There Is No Logical Negation Here, But There Are Alternatives: Modeling Conversational Negation with Distributional Semantics

Germán Kruszewski*
Center for Mind/Brain Sciences,
University of Trento

Denis Paperno
Center for Mind/Brain Sciences,
University of Trento

Raffaella Bernardi
Center for Mind/Brain Sciences,
University of Trento

Marco Baroni
Center for Mind/Brain Sciences,
University of Trento

Logical negation is a challenge for distributional semantics, because predicates and their negations tend to occur in very similar contexts, and consequently their distributional vectors are very similar. Indeed, it is not even clear what properties a "negated" distributional vector should possess. However, when linguistic negation is considered in its actual discourse usage, it often performs a role that is quite different from straightforward logical negation. If someone states, in the middle of a conversation, that "This is not a dog," *the negation strongly suggests a restricted set of alternative predicates that might hold true of the object being talked about. In particular, other canids and middle-sized mammals are plausible alternatives, birds are less likely, skyscrapers and other large buildings virtually impossible. Conversational negation acts like a graded similarity function, of the sort that distributional semantics might be good at capturing. In this article, we introduce a large data set of alternative plausibility ratings for conversationally negated nominal predicates, and we show that simple similarity in distributional semantic space provides an excellent fit to subject data. On the one hand, this fills a gap in the literature on conversational negation, proposing distributional semantics as the right tool to make explicit predictions about potential alternatives of negated predicates. On*

* Center for Mind/Brain Sciences, University of Trento, C.so Bettini 31, 38068 Rovereto (TN), Italy.
E-mail: german.kruszewski@unitn.it.

Submission received: 10 April, 2015; revised version received: 25 November, 2015; accepted for publication: 5 April, 2016.

doi:10.1162/COLI_a_00262

the other hand, the results suggest that negation, when addressed from a broader pragmatic perspective, far from being a nuisance, is an ideal application domain for distributional semantic methods.

1. Introduction

Distributional semantics (DS) derives vector-based representations of the meaning of words and other linguistic expressions by generalizing over the contexts in which such expressions occur in large text corpora (Turney and Pantel 2010; Erk 2012). By exploiting the rich commonsense knowledge encoded in corpora and a continuous notion of relatedness that is well-suited to capture the fuzzy nature of content-word semantics, DS representations can successfully model lexical aspects of meaning such as synonymy (Landauer and Dumais 1997), word analogy (Mikolov, Yih, and Zweig 2013), selectional preferences (Erk, Padó, and Padó 2010), and, to a certain extent, hypernymy (Roller, Erk, and Boleda 2014). There is, however, virtually no evidence that DS can capture the semantic properties of grammatical terms such as conjunctions, determiners, or adverbial particles. The very notion of continuous similarity that is so powerful in modeling lexical phenomena is problematic when it comes to capturing the discrete logical operations that are typically associated with the meaning of grammatical terms.

Adverbs and determiners expressing negation, such as English *no* and *not*, receive a very elegant treatment in logic-based approaches: if *dog* denotes the (appropriately indexed) set of all dogs, then *no dog* denotes the complement of the set. However, there is no straightforward "negation" operation that, when applied to the DS vector of *dog*, would derive a *no dog* vector capturing the same complement intuition in vector space. Moreover, negated elements tend to occur in the same contexts of their affirmative counterparts (cf.: *A dog was barking*, *No dog was barking*). Consequently, corpus-induced vectors of predicates and their negations are very similar. This "contextual invariance" of negation is indeed a well-known problem also in lexical semantics, where it has been observed that vectors of words and their (lexicalized) opposites tend to be extremely similar (Mohammad et al. 2013).

In this article, we argue that the problems with negation in DS arise because we are trying to capture a purely logical kind of negation that is neither well-suited to DS nor particularly useful for modeling real-life language usage. If we isolate *dogs* and *non-dogs* in the lab, the logical approach is very appealing: *non-dogs* include anything that is not a dog. However, consider which of the following two sentences is more likely to be uttered in a natural conversational context:

(1) a. This is not a dog… it is a wolf.

b. This is not a dog… it is a screwdriver.

If the negation of a predicate is just the complement of the corresponding set, then Examples (1a) and (1b) should be equally plausible. However, Example (1a) is clearly more natural than (1b).

Looking beyond the purely logical aspects of negation, a long tradition in formal semantics, pragmatics and psycholinguistics has stressed that, in actual conversational contexts, negation is not just excluding possible denotata of the predicates it takes scope over, but also suggesting the truth of an *alternative* assertion. Alternativehood (the possibility of an expression to constitute an alternative to a negated item) seems very well-suited to be modeled in DS. It is obviously similarity-based. It is, more specifically, tied to a *contextual* notion of similarity: We expect plausible alternatives to be objects

or events that tend to occur in contexts that are typical of the negated ones. Finally, alternativehood, just like many lexical properties successfully modeled in DS, is an inherently graded property. Consider:

(2) a. This is not a dog… it is a tarantula.

b. This is not a dog… it is a conference call.

Sentence (2a) is more surprising than (1a), but arguably less so than (1b). In turn, the contingencies in which the latter might be uttered, although undoubtedly bizarre, are still easier to conceive than those that would justify uttering Example (2b).

The first goal of the current article is then to introduce the computational linguistics community to the pragmatic, alternative-licensing view of negation, that we will call **conversational negation**.

Second, we illustrate how DS can contribute, from a new angle, to the literature on alternativehood under conversational negation. Thanks to its ability to automatically identify potential alternatives in a large vocabulary of linguistic expressions, DS allows us to make predictions about *which* elements fall into the (fuzzy) alternative set of a negated expression. This is a new contribution to studies on the semantics of alternatives (in negation or other domains), where authors rely instead on their intuition to pick a small number of candidate alternatives.

Our main empirical contributions are to provide a set of subject-rated negated-predicate/alternative statements, and to predict these ratings with DS. We collect alternativehood judgments in two (minimal) sentential contexts, and study how both sentential context and the negated-item/alternative relation affect the judgments. The most striking result of the computational simulations is how good simple distributional similarity is at predicting the plausibility of an alternative. This measure comes so close to an estimated upper bound that we can only improve over it by a small margin when we use compositional methods and supervision to take sentential context and the specifics of negation into account.

Finally, we present some conjectures on what a DS-based theory accounting for conversational negation could look like. We argue that negation should not be modeled as part of the static distributional representation of a single statement, but as a function that, given the negated predicate, produces a probability distribution over the predicates that are most likely to follow. This approach suggests, more generally, adopting a *dynamic* view of DS, not unlike the one that has been prominent for decades in other areas of semantics.

The rest of this article is structured as follows. Section 2 reviews attempts to model (logical) negation in DS. Section 3 surveys the literature on alternative-licensing conversational negation. Our data set containing subject plausibility ratings for negated-item/alternative pairs is introduced and analyzed in Section 4. In Section 5, we use DS to model the ratings in the data set. We conclude in Section 6 by looking at the theoretical implications of our work, as well as suggesting directions for further study.

2. Negation in Distributional Semantics

Because distributional semantics has traditionally focused on lexical aspects of meaning, negation has mostly been tackled, implicitly, as part of the study of opposites (*hot* and *cold*), or, more generally, "contrasting" words (*warm* and *cold*). A survey of the relevant DS literature is provided by Mohammad et al. (2013). The consensus view is that contrasting words tend to occur in similar contexts (Mohammad et al. even

propose a "distributional hypothesis of highly contrasting pairs" stating that highly contrasting pairs occur in similar contexts more often than non-contrasting word pairs). Thus, it is impossible to distinguish them from non-contrasting related words (e.g., synonyms) using standard distributional similarity measures, and ad hoc strategies must be devised.

Widdows (2003) presented a pioneering study of explicit negation in DS. The assumption of that work is that negated word meanings should be orthogonal, which is to say that they should not share any common feature. Specifically, Widdows proposes a binary negation operator, $NOT(A, B)$, which projects the vector representing A onto the orthogonal space of the B vector. In logical terms, this can be seen as conjunction with a negated predicate ($A \wedge \neg B$). The orthogonality assumption makes perfect sense for the information retrieval applications envisioned by Widdows (*web NOT internet*), but it is too strict to characterize the linguistic predicates of the relevant form in general (*Italian but not Roman* refers to somebody who shares many properties with Romans, such as that of speaking Italian).

Interest in grammatical words in general, and negation in particular, has recently risen thanks to the development of *compositional* DS models (Mitchell and Lapata 2010; Baroni 2013). A shared assumption within this framework is that the operation performed by negation on vectors should be defined a priori, attempting to mimic the logical properties of negation, rather than being induced from distributional data. Clark, Coecke, and Sadrzadeh (2008) explore the idea that sentences live in a space spanned by a single "truth-denoting" basis ($\vec{1}$), with the origin ($\vec{0}$) corresponding to "false." A sentence like *John likes Mary* is represented by $\vec{1}$ if the sentence is true, $\vec{0}$ otherwise. In this framework, further elaborated by Coecke, Sadrzadeh, and Clark (2010), negation is elegantly modeled as a swap matrix. Related approaches have been presented by Preller and Sadrzadeh (2011) and Grefenstette (2013). All this work, however, is purely theoretical, and it is not clear how the proposed models would be implemented in practice. Moreover, treating negation as a swapping operator only makes sense in the abstract scenario of a vector space representing truth values. If vectors of sentences and other expressions are instead distributional in nature (e.g., representing distributions over possible contexts), it is far from clear that swapped vectors would capture linguistic negation.

Indeed, Hermann, Grefenstette, and Blunsom (2013) argue that negation should not affect all dimensions in a vector, because a word, when negated, does not change its domain: The vector representation of *not blue* should still be close to that of other colors (Hovy [2010] also defends a similar view). Hermann and colleagues propose that vectors should include distinct "domain" and "value" features, and negation would modify (change sign and possibly rescale) only the latter. In this way, *not blue* would still be in the color domain, but its chromatic values would differ from those of *blue*. Extrapolating to nouns, under this proposal we might expect *not dog* to still belong to the canid domain, but with different values from those that specifically characterize dogs. This would capture the intuition we spelled out in the introduction that a wolf is a better non-dog than a screwdriver. However, the proposal of Hermann and colleagues is, again, purely theoretical, and we do not see how domain and value features with the desired properties could be induced from corpora on a large scale.

Such is the difficulty to model negation and other logical terms in DS that Garrette, Erk, and Mooney (2013) have proposed a division of labor between DS, handling lexical relations between content words, and first-order logic, accounting for the semantics of grammatical terms. In this framework, the issue of the distributional meaning of negation does not arise.

Finally, because of the theoretical nature of most of the relevant work, we lack benchmarks to evaluate empirical models of negation in DS.

3. Conversational Negation and Alternative Sets

As already pointed out by Horn (1972), all human languages have negation and no other animal communication system includes negative utterances. This alone makes linguistic negation intriguing and justifies the huge amount of literature dedicated to it. Furthermore, linguistic negation seems to play different roles and therefore constitutes a challenge for any formal theory. On the one hand, negation works as a truth-functional operator, and as such it has attracted philosophers and formal semanticists, for example, for its scope ambiguity behavior. However, linguistic negation also works as a conversational illocutionary marker, causing non-literal interpretations of propositions, and as such it has attracted the attention of linguists, psychologists, and epistemologists. Conversational negation is something different from a logical truth-functional operator that flips the values of its argument. In particular, in actual linguistic usage, the negation of a predicate often suggests that one of a set of alternatives might be holding.

The alternative set view of negation traces back, on the one hand, to Grice's conversational maxims (Grice 1975) and Horn's principle of alternate implicatures (Horn 1972), and, on the other, to the "alternative semantics" theory of focus (Rooth 1985).

Implicatures as studied in pragmatics are part of sentence interpretation on top of the strictly logical meaning. A commonly assumed mechanism for generating implicatures to a statement is to take alternatives of the statement as false. For instance, *Some dogs bark* implicates that the alternative *All dogs bark* is false, even though it is compatible with the literal interpretation of the sentence ("there are some dogs barking"). This derives pragmatic strengthening of a class of scalar elements (*some* $>$ *some but not all*, *can* $>$ *can but does not have to*, etc.).

Under the alternative semantics theory of focus (Rooth 1985), sentences are assigned, in addition to their usual semantic values, an alternative set. For a sentence *John likes JANE* with focus on the individual liked by the subject, the alternative set is the set of all propositions of the form $\{[\![\text{john likes } x]\!] | x \text{ an individual}\}$. Another analytical option is the so-called "structured meaning approach to focus," which considers alternatives only to parts of the sentence, for example, the noun *Jane*, but not to whole propositions (Krifka 1992). The two theories formalize the meaning of focus-sensitive operators differently, but, importantly for us, both assume the notion of semantic alternatives to be crucial for focus interpretation.

Perhaps the most widely known use of focus alternatives is the semantic analysis of focus particles such as *even* or *only*. But students of focus noticed that negation is also sensitive to them: *Not A* implies the truth of an alternative to *A*, for example, *John doesn't like JANE* suggests that John likes someone else. *This is not a boy* suggests an alternative assertion (*This is a girl*; *This is a man*); see Kratzer (1989, p. 646) for an explicit intensional analysis of negation in terms of focus alternatives. Negation, seen as a focus operator, is similar but logically opposite to *only*: whereas *John only likes JANE* means that John likes Jane but does not like anyone else, *John doesn't like JANE* means that John does not like Jane but likes someone else. Non-trivial interaction with focus alternatives is very typical for usage of negation in natural language, so negation is rarely used in a purely logical sense.

As usually assumed, the alternative set for a sentence is the set of semantic values resulting from replacing the negated element with arbitrary values of the right semantic type. However, it is clear that not all alternatives are created equal. The most plausible

ones are relevant across many varied contexts, whereas others require a heavy contextual pressure to become acceptable. For example, *boy* and *submarine* are of the same semantic type (that of unary predicates), but it requires a very unusual context for *This is not a boy* to suggest *This is a submarine*.[1] In most contexts a limited set of predicates (*girl*, *man*, etc.), all related to *boy*, constitute viable alternatives, and *submarine* is not one of them. So, although it is true that context affects the set of relevant alternatives, the most prominent ones are largely predictable from the propositional content of the utterance.

The plausibility of alternatives has been studied from a psychological angle, using reasoning tasks such as the construction of settings that verify or falsify a given rule (Evans 1972) or selection of information critical for determining the truth of a rule (Wason 1966). The alternative possibilities primed by negation have been said to be "the most likely or relevant members of the complement set" (Oaksford 2002, page 140). In particular, Oaksford and Stenning (1992) identify several mechanisms used for constraining contrast classes (viz. alternative sets): focus/intonation, high-level relations, and world knowledge. They take the sentence *Johnny didn't travel to Manchester by train* as an example. Different contrast classes will be built based on where the focus is put (Johnny vs. Manchester vs. train.). The high-level traveling schema relation imposes constraints for each slot (travelers, destinations, or mode of transportation). Other constraints are imposed by world knowledge. For example, if you are traveling from London, the vehicle of choice is unlikely to be a ship. In short, Oaksford and Stenning conclude that (emphasis ours) "the contrast-class members should be as *similar* as possible, in the relevant respects, to the negated constituent" (page 849).

Psychologists have also extensively investigated the issue of whether the presence of negation automatically evokes a set of alternatives ("search for alternatives" view) (Oaksford and Stenning 1992; Oaksford 2002) or whether the direct effect of negation is just one of information denial ("quasi-literal" view) (Evans, Clibbens, and Rood 1996). Even under this second view, alternatives can still be evoked and explored at a later interpretation stage. Indeed, based on behavioral evidence, Prado and Novek (2006, page 327) propose to reconcile the two views by concluding that "an initial reading of a negation will be narrow [viz. literal] and in some scenarios this might be enough. [..] A search for alternatives arises, but as part of a secondary effort to interpret the negation in the proposition." More recently, Ferguson, Sanford, and Leuthold (2008) presented eye-movement and ERP evidence in favor of the search for alternative view. Neither side of this debate denies that alternatives play an important role in the interpretation of negated statements, and we do not take a stand on it.

A phenomenon that is intuitively related to alternativehood is the degree of plausibility of a negated identity. Wason (1965, 1971) has claimed that the interpretation of negative statements of this kind is easier when the sentence negates a presupposition, something that is believed to be true. In this view, the sentence *The whale is not a fish* is easier to interpret than *The whale is not a motorcycle*. While negation underlines the difference between two terms, at the same time it presupposes that they are similar: It is pragmatically reasonable to negate that two terms are the same thing when they can be confused. An alternative approach to this view is proposed by Clark (1974), who claims that comprehending negation relies on detecting differences between the proposition negated and the actual state of affairs. This predicts that, when two things are dissimilar, it should be easier to perceive one of them as a negation of the other: *The*

1 For example, in a vision test, the patient might be asked to discriminate pictures of boys from pictures of submarines. In such a context, *This is not a boy* is uttered to imply *This is a submarine*.

whale is not a motorcycle should then be easier to interpret than *The whale is not a fish*. Cherubini and Mazzocco (2003), among others, have tested these theories, finding that similarity facilitates comprehension of negation across various experimental settings. These results establish a connection between alternativehood in psychological studies and the similarity relation captured by DS that we are going to investigate in the remainder of the article.

To conclude, the notion of conversational negation licensing an alternative set is useful to account for linguistic and psychological data. Moreover, alternatives appear to be linked to negated expressions by a relation of similarity. However, all studies we are aware of base their claims on a small number of hand-picked examples of felicitous or implausible alternatives. As a consequence, no model has been proposed that, given a negated element and an arbitrary predicate, makes explicit predictions about how plausible it is for the predicate to fall into the alternative set of the negated element. In the remainder of this article, we introduce a large data set of alternative plausibility ratings, and propose DS as a model to predict the plausibility of potential alternatives.

4. A Data Set of Alternative Plausibility Ratings

This section documents the creation and structure of our data set of plausibility ratings for alternatives to a negated predicate. We describe in turn the sentential frames we used, the sources of negated-predicate/alternative word pairs, the rating collection procedure, and its outcome.

4.1 Selecting Sentential Contexts

Semantic (or pragmatic) alternatives are defined for all types of interpreted constituents, from words to phrases to full sentences. Because our study is the first of its kind, it is best to start simple. We decided to focus on alternatives to common nouns, but, because a noun in isolation does not constitute a very natural utterance, we placed each noun in a minimal sentential context. The two simple options we adopt include using the noun predicatively in a classification statement (*This is (not) N*) (*IT* context) and having the noun existentially quantified (*There is (no) N here*) (*THERE* context) (of course, we do not claim such contexts to exhaust the spectrum of natural language negation usages). In both cases, context does not add much information, and the main burden lies on the noun itself.

Still, we are aware that even the simple constructions we are considering carry different pragmatic effects. The *IT* context has a "correction" reading whereas the *THERE* context has the flavor of an "at least" reading with the alternative compensating for the lack of the negated term. Moreover, *This is not N* suggests that the alternative should be a noun denoting something that could substitute or be taken for *N*. *There is no N*, on the other hand, suggests *situations* similar to ones in which there is *N*. Sets of salient alternatives in the two contexts need not be identical. For example, *There is no piano here...* might be plausibly continued with *...but there is music*, since pianos and music appear in similar situations. On the other hand, *This is not a piano, it is music* sounds odd, because pianos and music are very different things.

We consider the presence of these subtle differences a positive aspect of our setting: If we find that ratings are essentially comparable across contexts, then we have some evidence that we are getting stable and general alternativehood ratings. Should, instead, systematic differences emerge, we could gain preliminary empirical insights on how

sentential, as opposed to purely lexical, factors affect alternatives. We will briefly come back to this point in the analysis of subject ratings.

Asking participants to provide direct judgment on alternatives would require more explicit understanding of linguistic pragmatics than one can expect from a naive speaker, even with some training. Instead, we ask subjects to rate the explicit conjunction of the negated predicate and the alternative, as a close proxy for alternativehood. So for instance, instead of asking "Does *This is not a dog* plausibly evoke the alternative *This is a cat*?," we ask the more intuitive question: "How plausible is the sentence *This is not a dog, it is a cat*?" The corresponding THERE context is *There is no dog here, but there is a cat*, with *but* added as it makes the sentence more natural.

4.2 Selecting Potential Alternatives

As just discussed, our stimuli contain exactly two content words—a noun either in a predicative or in an existential negated position, and its potential alternative in a comparable position. To construct the noun pairs, we took as negated elements 50 randomly selected items from BLESS (Baroni and Lenci 2011), and paired them with potential alternatives from several sources. We picked alternatives from different sources in order to take a variety of relations that might affect alternativehood into account. However, we make no claim about the exhaustiveness of the phenomena we are considering, and we do not present theoretical conjectures about how lexical relations affect the likelihood of being a plausible alternative.

One of the sources was WordNet,[2] from which we extracted lists of cohyponyms, hyponyms, and hypernyms. For the 50 negated items, there are almost 4K WordNet cohyponyms, many of them based on very rare senses (*library* vs. *battery*). We picked the first 10 cohyponym synsets of each noun (as more common senses are typically listed earlier) and filtered out the ones expressed by phrases, ending up with a total of 534 negated-item/hyponym pairs (e.g., *deer/pollard*, *falcon/buteonine*, *chair/academicianship*, *bag/bread-bin*). We also covered hyponyms (WordNet subcategories), with a total of 314 distinct pairs (e.g., *truck/lorry*), and hypernyms, including 32 category names from the norms of Van Overschelde, Rawson, and Dunlosky (2004) and all WordNet supercategories, for a total of 216 distinct pairs (e.g., *garlic/seasoner*).

Although taxonomic relations are obviously important for determining a word's alternatives, it would be wrong to miss non-taxonomic relations, which we extracted from various other sources. We included nine nearest neighbors per item from the best "count" distributional semantic model of Baroni, Dinu, and Kruszewski (2014), functionally similar items (nouns that share the UsedFor relations) from ConceptNet[3] (408), and visually similar nouns from Silberer and Lapata (2012). From the latter, we included all pairs with more than minimal visual similarity (> 1), a total of 525 pairs (e.g., *bus/cabin*, *dress/trousers*). Further pairs were extracted from the University of South Florida Word Association Norms (Nelson, McEvoy, and Schreiber 1998). We only picked nouns, for a total of 492 free associates (e.g., *giraffe/trees*). To construct unrelated pairs, we randomly matched our nouns with ten frequent nouns each (e.g., *poplar/fuel*, *broccoli/firm*, *truck/world*, *falcon/guidance*). Finally, in an attempt to provide quantitative data to support the intuition that a concept cannot be an alternative to itself (**This is not a dog, it is a dog!*), we paired each word with itself in a set of "identity" pairs.

2 http://wordnet.princeton.edu.
3 http://conceptnet5.media.mit.edu/.

We removed from the resulting list pairs containing words attested less than 100 times in our source corpus (see Section 5.1.1). We also removed some obviously mistagged items (*th*, *these*, *others* incorrectly tagged as nouns), and potentially offensive materials. No other filtering was performed on the data. We were left with 2,649 pairs (50 identity pairs and 2,599 pairs for the other relations).[4]

For each pair, we generated *This is not an X, it is a Y* and *There is no X here, but there is a Y* sentences. We manually corrected determiners where needed (*a broccoli>broccoli*), number agreement in sentences with plural terms (*There is jeans>There are jeans*), and we adjusted capitalization (*pbs>PBS*).

4.3 Collecting Human Ratings

We used the Crowdflower service[5] to collect plausibility ratings on a 1–5 Likert scale for the 2,649 negated-item/alternative pairs embedded in both THERE and IT contexts, collecting at least 10 judgments per pair. Following the standard practice in crowdsourcing experiments, we added 42 manually crafted sentences that were obviously (un)acceptable (40 in the THERE setting since two pairs were more ambiguous in this context). Participants who did not rate these pairs within the expected ranges were excluded from the survey.

Subjects were asked to judge the plausibility of each sentence. To this end, they were told to think whether in an ordinary real-life situation (not "fairy-tale" circumstances) the sentence could be reasonably uttered. Furthermore, they were told that, in case of ambiguity, they had to choose, if available, the (sufficiently common) sense of a word that would make the sentence more plausible. Finally, participants were instructed to mentally add, remove, or change the article in each example if that made it more natural (e.g., *This is not a planet, it is a sun* could be changed to the more natural *This is not a planet, it is the sun*; *There is no subway here, but there is a bus*, could be changed to the more natural *There is no subway here, but there is the bus*). No further definition of the phenomena involved in the sentences were given. Subjects were instead provided with examples (fully reported in Tables 1 and 2).

Judging alternatives based on short sentential context only is a challenging task, and we decided to exclude from further analysis those pairs that were not rated consistently by the subjects. In particular, we discarded all pairs whose inter-subject variance was not significantly below chance at $\alpha = 0.1$, based on simulated chance variance distributions. We discarded all identity pairs because the overwhelming majority was characterized by very high variance. Evidently, the subjects found statements such as *This is not a cat, it is a cat* too nonsensical to even parse them coherently. After filtering, we were left with 1,231 IT pairs and 1,203 THERE pairs.[6]

4 Out of the 2,599 pairs, 241 belong to more than one relation: For example, not surprisingly, some free associates are also cohyponyms (e.g., *hawk/eagle*). For the ease of analysis, we assigned such items to the relation with the lower cardinality.

5 `http://www.crowdflower.com/`.

6 Because ratings are bound to the 1–5 interval, all else being equal, variance will tend to be higher for pairs that receive intermediate ratings than for extreme ones. Our exclusion criterion might thus have been too conservative for intermediate cases. We leave it to future research to devise experimental set-ups that would allow us to better distinguish between noise and genuine middle-ground cases, possibly switching to a comparative elicitation design, such as pairwise comparison or the MaxDiff method suggested by a reviewer.

Table 1
This is not an X, it is a Y. Examples provided to subjects.

This is not a horse, it is a donkey:
In this case, you should assign a very high rating since one can easily conceive a real-life situation in which the sentence is uttered; I see an animal, I think it's a horse, but someone makes me notice it's a donkey, instead.

This is not beans, it is corn:
In this case, you should assign a very high rating since one can easily conceive real-life situations in which the sentence is uttered; I am preparing a soup, I take out a can thinking it is beans, but then I or someone else notices it's actually corn.

This is not a boy, it is a girl:
In this case, you should assign a very high rating since the sentence sounds plausible. It is easy to imagine real-life situations in which the sentence is uttered; e.g., I see a kid, I think it's a boy, but someone makes me realize it's a girl.

This is not a fact, it is a hypothesis:
In this case, you should assign a very high rating since this sentence is plausible in virtually any written argument.

This is not wine, it is a score:
The sentence doesn't really make sense. It is implausible that anyone would ever utter it, so you should assign it a very low rating.

This is not a league, it is a floor:
The sentence doesn't really make sense. It is implausible that anyone would ever utter it, so you should assign it a very low rating.

This not an agreement, it is a vehicle:
The sentence doesn't really make sense. It is implausible that anyone would ever utter it, so you should assign it a very low rating.

This is not a bomb, it is a house:
... might make sense in the context of the Wizard of Oz, or as a correcting remark during a vision test of a near-blind person who confuses unrelated images, but not under ordinary circumstances. Therefore, you should assign it a low score.

This is not a club, it is a pole:
You should read *club* and *pole* in the sense of physical objects, and not in their respective *association* and *extreme point* senses, given that the physical object interpretations make the sentence more plausible. Therefore, the sentence should not receive a low rating.

4.4 Distribution of Alternative Plausibility Ratings

Table 3 reports summary statistics for the reliably rated items. As the table shows, excluding high-variance pairs still leaves us with relatively many examples of the non-identity relations in both sentential settings.

Not surprisingly, we find that *unrelated* alternatives received the lowest ratings in both IT and THERE contexts. It is also not too surprising that, for all other relations except hyponyms, the THERE ratings are higher than the IT ratings. As we observed earlier, the IT context suggests that the alternative should be something highly similar to the negated item, whereas in the THERE context, it is the situations in which the negated item might occur that must be shared with the alternative. Figure 1 reports the pairs where the positive difference between the THERE and IT mean rating is largest (the opposite happens very rarely). The case of *castle* and *prince* is exemplary: a castle

Table 2
There is no X here, but there is Y. Examples provided to subjects.

There is no horse here, but there is a donkey: In this case, you should assign a very high rating since one can easily conceive a real-life situation in which the sentence is uttered; I am looking for a horse and someone tells me that there is no horse here but there is a donkey, instead.
There are no beans here, but there is corn: In this case, you should assign a very high rating since one can easily conceive real-life situations in which the sentence is uttered; I am preparing a soup and I ask my friend for beans; she tells me that there are no beans, but I could use corn instead.
There is no boy here, but there is a girl: In this case, you should assign a very high rating since the sentence sounds plausible. It is easy to imagine real-life situations in which the sentence is uttered; e.g., I am casting a kid for a show, hoping there would be a boy that fits, but someone tells me there is a girl instead.
There is no fact here, but there is a hypothesis: In this case, you should assign a very high rating since this sentence is plausible in a review of some written argument.
There is no wine here, but there is a score: The sentence doesn't really make sense. It is hard to think of a situation in which a score could be proposed as a reasonable alternative to wine, so you should assign the sentence a very low rating.
There is no league here, but there is a floor: The sentence doesn't really make sense. Again, it is hard to think of a context involving a floor as an alternative to a league, so you should assign the sentence a very low rating.
There is no agreement here, but there is a vehicle: The sentence doesn't really make sense, for the same reasons as the previous ones, so you should assign it a very low rating.
There is no bomb here, but there is a house: . . . might make some sense in the context of the Wizard of Oz, or as a correcting remark during a vision test of a near-blind person who confuses unrelated images, but not under ordinary circumstances. Therefore, you should assign it a low score.
There is no club here, but there is a pole: You should read *club* and *pole* in the sense of physical objects, and not in their respective *association* and *extreme point* senses, given that the physical object interpretations make the sentence more plausible. Therefore, the sentence should not receive a low rating.

is a very different entity from a prince (hence, the oddness of **This is not a castle, it is a prince*), but the presence of a prince suggests that we should be at least surprised by the absence of a castle, and so *There is no castle here, but there is a prince* sounds like a reasonable observation. In this perspective, it makes perfect sense that the alternatives with the largest relation-wise positive THERE–IT difference are free associates and visually related items (items that might look similar, but can be ontologically quite different).

Perhaps the most surprising result of the survey is that hyponyms receive high ratings in both contexts, and they are indeed the only relation showing a slight preference for the IT context. We expected hyponyms to be treated as implausible alternatives, because they should lead to contradictory sentences (**This is not a vegetable, it's a potato*). By inspecting the highly rated hyponyms, we conclude that the unexpected pattern is almost entirely due to the following artifact: All our target negated items are base-level concepts from the BLESS resource, and not general category names. Consequently, the

Table 3
Summary of alternative plausibility distributions: Number of reliably rated pairs, medians of mean pair rating, and pair rating variance.

	IT			THERE		
	N	Median mean rating	Median variance	N	Median mean rating	Median variance
cohyponym	126	1.8	0.63	133	3.3	0.71
distributional	98	4.1	0.90	126	4.3	0.71
free associate	141	1.6	0.71	124	4.1	0.60
functional	161	1.5	0.71	125	1.8	0.70
hypernym	78	1.6	0.71	88	2.0	0.70
hyponym	85	4.1	0.71	67	3.8	0.63
unrelated	412	1.2	0.18	373	1.2	0.35
visual	130	1.6	0.69	167	3.8	0.72
TOT	1,231	1.4	0.50	1,203	1.6	0.50

hierarchical distinction with their WordNet hyponyms is not very sharp, and indeed in most of the cases in which a negated-noun/hyponym pair receives a high rating, the two terms can be as easily interpreted as cohyponyms: *coat/jacket*, *truck/van*, *shirt/t-shirt*, *bag/rucksack*, *cat/panther*, and so on. Indeed, these are better cohyponym pairs than many of those we harvested through WordNet cohyponym links, that were often too distantly related (*bottle/bath* under the *vessel* category) or based on unusual senses of words (*goat/mug* under the *victim* category). As a result, cohyponyms received, on average, lower ratings compared with hyponyms (again, the THERE context, being more tolerant towards distant relations, affords higher cohyponym ratings).

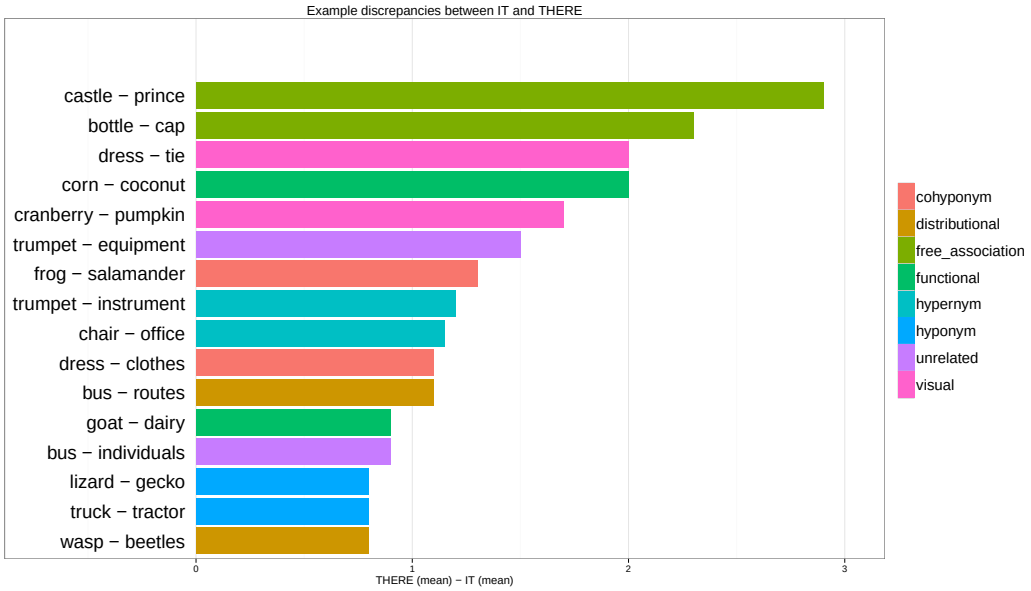


Figure 1
Pairs with largest positive THERE–IT mean rating difference.

Interestingly, in both contexts, distributionally harvested alternatives (nearest neighbors in a distributional semantic space) receive, on average, high ratings, comparable to those of hyponyms. By looking at top-rated distributional alternatives, we find that many of them are close cohyponyms (e.g., *lizard/iguana*, *poplar/elm*, *trumpet/saxophone*). So, we conclude that, as expected, *close* cohyponyms are very good alternatives (both in the IT and THERE contexts), and that following the WordNet *hyponym* link or harvesting near distributional neighbors are better ways to get at the close hyponyms than relying on the WordNet cohyponym relation.

Although hypernyms did not receive very high ratings in general, we observe once more a preference for them in the THERE context. This is at least in part because of the presence of *but* in the latter frame, which encourages a "contrastive" reading (cf. *There is no trumpet here, but there is an instrument* vs. *?This is not a trumpet, it is an instrument*).

The previous discussion has emphasized the differential effect of sentential contexts and distinct relations on the ratings. This analysis should, however, not obscure two important points illustrated by the scatterplots in Figure 2. First, correlations between IT and THERE ratings are uniformly extremely high, except for the unrelated pairs, where the (relatively) low correlation is simply an artifact of the lack of spread among

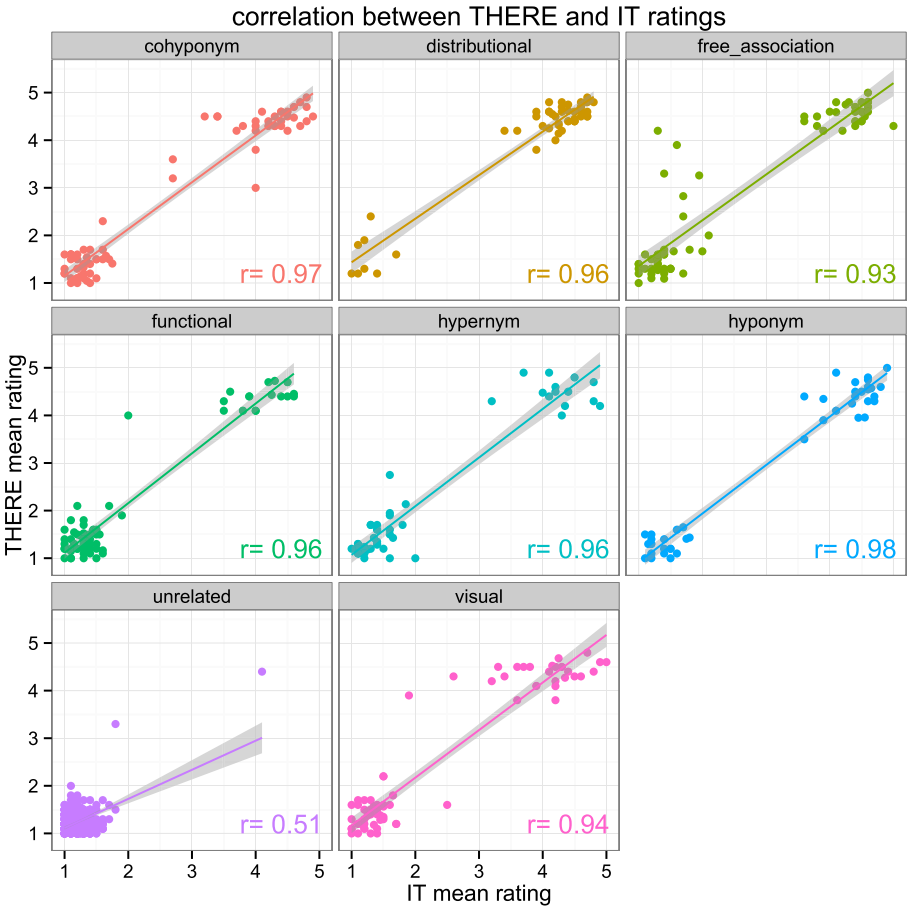


Figure 2
Scatterplots of IT and THERE mean per-pair ratings itemized by relation, with the respective Pearson correlation scores.

pair ratings (as nearly all pairs get very low scores). It seems safe to conclude, then, that the subtle pragmatic and semantic implications carried by the two contexts affected subjects' ratings only very marginally, and we are getting robust alternativehood intuitions across contexts. Consequently, we also expect that the same similarity measure might be able to approximate the ratings in either context reasonably well. Second, for all relations (except the one linking unrelated pairs), there is a good number of both plausible and implausible alternatives.[7] This shows that relation type does not suffice to account for the plausibility of an alternative, and we need a more granular similarity measure. We thus turn, in the next section, to various candidates that distributional semantics provides for such a measure.

5. Predicting Alternative Plausibility with Distributional Semantics

Predicting human ratings about the plausibility of alternatives under conversational negation with a DS model is straightforward. Following standard practice (Turney and Pantel 2010), we take the *cosine* of the angle between the distributional vectors of a negated noun and the potential alternative to be an estimate of their degree of semantic similarity. Because alternativehood obviously correlates with similarity, this simple method should provide good estimates of perceived alternative plausibility.

We further explore two independent ways to enhance prediction quality. First, because we saw that ratings are affected by sentential context, we exploit compositional DS methods to derive vector representations of negated items and their potential alternatives when embedded in the IT and THERE contexts. Second, because we know that generic similarity cannot be the whole story (for example, hyponyms are expected to be quite similar to their hypernyms, but our survey results suggest that hypernyms are not particularly felicitous alternatives), we feed rated pairs as training examples to a supervised algorithm (specifically, support vector regression). We hope, in this way, to tune generic DS similarity to the specific characteristics of alternativehood.

For purposes of model building and evaluation, we randomly split the reliable pairs in our data set into 563/550 IT/THERE training items, 186/187 development items, and 482/466 test items (the split is carried out so as to maximize IT/THERE overlap across the subsets).

5.1 Model Implementation

5.1.1 Distributional Semantic Space Construction. We extracted co-occurrence information from a corpus of about 2.8 billion words obtained by concatenating ukWaC,[8] Wikipedia,[9] and the British National Corpus.[10] Our vocabulary—the words we produce semantic vectors for—is composed of the 20K most frequent nouns in the corpus (in inflected form), plus those needed to have full coverage of the rated data set. We counted sentence-internal co-occurrences of these nouns with the top 20K most frequent inflected words, thus obtaining 20K-dimensional vectors. Following standard practice, we transformed raw counts into non-negative Pointwise Mutual Information scores (Evert 2005) and compressed the resulting vectors down to 300 dimensions using Singular Value Decomposition (Golub and Van Loan 1996). All the described parameters were

7 The fact that the ratings are polarized towards the extremes is partially an artifact of removing high-variance cases.

8 http://wacky.sslmit.unibo.it.

9 http://en.wikipedia.org.

10 http://www.natcorp.ox.ac.uk.

picked without tuning, based on our previous experience and insights from earlier literature (e.g, Bullinaria and Levy 2012). We exploited the development set to decide if we should rescale the Singular Value Decomposition vectors by the corresponding singular values (as suggested by mathematical considerations) or not (as recently recommended, based on empirical evidence, by Levy, Goldberg, and Dagan [2015]). We found that the latter option is better for our purposes.

We also used the development set to compare the standard "count" vectors we just described to the neural-language-model vectors shown to be at the state of the art in many semantic tasks by Baroni, Dinu, and Kruszewski (2014). Perhaps surprisingly, we found that, for our purposes, the count vectors are better.

5.1.2 Modeling Sentential Contexts with Composition Functions. In order to produce semantic representations for our target nouns in the relevant sentential contexts, we harness the compositional extension of DS. In particular, we adopt the functional model independently proposed by Coecke, Sadrzadeh, and Clark (2010) and Baroni, Bernardi, and Zamparelli (2014). In this model, certain linguistic expressions (e.g., verbs) are represented by linear functions (matrices or higher-order tensors) taking distributional vectors representing their arguments (e.g., nouns) as input, performing composition by matrix-by-vector multiplication (or the equivalent operation for tensors of larger arity), and returning output vectors representing the relevant composed expressions (e.g., sentences). We assume that the affirmed and negated IT and THERE contexts (*this/it is X*, *this/it is not X*, *there is X*, and *there is no X*) are functions (matrices) that take vector representations of the nouns of interest (e.g., *hawk*) as input and return vectors representing the corresponding phrases (e.g., *this/it is a hawk*) in output. We then use cosines between negated and alternative phrases to predict alternativehood plausibility.

We considered two other approaches to composition that we ruled out based on poor development set performance. These were: ignoring negation (e.g., using the same composition function for *this/it is a hawk* and *this/it is not a hawk*), and modeling negation as a separate composition step (e.g., deriving *this is not a hawk* from the application of two composition functions: *not*(*this_is*(*hawk*))). We did not attempt to model *here* and, more importantly, *but* in the THERE context (*there is no X* ***here, but*** *there is a Y*).

To estimate the function parameters (matrix weights), we followed the approach of Guevara (2010), Baroni and Zamparelli (2010), and Dinu, Pham, and Baroni (2013), as implemented in the DISSECT toolkit.[11] Specifically, we extracted from our corpus distributional vectors for sufficiently frequent phrases matching the target template (e.g., *this/it is cat*), and estimated the corresponding function by optimizing (in the least-squares sense) the mapping from the vectors of the nouns in these phrases (*cat*) onto the corpus-extracted phrase vectors. Theoretical and empirical justifications for this method can be found in Baroni, Bernardi, and Zamparelli (2014), or any of the articles mentioned here. We relied on a MaltParser[12] dependency parse of our corpus to identify the target phrases, and treated other words occurring in the same sentences as contexts (e.g., from *I think that this is a very sly cat*, we would have extracted *I*, *think*, *that*, *a*, *very*, and *sly* as contexts). The phrase vectors were assembled and transformed using the same parameters we used for nouns (see Section 5.1.1). We built vectors for any phrase that (i) matched one of the templates of interest; (ii) contained a noun from the semantic space vocabulary; (iii) and occurred with at least 100 distinct collocates in the corpus.

11 http://clic.cimec.unitn.it/composes/toolkit/.
12 http://www.maltparser.org/.

These constraints left us with 8,437 training examples for *this/it is X*, 3,071 for *this/it is not X*, 7,350 for *there is X*, and 3,220 for *there is no X*.

To gain further insight into the properties of the resulting phrase vectors, we randomly selected 500 pairs of words such that they would have a corresponding *there is X* vector. Then, we calculated their respective cosine similarity, obtaining a mean score of 0.02, indicating that words are quite dissimilar on average. Next, we replaced one of the words in each of the pairs with its corresponding phrase vector and computed the cosine score again. Interestingly, the average cosine score between phrase and word vectors is 0.06, which is significantly higher, confirming that phrase representations obtained with the method outlined here live in the same sub-space as words do. Moreover, we expect that, even if *X* and *Y* are unrelated concepts, it is likely that the utterances *there is X* and *there is Y* should be somewhat similar in meaning, since they are expressing related "existential" claims. Indeed, if we compare phrase vectors corresponding to the same pairs of words we used earlier, we find that the cosine score increases from 0.02 to 0.54. We can conclude, then, that word and phrase vectors co-exist in the same space and, at the same time, that they have, sensibly, different distributions: phrase vectors, because of their shared semantics, tend to be more similar to each other than word vectors are.

5.1.3 Supervised Regression. The distributional vectors of negated items and their alternatives (either noun or composed phrase vectors) were also fed to a supervised regression algorithm, in order to tune similarity to the specific factors that determine felicitous alternativehood. We explored all the neural-network techniques of Kruszewski and Baroni (2015) across a wide range of hyperparameters, but found, on the development set, that it was best to use support vector regression (Drucker et al. 1996) with an RBF kernel. We also exploited the development set to tune the hyperparameters of this model through a broad grid search, conducted separately for noun vs. phrase inputs and IT vs. THERE ratings.

We had to decide how to merge the vectors representing a negated item and its candidate alternative in order to feed them together to the supervised regression algorithm. Following recent work that addressed the same problem in the context of supervised entailment detection with distributional vectors (Roller, Erk, and Boleda 2014; Weeds et al. 2014), we explored the options of concatenating and subtracting the input vectors. We picked the second strategy as it consistently produced better development set results across all settings. Note that subtraction, unlike concatenation, explicitly encodes the knowledge that we are comparing two feature vectors living in the same space, and what matters should be how the two vectors differ on a feature-by-feature basis.

Furthermore, because supervision should zero in on the information that is not already captured by direct distributional similarity, we explored the possibility to train supervised regression on the residuals of cosine-based rating predictions. Concretely, we first trained a simple single-variable linear regression model using the cosine scores to predict the human ratings. Then, we trained a support vector regression model to predict the difference between human ratings and the output of the cosine-to-ratings model from distributional representations. The final rating prediction for a target pair was obtained by first feeding the corresponding cosine score to the cosine-to-ratings linear function and further incrementing or decrementing the resulting value by the residual produced by the support vector regression model we just described for the pair. Compared with learning the regression parameters by directly predicting the original ratings, this strategy produced better development set results across the board, and we thus adopted it on the test set.

Table 4
Percentage Pearson correlations between model-produced scores and mean human ratings of test set pairs.

Model	IT	THERE
nouns		
unsupervised	86.0	89.3
supervised	86.4	89.7
composed phrases		
unsupervised	70.3	77.5
supervised	87.2	90.2
experts	88.5	92.8

5.2 Results

Table 4 reports Pearson correlations with mean human plausibility ratings on the test set pairs. The *unsupervised* results are obtained by directly correlating vector cosines, the *supervised* ones by correlating scores obtained by feeding the relevant distributional vectors to the trained regression algorithm, as detailed in Section 5.1.3. In the *nouns* setting, the vectors represent the negated and alternative nouns, whereas in the *composed phrases* setting, the vectors are compositionally derived by applying the relevant phrase functions to the nouns. To put the results into perspective, the table also contains an upper bound (*experts* row) that we estimated by rating the test set pairs ourselves, and correlating our averaged ratings with the ones elicited from subjects.

The first and most important result is that simple unsupervised word-based DS similarity is *very* good at predicting alternativehood judgments. Both the IT and THERE correlations obtained in this way are just a few points below the upper bound. Although the underlying benchmarks are not directly comparable, the correlations are as high or higher than those of state-of-the-art systems on widely used semantic similarity data sets (see, e.g., Baroni, Dinu, and Kruszewski 2014), suggesting alternativehood judgments as the ideal task for DS.[13]

Supervision alone has virtually no effect on performance, and composition alone has a strong *negative* effect. However, by combining composition and supervision, we obtain an increase in correlation of about 1% for both IT and THERE. Because we are very close to the upper bound, we cannot hope for much more than tiny improvements of this sort. Still, the correlations between unsupervised noun-based cosines and supervised phrase-based scores are extremely high (99% for IT and 98% for THERE), making it pointless to search for interesting differences between the approaches. We thus focus the rest of the analysis on the simple unsupervised word-based model.

Figure 3 reports the correlations of unsupervised noun-based DS similarity with IT and THERE ratings itemized by relation, illustrating how the results are consistently

13 One possible concern is that distributional models are favored by the fact that we used distributional neighbors as one of the sources in constructing the data set. However, when evaluating the various approaches after removing the distributional class, we obtain results very similar to those reported in Table 4 (larger than 1.5% drops in correlation only in the sub-optimal unsupervised composed-phrase setting).

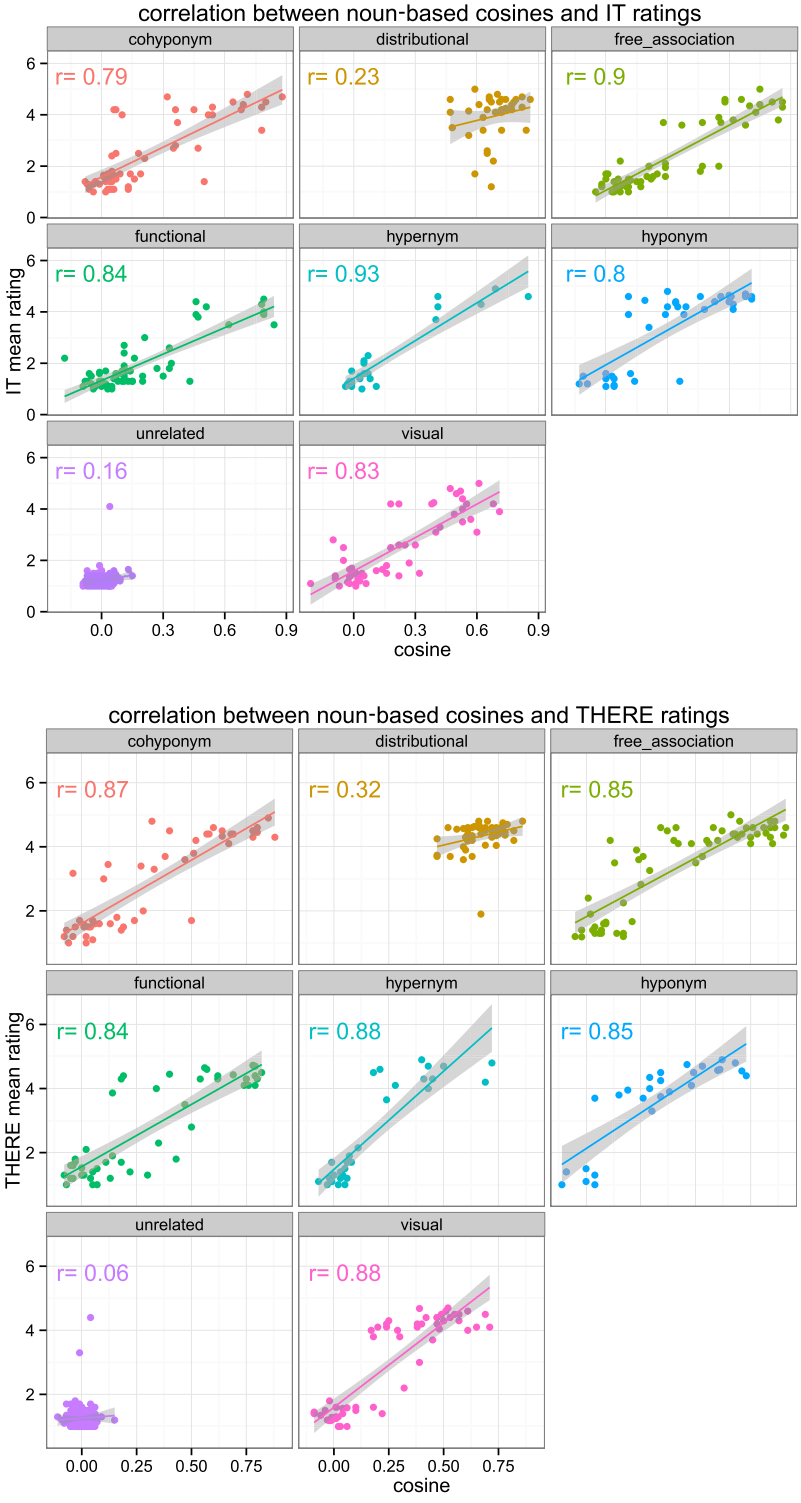


Figure 3
Scatterplots of negated-item/alternative cosines and mean human IT (top) and THERE (bottom) ratings, itemized by relation, with the respective Pearson correlation scores.

high. As the scatterplots show, the low *unrelated* and *distributional* correlations are an artifact of the lack of spread of the corresponding values (not surprisingly, cosines of distributional neighbors are uniformly high). Besides these trivial effects, the data suggest that a single similarity measure is a good predictor of alternativehood, with no strong impact of relation or context type.

On a more qualitative level, the only (weak but) systematic trend we observe when looking at the largest discrepancies between noun cosines and IT/THERE ratings is the following: Many pairs with a large positive difference in favor of cosines are only "topically related," and thus do not make for good alternatives. Examples include: *television/airing* (IT), *shirt/flannel*, *lizard/tongue* (THERE), *television/ABC*, *television/rediffusion*, *cranberry/soda*, and *dagger/diamond* (both). This discrepancy might easily be addressed by building DS models based on narrower contexts, which are known to produce similarity estimates that are more ontologically tight, and less broadly topic-based (Sahlgren 2006). Other discrepancies are due to idiosyncratic properties of specific words. For example, among the pairs with the largest positive difference between (scaled) IT ratings and cosines, we find *cat/cougar*, *cat/panther*, and other *cat/big-cat* pairs, where the corpus-based distributional model is strongly influenced by the pet function of domestic cats, and thus underestimates their similarity to large felines. Similarly, subjects assigned a high THERE rating to *bottle/cup*, whereas the distributional representation of *cup* might be too dominated by its "prize" sense, making it less similar to *bottle*. In a realistic scenario, the relevant statements would be produced in context (*There is no bottle here, but there is a cup* would presumably be uttered as part of a discussion of liquid containers, rather than sports events). Word ambiguity effects could then be addressed by adapting the DS representations of the target terms to the broader context (e.g., emphasizing the container components of *cup*) by means of standard word-meaning-in-context methods (Erk and Padó 2008).

6. Discussion and Future Work

This special issue addresses the integration of formal and distributional models. This is a challenging task, because the two strands of research rely on different research cultures, methodologies, and simplifying assumptions inevitable in any scientific work. Some studies on finding a unifying perspective have been carried out on both empirical and theoretical fronts (Garrette, Erk, and Mooney 2013; Grefenstette 2013; Lewis and Steedman 2013; McNally and Boleda, to appear). We started our work on this article with the firm conviction that distributional models can do more, inspiring and helping to tackle new tasks that are both theoretically interesting and empirically challenging. The pragmatic contribution of negation is just such a task. Our success at modeling alternatives under conversational negation shows that distributional models are immediately applicable to theoretically interesting problems beyond arguably trivial ones, such as identifying near-synonyms.

Cosine in a distributional semantic space turned out to be a very good indicator of alternativehood, suggesting that distributional similarity captures a notion of substitutability in context that is very much in line with the idea of an alternative. Alternatives, in turn, play an important role in many more linguistic phenomena than just conversational negation, and DS, as already conjectured by Baroni, Bernardi, and Zamparelli (2014), might just be the right tool for a large-scale, explicit approach to predicting the set of possible alternatives that are salient at a given point in a conversation. In a broader theoretical perspective, this also resonates with the view from cognitive

science of the brain as an "analogy machine," constantly generating rough predictions about the next state of the world (or of a conversation), based on similarity matching between current information and representations in memory (Bar 2007; Hofstadter and Sander 2013).

In further research, we would like first of all to enlarge our human rating data set in order to make it more challenging to the plain DS similarity approach. Crucially, the current set lacks a sufficient number of synonyms, which will have very high DS similarity but should not make for plausible alternatives (*?This is not an automobile, it is a car*). Moreover, alternativehood is not necessarily symmetric (*There is no cat here, but there is an animal* is plausible, but *?There is no animal here, but there is a cat* sounds odd). Inverted pairs will obviously be a challenge to the symmetric cosine measure. We expect that, once we enrich the data set with such cases, plain DS similarity will no longer suffice, and we will have to rely on supervised approaches that currently seem almost superfluous.

Similarly, compositional methods were only very moderately useful to account for our current skeletal sentential contexts. Compositional, or, more generally, word-meaning-in-context approaches (Erk 2010) will have a better chance to prove their worth if we extend the empirical base to include more informative contexts, and let longer constituents be under the scope of negation and/or in the alternative set (cf., *There is no bachelor here, but there is a... married man, ??unmarried man*).

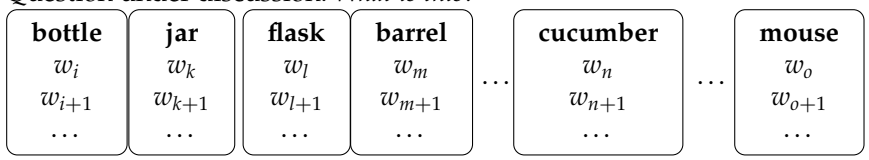
From an empirical point of view, it will of course also be important to study conversational negations in more natural set-ups than the highly controlled experiment we designed.

From a theoretical point of view, the most pressing and interesting issue pertains to integrating what we observed empirically into a formal model. We think that, just like in standard approaches to semantics and discourse (Kamp and Reyle 1993; Ginzburg 2012), and as also recently advocated by McNally and Boleda (to appear), meaning in DS must be modeled in *dynamic* terms, that is, by considering how each utterance affects shared informational content as a coherent discourse unfolds. In particular, we assume that part of the meaning of a linguistic expression is given by a probability distribution over possible linguistic expressions that might follow. Negation, like other dynamic operators, changes this distribution by updating probabilities, according to a similarity function between the distributional representation of the negated expression and the vectors of possible continuations. Appealingly, distributional vectors, especially when contextualized with appropriate techniques, can account at once for generic conceptual aspects of meaning (if we are talking about dogs, other animals are a priori more relevant alternatives) as well as pragmatic and episodic factors (if the topic at hand is therapies for depression, pills might be perfectly coherent alternatives to dogs). Ideally, specific properties of negation and other operators should be induced from corpus data. In particular, the logical ("complement") constraint on negation might be captured by a non-linear scoring function that assigns very low probabilities to vectors that are too similar to the one of the negated element.

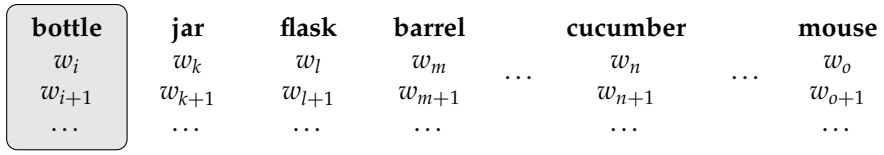
A formal-distributional unification strategy could build on the inquisitive approach to utterance meaning. Inquisitive semantics, which ultimately stems from the analysis of questions by Groenendijk and Stokhof (1984), is devised as a logical system able to characterize (some) conversation moves, and it recently received a complete axiomatization (Ciardelli and Roelofsen 2011). In particular, inquisitive semantics interprets a question as a partition of the set of possible worlds $\{w_1, w_2 \dots\}$ according to potential answers to the question, cf. Example (3a), where the relevant answer fragments are given in bold. A complete answer selects one of the sets in the partition (3b), while a partial

one, e.g., Example (3c), just eliminates some of the possibilities. However, one could imagine an arguably more realistic system where negation does not just exclude one of the possibilities in question (as it does standardly in inquisitive semantics) but assigns different weights or probabilities to the remaining ones, as illustrated schematically in Example (3d), where darker shades of gray correspond to greater weight (probability) of the alternative proposition. As our article suggests, distributional models provide very strong cues for such weights.

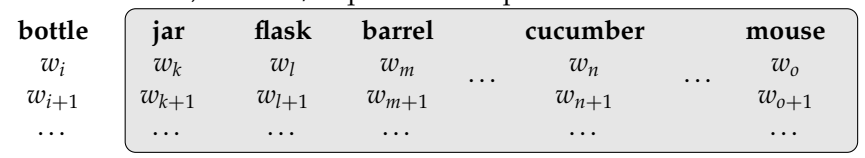
(3) a. Question under discussion: *What is this?*



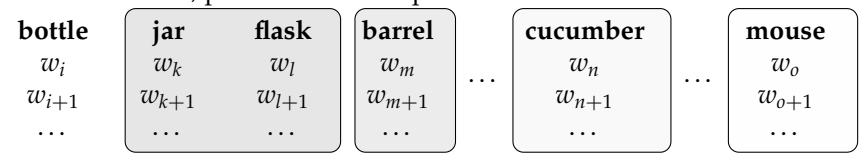
b. Affirmative statement: *This is a bottle.*



c. *This is not a bottle*, classical/inquisitive interpretation.



d. *This is not a bottle*, probabilistic interpretation.



From the point of view of linguistic analysis, a very important issue that remains to be settled is whether linguistic negation always evokes alternatives, and whether the latter are always determined by similarity. We conjecture this to be the case but also that, when negation takes wider scope, the range of alternatives becomes much broader and so does the notion of similarity that we must adopt. Consider:

(4) a. I do not want to watch a movie this afternoon...

b. ... I want to study.

c. ... I want to become an architect.

Example (4b) is a more coherent continuation for (4a) than (4c), and a very broad notion of similarity seems to be at work here as well (roughly, plausible alternatives to Example (4a) must involve activities that can be carried out in the span of an afternoon). It remains to be seen whether compositional distributional semantics is up to the task of modeling alternatives at this level of abstractness.

From an applied perspective, an intriguing application of alternative prediction, suggested to us by Mark Steedman, would be to collect evidence about factoids from implicit contexts. Consider, for example, an information extraction system that is harvesting corpus data supporting the factoid *Pablo Picasso owns poodle*. Obviously, sentences directly stating this fact are the most informative. However, our results suggest that even *Pablo Picasso did not own a chihuahua* should bring (probabilistic) support to the given factoid, to the extent that *poodle* is a plausible alternative for *chihuahua*.

We recognize that the discussion in this conclusion is very speculative in nature, and much empirical and theoretical work remains to be done. However, we hope to have demonstrated that accounting for negation, far from being one of the weak points of this formalism, is one of the most exciting directions in the development of a fully linguistically motivated theory of distributional semantics.

Finally, we also hope that the community will not only be inspired by our modeling results, but will benefit from the data set of alternatives that we collected for this article.[14] For instance, an immediate use of the data set would be in constructing natural examples involving alternatives for theoretical or experimental research in semantics, pragmatics, and reasoning.

Acknowledgments

Jacopo Romoli and Roberto Zamparelli first made us aware of the possible link between alternative semantics and DS. We got the inspiration to work on whether DS can capture alternatives under conversational negation from an informal talk given by Mark Steedman at the 2013 Dagstuhl Seminar on Computational Models of Language Meaning in Context, and from some remarks made by Hans Kamp in the same occasion. Hinrich Schütze first introduced the idea of *dynamic* DS at the same seminar. We thank Uri Hasson and Raquel Fernandez for important bibliographic advice. We had many illuminating discussions on related topics with Roberto Zamparelli, Gemma Boleda, Nick Asher, the Rovereto Composers, and the members of the intercontinental FLOSS reading group. Finally, we thank the reviewers for helpful comments. Our work is funded by ERC 2011 Starting Independent Research Grant no. 283554 (COMPOSES).

References

Bar, Moshe. 2007. The proactive brain: Using analogies and associations to generate predictions. *Trends in Cognitive Science*, 11(7):280–289.

Baroni, Marco. 2013. Composition in distributional semantics. *Language and Linguistics Compass*, 7(10):511–522.

Baroni, Marco, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, 9(6):5–110.

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

Baroni, Marco and Alessandro Lenci. 2011. How we BLESSed distributional semantic evaluation. In *Proceedings of the EMNLP GEMS Workshop*, pages 1–10, Edinburgh.

Baroni, Marco and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of EMNLP*, pages 1183–1193, Boston, MA.

Bullinaria, John and Joseph Levy. 2012. Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming and SVD. *Behavior Research Methods*, 44:890–907.

Cherubini, Paolo and Alberto Mazzocco. 2003. L'effetto della somiglianza nella comprensione della negazione. *Giornale Italiano di Psicologia*, 2:327–353.

14 The data set is available from the site of the Composes project: `http://clic.cimec.unitn.it/composes`.

Ciardelli, Ivano and Floris Roelofsen. 2011. Inquisitive logic. *Journal of Philosophical Logic*, 40(1):55–94.

Clark, Herbert. 1974. Semantics and comprehension. In *Current Trends in Linguistics*, volume 12. Mouton, pages 1291–1428.

Clark, Stephen, Bob Coecke, and Mehrnoosh Sadrzadeh. 2008. A compositional distributional model of meaning. In *Proceedings of the Second Symposium on Quantum Interaction*, pages 133–140, Oxford.

Coecke, Bob, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36:345–384.

Dinu, Georgiana, Nghia The Pham, and Marco Baroni. 2013. General estimation and evaluation of compositional distributional semantic models. In *Proceedings of ACL Workshop on Continuous Vector Space Models and their Compositionality*, pages 50–58, Sofia.

Drucker, Harris, Christopher Burges, Linda Kaufman, Alex Smola, and Vladimir Vapnik. 1996. Support vector regression machines. In *Proceedings of NIPS*, pages 155–161, Denver, CO.

Erk, Katrin. 2010. What is word meaning, really? (And how can distributional models help us describe it?). In *Proceedings of the EMNLP GEMS Workshop*, pages 17–26, Uppsala, Sweden.

Erk, Katrin. 2012. Vector space models of word meaning and phrase meaning: A survey. *Language and Linguistics Compass*, 6(10):635–653.

Erk, Katrin and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of EMNLP*, pages 897–906, Honolulu, HI.

Erk, Katrin, Sebastian Padó, and Ulrike Padó. 2010. A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, 36(4):723–763.

Evans, Jonathan. 1972. Interpretation and 'matching bias' in a reasoning task. *Quarterly Journal of Experimental Psychology*, 24:193–199.

Evans, Jonathan, John Clibbens, and Benjamin Rood. 1996. The role of implicit and explicit negation in conditional reasoning bias. *Journal of Memory & Language*, 35:392–409.

Evert, Stefan. 2005. *The Statistics of Word Cooccurrences*. Ph.D dissertation, Stuttgart University.

Ferguson, Heather, Anthony Sanford, and Hartmut Leuthold. 2008. Eye-movements and ERPs reveal the time course of processing negation and remitting counterfactual worlds. *Brain Research*, 1236:113–125.

Garrette, Dan, Katrin Erk, and Ray Mooney. 2013. A formal approach to linking logical form and vector-space lexical semantics. In H. Bunt, J. Bos, and S. Pulman, editors, *Computing Meaning, Vol. 4*. Springer, Berlin, pages 27–48.

Ginzburg, Jonathan. 2012. *The Interactive Stance*. Oxford University Press.

Golub, Gene and Charles Van Loan. 1996. *Matrix Computations (3rd ed.)*. JHU Press, Baltimore, MD.

Grefenstette, Edward. 2013. Towards a formal distributional semantics: Simulating logical calculi with tensors. In *Proceedings of *SEM*, pages 1–10, Atlanta, GA.

Grice, Paul. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics III: Speech Acts*. Academic Press, New York, pages 41–58.

Groenendijk, Jeroen and Martin Stokhof. 1984. *Studies on the Semantics of Questions and the Pragmatics of Answers*. Ph.D. thesis, Universiteit van Amsterdam.

Guevara, Emiliano. 2010. A regression model of adjective-noun compositionality in distributional semantics. In *Proceedings of the EMNLP GEMS Workshop*, pages 33–37, Uppsala.

Hermann, Karl Moritz, Edward Grefenstette, and Phil Blunsom. 2013. "Not not bad" is not "bad": A distributional account of negation. In *Proceedings of ACL Workshop on Continuous Vector Space Models and their Compositionality*, pages 74–82, Sofia.

Hofstadter, Douglas and Emmanuel Sander. 2013. *Surfaces and Essences: Analogy as the Fuel and Fire of Thinking*. Basic Books, New York.

Horn, Laurence. 1972. *On the Semantics Properties of the Logical Operators in English*. Ph.D. dissertation, Indiana University Linguistics Club.

Hovy, Ed. 2010. Negation and modality in distributional semantics. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, page 50, Uppsala.

Kamp, Hans and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Model-Theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.

Kratzer, Angelika. 1989. An investigation of the lumps of thought. *Linguistics and Philosophy*, 12(5):607–653.
Krifka, Manfred. 1992. A compositional semantics for multiple focus constructions. In *SALT*, pages 127–158, Ithaca, NY.
Kruszewski, Germán and Marco Baroni. 2015. So similar and yet incompatible: Toward automated identification of semantically compatible words. In *Proceedings of NAACL*, pages 64–969, Denver, CO.
Landauer, Thomas and Susan 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.
Levy, Omer, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, 3:211–225.
Lewis, Mike and Mark Steedman. 2013. Combined distributional and logical semantics. *Transactions of the Association for Computational Linguistics*, 1:179–192.
McNally, Louise and Gemma Boleda. To appear. Conceptual vs. referential affordance in concept composition. In Yoad Winter & James Hampton, editors, *Concept Composition and Experimental Semantics/Pragmatics*. Springer.
Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of NAACL*, pages 746–751, Atlanta, GA.
Mitchell, Jeff and Mirella Lapata. 2010. Composition in distributional models of semantics. *Cognitive Science*, 34(8): 1388–1429.
Mohammad, Saif, Bonnie Dorr, Graeme Hirst, and Peter Turney. 2013. Computing lexical contrast. *Computational Linguistics*, 39(3):555–590.
Nelson, Douglas, Cathy McEvoy, and Thomas Schreiber. 1998. The University of South Florida word association, rhyme, and word fragment norms. `http://www.usf.edu/FreeAssociation/`.
Oaksford, Mike. 2002. Contrast classes and matching bias as explanations of the effects of negation on conditional reasoning. *Thinking & Reasoning*, 8(2):135–151.
Oaksford, Mike and Keith Stenning. 1992. Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18(4):835.
Prado, Jérôme and Ira Novek. 2006. How reaction times can elucidate matching effects and the processing of negation. *Thinking & Reasoning*, 12(3):309–328.
Preller, Anne and Mehrnoosh Sadrzadeh. 2011. Bell states and negative sentences in the distributed model of meaning. *Electronic Notes in Theoretical Computer Science*, 270(2):141–153.
Roller, Stephen, Katrin Erk, and Gemma Boleda. 2014. Inclusive yet selective: Supervised distributional hypernymy detection. In *Proceedings of COLING*, pages 1025–1036, Dublin.
Rooth, Mats. 1985. *Association with Focus*. Ph.D. dissertation, GLSA, Department of Linguistics, University of Massachusetts, Amherst.
Sahlgren, Magnus. 2006. *The Word-Space Model*. Ph.D. dissertation, Stockholm University.
Silberer, Carina and Mirella Lapata. 2012. Grounded models of semantic representation. In *Proceedings of EMNLP*, pages 1423–1433, Jeju Island.
Turney, Peter and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.
Van Overschelde, James, Katherine Rawson, and John Dunlosky. 2004. Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, 50:289–335.
Wason, Peter. 1965. The contexts of plausible denials. *Journal of Verbal Learning and Verbal Behavior*, 4:7–11.
Wason, Peter. 1966. Reasoning. In B. M. Foss, editor, *New Horizons in Psychology*, volume 1. Penguin, Harmondsworth, UK, pages 135–151.
Wason, Peter. 1971. In real life negatives are false. *Communication and Cognition*, 4:239–253.
Weeds, Julie, Daoud Clarke, Jeremy Reffin, David Weir, and Bill Keller. 2014. Learning to distinguish hypernyms and co-hyponyms. In *Proceedings of COLING*, pages 2249–2259, Dublin.
Widdows, Dominic. 2003. Orthogonal negation in vector spaces for modelling word meanings and document retrieval. In *Proceedings of ACL*, pages 136–143, Sapporo.