

# Distinguishing between paradigmatic semantic relations across word classes: human ratings and distributional similarity

*Sabine Schulte im Walde*  
Universität Stuttgart

## ABSTRACT

This article explores the distinction between paradigmatic semantic relations, both from a cognitive and a computational linguistic perspective. Focusing on an existing dataset of German synonyms, antonyms and hypernyms across the word classes of nouns, verbs and adjectives, we assess human ratings and a supervised classification model using window-based and pattern-based distributional vector spaces. Both perspectives suggest differences in relation distinction across word classes, but easy vs. difficult class–relation combinations differ, exhibiting stronger ties between ease and naturalness of class-dependent relations for humans than for computational models.

In addition, we demonstrate that distributional information is indeed a difficult starting point for distinguishing between paradigmatic relations but that even a simple classification model is able to manage this task. The fact that the most salient vector spaces and their success vary across word classes and paradigmatic relations suggests that combining feature types for relation distinction is better than applying them in isolation.

*Keywords:*  
*semantic relations, human ratings, distributional semantics, automatic classification*

Paradigmatic semantic relations such as synonymy, antonymy, hypernymy and (co-)hyponymy define relations between words that can be found in the same position in a syntagma (de Saussure 1916). They are central to the organisation of the mental lexicon (Deese 1965; Miller and Fellbaum 1991; Murphy 2003), by providing a structure for the lexical concepts that words express. According to Miller and Fellbaum (1991), this relational structure differs across word classes, as *“no single set of semantic relations [...] is adequate for the entire lexicon: nouns, adjectives, and verbs each have their own semantic relations and their own organisation determined by the role they must play in the construction of linguistic messages”*. For example, while hypernymy is considered a natural relation for organising the noun lexicon, it is regarded as less important for organising the verb lexicon, and as rather unnatural for organising the adjective lexicon. In contrast, antonymy is taken to represent the core relation for organising the adjective lexicon, and next to hypernymy, synonymy and entailment, antonymy also plays an important role in the mental lexicon for verbs.

From a computational point of view, modelling paradigmatic semantic relations is important for any application in natural language processing (NLP) such as machine translation and textual entailment, which go beyond the general notion of semantic relatedness and require distinguishing between specific semantic relations. Distributional semantic spaces (also known as vector space models) present a method of determining the meaning and the semantic relatedness between target words within a geometric setting (Budanitsky and Hirst 2006; Turney and Pantel 2010). These models rely on the Distributional Hypothesis and exploit corpus co-occurrences in vector space models to describe and compare the meanings of linguistic units such as words, phrases and sentences (Harris 1954; Firth 1957). Paradigmatic relations are notoriously difficult to be distinguished by standard distributional models, however, because the first-order co-occurrence distributions of the related words tend to be very similar across the relations. For example, in the sentence variants *“The boy/girl/person loves/hates the cat”*, the nominal

co-hyponyms *boy* and *girl* and their hypernym *person*, as well as the verbal antonyms *love* and *hate*, occur in identical contexts, respectively.

Our research presented in this article brings together perspectives from cognitive semantics and distributional semantics, and explores and compares the distinction of three major paradigmatic semantic relations across the three word classes of nouns, verbs and adjectives. We deliberately chose synonymy, antonymy and hypernymy as our target relations because (a) as illustrated above, they play a major role in the organisation of the mental lexicon but nevertheless differ in how natural and important they are for the organisation of the lexica across word classes, and because (b) they are notoriously difficult to be distinguished by distributional models. The questions we address in the current study are the following:

- Can humans and distributional approaches reliably distinguish between synonyms, antonyms and hypernyms across word classes?
- Which class–relation combinations are easy/difficult for humans and which are easy/difficult for distributional approaches?
- Does the ease in relation distinction reflect the naturalness of a relation type for a word class?

We expected that differences in the naturalness of relations across word classes should be reflected by (a) how humans perceive and distinguish semantic relatedness, and by (b) how successful standard distributional approaches are in modelling semantic relatedness.

For the cognitive perspective, we rely on an existing dataset of paradigmatic semantic relation pairs for German (Scheible and Schulte im Walde 2014). Most crucially, the dataset contains ratings of relation strength provided by human judges, for positive as well as for negative relation instances; in addition, the selection of relation pairs across word classes in the dataset is balanced for the number of positive and negative instances, semantic class, frequency and polysemy. For the computational perspective, we rely on distributional similarity scores from standard vector space models as obtained from a large

web corpus, and a simple supervised classification model.<sup>1</sup> Our study demonstrates that the reliable distinction between relations indeed depends on word classes, both for humans and for distributional approaches. Easy vs. difficult class–relation combinations however differ for humans vs. computational models, with stronger ties between ease and naturalness of class-dependent relations for humans.

More specifically regarding our distributional approaches, we demonstrate not only that (a) the models behave differently across word classes, but also that (b) distributional similarity by itself is indeed a difficult starting point for distinguishing paradigmatic relations; nevertheless, (c) even a simple classification model is able to distinguish between relations. Last but not least, we demonstrate that the distributional feature types in the computational models have different strengths and weaknesses in distinguishing between specific paradigmatic relations for specific word classes, which is why exploring feature variants is still a worthwhile subtask in this line of research.

In the remainder of this article we first provide an in-depth overview of previous work on paradigmatic semantic relations in the (mental) lexicon as well as variants of human rating collections and computational approaches regarding paradigmatic relation distinction (Section 2). In Section 3 we describe the human ratings and the distributional information underlying our analyses and classification experiments in the main body of this article (Section 4).

## 2

## RELATED WORK

### 2.1

### *Paradigmatic semantic relations in the lexicon*

The term ‘paradigmatic’ goes back to de Saussure (1916), who introduced a distinction between linguistic elements based on their position relative to each other. This distinction derives from the linear

---

<sup>1</sup>Note that in this study we do not aim to offer in-depth comparisons of multiple distributional representations and algorithms but rather focus on simple standard approaches, given that our goal is not an optimisation of representations and algorithms but exploring the ground distributional information.

nature of linguistic elements, which is reflected in the fact that speech sounds follow each other in time. Saussure refers to successive linguistic elements that combine with each other as ‘syntagma’, and thus the relation between these elements is called ‘syntagmatic’. On the other hand, elements that can be found in the same position in a syntagma, and which could be substituted for each other, are in a ‘paradigmatic’ relationship. While syntagmatic and paradigmatic relations can occur between a variety of linguistic units (such as phonemes, morphemes, words, clauses, sentences), the focus of this research is on paradigmatic relations between words.

A long-standing methodology to explore semantic relations in the mental lexicon makes use of free association norms: Researchers have analysed the (semantic) relationships between target stimuli and their associations, where participants were requested to provide the first word(s) that came to mind when presented with the stimuli. Depending on the collected norms, the stimuli were drawn from just one or across several word classes. Given that the provided associations also vary across word classes, association norms provide a means to investigate the relationships between the stimuli and their associations, among which paradigmatic relations represent a dominant role. In this vein, we provide a brief overview of prominent association norms and relevant semantic analyses.

Following an idea originally suggested by Francis Galton in 1880, the first association norms were collected by Kent and Rosanoff (1910), for 100 English noun and adjective stimuli. The Kent and Rosanoff stimuli were translated into German, allowing for the collection of parallel association norms in German (Russell and Meseck 1959; Russell 1970). Another well-studied collection was assembled by Palermo and Jenkins (1964), comprising associations for 200 words across various parts-of-speech. The Edinburgh Association Thesaurus (Kiss *et al.* 1973) was a first attempt to collect association norms on a larger scale, and to create a network of stimuli and associates, starting from a small set of stimuli derived from the Palermo and Jenkins norms. On a much larger scale, the association norms from the University of South Florida (Nelson *et al.* 1998) were compiled over the course of more than 20 years. More than 6,000 participants produced nearly three-quarters of a million responses to 5,019 stimulus words. The currently largest-scale norms are being collected by de Deyne and colleagues,

who run an online<sup>2</sup> collection of associations across 13 languages, containing already >10 million stimulus-associate pairs (de Deyne *et al.* 2013).

A major line of research has relied on association norms to investigate the relations between the stimuli and their associations. Regarding paradigmatic and syntagmatic relations, Clark (1971) categorised stimulus-association relations into sub-categories by establishing rules, such as the paradigmatic *minimal-contrast rule* asserting that humans produce associations which are antonymous to the stimuli across word classes, and the syntagmatic *selectional feature realisation rule* asserting that humans produce selectionally preferred complements, also across word classes. Heringer (1986) focused on syntagmatic associations to a small selection of 20 German verbs. He asked his subjects to provide question words as associations (e.g., *wer* ‘who’, *warum* ‘why’), and used the responses to investigate the valency behaviour of the verbs. Bagger Nissen and Henriksen (2006) systematically distinguished between syntagmatic and paradigmatic relations across word classes when comparing associations to nouns, verbs and adjectives for English L1 and L2 adult speakers. They observed different response patterns across the word classes: Regarding paradigmatic relations, for both L1 and L2 they found more paradigmatic responses for nouns than for adjectives, and more for adjectives than for verbs.

To the best of our knowledge, only a small number of investigations distinguished *between* paradigmatic relations in association norms. Schulte im Walde *et al.* (2008) collected and analysed free associations to 409 German nouns and 330 German verbs. They performed detailed analyses at the syntax-semantics interface, and quantified the part-of-speech categories of the associate responses, the syntagmatic co-occurrences, and the syntagmatic and paradigmatic relationships between the stimuli and the associations. Regarding paradigmatic relations, they relied on *GermaNet* (Hamp and Feldweg 1997; Kunze 2000), the German equivalent of *WordNet* (Fellbaum 1998b), where they found paradigmatic relationships for 47% of the verb-verb stimuli-associate tokens and for 17% of the noun-noun stimuli-associate tokens. Most of the verb-verb pairs were in some hypernymy relation (43% co-hyponymy, 26% hyponymy, 21% hypernymy);

---

<sup>2</sup><https://smallworldofwords.org/en/project/stats>

ditto for the noun-noun pairs (47%, 6%, 29%, respectively). Guida and Lenci (2007) replicated most of their analyses on verb association norms for 312 Italian verbs. They found a much larger proportion of verb-verb synonymy (38.3%) and antonymy (4.5%) and a smaller number of hypernymy relations (11.7% co-hyponymy, 5.9% hyponymy, 22.8% hypernymy).

Apart from research on paradigmatic relations that relied on word association norms, there is an enormous body of work that provides theoretical conceptualisations of these relations in the mental lexicon. A seminal description of lexical relations (with a strong focus on antonymy) can be found in Cruse (1986). He states that paradigmatic relations “*reflect the way infinitely and continuously varied experienced reality is apprehended and controlled through being categorised, subcategorised and graded along specific dimensions of variation*”. Cruse describes and exemplifies types and sub-types of paradigmatic relations across word classes. Murphy (2003) focuses on the representation of paradigmatic relations in the lexicon, discussing synonymy, antonymy, contrast, hyponymy and meronymy, also across word classes. In her view, antonymy is a sub-type of contrast within a binary paradigm, and as in Cruse (1986) her analyses on antonymy are “*over-represented, since it is the most controversial semantic relation in terms of whether it is an arbitrary relation among words or a predictable relation among word meanings or concepts*”. Most of her discussions concern linguistic vs. meta-linguistic representations of relations, reference of relations to words vs. concepts, and lexicon storage.

In addition, a series of linguistic and psycholinguistic studies in the 1980s and 1990s investigated paradigmatic relations, typically restricted to either nouns or adjectives, and to a selection of relations. For example, Lehrer and Lehrer (1982), Charles and Miller (1989), Gross *et al.* (1989), Justeson and Katz (1991, 1992) and Murphy and Andrew (1993) studied antonymy and synonymy of adjectives. Chaffin and Herrmann (1981, 1984) looked at various relations mainly for nouns and adjectives, and a selection of syntagmatic verb-noun relations. Winston *et al.* (1987) developed a taxonomy for nominal meronymy, and Chaffin and Glass (1990) explored reading time differences for nominal hypernyms vs. synonyms.

Closest to our work and, as far as we know, the only studies that systematically explored and compared types of paradigmatic relations

across word classes, are those related to the organisation of the Princeton *WordNet*. While the most detailed descriptions are available from a special issue in the *Journal of Lexicography* (Miller *et al.* 1990; Gross and Miller 1990; Fellbaum 1990), Miller and Fellbaum (1991) provide a meta-level summary of relational structures and decisions across word classes. As basis for the *WordNet* organisation, Miller and Fellbaum state that “*the mental lexicon is organised by semantic relations. Since a semantic relation is a relation between meanings, and since meanings can be represented by synsets, it is natural to think of semantic relations as pointers between synsets*”. The semantic relations in *WordNet* include the paradigmatic relations synonymy, hypernymy/hyponymy, antonymy, and meronymy. Because “*no single set of semantic relations [...] is adequate for the entire lexicon: nouns, adjectives, and verbs each have their own semantic relations and their own organisation determined by the role they must play in the construction of linguistic messages*”, these paradigmatic relations are instantiated across word classes to various degrees. For nouns, *WordNet* implements a hierarchical organisation of synsets (i.e., sets of synonymous word meanings) relying on hypernymy relations, and it also provides meronymy relations. For adjectives, Miller and Fellbaum regard antonymy as the central organisational relation. Verbs are considered the most complex and polysemous word class. They are organised on a verb-specific variant of hypernymy, i.e., *troponymy*:  $v_1$  is to  $v_2$  in some manner, that operates on semantic fields which are instantiated as synsets. Troponymy itself is conditioned on entailment and temporal inclusion. In addition to synonymy and troponymy, antonymy is also considered an important relation for verbs. Overall, the *WordNet* specifications for paradigmatic relation between word classes – which themselves rely on a large body of earlier explorations – are taken as the theoretical basis for our work.

## 2.2

### *Human ratings of paradigmatic relations*

Over the years a number of datasets have been made available for studying and assessing semantic relatedness. Regarding the most famous judgements on *similarity*, Rubenstein and Goodenough (1965) obtained data from 51 subjects on 65 English noun pairs, a seminal study which was later replicated by Miller and Charles (1991)

and Resnik (1995). Finkelstein *et al.* (2002) created *WordSim353*, a set of 353 English noun-noun pairs rated by 16 subjects according to their semantic relatedness on a scale from 0 to 10. For German, Gurevych (2005) replicated Rubenstein and Goodenough's experiments after translating the original 65 word pairs into German. Schmidt *et al.* (2011) translated a subset of 280 target pairs from *WordSim353* into German, however keeping the ratings from the English source.

*TOEFL (Test of English as a Foreign Language)* is a common dataset for distinguishing *synonymy* from other relations. Each similarity question represents a multiple choice, with four alternatives for a given stem. Landauer and Dumais (1997) collected 80 TOEFL questions for English; Mohammad *et al.* (2007) collected 426 questions for German.

*BLESS* (Baroni and Lenci 2011) represents one of the earliest collections containing several semantic relations. It focuses on nouns and includes 200 distinct English concrete nouns as target concepts, equally divided between living and non-living entities, and grouped into 17 broad classes. For each target concept, *BLESS* provides related concepts connected through a semantic relation (hypernymy, co-hyponymy, meronymy, attribute, event), or through a null-relation. A similar dataset, *EVALution*, was induced from *ConceptNet* and *WordNet* and subsequently filtered (Santus *et al.* 2015). The *SimLex-999* dataset (Hill *et al.* 2015) was one of the first collections containing information across word classes. It contains 999 word pairs (666 noun, 222 verb and 111 adjective pairs) and was explicitly built to test models on capturing similarity rather than relatedness or association.

While these collections represent state-of-the-art datasets of human ratings of semantic similarity or relatedness, we are interested in judgements on specific types of relatedness and across word classes, which is covered by none of the collections. *WordNet* represents the resource that is most strongly relevant for our purposes but heavily biased towards hypernymy, while synonymy – and even more so – antonymy are represented to a much smaller degree. In addition, the strength of related pairs in *WordNet* is not quantified. Therefore, we rely on a dataset where humans first generated and then rated noun, verb and adjective pairs for synonymy, antonymy and hypernymy.

Although not many approaches in NLP have explicitly addressed the distinction between several paradigmatic semantic relations, there is a rich tradition on identifying synonyms, antonyms or hypernyms, and on distinguishing between subsets of two paradigmatic relations.

Prominent work on identifying *synonyms* was conducted by Edmonds, who employed a co-occurrence network and second-order co-occurrence (Edmonds 1997, 1998, 1999; Edmonds and Hirst 2002), and Curran who explored word-based and syntax-based co-occurrence for thesaurus construction (Curran 2002, 2003)). Van der Plas and Tiedemann (2006) compared a standard distributional approach against cross-lingual alignment; Erk and Padó (2008) defined a vector space model for word meaning in context, to identify synonyms and the substitutability of verbs.

Most computational work addressing *hypernyms* was performed for nouns, cf. the lexico-syntactic patterns by Hearst (1992) and an extension of the patterns by dependency paths (Snow *et al.* 2004). Weeds *et al.* (2004), Lenci and Benotto (2012), Santus *et al.* (2014a), Levy *et al.* (2015), Shwartz *et al.* (2016) and Nguyen *et al.* (2017) represent systems that identify hypernyms in distributional spaces. Examples of approaches that addressed the automatic construction of a hypernym hierarchy (for nouns) are Caraballo (2001), Velardi *et al.* (2001), Cimini *et al.* (2004) and Snow *et al.* (2006). Hypernymy between verbs was discussed by Fellbaum (1990), Fellbaum and Chaffin (1990) and Fellbaum (1998a).

There are comparably few approaches to the automatic induction of *antonyms*. A number of studies in the early 90s tested the co-occurrence hypothesis, e.g., Charles and Miller (1989), Justeson and Katz (1991), Fellbaum (1995), and another set of approaches in the last decade elaborated on the distributional properties of antonyms regarding syntagmatic co-occurrence, their discourse functions, and their canonicity (Paradis *et al.* 2009; Jones *et al.* 2012; Paradis 2016). In natural language processing, approaches to antonymy were to a large extent driven by text understanding efforts, or embedded in a larger framework aiming to identify contradiction (Lucerto *et al.* 2004; Harabagiu *et al.* 2006; Mohammad *et al.* 2008; de Marneffe *et al.* 2008).

A main emphasis regarding the distinction *between* paradigmatic semantic relations has been on systems addressing **synonyms vs. antonyms**. Lin *et al.* (2003) used patterns and bilingual dictionaries to retrieve distributionally similar words, and relied on clear antonym patterns such as ‘either X or Y’ in a post-processing step to distinguish synonyms from antonyms. Yih *et al.* (2012) developed a Latent Semantic Analysis (LSA) approach incorporating a thesaurus. Chang *et al.* (2013) extended this approach to induce vector representations that can capture multiple relations. The study by Mohammad *et al.* (2013) evaluated a thesaurus-based approach, where word pairs that occurred in the same thesaurus category were assumed to be close in meaning and marked as synonyms, while word pairs occurring in contrasting thesaurus categories or paragraphs were marked as opposites. Whereas the above-mentioned approaches rely on additional knowledge sources, Turney (2008) developed a corpus-based approach to model relational similarity, addressing (among other tasks) the distinction between synonyms and antonyms. In a similar vein, Scheible *et al.* (2013) showed that with the use of appropriate features, the distributional difference between adjectival antonyms and synonyms can be identified via a simple word space model, and Santus *et al.* (2014c,b) used average precision to distinguish between antonyms and synonyms in standard vector spaces.

Most recently, the problem of synonym/antonym distinction has also been addressed with word embedding models. Adel and Schütze (2014) integrated coreference chains extracted from large corpora into a skip-gram model to create word embeddings that identified antonyms. Ono *et al.* (2015) proposed using thesaurus-based word embeddings to detect antonyms. They suggested the implementation of a model that trains word embeddings on thesaurus information, and one model that incorporated distributional information into the thesaurus model. Pham *et al.* (2015) introduced a multitask lexical contrast model by incorporating WordNet into a skip-gram model to train semantic vectors to predict contexts. Nguyen *et al.* (2016a) proposed two approaches that make use of lexical contrast information in distributional standard vs. word embeddings vector spaces. One approach strengthened word features that were most salient for determining word relatedness, assuming that feature overlap in synonyms is stronger than feature overlap in

antonyms; the other model was an extension of a skip-gram model with negative sampling (Mikolov *et al.* 2013) that integrated the lexical contrast information into the objective function. Nguyen *et al.* (2016b) presented a neural network model that exploited lexico-syntactic patterns from syntactic parse trees and in addition integrated the distance between the related words along the syntactic path as a feature.

Regarding pattern-based approaches to identify and distinguish lexical semantic relations in more general terms, Hearst (1992) was the first to propose lexico-syntactic patterns as empirical pointers towards relation instances, focusing on hyponymy. Girju (2003) applied a single pattern to distinguish pairs of nouns that are in a causal relationship from those that are not, and Girju *et al.* (2006) extended the work towards part-whole relations, applying a supervised, knowledge-intense approach. Chklovski and Pantel (2004) were the first to apply pattern-based relation extraction to verbs, distinguishing five non-disjoint relations (similarity, strength, antonymy, enablement, happens-before). Pantel and Pennacchiotti (2006) developed *Espresso*, a weakly-supervised system that exploits patterns in large-scale web data to distinguish between five noun-noun relations (hyponymy, meronymy, succession, reaction, production). Similarly to Girju *et al.* (2006), they used generic patterns, but relied on a bootstrapping cycle combined with reliability measures, rather than manual resources.

Whereas each of the aforementioned approaches considered maximally two paradigmatic relations and one word class, only a small number of approaches were systematically explored across these relations and classes: Yap and Baldwin (2009) employed syntactic pre-processing and an SVM-based classifier, and experimented with different corpora, to distinguish antonymy, hyponymy and synonymy, while focusing on English nouns. Schulte im Walde and Köper (2013) relied on standard corpus-based patterns to distinguish between the same three paradigmatic relations, proposing a unified framework for German nouns, verbs and adjectives. Roth and Schulte im Walde (2014) extended the pattern-based approach by incorporating discourse markers and applied their model across the same relations and the three word classes, both for English and for German.

## DATA

3

The following two subsections describe the two types of data our explorations rely on: the cognitive resource with human ratings of paradigmatic relations (Section 3.1), and the distributional information used in the computational models (Section 3.2).

### *Human ratings of paradigmatic relations*

3.1

Our database of semantic relations for German adjectives, nouns and verbs focuses on the three types of paradigmatic relations referred to as *sense-relations* by Lyons (1968, 1977): synonymy, antonymy, and hypernymy. For the collection of the database, we implemented two experiments involving human participants (Scheible and Schulte im Walde 2014). Starting with a set of target words, in the first experiment participants were asked to propose suitable synonyms, antonyms, and hypernyms for each of the targets. For example, for the target verb *befehlen* ('to command'), participants proposed synonyms such as *anordnen* ('to order'), antonyms such as *gehörchen* ('to obey'), and hypernyms such as *sagen* ('to say'). In the second experiment, participants were asked to rate the strength of a given semantic relation with respect to a word pair on a given scale. For example, participants would be presented with the pair *befehlen*–*gehörchen* and asked to rate the strength of antonymy between the two words. All word pairs were assessed with respect to all three relation types.

In the following, Section 3.1.1 provides an overview of GermaNet, from which the set of target words was drawn. Section 3.1.2 introduces the platform used to implement the experiments, Amazon Mechanical Turk. Sections 3.1.3 and 3.1.4 then describe the two experiments to collect the human rating data. The dataset is publicly available at <http://www.ims.uni-stuttgart.de/data/sem-rel-database>.

#### Target source: GermaNet

3.1.1

GermaNet is a lexical-semantic word net that provides information on semantic relations for German nouns, verbs, and adjectives. GermaNet has been modelled along the lines of the Princeton WordNet for English (Miller *et al.* 1990; Fellbaum 1998b) and shares its general design principles (Hamp and Feldweg 1997; Kunze and Wagner

1999; Lemnitzer and Kunze 2007). For example, lexical units denoting the same concept are grouped into synonym sets ('synsets'). These are in turn interlinked via conceptual-semantic relations (such as hypernymy) and lexical relations (such as antonymy). For each of the major word classes, the databases further take a number of semantic categories into consideration, expressed with top-level nodes in the semantic network (such as *Artefakt* 'artifact', *Geschehen* 'event', *Gefühl* 'feeling'). In contrast to WordNet, GermaNet also includes so-called 'artificial concepts' to fill lexical gaps and thus enhance network connectivity, and to avoid unsuitable co-hyponymy (e.g. by providing missing hypernyms or hyponyms). GermaNet also differs from WordNet in the way in which it handles parts-of-speech. For example, while WordNet employs a clustering approach for structuring adjectives, GermaNet uses a hierarchical structure similar to the one employed for the noun and verb hierarchies. Finally, WordNet and GermaNet also differ in size: While WordNet 3.0 contains a total of 117,659 synsets and 155,287 lexical units, the respective numbers for GermaNet 6.0 (which we used in the current study) are considerably smaller, with 69,594 synsets and 93,407 lexical units.

Since GermaNet is the largest database of its kind for German, and given that it encodes all types of relations that are of interest for us (synonymy, antonymy, and hypernymy), it represented a suitable starting point for our purposes.<sup>3</sup> Relying on GermaNet version 6.0<sup>4</sup> and the respective *JAVA API*, we used a stratified sampling technique to randomly select 99 nouns, 99 adjectives and 99 verbs from the GermaNet files. The random selection was balanced for:

1. the *size of the semantic classes*,<sup>5</sup> accounting for the 16 semantic adjective classes and the 23 semantic classes each for nouns and verbs, as represented by the file organisation;

---

<sup>3</sup>For reasons why we did not use GermaNet to directly extract relation pairs (i.e., it is unbalanced regarding relation types; does not contain relation quantification or negative evidence; etc.), see the end of Section 2.2.

<sup>4</sup>When we started the collection, GermaNet 6.0 represented the latest version. Information about current statistics can be found at <http://www.sfs.uni-tuebingen.de/GermaNet/>.

<sup>5</sup>For example, if an adjective GermaNet class contained 996 word types, and the total number of adjectives over all semantic classes was 8,582, and with

2. **three polysemy classes** according to the number of GermaNet senses: I) monosemous, II) two senses and III) > two senses;
3. **three frequency classes** according to the type frequency in the German web corpus *SdeWaC* (Faaß and Eckart 2013), which contains approx. 880 million words: I) *low* (200–2,999), II) *mid* (3,000–9,999) and III) *high* ( $\geq 10,000$ ).

The total number of 99 targets per word class resulted from distinguishing 3 polysemy classes and 3 frequency classes,  $3 \times 3 = 9$  categories, and selecting 11 instances from each polysemy–frequency category, in proportion to the semantic class sizes.

#### Experimental platform: Mechanical Turk

#### 3.1.2

The experiments described below were implemented in Amazon Mechanical Turk (AMT),<sup>6</sup> a web-based crowdsourcing platform which allows simple tasks (so-called HITs) to be performed by a large number of people in return for a payment. In our first experiment, human associations were collected for different semantic relation types, where AMT workers were asked to propose suitable synonyms, antonyms, and hypernyms for each of the targets. The second experiment was based on a subset of the generated synonym/antonym/hypernym pairs and asked the participants to rate each pair for the strength of synonymy, antonymy, and hypernymy between them, on a scale between 0 (minimum strength) and 5 (maximum strength). To control for non-native speakers of German and spammers, each batch of HITs included two examples of ‘non-words’ (i.e., invented words following German morphotactics) in random positions. If participants did not recognise the invented words, we excluded all their ratings from consideration. While we encouraged participants to complete all HITs in a given batch, we also accepted a smaller number of submitted HITs, as long as the workers had a good overall feedback score.

---

99 stimuli collected in total, we wanted that proportion out of 99 stimuli that corresponded to the proportion of the class size relative to the total number of adjectives  $996/8,582$  and thus randomly selected 11 adjectives from this class:  $99 * 996/8,582 \approx 11.49$ .

<sup>6</sup><https://www.mturk.com>

3.1.3

Generation experiment

The goal of the generation experiment was to collect human associations for the semantic relation types synonymy, antonymy, and hypernymy. For each of our  $3 \times 99$  adjective, noun, and verb targets, we asked 10 participants to propose a suitable synonym, antonym, and hypernym. Targets were bundled randomly in 9 batches per word class, each including 9 targets plus two invented words. The experiment consisted of separate runs for each relation type to avoid confusion between them, with participants first generating synonyms, then antonyms, and finally hypernyms for the targets, resulting in  $3 \text{ word classes} \times 99 \text{ targets} \times 3 \text{ relations} \times 10 \text{ participants} = 8,910$  target–response pairs. Table 1 provides some examples of the generated target–response pairs for each word class and each paradigmatic relation, accompanied by the number of times a specific response was given (with a maximum of 10 responses).

3.1.4

Rating experiment

In the second experiment, Mechanical Turk workers were asked to rate a given semantic relation with respect to a word pair on a 6-point scale between 0 (minimum strength) and 5 (maximum strength). The main purpose of this experiment was to identify and distinguish between “strong” and “weak” examples for specific relations across word classes. The number of times a specific response was given in the generation experiment does not necessarily indicate the strength of the relation. This is especially true for responses that were suggested by only one or two participants, where it is difficult to tell if the response is an error, or if it relates to an idiosyncratic sense of the target word that the other participants did not think of in the first instance. Crucially, in the rating experiment all word pairs were assessed with respect to all three relation types, thus asking not only for positive but also for negative evidence of semantic relation instances.

The set of word pairs used as an input was a carefully selected subset of responses acquired in the generation experiment. For each of the 99 targets and each of the semantic relations (antonymy, synonymy, and hypernymy), we included two responses: the *response with the highest frequency* (random choice if several available) and a *response*

Table 1: Examples of generated target–response pairs, and their strengths

	ANT		SYN		HYP	
NOUN	<i>Bein/Arm</i> (leg/arm)	10	<i>Killer/Mörder</i> (killer)	8	<i>Ekel/Gefühl</i> (disgust/feeling)	7
VERB	<i>Zeit/Raum</i> (time/space)	3	<i>Gerät/Apparat</i> (device)	3	<i>Arzt/Beruf</i> (doctor/profession)	5
	<i>verbieten/erlauben</i> (forbid/allow)	10	<i>üben/trainieren</i> (practise)	6	<i>trampeln/gehen</i> (lumber/walk)	6
	<i>setzen/stehten</i> (sit/stand)	4	<i>setzen/platzieren</i> (place)	3	<i>welten/bewegen</i> (wave/move)	3
ADJ	<i>dunkel/hell</i> (dark/light)	10	<i>mild/sanft</i> (smooth)	9	<i>grün/farbig</i> (green/colourful)	5
	<i>heiter/trist</i> (cheerful/sad)	2	<i>bekannt/vertraut</i> (familiar)	4	<i>heiter/hell</i> (bright/light)	1

Table 2: Examples of mean target–response ratings and mean differences

	Target Pair	Generation	ANT	SYN	HYP	Difference
NOUN	<i>Arzt/Beruf</i> (doctor/profession)	HYP: 5	0.8	1.1	4.7	HYP–SYN 3.6
	<i>Verhandlung/Gespräch</i> (negotiation/conversation)	HYP: 4	0.6	2.8	4.0	HYP–SYN 1.2
VERB	<i>befehlen/gehorchen</i> (command/obey)	ANT: 6	4.4	0.3	0.1	ANT–SYN 4.1
	<i>schmierien/streichen</i> (grease/paint)	SYN: 4	0.9	2.2	3.3	SYN–HYP –1.1
ADJ	<i>faul/fleißig</i> (lazy/diligent)	ANT: 8	5.0	0.5	0.0	ANT–SYN 4.5
	<i>gewitzt/naiv</i> (smart/naive)	ANT: 3	3.0	0.3	0.4	ANT–SYN 2.7

with a low frequency (2, if available, otherwise 1; random choice if several available). Multi-word responses and blanks were excluded.

In theory, each target should have 6 associated pairs ( $2 \times \text{ANT}$ ,  $2 \times \text{HYP}$ ,  $2 \times \text{SYN}$ ). In practice, there are sometimes fewer than 6 pairs per target in the dataset, because (a) for some targets, only one response was available for a given relation (e.g., if all 10 participants provided the same response), or (b) no valid response of the required frequency type was available. The resulting dataset includes 1,684 target–response pairs altogether, 546 of which are adjective pairs, 574 noun pairs, and 564 verb pairs. To avoid confusion, the ratings were collected in separate experimental settings, i.e., for each word class and each relation type, all generated pairs were first evaluated for the strength of one relation, and then for the strength of another relation. Table 2 provides some examples of mean ratings for target–response pairs and the three semantic relations, together with the original relation (see column *Generation*) and the strength of generation (1–10).

### 3.2

#### *Corpora and distributional information*

As a corpus for our distributional models we relied on one of the currently largest German web corpora, *DECOW14AX*, with approx. 12 billion words (Schäfer and Bildhauer 2012). It was already lemmatised and assigned part-of-speech tags by the *Tree Tagger* (Schmid 1994).

We induced two types of distributional information from the web corpus in order to create two types of vector space models (Bullinaria and Levy 2007; Turney and Pantel 2010), one using window co-occurrence and one using lexico-syntactic patterns. Regarding *window co-occurrence*, we created a standard vector space for all target and response words that were part of our relation pairs. We relied on co-occurrence frequencies from a sentence-internal 20-word window (i.e., 20 words to the left and 20 words to the right of a word in the corpus but not going beyond sentence borders, as sentences in *DECOW14AX* are scrambled) to determine the co-occurring content words and the strengths of co-occurrence. For example, if *schnurren* ‘to purr’ occurred a total of 235 times in the context of *Katze* ‘cat’ – where the context of *Katze* is defined as the 20 preceding and the 20 following words – then the dimension *schnurren* for the target word

*Katze* in the vector space was assigned the frequency 235. To compare different windows sizes and vector space strengths, we also used co-occurrence information from a 5-word window, and we also compared co-occurrence frequencies with *local mutual information (lmi)* scores, cf. Evert (2005), which often provide better estimates for word co-occurrence strength. The window co-occurrence information refers to words (i.e., it provides co-occurrence vectors for target or response words such as *Katze* ‘cat’) rather than to target–response word pairs (such as *Katze–Tier* ‘cat–animal’), so Section 4.2 will explain how to induce vectors for word pairs from the vectors of individual words.

Regarding *lexico-syntactic patterns*, we directly induced a vector space for the word-relation pairs (Hearst 1992; Chklovski and Pantel 2004, i.a.). I.e., we relied on the linear word sequences  $l_1 \dots l_n$  in the corpus between any two related words  $w_i$  and  $w_j$  (representing synonyms, antonyms or hypernyms) to initiate the vector space dimensions for the relation pair  $w_i-w_j$ . For example, if we saw the hypernymy pair *Katze–Tier* ‘cat–animal’ in the token sequence “... **Tier** wie *Huhn, Taube, Katze* ...”, the respective lexico-syntactic pattern (and, correspondingly, one dimension in the vector space) was the intermediate sequence “*wie Huhn, Taube,*”. We distinguished between two sub-types of patterns in our vector representations, those taking into account the linear order of the words  $w_i$  and  $w_j$  (i.e., patterns distinguishing between  $w_i l_1 l_2 \dots l_n w_j$  and  $w_j l_1 l_2 \dots l_n w_i$ ), and those without taking the direction into account.

## DISTINGUISHING PARADIGMATIC RELATIONS

4

As outlined in the Introduction, our research brings together perspectives from cognitive lexical semantics and distributional semantics, and compares the distinction of paradigmatic semantic relations for German across the three word classes of nouns, verbs and adjectives:

- Can humans and distributional approaches reliably distinguish between synonyms, antonyms and hypernyms across word classes?

- Which class–relation combinations are easy/difficult for humans and which are easy/difficult for distributional approaches?
- Does the ease in relation distinction reflect the naturalness of a relation type for a word class?

We expect that differences in the naturalness of paradigmatic relations across word classes are reflected in how humans perceive and distinguish semantic relatedness (Section 4.1), and in the performance of corpus-based distributional approaches (Section 4.2).

## 4.1

*Human distinction*

For the cognitive perspective, we rely on the dataset of human-generated paradigmatic semantic relation pairs rated for their relation strength as described in the rating experiment in Section 3.1.4. We disregarded relation pairs that were originally generated only once, and we also disregarded ambiguous pairs, i.e. pairs that were generated for more than one relation type. For example, the noun *Erde* ('soil') was generated both as synonym (3 times) and as hypernym (twice) for the target noun *Torf* ('peat').

Table 3 shows the numbers of relation pairs across word classes and relation types with respect to the originally generated relation. The table also compares pairs excluding vs. including ambiguity (–/+*amb*, respectively). It is already interesting to observe that for relation pairs involving hypernymy and synonymy (HYP and SYN) there was considerably more ambiguity among the generated relation pairs than for antonymy (ANT): For verbs and adjectives, for which hypernymy represents a less natural semantic relation than for nouns, only 29.3–34.2% of the considered generated pair types

Table 3:  
Number of relation pairs  
in the dataset

		ANT	HYP	SYN	<i>all</i>
NOUN	– <i>amb</i>	101	91	82	274
	+ <i>amb</i>	118	159	151	428
VERB	– <i>amb</i>	122	66	63	251
	+ <i>amb</i>	132	193	193	518
ADJ	– <i>amb</i>	127	54	58	239
	+ <i>amb</i>	133	184	189	506

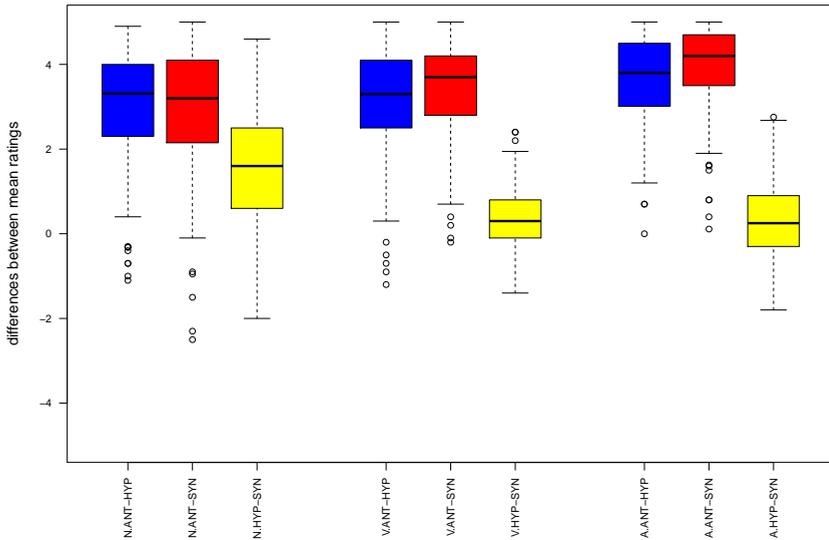
were unambiguous, while for nouns the unambiguous pairs correspond to  $\approx 55\%$ . As mentioned above, the *-amb* dataset represents the basis for exploring differences in relation distinction across word classes by humans. For completeness, Appendix A.1 provides the human distinction results for ambiguous pairs, in comparison to those for unambiguous pairs.

In order to assess how well the experiment participants could distinguish between the paradigmatic relations, we calculated the differences in mean ratings for a specific relation pair. For example, we obtained a mean rating of 4.4 on our scale 0–5 from the experiment participants for the antonym pair *befehlen–gehorschen* ('command–obey') regarding antonymy, and we obtained a mean rating of only 0.3 for this pair regarding hypernymy, so the difference in the mean ratings was 4.1. Obviously, the experiment participants were rather sure that the target pair represented antonymy, and they were also rather sure that the target pair did not represent hypernymy. In contrast, the difference in mean ratings for the antonym pair *bedürfen–verzichten* ('require–abstain') regarding antonymy vs. hypernymy ratings was only 2.1, demonstrating that the latter antonym pair represented a weaker instance of antonymy for the experiment participants. Table 2 provides differences in mean ratings for further example target–response pairs (see column 'Difference').

Figures 1 and 2 present these mean differences for each word class and across all relation pairings. Figure 1 provides a coarse view on relation distinction and does not tell us which relation was the original relation and which was the rated relation (e.g., whether a pair has been generated as a synonym pair and then rated for synonymy vs. antonymy, or whether a pair has been generated as an antonym pair and then rated for antonymy vs. synonymy); Figure 2 then incorporates this distinction.

The figures illustrate that the experiment participants found it easier to distinguish between antonyms and hypernyms (ANT–HYP, blue boxes) as well as between antonyms and synonyms (ANT–SYN, red boxes), where the differences in mean ratings between the original and the rated relation are larger, in comparison to distinguishing between hypernyms and synonyms (HYP–SYN, yellow boxes), where the differences in mean ratings for the two relations are smaller. These findings hold across word classes, but we can also see that the ten-

Figure 1:  
Human  
distinction of  
paradigmatic  
relations  
(coarse)



density is stronger for adjectives and verbs (in comparison to nouns) where the differences are  $\approx 0$ , i.e., the mean ratings for synonyms and hypernyms regarding a specific word pair were nearly identical.

The fine-grained analysis in Figure 2 in addition demonstrates that adjectival HYP–ANT is more difficult for the humans than adjectival ANT–HYP, and that adjectival HYP–SYN is more difficult than adjectival SYN–HYP (in both cases the boxes do not even overlap); to a lesser degree we find the same dispute between verbal HYP–ANT and ANT–HYP (where the median of the former is outside the box of the latter). Interestingly, in all these three cases the differences between mean ratings were lower when the original relation was hypernymy, which represents a less natural semantic relation for verbs and adjectives than for nouns, i.e. the experiment participants did not perceive the generated hypernyms as strong instances of that relation type in comparison to the respective other paradigmatic relation.

Overall, the differences in mean ratings suggest (a) that humans clearly distinguish antonyms from synonyms and also from hypernyms, but have more difficulties in distinguishing between synonyms and hypernyms, and (b) that distinguishing hypernymy from the other two relations is more difficult for adjectives and verbs (in comparison to nouns), for which hypernymy represents a less natural semantic relation. The boxplots in Appendix A.1 – which compare the coarse- and

## Distinguishing paradigmatic semantic relations

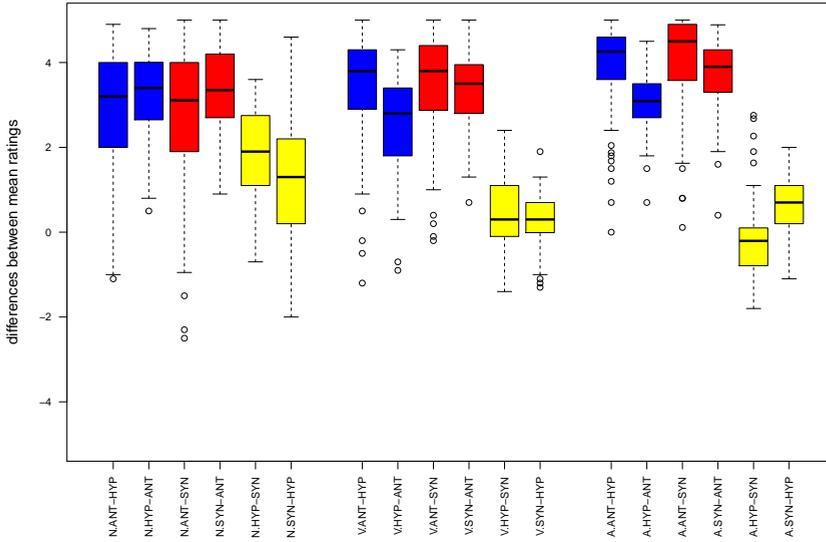


Figure 2:  
Human  
distinction of  
paradigmatic  
relations (**fine**)

fine-grained analyses in Figures 1 and 2 against the respective analyses on relation pairs including ambiguity – confirm these insights.

### *Distributional classification models*

4.2

For the computational perspective, we explore two levels of processing the distributional co-occurrence information in the standard vector space models introduced in Section 3.2. We start out with cosine distances between any two word pairs within the set of target–response pairs, in order to illustrate the difficult basis of a distributional model for distinguishing between paradigmatic relations (Section 4.2.1). In a series of supervised classification experiments we then present the results of automatically categorising the target–response pairs into semantic relations (Section 4.2.2).

#### Cosine similarities between relation pairs

4.2.1

As explained in Section 3.2, we rely on corpus co-occurrences to activate and quantify dimensions in word vectors (Bullinaria and Levy 2007; Turney and Pantel 2010). The geometric distance between two word vectors then determines the distance between the two words. The closer two vectors are in the vector space, the more semantically

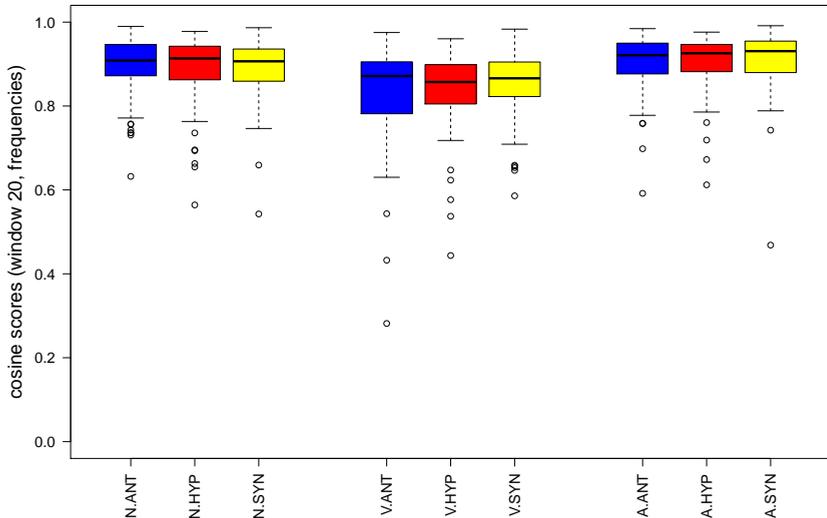
related we expect the represented words to be, based on the Distributional Hypothesis (Harris 1954; Firth 1957).

Regarding paradigmatic semantic relations, the generally agreed upon assumption is that the related word pairs are relatively close to each other in word space across the relation types, because for all paradigmatic relations the related words are distributionally similar to each other. In the following, we explore this assumption for our dataset.

We calculated the cosine scores between the target words and the response words for each target–response pair. The cosine score specifies the angle between two vectors, with 1 indicating minimal distance (i.e., identity, and therefore maximal relatedness) between the vectors. We used the same set of unambiguous rated pairs as exploited by Figures 1 and 2, together with the respective co-occurrence vector spaces. Figures 3 and 4 present boxplots of cosine scores for all word pairs across word classes and semantic relations, relying on co-occurrence frequencies within 20-word windows, and on the corresponding vectors with lmi scores.

The plots illustrate that the cosine values are indeed very similar across our three paradigmatic relations for a specific word class, with slightly lower scores for verb relatedness. The lmi scores obviously influence the magnitudes of the cosine scores, and they manage

Figure 3:  
Boxplots of cosine scores across classes and relations (window 20, frequencies)



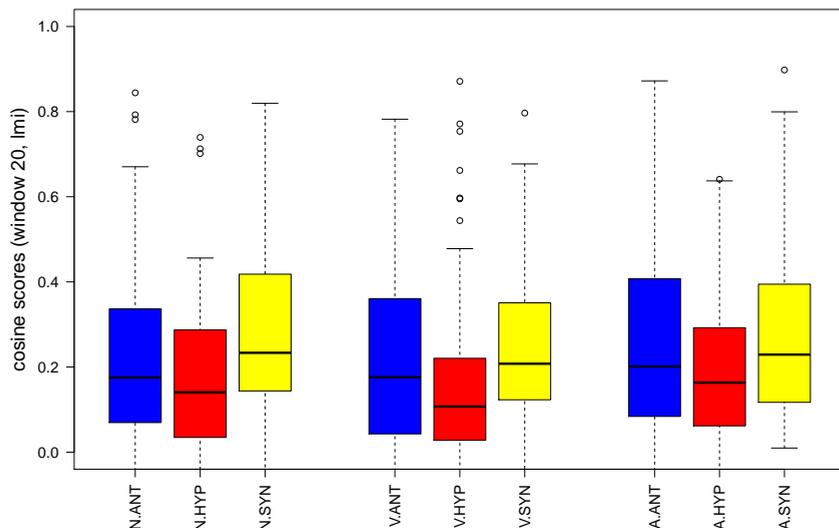


Figure 4:  
Boxplots of  
cosine scores  
across classes  
and relations  
(window 20,  
**lmi** scores)

to disperse them. Appendix A.2 illustrates that the same tendencies can be observed for 5-word co-occurrence windows, and also when extending the underlying dataset with ambiguous word pairs.

#### Automatic classification of relation pairs

#### 4.2.2

In a series of classification experiments relying on the distributional word spaces we explored whether automatic approaches are able to categorise word pairs according to their paradigmatic semantic relations, even though the vectors of the word pairs are all very close in vector space. In the following, we present classification results of a simple *nearest-centroid classifier* (also known as *Rocchio classifier*, cf. Manning *et al.* 2008) that compares window-based co-occurrence features against pattern-based co-occurrence features. A subset of the classification experiments was previously described by Schulte im Walde and Köper (2013) and David (2014), but was re-implemented and re-run for the current article to ensure the same underlying target pairs and corpus data across approaches.

The classification was done as follows. For each word class separately, we calculated three mean vectors: one for each lexical semantic relation (synonymy, antonymy, hypernymy), as based on a set of training pairs. We then predicted the semantic relation for a set of test pairs, by choosing for each test pair the most similar class mean vector

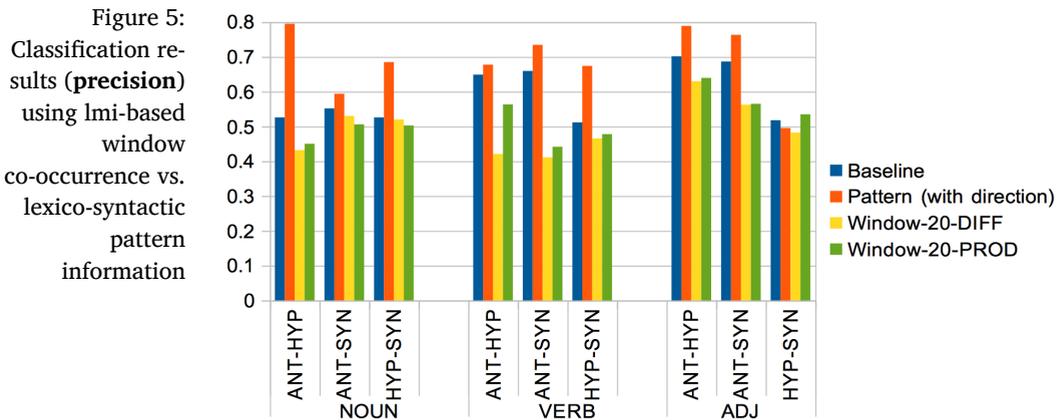
as determined by the respective cosine scores. Across the experiments, we used 5-fold cross-validation for training and testing.

The classification setup for the pattern-based vectors is straightforward, because a pattern vector represents a word pair. The window co-occurrence vectors however represent words and not word pairs and thus require a preprocessing step to obtain vectors for word pairs. We applied two variants to initiate window co-occurrence vectors for the target–response pairs, as based on their individual word vectors:

**WINDOW-DIFF:** For each target–response word pair, we calculated the difference vector between the two involved word vectors, i.e., the value of each dimension in the difference vector is computed as the absolute difference between the respective values in the two word vectors. The centroids of the relation classes correspond to mean difference vectors.

**WINDOW-PROD:** For each target–response word pair, we calculated the product vector for the two involved word vectors, i.e., the value of each dimension in the product vector is computed as the product of the respective values in the two word vectors. The centroids of the relation classes correspond to mean product vectors.

Figures 5 and 6 present the results of the nearest-centroid classifier across word classes, relations, and types of distributional information. While Figure 5 shows the results in terms of precision (i.e., the average proportion of correct class assignments among all classified instances of relation pairs), Figure 6 shows the results in terms



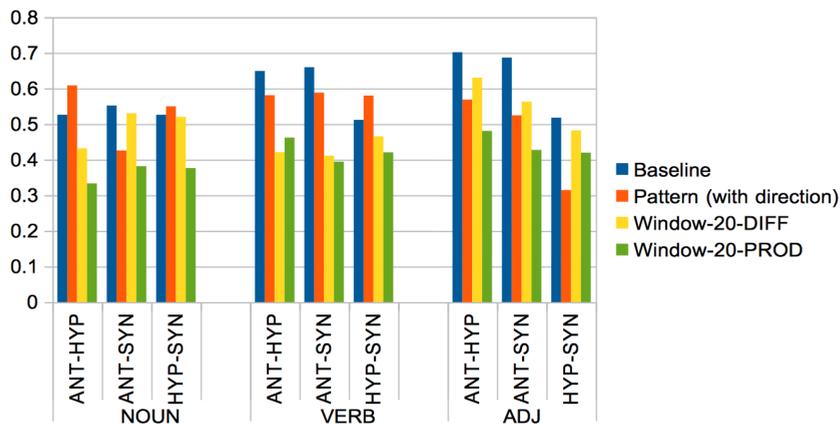


Figure 6: Classification results (**accuracy**) using lmi-based window co-occurrence vs. lexico-syntactic pattern information

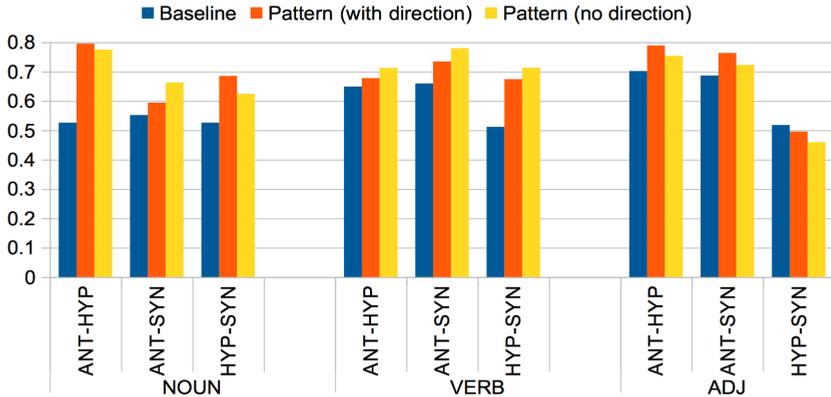
of accuracy (i.e., the average proportion of correct class assignments among all existing instances of relation pairs).

Both Figures 5 and 6 compare vector spaces with lmi scores for pattern-based features with direction (i.e., patterns distinguishing between  $w_i \langle pattern \rangle w_j$  and  $w_j \langle pattern \rangle w_i$ ), and window-based features relying on a 20-word co-occurrence window. We decided in favour of lmi-score vector spaces rather than frequency vector spaces, because our analyses in Section 4.2.1 indicated that lmi scores disperse the cosine scores in the vectors. Results of other classification variants (i.e., relying on frequencies; pattern-based features without direction information; window-based features relying on a 5-word co-occurrence window) are described in Appendix A.3.

Figure 5 shows that – regarding precision – pattern information in most cases outperforms not only the respective majority baseline but also the two variants of window information. The only exception takes place for distinguishing between adjectival HYP–SYN. And also except for this very case, classification based on window features is consistently worse than the baselines. WINDOW-DIFF and WINDOW-PROD results do not show consistent differences, except for verbal ANT–HYP, for which WINDOW-PROD clearly outperforms WINDOW-DIFF. Figures 14 and 15 in Appendix A.3 illustrate that the same tendencies are found for frequency-based vector spaces, which are however overall worse than lmi-based vector spaces.

Figure 6 shows that – regarding accuracy – most of the classification results are below the majority baseline. Pattern-based results

Figure 7:  
Classification  
results  
(precision)  
using lmi-based  
pattern  
information



are only above baseline results for nominal ANT–HYP and HYP–SYN as well as for verbal HYP–SYN distinctions; window-based results do not outperform the baselines in any of the scenarios. In cases where both pattern-based and window-based results are below the baselines, pattern-based results outperform window-based results for all verbal relation distinctions; window-based results outperform pattern-based results for all adjectival relation distinctions, and for nominal ANT–SYN. In most cases, WINDOW-DIFF clearly outperforms WINDOW-PROD.

Figure 7 provides a view that is quite alike Figure 5, zooming into the overall best results<sup>7</sup> when using pattern-based information. First of all, Figure 7 compares the classification results with/without using pattern direction information. We can see that there are no consistent differences between the two representations: the patterns without directional information are slightly better for verbs; and the patterns with directional information are slightly better for adjectives. The results for nouns depend on the relation types. Appendix A.3 provides additional information illustrating in the same manner that lmi-based patterns in general outperform lmi-based window information, both for a 20-word and a 5-word window.

Moreover, comparing our pattern-based classification results in Figure 7 with the coarse view on human relation distinction in Figure 1,

<sup>7</sup> For the remainder of the paper, we will explore precision rather than accuracy results because we are interested in the qualitative feature potential, disregarding data sparsity issues.

we do not see much overlap in general tendencies. In relation to the respective baselines, nominal ANT–HYP and nominal and verbal HYP–SYN distinctions are handled particularly well in the automatic classifications; adjectival HYP–SYN distinction is particularly bad. This provides a very different story than the human distinctions, where HYP–SYN were consistently distinguished more poorly than the other relation combinations, across word classes.

Tables 4–9 provide confusion matrices for a more detailed view on correct and wrong relation classifications. Here we took into account all class assignments of relation pairs in the respective 5-fold cross-validation, a total of  $N = 1,528$  across word classes and relation combinations. For each word class and relation, we calculated the number of pairs classified correctly/wrongly, or not at all.<sup>8</sup> The diagonal numbers in bold font indicate the correct class assignments, and the accuracy  $acc_N$  indicates the proportion of those correct classifications regarding  $N$ .

Comparing the lmi-based Tables 4–6, the  $acc_N$  scores confirm that pattern-based information outperforms both variants of window-based information. We can also observe differences across word classes and relation types. For example, the patterns are extremely useful for identifying verbal antonyms, while WINDOW-DIFF is crucial for discover-

<sup>8</sup>A word pair was not classified at all if all vector feature values of at least one of the words were zero. This happened if one or both of the words did not occur in the corpus, or if the words did not co-occur (in the case of patterns), or after multiplying feature values.

		ANT	HYP	SYN	NONE	<i>all</i>
NOUN	ANT	<b>101</b>	21	24	56	202
	HYP	10	<b>135</b>	9	28	182
	SYN	33	34	<b>54</b>	43	164
VERB	ANT	<b>152</b>	31	18	43	244
	HYP	21	<b>90</b>	15	6	132
	SYN	21	21	<b>51</b>	33	126
ADJ	ANT	<b>139</b>	21	20	74	254
	HYP	7	<b>44</b>	25	32	108
	SYN	11	11	<b>52</b>	42	116
		$acc_N = 0.5353$			$N = 1,528$	

Table 4:  
Confusion matrix for class assignment using **lmi-based pattern** features (with direction)

Table 5:  
Confusion matrix  
for class assignment using  
**lmi-based WINDOW-DIFF**  
features

		ANT	HYP	SYN	NONE	<i>all</i>
NOUN	ANT	<b>118</b>	77	7	0	202
	HYP	32	<b>145</b>	5	0	182
	SYN	79	78	<b>7</b>	0	164
VERB	ANT	<b>73</b>	91	80	0	244
	HYP	18	<b>91</b>	23	0	132
	SYN	29	46	<b>51</b>	0	126
ADJ	ANT	<b>189</b>	25	40	0	254
	HYP	42	<b>27</b>	39	0	108
	SYN	41	19	<b>56</b>	0	116
$acc_N = 0.4954$						$N = 1,528$

Table 6:  
Confusion matrix  
for class assignment using  
**lmi-based**  
**WINDOW-PROD** features

		ANT	HYP	SYN	NONE	<i>all</i>
NOUN	ANT	<b>79</b>	51	20	52	202
	HYP	27	<b>97</b>	16	42	182
	SYN	47	49	<b>23</b>	45	164
VERB	ANT	<b>117</b>	29	65	33	244
	HYP	38	<b>29</b>	38	27	132
	SYN	27	21	<b>68</b>	10	126
ADJ	ANT	<b>135</b>	21	34	64	254
	HYP	28	<b>33</b>	23	24	108
	SYN	28	18	<b>45</b>	25	116
$acc_N = 0.4097$						$N = 1,528$

Table 7:  
Confusion matrix  
for class assignment using  
**frequency-based pattern**  
features (with direction)

		ANT	HYP	SYN	NONE	<i>all</i>
NOUN	ANT	<b>132</b>	19	17	34	202
	HYP	17	<b>136</b>	7	22	182
	SYN	46	47	<b>43</b>	28	164
VERB	ANT	<b>135</b>	47	32	30	244
	HYP	18	<b>94</b>	18	2	132
	SYN	26	27	<b>51</b>	22	126
ADJ	ANT	<b>140</b>	34	33	47	254
	HYP	7	<b>67</b>	31	3	108
	SYN	10	25	<b>62</b>	19	116
$acc_N = 0.5628$						$N = 1,528$

		ANT	HYP	SYN	NONE	<i>all</i>
NOUN	ANT	<b>121</b>	42	39	0	202
	HYP	36	<b>121</b>	25	0	182
	SYN	51	47	<b>66</b>	0	164
VERB	ANT	<b>136</b>	58	50	0	244
	HYP	36	<b>65</b>	31	0	132
	SYN	32	30	<b>64</b>	0	126
ADJ	ANT	<b>151</b>	41	62	0	254
	HYP	39	<b>34</b>	35	0	108
	SYN	33	16	<b>67</b>	0	116

$acc_N = 0.5399$        $N = 1,528$

Table 8:  
Confusion matrix  
for class assignment using  
**frequency-based**  
**WINDOW-DIFF** features

		ANT	HYP	SYN	NONE	<i>all</i>
NOUN	ANT	<b>109</b>	43	32	18	202
	HYP	40	<b>101</b>	29	12	182
	SYN	42	43	<b>61</b>	18	164
VERB	ANT	<b>44</b>	99	97	4	244
	HYP	10	<b>103</b>	19	0	132
	SYN	14	43	<b>67</b>	2	126
ADJ	ANT	<b>108</b>	62	62	22	254
	HYP	22	<b>57</b>	29	0	108
	SYN	18	32	<b>60</b>	6	116

$acc_N = 0.4647$        $N = 1,528$

Table 9:  
Confusion matrix  
for class assignment using  
**frequency-based**  
**WINDOW-PROD** features

ing verbal hypernyms. WINDOW-PROD seems to overall classify more poorly than the other two feature types; it slightly outperforms them in only one case, for verbal synonyms. WINDOW-DIFF has a particular strength in that it classifies all  $N$  relation pairs (NONE = 0 for all class–relation combinations). Obviously the vectors are less sparse than for the patterns, and they do not become more sparse in the vector pair creation, differently to the WINDOW-PROD vectors.

Looking at the frequency-based Tables 7–9, we find the same tendencies regarding  $acc_N$  as for Tables 4–6, but the frequency-based  $acc_N$  values are consistently higher than the respective lmi-based  $acc_N$  values. This is in contrast to what the precision results presented in Appendix A.3 show, where the frequency-based precision results for the patterns are worse than the respective lmi-based precision results, and the results for the window-based vector spaces vary. Comparing

Tables 7 and 8, we can observe that both patterns and WINDOW-DIFF are strong in identifying antonyms across word classes; that the patterns are also strong in identifying hypernyms (and WINDOW-DIFF is less strong) across word classes; and that WINDOW-DIFF is strong in identifying synonyms (and the patterns are less strong) across word classes. Thus, the confusion matrices demonstrate in more detail than the plots that the most successful vector spaces each have their strengths and weaknesses regarding specific relation types.

5

CONCLUSION

In this article, we explored the distinction between the three paradigmatic semantic relations of synonymy, antonymy, and hypernymy, both from a cognitive linguistic perspective and a computational linguistic perspective. We expected differences in how natural relations are across word classes to be reflected in how humans perceive and distinguish semantic relatedness, and in the extent that corpus-based distributional approaches are successful in modelling semantic relatedness. More specifically, we addressed the following questions in this study:

- Can humans and distributional approaches reliably distinguish between synonyms, antonyms and hypernyms across word classes?
- Which class–relation combinations are easy/difficult for humans and which are easy/difficult for distributional approaches?
- Does the ease in relation distinction reflect the naturalness of a relation type for a word class?

Regarding the human distinction between the three paradigmatic relations, we first of all observed that among the human-generated relation pairs involving hypernymy and synonymy there was considerably more ambiguity than for antonymy. Especially for verbs and adjectives, for which hypernymy represents a less natural semantic relation than for nouns, a large proportion of the considered generated pair types were ambiguous between hypernymy and synonymy.

In addition, when looking at the differences in mean relation ratings we found (a) that humans clearly distinguished antonyms from synonyms and also from hypernyms, but had more difficulties in distinguishing between synonyms and hypernyms, and (b) that distinguishing hypernymy from the other two relations was more difficult for adjectives and verbs (in comparison to nouns), for which hypernymy represents a less natural semantic relation.

When comparing our best automatic classification results with human relation distinction, we did not find much overlap in general tendencies. Distinguishing between hypernyms and antonyms/synonyms for nouns worked particularly well, just as distinguishing hypernyms and synonyms for verbs. Overall, this provides a very different story than in the case of human distinctions, where hypernyms and synonyms were consistently distinguished more poorly than the other relation combinations across word classes.

The most interesting insights from the computational perspective arose from comparing the various feature types, where each of them showed rather different strengths and weaknesses. Overall – regarding precision – the pattern-based vector spaces clearly outperformed not only the respective majority baselines but also the two variants of window information (WINDOW-DIFF and WINDOW-PROD) for both 20-word and 5-word windows and across almost all class–relation combinations. When taking a more fine-grained look at the confusion matrices for all 1,528 individual class assignments of relation pairs, the picture was more diverse: The patterns were extremely useful in identifying verbal antonyms, while WINDOW-DIFF was crucial in discovering verbal hypernyms. WINDOW-PROD seemed to generally classify more poorly than the other two feature types; it slightly outperformed them in only one case, for verbal synonyms. WINDOW-DIFF showed a particular strength in that it classified all relation pairs; obviously the vectors were less sparse than for the patterns, and they did not become more sparse in the vector pair creation, differently to WINDOW-PROD vectors.

Overall, even though distributional similarity per se represents a difficult starting point for distinguishing paradigmatic relations (which we illustrated for our dataset), our computational explorations demonstrated that distributional classification models successfully distinguish between them. The most salient feature types

and their success varied across word classes and paradigmatic relation types.

So both for humans and for the automatic approaches, the reliable distinction between relations depends on the specific class–relation combinations. However, easy vs. difficult class–relation combinations differ for humans and computational models, exhibiting stronger ties between ease and naturalness of class-dependent relations for humans than for computational models on the one hand, and strong ties between vector space parameters and relation types on the other hand. For future work on automatic relation distinction, the latter suggests combining feature types (for example, in an ensemble) rather than applying them in isolation.

## A

## APPENDIX

### A.1 *Human distinction of relation pairs in-/excluding ambiguity*

Figures 8 and 9 compare human distinctions of relation pairs excluding ambiguity (left panels, identical to Figures 1 and 2) against human distinctions of relation pairs including ambiguity (right panels). The plots suggest that our conclusions for relation distinction regarding relation pairs excluding ambiguity (cf. Section 4.1) apply similarly to relation pairs including ambiguity: (a) humans clearly distinguish antonyms from synonyms and also from hypernyms, but have more difficulties in distinguishing between synonyms and hypernyms, and (b) distinguishing hypernymy from the other two relations is more difficult for adjectives and verbs (in comparison to nouns), for which hypernymy represents a less natural semantic relation.

### A.2 *Cosine similarities between relation pairs*

Figures 10 to 13 illustrate that neither (a) relying on 5-word instead of 20-word window co-occurrences nor (b) relying on lmi scores instead of co-occurrence frequencies nor (c) including ambiguous rela-

*Distinguishing paradigmatic semantic relations*

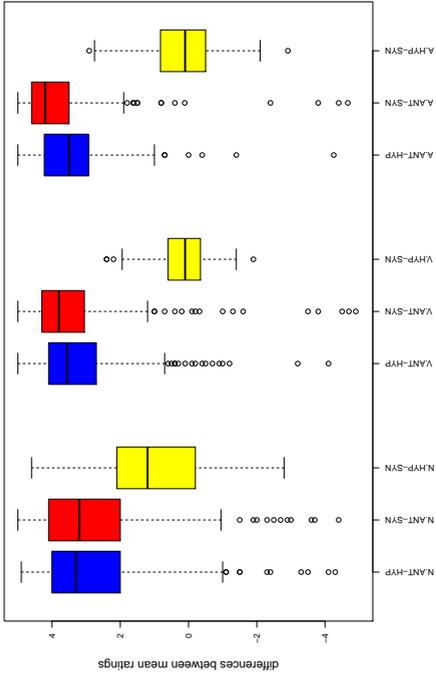


Figure 8: Human distinction of paradigmatic relations (coarse): including vs. excluding ambiguity

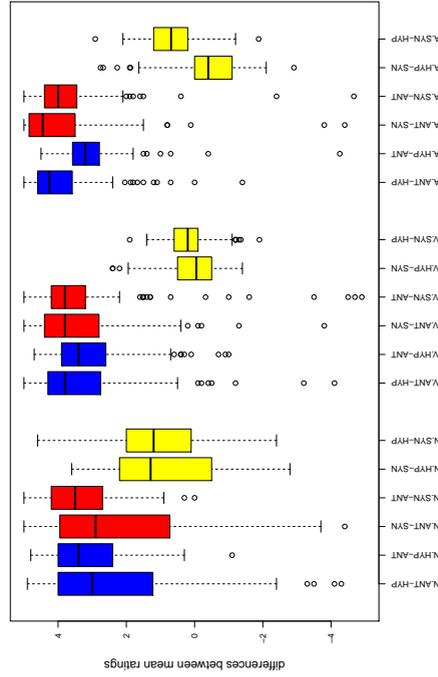


Figure 9: Human distinction of paradigmatic relations (fine): including vs. excluding ambiguity

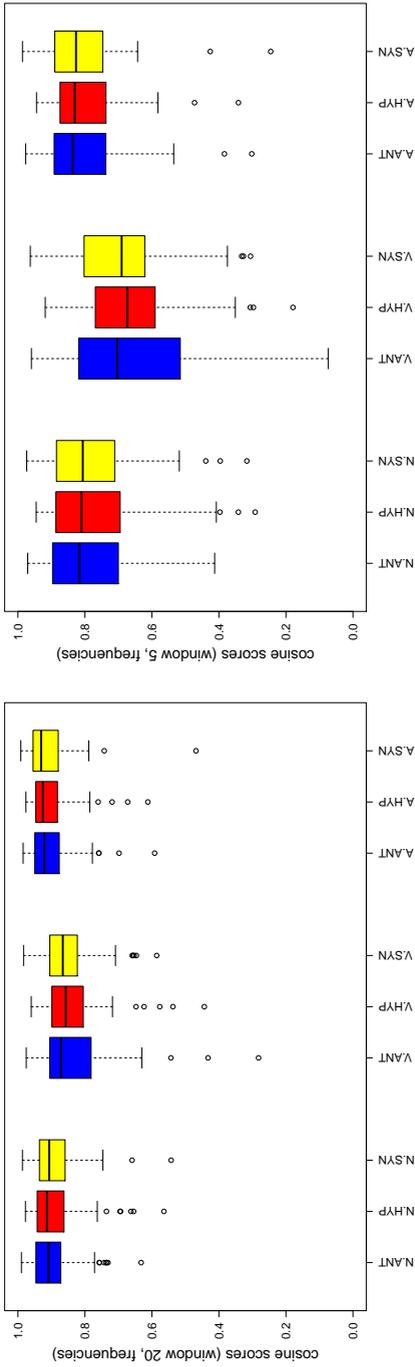


Figure 10: Boxplots of cosine scores across classes and relations (windows: 20 (left) vs. 5, frequencies)

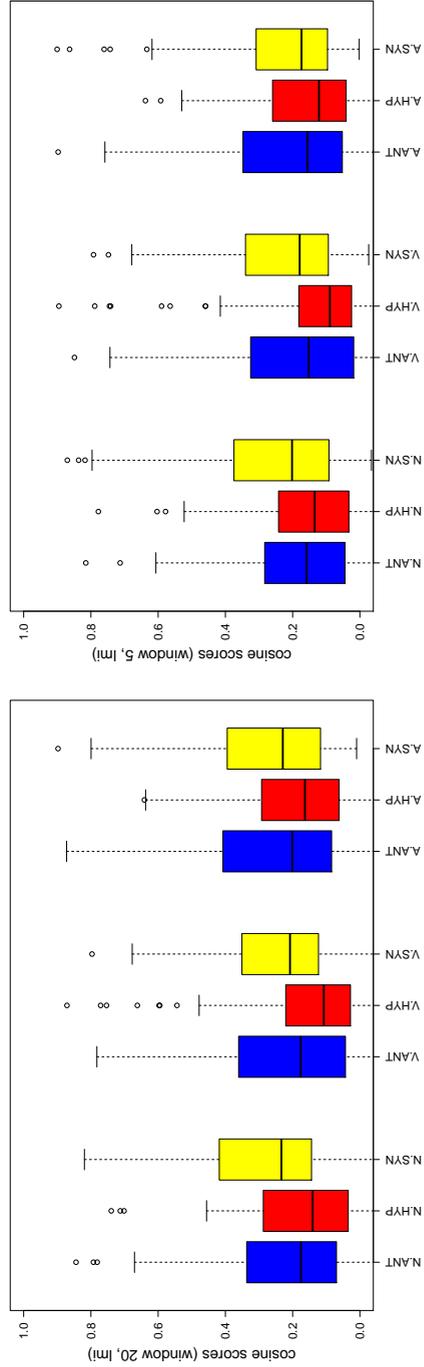


Figure 11: Boxplots of cosine scores across classes and relations (windows: 20 (left) vs. 5, lmi scores)

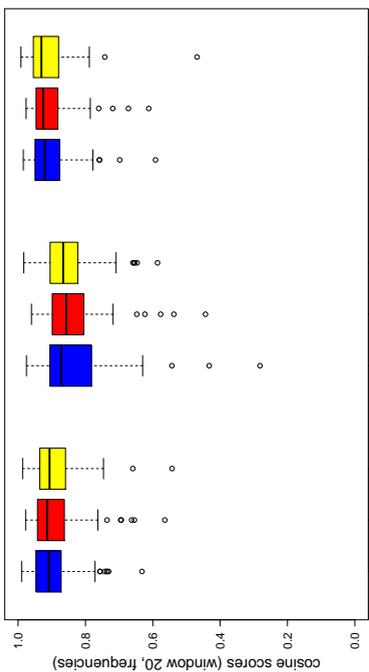
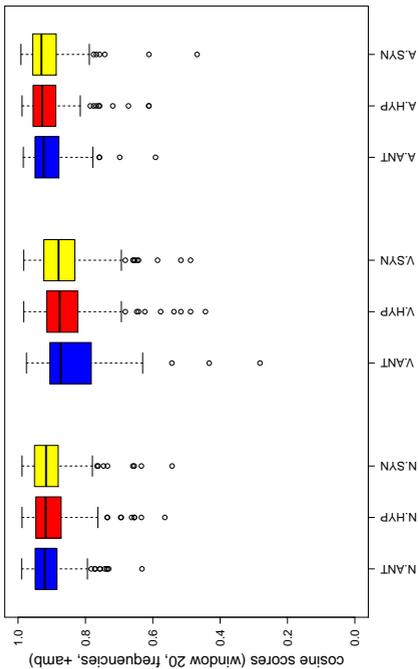


Figure 12: Boxplots of cosine scores across classes and relations (window 20, frequencies, excluding (left) vs. including ambiguity)

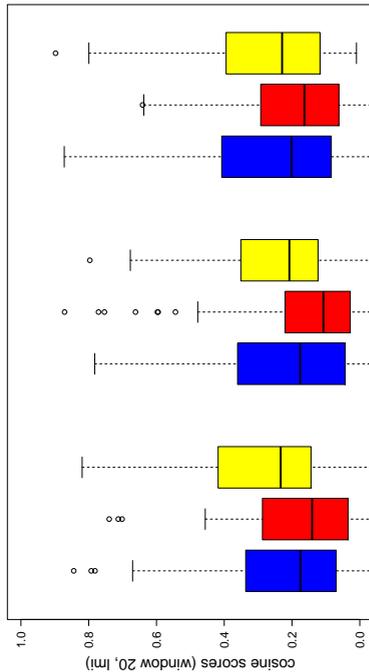
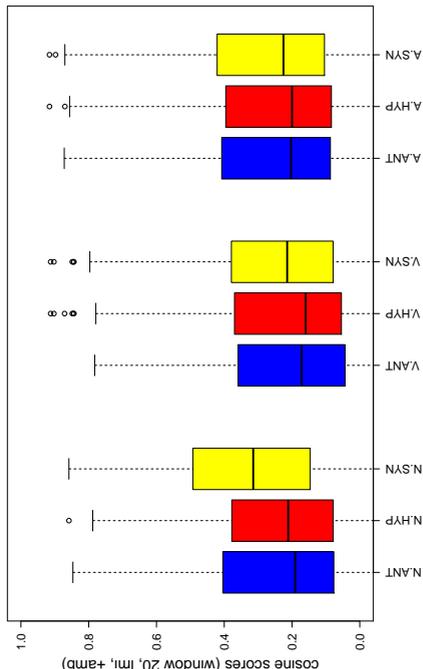
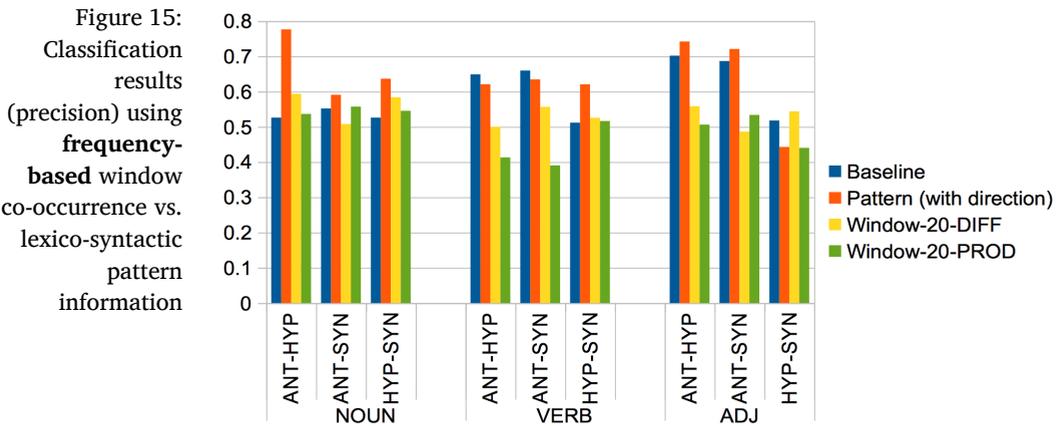
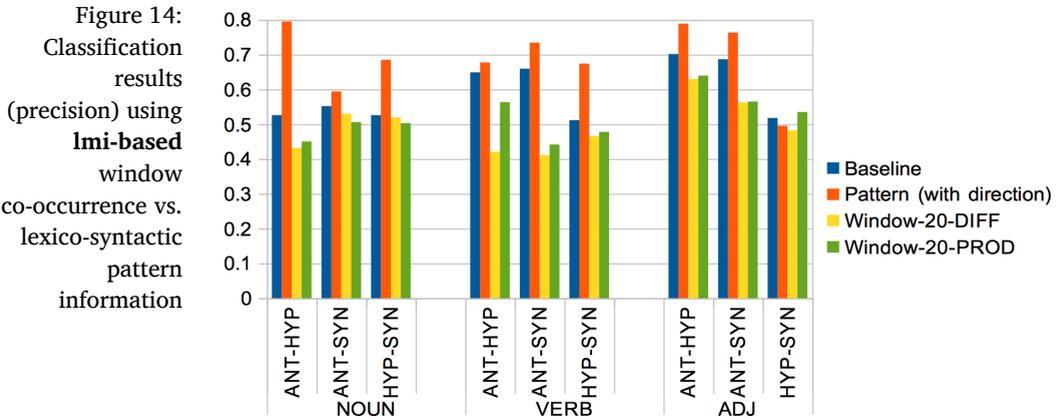


Figure 13: Boxplots of cosine scores across classes and relations (window 20, lmi scores, excluding (left) vs. including ambiguity)

tion pairs changes the overall picture that cosine scores are indeed very similar across our three target paradigmatic relations for a specific word class, cf. our conclusions in Section 4.2.1.

A.3 Automatic classification of relation pairs

Figures 14 and 15 illustrate the differences in classification results when relying on vector spaces with lmi scores (Figure 14, identical to Figure 5) vs. raw frequencies (Figure 15). Using pattern-based features, the plots clearly show consistently better results when using lmi scores in comparison to frequencies. Using window-based features, the results differ more strongly: the WINDOW-20-DIFF results



are better for frequency-based vector spaces than for lmi-based vector spaces, and while they are rather similar to the WINDOW-20-PROD results in the lmi-based spaces, they generally outperform them in the frequency-based spaces.

Figures 16–18 compare lmi-based pattern and window spaces. They once more illustrate that the patterns in Figure 16 (identical to Figure 7) outperform window information, both for a 20-word and a 5-word window. Comparing Figures 17 and 18, we can also see that there are no strong differences regarding the window sizes (20 vs. 5).

The 5-word windows relying on frequencies (right panel in Figure 10) slightly lower the range of the cosine scores, and enlarge the second and third quartiles while the medians stay highly similar, when comparing against the corresponding 20-word windows relying on

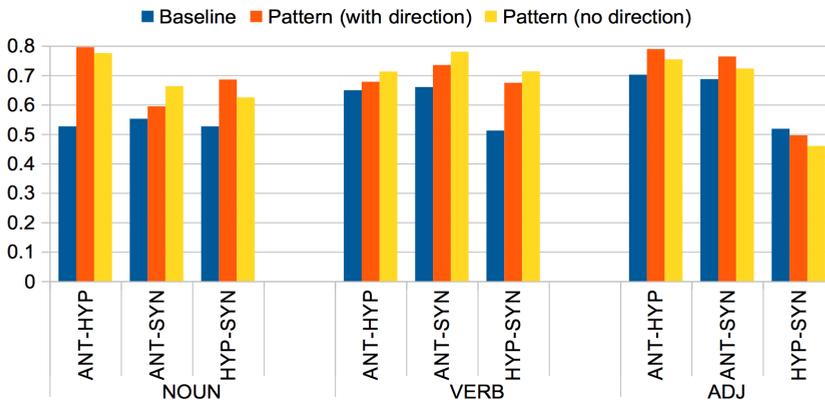


Figure 16: Classification results (precision) using lmi-based **pattern** information

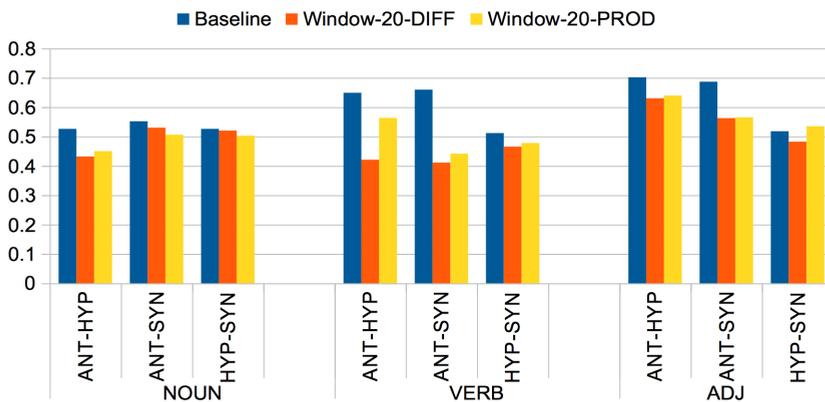
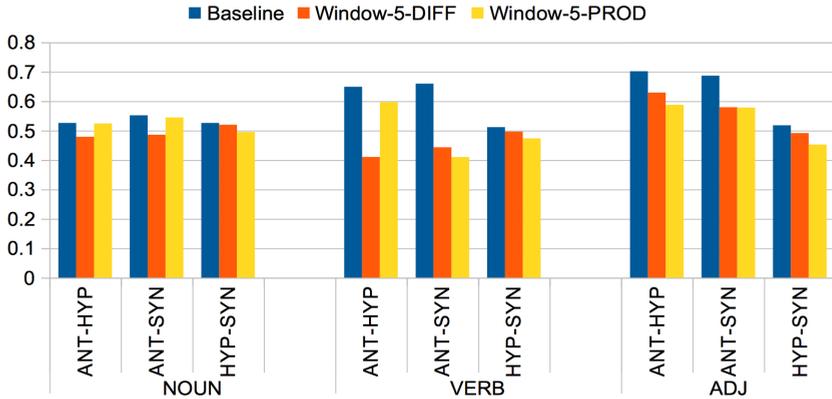


Figure 17: Classification results (precision) using lmi-based **window-20** information

Figure 18:  
Classification  
results  
(precision) using  
lmi-based  
**window-5**  
information



frequencies (left panel in Figure 10). The lmi scores in comparison to the frequencies strongly influence the magnitudes of the cosine scores, and slightly disperse them (see left and right panels in Figure 11 in comparison to the corresponding ones in Figure 10).

Figures 12 and 13 show for 20-word windows relying on frequencies and lmi scores, respectively, that including ambiguous relation pairs (right panels) hardly changes the overall picture at all, in comparison to the left panels which are identical to those in Figures 10 and 11 and exclude ambiguous pairs.

## ACKNOWLEDGEMENTS

The research was supported by the DFG Heisenberg Fellowship SCHU-2580/1, the DFG Research Grant SCHU 2580/2 and the DFG Collaborative Research Centre SFB 732.

## REFERENCES

Heike ADEL and Hinrich SCHÜTZE (2014), Using mined coreference chains as a resource for a semantic task, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1447–1452, Doha, Qatar.

Henriette BAGGER NISSEN and Birgit HENRIKSEN (2006), Word class influence on word association test results, *International Journal of Applied Linguistics*, 16(3):389–408.

Marco BARONI and Alessandro LENCI (2011), How we BLESSED distributional semantic evaluation, in *Proceedings of the EMNLP Workshop on Geometrical Models for Natural Language Semantics*, pp. 1–10, Edinburgh, UK.

Alexander BUDANITSKY and Graeme HIRST (2006), Evaluating WordNet-based measures of lexical semantic relatedness, *Computational Linguistics*, 32(1):13–47.

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

Sharon A. CARABALLO (2001), *Automatic acquisition of a hypernym-labeled noun hierarchy from text*, Ph.D. thesis, Brown University.

Roger CHAFFIN and Arnold GLASS (1990), A comparison of hyponym and synonym decisions, *Journal of Psycholinguistic Research*, 19(4):265–280.

Roger CHAFFIN and Douglas HERRMANN (1981), Comprehension of semantic relationships and the generality of categorization models, *Bulletin of the Psychonomic Society*, 17(2):69–72.

Roger CHAFFIN and Douglas HERRMANN (1984), The similarity and diversity of semantic relations, *Memory and Cognition*, 12(2):134–141.

Kai-Wei CHANG, Wen-tau YIH, and Christopher MEEK (2013), Multi-Relational Latent Semantic Analysis, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1602–1612, Seattle, WA, USA.

Walter CHARLES and George MILLER (1989), Contexts of antonymous adjectives, *Applied Psycholinguistics*, 10:357–375.

Timothy CHKLOVSKI and Patrick PANTEL (2004), VerbOcean: mining the Web for fine-grained semantic verb relations, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 33–40, Barcelona, Spain.

Philipp CIMIANO, Lars SCHMIDT-THIEME, Aleksander PIVK, and Steffen STAAB (2004), Learning taxonomic relations from heterogeneous evidence, in *Proceedings of the ECAI Workshop on Ontology Learning and Population*, Valencia, Spain.

Herbert H. CLARK (1971), Word associations and linguistic theory, in John LYONS, editor, *New Horizon in Linguistics*, chapter 15, pp. 271–286, Penguin.

D. Allan CRUSE (1986), *Lexical semantics*, Cambridge Textbooks in Linguistics, Cambridge University Press, Cambridge, UK.

James CURRAN (2002), Ensemble methods for automatic thesaurus extraction, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 222–229, Philadelphia, PA.

- James CURRAN (2003), *From distributional to semantic similarity*, Ph.D. thesis, Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh.
- Benjamin DAVID (2014), *Comparison and combination of feature types for the automatic classification of semantic relation pairs*, Diplomarbeit, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.
- Simon DE DEYNE, Daniel J. NAVARRO, and Gert STORMS (2013), Better explanations of lexical and semantic cognition using networks derived from continued rather than single word associations, *Behavior Research Methods*, 45(2):480–498.
- Marie-Catherine DE MARNEFFE, Anna N. RAFFERTY, and Christopher D. MANNING (2008), Finding contradictions in text, in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 1039–1047, Columbus, OH.
- Ferdinand DE SAUSSURE (1916), *Cours de linguistique générale*, Payot.
- James DEESE (1965), *The structure of associations in language and thought*, The John Hopkins Press, Baltimore, MD.
- Philip EDMONDS (1997), Choosing the word most typical in context using a lexical co-occurrence network, in *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pp. 507–509, Madrid, Spain.
- Philip EDMONDS (1998), Translating near-synonyms: possibilities and preferences in the interlingua, in *Proceedings of the AMTA/SIG-IL 2nd Workshop on Interlinguas*, pp. 23–30, Langhorne, PA.
- Philip EDMONDS (1999), *Semantic representations of near-synonyms for automatic lexical choice*, Ph.D. thesis, Department of Computer Science, University of Toronto, published as technical report CSRI-399.
- Philip EDMONDS and Graeme HIRST (2002), Near-synonymy and lexical choice, *Computational Linguistics*, 28(2):105–144.
- Katrin ERK and Sebastian PADÓ (2008), A structured vector space model for word meaning in context, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 897–906, Waikiki, Hawaii, USA.
- Stefan EVERT (2005), *The statistics of word co-occurrences: word pairs and collocations*, Ph.D. thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.
- Gertrud FAAß and Kerstin ECKART (2013), SdeWaC – a corpus of parsable sentences from the Web, in *Proceedings of the International Conference of the German Society for Computational Linguistics and Language Technology*, pp. 61–68, Darmstadt, Germany.

- Christiane FELLBAUM (1990), English verbs as a semantic net, *Journal of Lexicography*, 3(4):278–301.
- Christiane FELLBAUM (1995), Co-occurrence and antonymy, *Lexicography*, 8(4):281–303.
- Christiane FELLBAUM (1998a), A semantic network of English verbs, in Fellbaum (1998b), pp. 69–104.
- Christiane FELLBAUM, editor (1998b), *WordNet – an electronic lexical database*, Language, Speech, and Communication, MIT Press, Cambridge, MA, USA.
- Christiane FELLBAUM and Roger CHAFFIN (1990), Some principles of the organization of verbs in the mental lexicon, in *Proceedings of the 12th Annual Conference of the Cognitive Science Society of America*, pp. 420–427.
- Lev FINKELSTEIN, 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.
- John R. FIRTH (1957), *Papers in Linguistics 1934-51*, Longmans, London, UK.
- Roxana GIRJU (2003), Automatic detection of causal relations for question answering, in *Proceedings of the ACL Workshop on Multilingual Summarization and Question Answering – Machine Learning and Beyond*, pp. 76–83, Sapporo, Japan.
- Roxana GIRJU, Adriana BADULESCU, and Dan MOLDOVAN (2006), Automatic discovery of part-whole relations, *Computational Linguistics*, 32(1):83–135.
- Derek GROSS, Ute FISCHER, and George A. MILLER (1989), Antonymy and the representation of adjectival meanings, *Memory and Language*, 28(1):92–106.
- Derek GROSS and Katherine J. MILLER (1990), Adjectives in WordNet, *International Journal of Lexicography*, 3(4):265–277.
- Annamaria GUIDA and Alessandro LENCI (2007), Semantic properties of word associations to Italian verbs, *Italian Journal of Linguistics*, 19(2):293–326.
- Iryna GUREVYCH (2005), Using the structure of a conceptual network in computing semantic relatedness, in *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, pp. 767–778, Jeju Island, Korea.
- Birgit HAMP and Helmut FELDWEG (1997), GermaNet – a lexical-semantic net for German, in *Proceedings of the ACL Workshop on Automatic Information Extraction and Building Lexical Semantic Resources for NLP Applications*, pp. 9–15, Madrid, Spain.
- Sanda M. HARABAGIU, Andrew HICKL, and Finley LACATUSU (2006), Negation, contrast and contradiction in text processing, in *Proceedings of the 21st National Conference on Artificial Intelligence*, pp. 755–762, Boston, MA, USA.
- Zellig HARRIS (1954), Distributional Structure, *Word*, 10(23):146–162.

- Marti HEARST (1992), Automatic acquisition of hyponyms from large text corpora, in *Proceedings of the 14th International Conference on Computational Linguistics*, pp. 539–545, Nantes, France.
- Hans Jürgen HERINGER (1986), The verb and its semantic power: association as the basis for valence, *Journal of Semantics*, 4:79–99.
- Felix HILL, Roi REICHART, and Anna KORHONEN (2015), SimLex-999: evaluating semantic models with (genuine) similarity estimation, *Computational Linguistics*, 41(4):665–695.
- Steven JONES, M. Lynne MURPHY, Carita PARADIS, and Caroline WILLNERS (2012), *Antonyms in English: construals, constructions and canonicity*, Studies in English Language, Cambridge University Press, Cambridge, UK.
- John S. JUSTESON and Slava M. KATZ (1991), Co-occurrence of antonymous adjectives and their contexts, *Computational Linguistics*, 17:1–19.
- John S. JUSTESON and Slava M. KATZ (1992), Redefining antonymy: the textual structure of a semantic relation, *Literary and Linguistic Computing*, 7(3):176–184.
- Grace Helen KENT and Aaron J. ROSANOFF (1910), A study of association in insanity, *American Journal of Insanity*, 67(37–96):317–390.
- George R. KISS, Christine ARMSTRONG, Robert MILROY, and James PIPER (1973), An associative thesaurus of English and its computer analysis, in *The computer and literary studies*, Edinburgh University Press, <http://www.eat.rl.ac.uk/>.
- Claudia KUNZE (2000), Extension and use of GermaNet, a lexical-semantic database, in *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pp. 999–1002, Athens, Greece.
- Claudia KUNZE and Andreas WAGNER (1999), Integrating GermaNet into EuroWordNet, a multilingual lexical-semantic database, *Sprache und Datenverarbeitung*, 23(2):5–19.
- Thomas K. LANDAUER and Susan T. DUMAIS (1997), A solution to Plato's Problem: the latent semantic analysis theory of acquisition, induction and representation of knowledge, *Psychological Review*, 104(2):211–240.
- Adrienne LEHRER and Keith LEHRER (1982), Antonymy, *Linguistics and Philosophy*, 5:483–501.
- Lothar LEMNITZER and Claudia KUNZE (2007), *Computerlexikographie*, Gunter Narr Verlag, Tübingen, Germany.
- Alessandro LENCI and Giulia BENOTTO (2012), Identifying hypernyms in distributional semantic spaces, in *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics*, pp. 75–79, Montréal, Canada.

Omer LEVY, Steffen REMUS, Chris BIEMANN, and Ido DAGAN (2015), Do supervised distributional methods really learn lexical inference relations?, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 970–976.

Dekang LIN, Shaojun ZHAO, Lijuan QIN, and Ming ZHOU (2003), Identifying synonyms among distributionally similar words, in *Proceedings of the International Conferences on Artificial Intelligence*, pp. 1492–1493, Acapulco, Mexico.

Cupertino LUCERTO, David PINTO, and Héctor JIMÉNEZ-SALAZAR (2004), An automatic method to identify antonymy relations, in *Proceedings of the IBERAMIA Workshop on Lexical Resources and the Web for Word Sense Disambiguation*, pp. 105–111, Puebla, Mexico.

John LYONS (1968), *Introduction to theoretical linguistics*, Cambridge University Press, Cambridge, England.

John LYONS (1977), *Semantics*, Cambridge University Press, Cambridge, UK.

Christopher D. MANNING, Prabhakar RAGHAVAN, and Hinrich SCHÜTZE (2008), *Introduction to information retrieval*, Cambridge University Press, Cambridge, UK.

Tomas MIKOLOV, Wen TAU YIH, and Geoffrey ZWEIG (2013), Linguistic regularities in continuous space word representations, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, Atlanta, GA, USA.

George A. MILLER, Richard BECKWITH, Christiane FELLBAUM, Derek GROSS, and Katherine J. MILLER (1990), Introduction to WordNet: an on-line lexical database, *International Journal of Lexicography*, 3(4):235–244.

George A. MILLER and Walter G. CHARLES (1991), Contextual correlates of semantic similarity, *Language and Cognitive Processes*, 6(1):1–28.

George A. MILLER and Christiane FELLBAUM (1991), Semantic networks of English, *Cognition*, 41:197–229.

Saif MOHAMMAD, Bonnie DORR, and Graeme HIRST (2008), Computing word-pair antonymy, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 982–991, Waikiki, Hawaii, USA.

Saif M. MOHAMMAD, Bonnie J. DORR, Graeme HIRST, and Peter D. TURNEY (2013), Computing lexical contrast, *Computational Linguistics*, 39(3).

Saif M. MOHAMMAD, Iryna GUREVYCH, Graeme HIRST, and Torsten ZESCH (2007), Cross-lingual distributional profiles of concepts for measuring semantic distance, in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 571–580, Prague, Czech Republic.

Gregory L. MURPHY and Jane M. ANDREW (1993), The conceptual basis of antonymy and synonymy in adjectives, *Journal of Memory and Language*, 32(3):1–19.

M. Lynne MURPHY (2003), *Semantic relations and the lexicon*, Cambridge University Press, Cambridge, UK.

Douglas L. NELSON, Cathy L. MCEVOY, and Thomas A. SCHREIBER (1998), The University of South Florida word association, rhyme, and word fragment norms, <http://www.usf.edu/FreeAssociation/>.

Kim Anh NGUYEN, Maximilian KÖPER, Sabine SCHULTE IM WALDE, and Ngoc Thang VU (2017), Hierarchical embeddings for hypernymy detection and directionality, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 233–243, Copenhagen, Denmark.

Kim-Anh NGUYEN, Sabine SCHULTE IM WALDE, and Thang VU (2016a), Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 454–459, Berlin, Germany.

Kim-Anh NGUYEN, Sabine SCHULTE IM WALDE, and Thang VU (2016b), Neural-based noise filtering from word embeddings, in *Proceedings of the 26th International Conference on Computational Linguistics*, pp. 2699–2707, Osaka, Japan.

Masataka ONO, Makoto MIWA, and Yutaka SASAKI (2015), Word embedding-based antonym detection using thesauri and distributional information, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 984–989, Denver, Colorado, USA.

David S. PALERMO and James J. JENKINS (1964), *Word association norms: grade school through college*, University of Minnesota Press, Minneapolis, USA.

Patrick PANTEL and Marco PENNACCHIOTTI (2006), Espresso: leveraging generic patterns for automatically harvesting semantic relations, in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 113–120, Sydney, Australia.

Carita PARADIS (2016), Corpus methods for the investigation of antonyms across languages, in Päivi v. JUVONEN and Maria KOPTJEVSKAJA-TAMM, editors, *The lexical typology of semantic shifts*, volume 58 of *Cognitive Linguistics Research*, pp. 131–156, Mouton de Gruyter.

Carita PARADIS, Caroline WILLNERS, and Steven JONES (2009), Good and bad opposites: using textual and experimental techniques to measure antonym canonicity, *The Mental Lexicon*, 4(3):380–429.

Nghia The PHAM, Angeliki LAZARIDOU, and Marco BARONI (2015), A multitask objective to inject lexical contrast into distributional semantics, in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, pp. 21–26, Beijing, China.

Philip RESNIK (1995), Using information content to evaluate semantic similarity in a taxonomy, in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453, San Francisco, CA, USA.

Michael ROTH and Sabine SCHULTE IM WALDE (2014), Combining word patterns and discourse markers for paradigmatic relation classification, in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pp. 524–530, Baltimore, MD, USA.

Herbert RUBENSTEIN and John B. GOODENOUGH (1965), Contextual correlates of synonymy, *Communications of the ACM*, 8(10):627–633.

Wallace A. RUSSELL (1970), The complete German language norms for responses to 100 words from the Kent-Rosanoff Word Association Test, in Leo POSTMAN and Geoffrey KEPPEL, editors, *Norms of word association*, pp. 53–94, Academic Press, New York, USA.

Wallace A. RUSSELL and O.R. MESECK (1959), Der Einfluss der Assoziation auf das Erinnern von Worten in der deutschen, französischen und englischen Sprache, *Zeitschrift für Experimentelle und Angewandte Psychologie*, 6:191–211.

Enrico SANTUS, Alessandro LENCI, Qin LU, and Sabine SCHULTE IM WALDE (2014a), Chasing hypernyms in vector spaces with entropy, in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 38–42, Gothenburg, Sweden.

Enrico SANTUS, Qin LU, Alessandro LENCI, and Chu-Ren HUANG (2014b), Taking antonymy mask off in vector space, in *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computation*, Phuket, Thailand.

Enrico SANTUS, Qin LU, Alessandro LENCI, and Chu-Ren HUANG (2014c), Unsupervised antonym-synonym discrimination in vector space, in *Atti della Conferenza di Linguistica Computazionale Italiana*, Pisa, Italy.

Enrico SANTUS, 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*, pp. 64–69, Beijing, China.

Roland SCHÄFER and Felix BILDHAUER (2012), Building large corpora from the Web using a new efficient tool chain, in *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pp. 486–493, Istanbul, Turkey.

Silke SCHEIBLE and Sabine SCHULTE IM WALDE (2014), A database of paradigmatic semantic relation pairs for German nouns, verbs and adjectives, in

*Proceedings of the COLING Workshop on Lexical and Grammatical Resources for Language Processing*, pp. 111–119, Dublin, Ireland.

Silke SCHEIBLE, Sabine SCHULTE IM WALDE, and Sylvia SPRINGORUM (2013), Uncovering distributional differences between synonyms and antonyms in a word space model, in *Proceedings of the 6th International Joint Conference on Natural Language Processing*, pp. 489–497, Nagoya, Japan.

Helmut SCHMID (1994), Probabilistic part-of-speech tagging using decision trees, in *Proceedings of the 1st International Conference on New Methods in Language Processing*.

Sebastian SCHMIDT, Philipp SCHOLL, Christoph RENSING, and Ralf STEINMETZ (2011), Cross-lingual recommendations in a resource-based learning scenario, in *Proceedings of the 6th European Conference on Technology Enhanced Learning*, pp. 356–369, Palermo, Italy.

Sabine SCHULTE IM WALDE and Maximilian KÖPER (2013), Pattern-based distinction of paradigmatic relations for German nouns, verbs, adjectives, in *Proceedings of the 25th International Conference of the German Society for Computational Linguistics and Language Technology*, pp. 184–198, Darmstadt, Germany.

Sabine SCHULTE IM WALDE, Alissa MELINGER, Michael ROTH, and Andrea WEBER (2008), An empirical characterisation of response types in German association norms, *Research on Language and Computation*, 6(2):205–238.

Vered SHWARTZ, Yoav GOLDBERG, and Ido DAGAN (2016), Improving hypernymy detection with an integrated path-based and distributional method, in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 2389–2398, Berlin, Germany.

Rion SNOW, Daniel JURAFSKY, and Andrew Y. NG (2004), Learning syntactic patterns for automatic hypernym discovery, *Advances in Neural Information Processing Systems*, 17:1297–1304.

Rion SNOW, Daniel JURAFSKY, and Andrew Y. NG (2006), Semantic taxonomy induction from heterogenous evidence, in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 801–808, Sydney, Australia.

Peter D. TURNEY (2008), A uniform approach to analogies, synonyms, antonyms, and associations, in *Proceedings of the 22nd International Conference on Computational Linguistics*, pp. 905–912, Manchester, UK.

Peter D. TURNEY and Patrick PANTEL (2010), From frequency to meaning: vector space models of semantics, *Journal of Artificial Intelligence Research*, 37:141–188.

Lonneke VAN DER PLAS and Jörg TIEDEMANN (2006), Finding synonyms using automatic word alignment and measures of distributional similarity, in *Proceedings of the 21st International Conference on Computational Linguistics and*

*Distinguishing paradigmatic semantic relations*

*the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 866–873, Sydney, Australia.

Paola VELARDI, Paolo FABRIANI, and Michele MISSIKOFF (2001), Using text processing techniques to automatically enrich a domain ontology, in *Proceedings of the International Conference on Formal Ontology in Information Systems*, pp. 270–284, Ogunquit, ME.

Julie WEEDS, David WEIR, and Diana MCCARTHY (2004), Characterising measures of lexical distributional similarity, in *Proceedings of the 20th International Conference of Computational Linguistics*, pp. 1015–1021, Geneva, Switzerland.

Morton E. WINSTON, Roger CHAFFIN, and Douglas HERRMANN (1987), A taxonomy of part-whole relations, *Cognitive Science*, 11:417–444.

Willy YAP and Timothy BALDWIN (2009), Experiments on pattern-based relation learning, in *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pp. 1657—1660, Hong Kong.

Wen-Tau YIH, Geoffrey ZWEIG, and John C. PLATT (2012), Polarity inducing latent semantic analysis, in *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1212–1222, Jeju Island, Korea.

*Sabine Schulte im Walde*

© 0000-0002-8975-6255

schulte@ims.uni-stuttgart.de

Institut für Maschinelle Sprachverarbeitung  
Universität Stuttgart  
Pfaffenwaldring 5B  
70569 Stuttgart  
Germany

Sabine Schulte im Walde (2020), *Distinguishing between paradigmatic semantic relations across word classes: human ratings and distributional similarity*, *Journal of Language Modelling*, 8(1):53–101

doi <https://dx.doi.org/10.15398/jlm.v8i1.199>

This work is licensed under the *Creative Commons Attribution 4.0 Public License*.

cc  <http://creativecommons.org/licenses/by/4.0/>