA.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: A core of semantic knowledge. In *Proceedings of WWW*, pages 697–706, Banff.

Tanon, Thomas Pellissier, Denny Vrandečić, Sebastian Schaffert, Thomas Steiner, and Lydia Pintscher. 2016. From Freebase to Wikidata: The great migration. In *Proceedings of WWW*, pages 1419–1428, Montreal.

Taylor, John R. 2003. *Linguistic Categorization*. Oxford University Press.

Tsvetkov, Yulia, Manaal Faruqui, Wang Ling, Guillaume Lample, and Chris Dyer. 2015. Evaluation of word vector representations by subspace alignment. In *Proceedings of EMNLP*, pages 2049–2054, Lisbon.

Turney, Peter D. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416.

Turney, Peter D. and Saif M. Mohammad. 2015. Experiments with three approaches to recognizing lexical entailment. *Natural Language Engineering*, 21(3):437–476.

Turney, Peter D. and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Vendrov, Ivan, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2016. Order-embeddings of images and language. In *Proceedings of ICLR*, San Juan, PR.

Vilnis, Luke and Andrew McCallum. 2015. Word representations via Gaussian embedding. In *Proceedings of ICLR*, San Diego, CA.

Šuster, Simon, Ivan Titov, and Gertjan van Noord. 2016. Bilingual learning of multi-sense embeddings with discrete autoencoders. In *Proceedings of NAACL-HLT*, pages 1346–1356, San Diego, CA.

Vulić, Ivan and Anna Korhonen. 2016. Is "universal syntax" universally useful for learning distributed word representations? In *Proceedings of ACL*, pages 518–524, Berlin.

Vulić, Ivan, Nikola Mrkšić, Roi Reichart, Diarmuid Ó Séaghdha, Steve Young, and Anna Korhonen. 2017. Morph-fitting: Fine-tuning word vector spaces with simple language-specific rules. In *Proceedings of ACL*, pages 56–68, Vancouver.

Vylomova, Ekaterina, Laura Rimell, Trevor Cohn, and Timothy Baldwin. 2016. Take and took, gaggle and goose, book and read: Evaluating the utility of vector differences for lexical relation learning. In *Proceedings of ACL*, pages 1671–1682, Berlin.

Weeds, Julie, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING*, pages 2249–2259, Dublin.

Weeds, Julie and David Weir. 2003. A general framework for distributional similarity. In *Proceedings of EMNLP*, pages 81–88, Sapporo.

Weeds, Julie, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of COLING*, pages 1015–1021, Geneva.

Wieting, John, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2015. From paraphrase database to compositional paraphrase model and back. *Transactions of the ACL*, 3:345–358.

Wu, Wentao, Hongsong Li, Haixun Wang, and Kenny Q. Zhu. 2012. Probase: A probabilistic taxonomy for text understanding. In *Proceedings of SIGMOD*, pages 481–492, Scottsdale, AZ.

Wu, Zhibiao and Martha Palmer. 1994. Verb semantics and lexical selection. In *Proceedings of ACL*, pages 133–138, Las Cruces, NM.

Yaghoobzadeh, Yadollah and Hinrich Schütze. 2016. Intrinsic subspace evaluation of word embedding representations. In *Proceedings of ACL*, pages 236–246, Berlin.

Yih, Scott Wen-Tau, Geoffrey Zweig, and John C. Platt. 2012. Polarity inducing latent semantic analysis. In *Proceedings of EMNLP*, pages 1212–1222, Jeju.

Young, Peter, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the ACL*, 2:67–78.

Yu, Mo and Mark Dredze. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of ACL*, pages 545–550, Baltimore, MD.

Zhila, Alisa, Wen-tau Yih, Christopher Meek, Geoffrey Zweig, and Tomas Mikolov. 2013. Combining heterogeneous models for measuring relational similarity. In *Proceedings of NAACL-HLT*, pages 1000–1009, Atlanta, GA.

Zhitomirsky-Geffet, Maayan and Ido Dagan. 2009. Bootstrapping distributional feature vector quality. *Computational Linguistics*, 35(3):435–461.