COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2016) 1–35
Copyright © 2016 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12330

Spicy Adjectives and Nominal Donkeys: Capturing Semantic Deviance Using Compositionality in Distributional Spaces

Eva M. Vecchi,[a,b] Marco Marelli,[b] Roberto Zamparelli,[c] Marco Baroni[d]

[a]*Computer Laboratory, University of Cambridge*
[b]*Center for Mind/Brain Sciences, University of Trento*
[c]*Department of Psychology and Cognitive Science, University of Trento*
[d]*Department of Information Engineering and Computer Science, University of Trento*

Received 16 April 2013; received in revised form 30 September 2015; accepted 1 October 2015

Abstract

Sophisticated senator and *legislative onion*. Whether or not you have ever heard of these things, we all have some intuition that one of them makes much less sense than the other. In this paper, we introduce a large dataset of human judgments about novel adjective-noun phrases. We use these data to test an approach to semantic deviance based on phrase representations derived with compositional distributional semantic methods, that is, methods that derive word meanings from contextual information, and approximate phrase meanings by combining word meanings. We present several simple measures extracted from distributional representations of words and phrases, and we show that they have a significant impact on predicting the acceptability of novel adjective-noun phrases even when a number of alternative measures classically employed in studies of compound processing and bigram plausibility are taken into account. Our results show that the extent to which an attributive adjective alters the distributional representation of the noun is the most significant factor in modeling the distinction between acceptable and deviant phrases. Our study extends current applications of compositional distributional semantic methods to linguistically and cognitively interesting problems, and it offers a new, quantitatively precise approach to the challenge of predicting when humans will find novel linguistic expressions acceptable and when they will not.

Keywords: Distributional models; Semantic spaces; Compositionality; Meaning representation; Semantic deviance

Correspondence should be sent to Eva Maria Vecchi, Computer Laboratory, University of Cambridge, 15 JJ Thomson Ave, Cambridge CB3 0FD, UK. E-mail: evamariavecchi@gmail.com

1. Introduction

According to the “distributional hypothesis” (Harris, 1968), words that are similar in meaning tend to have similar distributions; that is, they tend to occur in the presence of similar words. This observation has led to the development of *distributional semantics*, a prominent approach in computational linguistics and cognitive science that approximates a word’s meaning through a numerical vector coding its pattern of co-occurrence with other expressions in a large corpus of language (Sahlgren, 2006; Turney & Pantel, 2010). The meaning of the word *painting*, for instance, could be characterized by a vector recording its co-occurrence with *artist*, *museum*, *colorful*, *abstract*, etc. Meaning relations can then be precisely characterized in geometric terms, since vectors can be treated as points in a multi-dimensional semantic space. Assuming that similar words tend to occur in similar contexts, the distributional vectors of these words will point in similar directions; therefore, geometric distance approximates similarity in meaning (Bullinaria & Levy, 2007; Grefenstette, 1994; Padó & Lapata, 2007; Schütze, 1997).

Examples of implementations of distributional semantics include the hyperspace analog to language model (HAL; Lund & Burgess, 1996), latent semantic analysis (LSA; Landauer & Dumais, 1997), and bound encoding of aggregate language environment (BEAGLE; Jones & Mewhort, 2007). In HAL, each word is represented by a vector where each element corresponds to a weighted co-occurrence count of that word with some other word. LSA also derives a high-dimensional space for words, but it uses co-occurrence information between words and the passages they occur in. BEAGLE incrementally builds reduced-dimensionality representations encoding both semantic information and word order.

Distributional semantic models induce meaning on a large scale from naturally occurring data with little or no supervision, and interesting connections to human language acquisition have indeed been drawn (Landauer & Dumais, 1997). Moreover, they are *general-purpose* approaches, since a model extracted once from a corpus (as a co-occurrence matrix) can be used to capture a large variety of different lexical semantics phenomena (Baroni & Lenci, 2010), thus simulating the flexibility and breadth of human semantic knowledge. Distributional semantic models (and their extensions, such as probabilistic topic models; Blei, Ng, & Jordan, 2003; Griffiths, Steyvers, & Tenenbaum, 2007) have also proved successful at simulating a wide range of psycholinguistic data, for example semantic priming (Griffiths et al., 2007; Landauer & Dumais, 1997; Lund & Burgess, 1996), word categorization (Laham, 2000), reading times (Griffiths et al., 2007; McDonald, 2000), and judgments of semantic similarity (McDonald, 2000) and association (Denhiére & Lemaire, 2004; Griffiths et al., 2007).[1]

Until recently, however, distributional semantics had not seriously addressed the problem of *compositionality* (Frege, 1892; Partee, 2004), the crucial property of natural language that allows speakers to derive the meaning of a complex linguistic constituent from the meaning of its immediate syntactic sub-constituents. Together with general syntactic combination processes, this principle is responsible for the productivity of natural language, which allows speakers to produce and understand sentences they have never encountered before.

To address this serious shortcoming, several recent proposals have strived to extend distributional semantics with methods to derive vectors for complex linguistic constituents, using compositional operations in the vector space (Baroni & Zamparelli, 2010; Coecke, Sadrzadeh, & Clark, 2010; Grefenstette & Sadrzadeh, 2011; Guevara, 2010; Mitchell & Lapata, 2010; Socher, Huval, Manning, & Ng, 2012; Zanzotto, Korkontzelos, Falucchi, & Manandhar, 2010). Most approaches derive distributional semantics representations for novel phrases from the corpus-extracted vectors of their lexical constituents. Since their output is naturally graded, these methods also promise to address the fact that compositionality is a matter of degree (Nunberg, Sag, & Wasow, 1994), ranging from fully compositional cases, as in those attributive adjective-noun phrases whose meaning is the intersection of the meaning of the noun and adjective (e.g., *rented car*, *wooden spoon*), to syntactically fixed expressions such as *take advantage* or *cut a deal*, where the meaning of some of the subparts can still be recognized in the final meaning, to idioms and multi-word expressions (*kick the bucket*, *red herring*, *by and large*), whose meaning cannot be distributed at all across the constituents. Indeed, distributional semantics has already been used to quantify degrees of compositionality (Baldwin, Bannard, Tanaka, & Widdows, 2003; Katz & Giesbrecht, 2006; Schone & Jurafsky, 2001) and recently compositional methods have been applied to this task (Kiela & Clark, 2013).

Multi-word expressions notwithstanding, language is still largely compositional, providing an open space for speakers to create novel but understandable complex linguistic expressions. Yet linguistic creativity has its limits: As native speakers, we have the clear intuition that not all of the infinitely many possible syntactically well-formed strings are equally semantically acceptable. Chomsky's classic "*colorless green ideas sleep furiously*" was devised precisely to show that syntax and semantics can diverge. Our knowledge of compositionality tells us that here the lexical semantics of the words *colorless, green*, and *ideas* do not combine properly. The result is a semantically deviant phrase which cannot be used in "normal" contexts (e.g., non-metalinguistic ones—see below for some qualifications), and therefore it will not be found in corpora, not even very large ones, since corpora largely document actual, normal language use.

Of course, the fact that a complex expression is not found in a corpus can be due to a variety of reasons, which can be quite difficult to tell apart: pure chance, the fact that the expression, though understandable, is ungrammatical, that it uses a rare or very complex structure, that it describes false facts or non-existent entities, or, finally, that it is nonsensical. One criticism aimed at corpus linguistics from the generative linguistic community was precisely that (crude) statistical approaches could not distinguish between these various possibilities (cf. Chomsky's famous remark that "I live in Dayton, Ohio" is not less grammatical, nor indeed, less meaningful, than "I live in New York," despite being far less frequent).

To make the problem more concrete, consider the difference between two adjective-noun phrases which are not attested in a large corpus of English: *grooved tangerine* and *residential steak*. Although it may be the case that you have never considered that a tangerine could have grooves, such an object is easy to imagine and it can be understood in out-of-the-blue contexts. On the other hand, *residential steak* describes an object that is quite hard to imagine. In what sense can a steak be *residential*? Perhaps in none, perhaps in too many: in the context of a man who always and only eats steak when he is in his residence, *his usual residential steak* makes sense. Notice, however, that now the adjective is used only as a proxy for a larger description (*eaten when in residence*). Out of the blue, *residential steak* is semantically very odd, but *grooved tangerine* is not (though it might be factually strange, whence its absence).

Beyond these intuitions, we still do not have a precise linguistic account of what it means for a linguistic expression to be “nonsensical” or semantically deviant, nor a clear relation between this notion and that of being unattested in a corpus: Semantic deviance remains a difficult and understudied phenomenon. In formal denotation-based semantics, for instance, a “meaningless sentence” could perhaps be characterized as one which is false in any imaginable situation (say, in any epistemically accessible possible world). However, this approach would still be unable to determine the degree or even the motivation for the deviance, and it could not predict when a novel string will be deviant. Moreover, there are many necessarily false expressions such as “*17 is not a prime*” which do not feel nonsensical, but simply false. Thus, the task of distinguishing between unattested but *acceptable* and unattested but *semantically deviant* linguistic expressions is not only a way to address a criticism about the limits of corpus linguistics, but also an interesting linguistic task, whose solution could have an impact on our theoretical and psychological understanding of language as a whole, and shape our future treatments of semantic deviance.[2]

In this paper, we first introduce a large database of adjective-noun (AN) phrase acceptability judgments. None of the phrases in the dataset is attested in a very large corpus (about 3 billion words), so we can reasonably assume that subjects never encountered them before. Thus, these data allow us to focus on the challenging task of measuring semantic acceptability of phrases for which no direct corpus evidence is available (certainly for computational systems and very probably for human subjects). To the best of our knowledge, this is the first large dataset of this sort to have been created, and consequently this is also the first attempt to account for them.

By modeling these data, we show that it is possible to use compositional distributional methods to distinguish unattestedness due to semantic deviance from all the other cases, in the domain of simple noun phrases. Specifically, we show how some properties of composed vectors representing AN phrases that never occur in a large corpus (and were not seen by our systems in their training phase) predict the semantic acceptability of the phrases; in particular, vectors of deviant phrases tend to be more distant in distributional space from the vectors of the constituent noun. Moreover, we show that distributional semantic methods improve over shallow word-based measures such as word length and word frequency-derived measures, as well as over probabilistic predictions made by sophisticated statistical models of co-occurrence. We also demonstrate that our composed-phrase-based measures account for the data better than an alternative distributional semantic approach that quantifies the thematic fit of adjectives and nouns without explicitly performing composition. More in general, our results suggest that (compositional) distributional semantics is not just a useful computational method to harvest meaning surrogates, but it can fill an important gap in a general theory of semantics that, as we argued above, lacks convincing accounts of the deviance phenomenon.

This paper is structured as follows. In Section 2, we introduce relevant background on the topic of semantic deviance and review relevant literature across various fields of research. We describe our large dataset of human semantic acceptability judgments of unattested forms in Section 3. Section 4 describes the methodology of our computational simulations. Section 5 reports the results of the modeling experiments, and the Conclusion (Section 6) briefly looks at our current work from a more general perspective and suggests directions for further studies.

2. Background

2.1. Selectional restrictions and thematic fit

The question of when a complex linguistic expression is semantically deviant has been addressed since the 1950s in various areas of linguistics. In theoretical generative linguistics, the issue is part of an ongoing discussion on the boundaries between syntax and semantics. For instance, despite Chomsky's (1957) claim that "*colorless green ideas sleep furiously*" is syntactically flawless, the unacceptability of this case could also be regarded as a violation of very fine-grained syntactic *selectional restrictions* on the arguments of verbs or modifiers, on the model of **much computer* (arguably a failure of *much* to combine with a noun + COUNT).

Asher (2011) presents what is probably the most thorough account of deviance in the spirit of the selectional restriction approach. He proposes a detailed system of *semantic* types, far beyond individuals (*e*) and truth values (*t*). Unacceptable phrases like *residential steak* can then be excluded by type incompatibility. Reducing Asher's proposal to a "cartoon" version for illustration purposes, we might have types such as < *e*-that-are-dwellings > and < *e*-that-you-eat-cooked >. Defining *steak* and *residential* as in (1), *residential* would not accept *steak* as a possible input.

(1) a. **steak**: <*e*-that-you-eat-cooked,*t*>
b. **residential**: <<*e*-that-are-dwellings, *t*>, <*e*-that-are-dwellings,*t*>>

While very elegant, Asher's approach has to stipulate the very rich type system it assumes, with no account of how a learner could induce it from linguistic data. Moreover, the type violation approach predicts that semantic deviance judgments should always be sharp, whereas a cognitively plausible model should account for gradient acceptability. Consider, for instance, the expressions in (2), all of which are unattested in a large corpus, and which have received descending acceptability ratings in the crowdsourcing experiment described in Section 3.

(2) a. creative apprentice
b. ?nuclear seating
c. *careful dark

It is clear that while (2-a) and (2-c) represent the binary extremes of acceptability, (2-b) is neither here nor there; it is clearly an odd expression, yet we would not want to consider it as deviant as (2-c). Equally important, the type-based selectional restriction account has problems with the polysemous nature of meaning combination. Our subjects found *warm garlic* highly acceptable, but *warm equation* very deviant. In a type-based system, this might be accounted for by requiring *warm* to modify expressions with a type denoting concrete and heatable things. But subjects also found a *warm discourse* to be highly acceptable. Asher (2011) proposes a theory of *type coercion*, in which a particular interpretation of a word or phrase is coerced from the context, designed to account for such shifts in meaning, but such mechanisms weaken the predictive power of the approach and seem to miss the intuition that composition is a highly flexible and adaptive process.

Related to selectional restrictions, the *thematic fit* between a verb and its arguments plays a central role in sentence processing research (e.g., McRae, Spivey-Knowlton, & Tanenhaus, 1998; Trueswell, Tanenhaus, & Garnsey, 1994). For example, *policeman* as object of *arresting* causes processing difficulties since policemen are very atypical patients of the verb. Experiments exploring thematic fit typically take subject fitness ratings at face value and use them as independent variables, without attempting to model them in turn. However, recently an interesting computational model predicting thematic fit ratings and relying, like us, on distributional semantics has been proposed by Erk, Padó, and Padó (2010) (see also Erk, 2007; Padó, Padó, & Erk, 2007).[3] The intuition behind this approach is that the similarity of a noun vector with a "prototype" vector representing the typical filler of a verb argument determines the plausibility of the noun as a verb argument. In an example provided in Padó, Padó, and Erk (2007), in order to judge whether *hunter* is a plausible agent of the verb *shoot*, the vector representation for *hunter* is compared to an average of the vectors of common agents of *shoot* observed in the corpus. The proximity of *hunter* to these examples reinforces the possibility that it is an appropriate agent for the verb.

Similarly to Erk and colleagues, we propose an approach to quantifying semantic acceptability based on distributional semantics. However, their method is based on checking constituent compatibility, whereas we first construct a representation for the phrase as a whole, and then look at properties of that representation that might cue acceptability. This makes our approach both more linguistically plausible (semantic acceptability is not determined by an *ad hoc* checking mechanism, but it falls out directly from intrinsic characteristics of independently needed phrase representations) and more general. For example, we can look at the output representation and its neighbors to understand *what* causes the incompatibility of the constituents or to see if the phrase can have a creative or metaphorical interpretation. Moreover, constructing phrase meaning representations allows us to apply the process *recursively*: We can not only check the acceptability of, say, *red car*, but also that of *yellow red car* (see Vecchi, Zamparelli, & Baroni, 2013 for a first attempt in this direction).

Finally, Schmidt, Kemp, and Tenenbaum (2006) proposed a proper mathematical model to distinguish acceptable AN phrases from deviant ones (when combinations are equally unattested). The model (based on the "M constraint"; Sommers, 1971) assumes that properties of objects are organized in a learnable strict hierarchy that can be used to evaluate the meaningfulness of novel combinations. For example, since *soccer games* cannot be *blue* but can be *an hour long*, and *bicycles* can be *blue* but cannot be *an hour long*, any third object cannot sensibly have both properties. On this basis, and since *bananas* are more similar to *bicycles* than *soccer games*, a speaker will consider *blue banana* to be a sensible combination, whereas *hour-long banana* will be deemed meaningless. However, the model remains purely theoretical, it requires hand coding of the hierarchy, and it has yet to be applied to real-world datasets.

2.2. *Psycholinguistic studies of phrase processing*

Psycholinguists have traditionally studied the processing of word combinations by focusing on compound words with nominal constituents. Their studies have shown that constituent representations are accessed when a compound is read, and that many variables influence this process.

Most studies have demonstrated that word frequency is one of the most robust factors driving processing speed: Words with a high frequency of occurrence are processed faster and more accurately than words with a low frequency of occurrence (Gardner, Rothkopf, Lapan, & Lafferty, 1987; Gordon, 1983; Hasher & Zacks, 1984). In addition, the frequencies of occurrence of the constituents of complex words and compounds have been shown to have an effect on lexical processing (Andrews, Miller, & Rayner, 2004; Juhasz, Starr, Inhoff, & Placke, 2003; Pollatsek, Hyönä, & Bertram, 2000). Researchers have also explored the effect of *family size*, that is, the number of distinct phrase types of which the word can be part (for instance, the number of distinct head nouns a given modifier is observed with in a corpus). De Jong, Feldman, Schreuder, Pastizzo, and Baayen (2002) showed that constituent family-size facilitates the lexical processing of compounds in both Dutch and English: the higher the family size of a constituent, the easier it is to process the compound. These effects are not necessarily independent: Kuperman, Schreuder, Bertram, and Baayen (2009), for example, observed multiple interactions involving compound frequency, constituent frequencies, and family size.

String length has been known to influence word processing (Baayen, Feldman, & Schreuder, 2006; New, Ferrand, Pallier, & Brysbaert, 2006). A study carried out in Bertram and Hyönä (2003) provides evidence that string length modulates the access to constituents during the lexical processing of compound words. Specifically, the authors found that in the case of long compounds, it is more likely that the constituents are used for processing (possibly through a compositional procedure), while in the case of short compounds there is probably a direct access to the lexical representations of the compound.

All these studies have investigated the processing of familiar word combinations, while the problem of how novel word combinations are elaborated has been relatively overlooked. Most studies on the latter topic have focused on the role of relational information. For example, research with novel phrases indicates that the time required to interpret a modifier-noun phrase is affected by the availability of the relation used to link the two constituents, with a stronger effect when the relation was associated with the modifier (Devereux & Costello, 2006; Gagné, 2002; Gagné & Shoben, 2002). Length in letters was also recently shown to positively modulate the acceptability of novel noun-noun compounds (Graves, Binder, & Seidenberg, 2013).

These works on novel phrases have focused on novel noun-noun compounds. Few studies were dedicated to AN combinations. Murphy (1990), for example, showed that AN phrases are easier to process and to interpret than paired NN compounds. Further results by Mullaly, Gagné, Spalding, and Marchak (2010) clarified how alternative senses of ambiguous adjectives impact their interpretation and plausibility.

While many studies have provided evidence on how (novel) compounds are processed and how variables such as relational properties and family size play an important role in lexical processing, models predicting the acceptability of novel phrases are for the most part untested, providing little information as to which variables influence acceptability.

2.3. Probabilistic language models

Computational linguists have long been interested in the issue of how to estimate the probability of co-occurrence of bigrams (or longer sequences) that are not observed or are exceedingly rare in a source corpus, and sophisticated probabilistic *language models* have been developed to estimate such probabilities indirectly (Jurafsky & Martin, 2008, ch. 4). Since attested bigram probability in corpora correlates positively with subjective plausibility ratings (Lapata, McDonald, & Keller, 1999), these generalized estimates might provide a plausibility measure for bigrams that are absent from the corpus. In particular, Lapata, Keller, and McDonald (2001) and Keller and Lapata (2003) used *class-based* language models to predict degrees of plausibility of AN combinations. The idea is that, while, say, *blue dog* might never occur in the input corpus, we can approximate its joint probability by that of the combination of *color adjectives* followed by *animal nouns*. Under this view, acceptability judgments are essentially likelihood-of-co-occurrence judgments under a generalized notion of co-occurrence, and semantics is only playing an indirect role in determining the classes used to compute generalized co-occurrence.

Like the thematic fit and word- and frequency-based measures reviewed above, language model-based measures predict the acceptability of a phrase without producing a representation for it. Still, since sequence probability factors do likely play a role in judgments, in our experiments below we re-implement a class-based model akin to those of Lapata and colleagues, and we show that our semantic composition measures significantly improve the fit to human acceptability judgments even when class-based probabilities are taken into account.

3. Collection of semantic acceptability judgments

Our goal is to study whether estimated distributional representations of ANs that never occur in a very large corpus because they are semantically deviant can be recognized as such. In order to do this, we collected an evaluation dataset of human plausibility judgments through a crowdsourcing experiment on CrowdFlower (CF, http://www.crowdflower.com) (Callison-Burch & Dredze, 2010; Munro et al., 2010).

We reasoned that, if adjectives and nouns that are very common never form a phrase together, this should be due to one of the last two factors mentioned in the Introduction: either they describe objects that are odd, rare or nonexistent (say, *grooved tangerines*, *platinum screws* or *Martian senators*), or the combination of A and N does not yield a comprehensible meaning. We thus extracted the 700 most frequent adjectives and 4K most frequent nouns from our source corpus (see Section 4.2.1 below), and manually removed problematic items from the two lists (mostly, words that were assigned to the target categories due to errors of the automated part-of-speech tagging). About half of the combinations generated by the cross-product of the two lists did not occur in the source corpus, and we extracted a sample of about 25K of these *unattested* ANs for our survey.

Since any unattested AN made of a frequent adjective and noun is by definition unfamiliar or at the very least describing an unusual concept, if we were to ask participants to judge the acceptability of each AN using an absolute method such as a standard Likert scale (1–7), we might expect most ANs to remain at the lower end of the scale. Thus, we designed the task in such a way that the participants were forced to make a binary choice on which of two ANs presented together made more sense. This way, we were able to analyze which variables significantly affected the choice of a more acceptable AN.

We constructed a set of AN_x–AN_y pairs in which each of the unattested ANs were seen five times in position x and five times in position y without repetition of pairs, resulting, in theory, in a list of 125K pairs to be judged. However, as some of them were lost due to technical reasons, the final dataset employed contained about 115K pairs. The CF contributors were presented the AN_x–AN_y pairs and asked to decide which of the two AN phrases makes *more* sense in each pair; for example, given the ANs *exact egg* and *Danish workplace*, the contributors would probably select the latter as the phrase that makes more sense (cf. Appendix A for a preview of the task and instructions as presented to the contributors). Since the pairs were matched blindly, it is likely that pairings consisting of two strange or incomprehensible ANs could arise. To address this possibility, contributors were also explicitly told to at least mark the one AN that seemed *less* strange. In addition, we instructed them to judge each AN regardless of which noun may follow it, that is, as a complete phrase; for instance, *blind starch* would likely be judged unacceptable, regardless of the acceptability of *blind starch producer*.

Any distinct participant was allowed to answer to 1,000 pairs at most. In fact, a total of 898 workers took part in the study, each of them evaluating 127 pairs on average. We requested participants to be native speakers of English and only accepted judgments coming from an English-speaking country. Moreover, CF offers a system of quality control, called Gold Standard Data, to determine the accuracy and trustworthiness of the participants. By pre-establishing the correct answers to a small set of data prior to collecting judgments, the system can then calculate the quality of a participant's performance and reject them if their accuracy drops below 70%. This gold data act as hidden tests that are randomly shown to the participants as they complete the task. We included a total of 180 "gold standard" items consisting of an equal number of ANs that were judged clearly acceptable vs. deviant by expert linguists in the study of Vecchi, Baroni, and Zamparelli (2011). We included them in the CF test set in the format AN_x–AN_y, where each pair contained one acceptable and one deviant AN, in random order. Although we cannot guarantee that non-native English speakers did not take part in the study, this system tried to ensure that only the data of speakers with a good command of English and sufficient motivation were retained. Since each AN_x–AN_y pair was seen by one subject, we report the accuracy with respect to the gold items rather than the inter-rater agreement. The average rater accuracy was 97%.

We can quantify a general score of acceptability on an AN-by-AN basis in our dataset by computing how often the AN was chosen as the more acceptable phrase with respect to the number of times the AN was seen by participants. The general scores of acceptability are distributed as shown in Fig. 1.

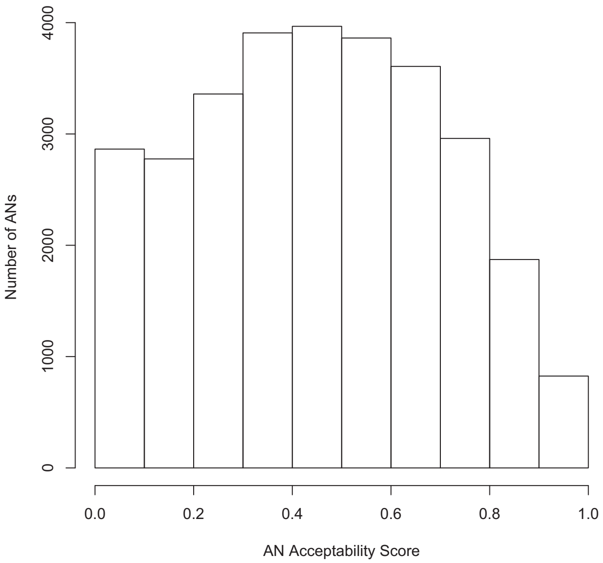


Fig. 1. Distribution of the average acceptability of AN phrases.

The full evaluation dataset is publicly available and can be downloaded from http://www.vecchi.com/eva/resources/an_data_cogsci.tar.gz.

4. Simulation methodology

4.1. Acceptability predictors

We considered a number of predictors that could explain the plausibility judgments we collected. In the present section, we first describe variables taken from the psycholinguistic and computational literature, and then introduce our original methods based on compositional distributional semantics.

4.1.1. Word-based measures

Inspired by the literature on compound processing reviewed above, we considered the effect of *family size*, defined here as the number of times any given adjective or noun is seen in distinct corpus-attested AN phrases. We hypothesize that adjectives and nouns occurring in many phrases, and thus with high family size, correspond to a more flexible semantics; as a result, they should be found more often with acceptable ANs.

A potential measure we also considered was the raw frequency of the component elements in the source corpus. However, the results when using raw frequency were similar to those seen with family size; in fact, the two measures turned out to correlated,[4] so for the experiments described here we only used family size.

Next, we considered the effect of *string length* of component adjectives and nouns for each AN, measured in letters. The hypothesis was that longer component adjectives and nouns should yield more acceptable ANs.

4.1.2. Language model-based measures

We implement a variant of the class-based *language model* (LANGMOD) of Lapata et al. (2001) and Keller and Lapata (2003). We first construct clusters of adjectives and nouns, and we use the co-occurrence counts of the corresponding clusters to estimate the joint probability of specific adjectives and nouns. Intuitively, to measure the plausibility of an unknown phrase such as *parliamentary tomato*, we estimate how likely it is that *parliamentary-like* properties are attributed to *tomato-like* things.

With regard to the clustering step, we implement the recent Affinity Propagation method (Frey & Dueck, 2007), which can automatically find an optimal number of meaningful clusters with stable performance. Clustering was done on constituent vectors, separately for adjectives and nouns, to cover the full target vocabulary of 8K nouns and 4K adjectives (cf. Section 4.2 below for details on vocabulary items and the vectors representing them). This process yielded 667 noun clusters and 420 adjective clusters in total.

The estimated probabilities of the target adjective (noun), given the target noun (adjective), is based on the estimated counts computed with respect to the clusters. We tested three possible approaches for estimating the counts of a given AN:

(i) count all occurrences of any element in *clust*(*A*) with the target *N*, for example, *legislative tomato*, *presidential tomato*, and *elected tomato*;
(ii) count all occurrences of any element in *clust*(*N*) with the target *A*, for example, *parliamentary spinach*, *parliamentary cucumber*, and *parliamentary lemon*;
(iii) count all occurrences of any element in the adjective cluster, *clust*(*A*), with any element in the noun cluster, *clust*(*N*), for example, *legislative cucumber*, *presidential spinach*, and *elected lemon*.

For each way to estimate the AN counts, we implemented three probability measures: joint ($P(A, B)$) and conditional in both directions ($P(A|N)$ and $P(N|A)$).

While we experimented with all possible combinations of counting method and probability measure, we only report in the analysis below the results we obtained by picking the most general counting method (iii) and using the conditional $P(N|A)$, since this was the approach that produced the best results. The LANGMOD score we discuss below is thus given by:

$$P(N|A) = \frac{C(clust(A), clust(N))}{C(clustA)} \tag{1}$$

We expect ANs with a lower estimated probability of N following A to be more deviant.

4.1.3. Thematic fit measures

The next method, while still not using compositional representations, is more directly grounded in distributional semantics and adapts the basic idea of Erk, Padó, and Padó (2010) and related earlier work to measure the fit of an adjective to a noun, or *vice versa*.

In order to compute the *thematic fit* (THEMEFIT) measure, we first construct a "typical adjective (noun)" vector for each noun (adjective) in our dataset by computing the average vector of the 20 most frequently co-occurring adjectives (nouns) with that element. We then determine the estimated appropriateness of the adjective (noun) by calculating the cosine score between the target noun (adjective) and the computed "typical adjective (noun)" vector. The closer a noun (adjective) is to the meaning of the typical noun (adjective) for the adjective (noun), the more acceptable the resulting phrase should be.

4.1.4. Phrase-based distributional semantic measures

We finally present three measures that rely on specific properties of distributional phrase vectors derived with compositional methods. We introduced the first two measures in Vecchi et al. (2011); the third is a variant of a measure proposed there.

Although a *marble iPad* might have lost some essential properties of iPads (it could, for example, be an iPad statue you cannot use as a tablet), to the extent that we can make sense of it, it must retain at least some characteristics of iPads (at the very least, it will be shaped like one). On the other hand, we probably cannot converge on one good interpretation for *legislative onion* (laws written in layers? legislations that make you weep? food prescribed by a vegetarian dictator?) and thus cannot attribute it even a subset of the regular onion properties. For these reasons, we hypothesize that model-generated vectors of less acceptable ANs will be farther from component Ns as represented in the semantic space, forming a wider angle with the component N vectors, thus corresponding to lower *cosine* scores for less acceptable ANs (cf. Fig. 2). The very same idea has been exploited in the literature on detecting idioms and other non-compositional phrases: *pickled herring* should be near *herring* in semantic space, but *red herring* should not (Katz & Giesbrecht, 2006). Obviously, non-compositional meanings can only be acquired for well-attested phrases: If a novel AN meaning is far from that of its head, no previously stored lexicalized interpretation can come to the rescue, and the phrase will be uninterpretable.

Next, we hypothesize that, since the values in the dimensions of a semantic space are a distributional proxy to the meaning of an expression, a meaningless expression should in general have low values across the semantic space dimensions (not being associated with any "meaning dimension" in particular). Consider the common interpretation of the dimensions (or clusters of dimensions) of a distributional semantic space as possible "topics" of discourse (Griffiths et al., 2007): A meaningful phrase such as *academic crusade* will be strongly associated to topics such as academia and politics, and thus have high values on the relevant dimensions. Out of context, it's hard to tell which topics are being addressed when an *academic bladder* enters the discourse, suggesting that this phrase semantic distribution should be low across all dimensions. We thus predict the *vector length* (VLENGTH) of a model-generated AN vector to be a significant factor in the choice of acceptable/unacceptable ANs: the shorter the vector the more likely the AN will be considered less acceptable (cf. Fig. 3).[5,6]

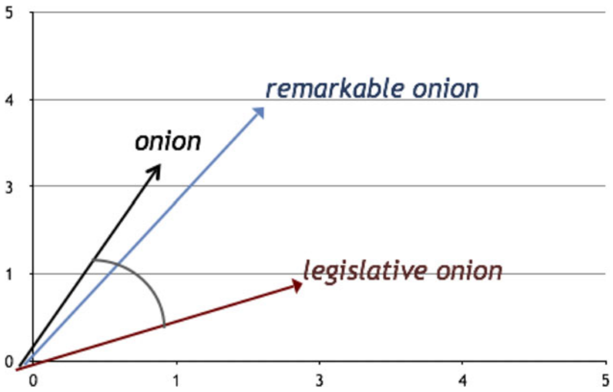


Fig. 2. Prediction for cosine.

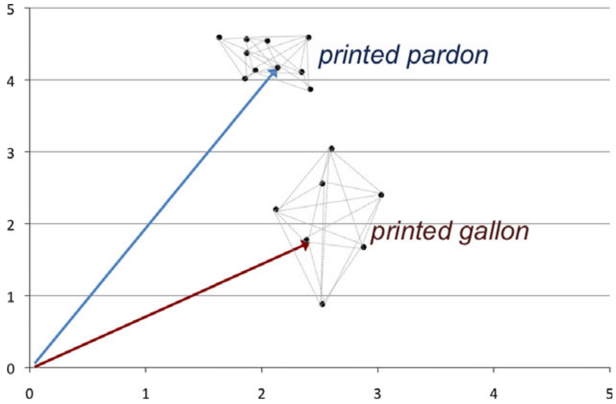


Fig. 3. Prediction for vector length.

In Vecchi et al. (2011), we proposed a measure that reflected neighborhood *isolation* (previously called, somewhat confusingly, “density”), based on the expectation that model-generated vectors of deviant ANs might have few neighbors in the semantic space, since our space is populated by nouns, adjectives, and ANs that are frequently attested in our corpus and should thus be meaningful, and senseless phrases should not be close in meaning to any sensible expression. This measure was calculated by simply taking the average of the cosines between the predicted AN vector and its (top 10) nearest neighbors, expecting deviant ANs to be more isolated than acceptable ANs, corresponding to a lower average cosine score. Indeed, *smooth insecurity*, *printed capitalist*, and *blind multiplier* were found in a more isolated neighborhood (average cosine score <0.55) than the more acceptable *cultural extremist*, *spectacular sauce*, and *coastal summit* (average cosine score >0.75).

While isolation clearly captures real semantic facts, it has some conceptual limits, in that it looks at the relation of an AN with its nearest neighbors but not at their internal coherence. Instead, the intuition we wanted to capture was that a meaningful area of semantic space should be populated by many related concepts, forming a coherent topic. A *cultural extremist* vector, for example, might be located in an area of semantic space pertaining to intellectual and political topics, so that its neighbors will in turn be strongly semantically related to each other. On the other hand, since we do not know what a *blind multiplier* is about, it’s unlikely that its (distant) neighbors will form a coherent set: Some might come from math, others from optometry, and so on. We thus conjectured that model-generated vectors for deviant ANs would share the neighborhood with elements that are not just few, but even dissimilar among themselves. We predicted that ANs with a higher average similarity between their neighbors, or a higher neighborhood *density*, would correspond to more acceptable ANs (cf. Fig. 4). We operationalize this notion by taking the average of the cosines between each element in the neighborhood, which includes the (top 10) nearest neighbors as well as the model-generated AN. Though in

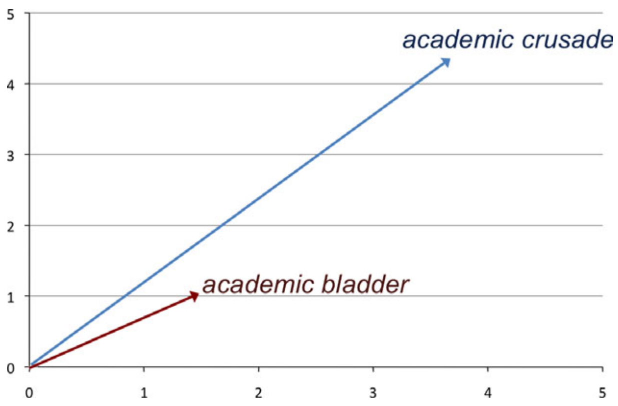


Fig. 4. Prediction for density.

theory the two measures are independent, in practice we found that the effects of the isolation and the density measures were highly correlated for all composition models.[7] Thus, we report only the results for the density measure introduced here, since it provides a more thorough characterization of the neighborhood structure.

4.2. Distributional semantic space

The distributional *semantic space* we use for our experiments consists of a matrix where each row represents the meaning of an adjective, noun, or AN as a distributional vector. We first introduce the source corpus, then the vocabulary of words and ANs that we represent in the space, and finally the procedure adopted to build the vectors representing the vocabulary items from corpus statistics, and obtain the semantic space matrix.

4.2.1. Source corpus

We use as our source corpus the concatenation of the Web-derived ukWaC corpus (http://wacky.sslmit.unibo.it/), a mid-2009 dump of the English Wikipedia (http://en.wikipedia.org) and the British National Corpus (http://www.natcorp.ox.ac.uk/). The corpus has been tokenized, POS-tagged, and lemmatized with the TreeTagger (http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/), and it contains about 2.8 billion tokens. We extract all statistics at the lemma level, meaning that we consider only the canonical form of each word ignoring inflectional information, such as pluralization and verb inflection.

4.2.2. Semantic space vocabulary

We first populated our semantic space with a core vocabulary containing the 4K most frequent adjectives and the 8K most frequent nouns from the corpus. The vocabulary was then extended to include a large set of ANs (119K cumulatively), for a total of 132K vocabulary items. The large majority of ANs was randomly selected among those that occurred at least 100 times in the corpus and were formed by the combination of one of the 700 most frequent adjectives with one of the 4K most frequent nouns. To add further variety to the semantic space, we sampled a less controlled set of 3.5K ANs randomly picked among those that are attested at least 100 times in the corpus.

4.2.3. Semantic space construction

For each of the items in our vocabulary, we first build 10K-dimensional vectors by recording the item's sentence-internal co-occurrence with the top 10K most frequent content words (nouns, adjectives, verbs, or adverbs) in the corpus, excluding the 300 most frequent ones (because they are borderline-grammatical elements with little discriminative power). Following standard practice, raw co-occurrence counts were transformed into local mutual information (LMI) scores, an association measure that closely approximates the commonly used log-likelihood ratio while being simpler to compute (Baroni & Lenci, 2010; Evert, 2005). Specifically, given a row element r (here, the adjectives, nouns or ANs in the semantic space), a column element c (in this case, the 10K most frequent content words), and a counting function $C(x, y)$, then

$$LMI = C(r,c) \cdot \log \frac{C(r,c)C(*,*)}{C(r,*)C(*,c)} \tag{2}$$

where $C(r, c)$ is how many times r co-occurs with c, $C(r, *)$ is the total count of r, $C(*, c)$ is the total count of c, and $C(*, *)$ is the cumulative co-occurrence count of any r with any c.

Next, we reduce the full co-occurrence matrix applying the singular value decomposition (SVD) operation, a technique of dimensionality reduction that approximates a sparse co-occurrence matrix with a denser lower-rank matrix, in our case reducing dimensions from 10K to 600.[8] The SVD technique is used in LSA and related distributional semantic methods because, besides easing computational load, there is extensive evidence that it improves semantic representations (Bullinaria & Levy, 2012; Landauer & Dumais, 1997; Rapp, 2003; Schütze, 1997; Turney & Pantel, 2010).

As a quality check, we verified that the vectors in our semantic space attain state-of-the-art-range performance on various benchmarks (cf. Appendix B).

4.3. Composition methods

Mitchell and Lapata (2008, 2009, 2010) explore a variety of composition strategies for distributional semantic models and find three simple methods to work quite well across the board.

Given two constituent vectors **u** and **v**, the (weighted) *additive* model (*add*) derives the phrase vector **p** by summing them:

$$\mathbf{p} = \alpha\mathbf{u} + \beta\mathbf{v} \tag{3}$$

The intuition here is that the phrase vector inherits all the active features (dimensions) of both constituents, possible in different proportions. The *multiplicative* (*mult*) approach uses instead component-wise multiplication, where the i-th component of the composed vector is given by

$$p_i = u_i v_i \tag{4}$$

Multiplication has a "feature intersection" effect, where only dimensions that have high values in both constituent vectors will be high in the phrase, whereas dimensions that are high for just one input will cancel out.

Another effective method introduced by Mitchell and Lapata is *dilation* (*dil*), defined as:

$$\mathbf{p} = (\mathbf{u} \cdot \mathbf{u})\mathbf{v} + (\lambda - 1)(\mathbf{u} \cdot \mathbf{v})\mathbf{u} \tag{5}$$

Under dilation, the phrase is obtained by "stretching" the **v** vector by a factor λ in the direction of **u**. The intuition is that the action of combining two words results in one making some semantic components more salient in the other. Dilation stretches **v** so as to emphasize the contribution of its components pointing in the direction of **u**. By looking at Eq. (5), observe that another intuition for the dilation method is that it is a variant of the additive model where the relative weight of **u** changes from phrase to phrase depending on how similar it is to **v**. In our experiment, we take **u** to be the noun and **v** the adjective, because this configuration worked best in the tuning experiments reported in Appendix C.

Mitchell and Lapata (2008, 2009, 2010) evaluate these models on a wide range of tasks ranging from paraphrasing to statistical language modeling to predicting similarity intuitions, obtaining good performances overall. The empirical effectiveness of their relatively simple models has also been confirmed by many later studies; see in particular Blacoe and Lapata (2012), who showed that they provide performance comparable or superior to sophisticated deep neural architecture methods.

In the *lexical function* (*lf*) approach first introduced in Baroni and Zamparelli (2010), attributive adjectives are treated as functions from noun meanings to noun meanings (see Coecke et al., 2010, for a closely related proposal). This is a standard approach in Montague semantics (Montague, 1974), except noun meanings here are distributional vectors, not denotations, and adjectives are (linear) functions, encoded in matrices whose weights are induced with regression methods from corpus-extracted examples of the phrases they produce. The formula to compose a phrase is then:

$$\mathbf{p} = \mathbf{U}\mathbf{v} \tag{6}$$

where **U** is the matrix representing a specific adjective, and **v** is a noun vector. Baroni and Zamparelli (2010) show that this model significantly outperforms other vector composition methods, including addition, multiplication, and the "full additive" method of Guevara (2010) and Zanzotto et al. (2010) on the task of approximating corpus-extracted ANs. More extensive comparative evaluations demonstrating the effectiveness of the lexical function model were performed by Dinu, Pham, and Baroni (2013) and Li, Baroni, and Dinu (2014).

Since the four methods we just reviewed have already been shown to be successful, specifically, in many tasks involving AN phrases (Boleda, Vecchi, Cornudella, & McNally, 2012; Dinu et al., 2013; Vecchi et al., 2011; Vecchi et al., 2013), we focus on them for the current study. We optimized their parameters as described in Appendix C.

5. Simulation results

The effect of each predictor of interest on the participants' judgments was estimated by means of mixed-effects logit models (Jaeger, 2008).[9] This was aimed at testing how much the different measures would increase the likelihood of choosing one AN over the other. For each AN–AN pair, the dependent dichotomous variable was whether a participant chose the first or the second element. As proposed by Baayen, Davidson, and Bates (2008), we introduced random intercepts of participants in order to account for the random variance associated to judges, and in particular their possible *a priori* tendency to pick the first or second element. Moreover, we introduced random intercepts associated to the constituents of both phrases (i.e., first adjective, first noun, second adjective, second noun).[10] This way, we expect to capture both the variance associated to the individual constituents and the one associated to the phrases the constituents are nested in, since each AN is univocally identified by the corresponding adjective + noun combination. This random structure was aimed at accounting for unexplained item variance, and in particular that associated to variation in pair difficulty: Every judgment was relative in its nature but, given that each AN was presented with several competitors, its likelihood of being chosen (*a priori*, in either positions) could be estimated and modeled using the constituent random intercepts, in turn assuring that the final results are not influenced by specific items being overly easy or difficult. See Appendix D and Section 5.3 for alternative analyses of the simulations, confirming the general results we report below.

We evaluate performance of the distributional semantic measures of semantic acceptability (our composed-phrase-based cues and the thematic fit method of Erk and colleagues) in the following challenging regime. We first establish a strong non-semantic baseline taking into account constituent length, family size, and co-occurrence likelihood as estimated by the best performing class-based language model measure (cf. Section 4.1.2). The baseline thus includes five fixed effects each for the left- and right-hand ANs: adjective and noun STRING LENGTH and FAMILY SIZE, and $P(N|A)$ (LANGMOD). We then added, in turn, each semantic measure of interest for both the left- and right-hand ANs as additional fixed effects to the baseline model, and tested whether they significantly improved model fit, that is, whether the result of the likelihood ratio test comparing the goodness-of-fit of the model before and after introducing the new parameters was significant (Baayen et al., 2008). We test 2 THEMEFIT factors (noun checked for fit to adjective or *vice versa*) and 12 phrase-based factors (cosine, vlength and density crossed with (4) composition methods), for a total of 14 variables.

5.1. Baseline measures

All baseline measures are significant factors when choosing which AN makes more sense; cf. Table 1. This is consistent with previous studies (see Section 2), thus confirming the reliability of the plausibility data we collected. The coefficients reported on the table indicate, for each unit change in the predictor, the corresponding change in the log-odds of choosing the right-hand as opposed to left-hand AN. Therefore, the polarity of the estimated coefficients indicates the likelihood of choosing the left-hand (*L*, negative) as opposed to the right-hand (*R*, positive) AN as more acceptable.

For both adjectives and nouns, longer words result in more acceptable ANs. This is in line with the study by Graves et al. (2013) indicating that longer phrases are deemed more acceptable. The present results suggest that the effect might depend on the length of the individual elements, rather than the word combination itself. A possible explanation is that longer component words are generally more abstract and may therefore be more flexible when integrating new modification. Denominal adjectives, for instance, are often relatively long and can be very unspecified with respect to the relation that connects the noun root they contain with the AN head (e.g., *industrial pollution* vs. *industrial site* vs. *industrial process*). An attention-capturing effect is, however, also possible: Longer words are perceptually more salient, and for this reason they may be more likely to be chosen in a forced-choice task when the answer is unclear. We also observe that the length effect is stronger for nouns than adjectives: We conjecture that this is due to nouns having a wider length range, with a larger maximum value (17, vs. 14 for adjectives).

Table 1
Baseline measures

Measure	Coefficient	*SE*	*z*-value	*p*
$\mathrm{A}_L$string length	−0.0696	.0003	−20.16	.0001
$\mathrm{A}_R$string length	0.0714	.0003	20.65	.0001
$\mathrm{N}_L$string length	−0.1073	.0003	−31.56	.0001
$\mathrm{N}_R$string length	0.1029	.0003	30.16	.0001
$\mathrm{A}_L$family size	−0.0003	.0001	−22.61	.0001
$\mathrm{A}_R$family size	0.0003	.0001	25.56	.0001
$\mathrm{N}_L$family size	−0.0018	.0001	−24.18	.0001
$\mathrm{N}_R$family size	0.0019	.0001	25.05	.0001
langmod$_L$	−28.2901	3.4241	−8.26	.0001
langmod$_R$	26.4301	3.4511	7.66	.0001

Notes. Results of the logit mixed effects model run on the CrowdFlower data using the baseline measures. The results include the effect of the constituent family size, the constituent string length, and class-based-language-model $P(N|A)$ (langmod) on the choice of acceptable ANs. Separate coefficients are estimated for the left- (*L*) and right-hand (*R*) phrases, and, when applicable, for the adjective (A) and the noun constituent (N).

For both adjectives and nouns, a higher family size yields more acceptable ANs. This is quite intuitive, since items with a high family size should be highly productive, and therefore less restrictive in what they combine with.

Finally, we find that the results for the language model-based measure are in line with our intuition and with previous studies: the higher the estimated probability of N given A, the more acceptable the phrase (although the effect is not as strong as for the arguably shallower STRING LENGTH and FAMILY SIZE measures).

5.2. *Improvement on baseline brought about by distributional semantic measures*

Table 2 shows the results of the likelihood ratio test comparing the goodness-of-fit of the model including the baseline measures before and after introducing each semantic measure.[11] Reported goodness-of-fit is measured in terms of both AIC and log-likelihood. Nearly all measures based on distributional semantic representations significantly improve the fit.

The non-compositional THEMEFIT measure, while significantly improving fit, is not as good as the best measure exploiting composed phrases, namely COSINE, as seen in the comparison of AIC scores in Table 2. This is an important finding, since the THEMEFIT method has been developed specifically to measure the thematic fit of constituents, and

Table 2
Performance of the semantic measures in predicting AN acceptability

Measure		*df*	AIC	logLik	Chisq	*p*
BASELINE		16	145008	−72488		
THEMEFIT	*adjective*	18	144396	−72180	615.43	.0001
	noun	18	144317	−72140	694.63	.0001
VLENGTH	*add*	18	143966	−71965	1,045.20	.0001
	mult	18	144754	−72359	257.84	.0001
	dil	18	144373	−72168	638.76	.0001
	lf	18	144906	−72435	105.57	.0001
COSINE	*add*	18	144085	−72024	926.58	.0001
	mult	18	144190	−72077	821.66	.0001
	dil	18	143175	−71569	1,836.60	.0001
	lf	18	143777	−71871	1,234.30	.0001
DENSITY	*add*	18	144922	−72443	89.31	.0001
	mult	18	145010	−72487	2.03	.3632
	dil	18	144619	−72291	392.61	.0001
	lf	18	144958	−72461	53.98	.0001

Notes. Results for logit mixed effects models including both baseline and semantic measures as opposed to baseline measures only. Goodness-of-fit is expressed in terms of AIC and log-likelihood; *p* values based on chi-square tests. Measures considered: thematic fit (THEMEFIT) for both *adjective* and *noun*, length of the composed phrase vector (VLENGTH), similarity between the noun component and the phrase vector (COSINE), density of the neighborhood of the phrase vector (DENSITY). Compositional models considered: additive (*add*), multiplicative (*mult*), dilation (*dil*), lexical function (*lf*). *df*, degrees of freedom.

does not provide a representation of the phrase. Our composition-based approach is at the same time more general and empirically more effective on the task at hand.

The goodness-of-fit improves most consistently across composition methods (in terms of difference in log-likelihood and AIC) with respect to the cosine measure, namely the proximity between the composed AN vector and its component noun. This confirms the intuition that a new phrase, to stay sensible, cannot stray too much from its head meaning, and it suggests that all considered composition methods naturally capture the desired effect of altering the vector direction more radically when a combination is not meaningful.

The *add* model seems particularly well suited for the vlength cue. Since a vector derived by addition will be longer when the input vectors are closer, a specific factor playing a role here might be that adjectives and nouns that are already similar in meaning will be more likely to be interpretable when combined. However, the significant effect of vlength irrespective of composition method is also supporting the more general intuition we motivated this cue with (vectors of meaningless phrases have low values across the board since they are not associated to any coherent topic of discourse).

At face value, the density measure also appears to be a good acceptability indicator (except when computed on *mult*-derived vectors). However, the effect is in all cases in the *opposite* direction with respect to our hypothesis: Unacceptable ANs have significantly denser neighborhoods than acceptable ones. To gain some insight into this surprising result, we inspected the nearest neighbors in semantic space of unacceptable ANs with high density, finding that they are often closer to the meaning of the component adjective than acceptable ANs with high density. The examples in (3) list the nearest neighbors in the semantic space for a set of ANs with high neighborhood density, based on the results from the *lf* composition method (here and below, we use asterisks to mark ANs with low acceptability scores; see (4) for additional examples).

(3) a. *angry lamp { *shocked*, *fearful*, *angry*, *defiant* }
b. *nuclear fox { *nuclear*, *nuclear arm*, *nuclear development*, *nuclear expert* }
c. warm garlic { *green salad*, *wild mushroom*, *sauce*, *green sauce* }
d. spectacular striker { *goal*, *crucial goal*, *famous goal*, *amazing goal* }

We see that the nearest neighbors for the high-density, semantically deviant ANs in (3-a,b) are more similar in meaning to the component adjectives than the neighbors of high-density, acceptable ANs in (3-c,d). Furthermore, we find that neighbors for acceptable ANs with high density are more often close to the component noun, while neighbors for unacceptable ANs do not maintain any meaning of the component noun. Now, by construction, our semantic space contains more ANs per adjective than per noun (on average, 162 vs. 30). Thus, if the meaning of the adjective overpowers the meaning of the AN in deviant cases, the composed meaning will likely occupy an area within the denser adjective neighborhoods. If this tentative analysis is correct, the results for the density measure are actually confirming the trend uncovered by the cosine heuristic, namely that unacceptable ANs are characterized by a strong pull out of the semantic domain of the noun.

Finally, all composition models behave quite similarly in quantitative terms (we should not blame *mult* for failing to reach significance in combination with the DENSITY measure, in light of the surprising behavior of the latter measure). Later in this section, we will take a qualitative look at the AN vectors generated by the various methods, where we will see some differences emerge.

5.3. Accuracy of semantic measures

We present here an alternative analysis of the results in terms of accuracy of the measures at predicting, for each of the 114.5K AN pairs in our evaluation set, the AN that was preferred by the subject who rated the pair. This analysis does not take into account the control factors we consider in the main analysis. In particular, it does not combine distributional semantic measures with form- and language model-based scores, and it does not attempt to control for random variance associated to items and subjects. On the other hand, it provides an intuitive way to assess how much better our measures perform with respect to a simple baseline consisting in always picking the second AN (the majority choice). In particular, for each measure, we report statistical significance of a two-tailed binomial exact test comparing the number of pairs correctly classified by the measure to the number of hits of the majority baseline.

Table 3
Accuracy of the semantic measures in predicting chosen AN in a pair

Measure		Accuracy	*p*
MAJORITY BASELINE		0.515	NA
THEMEFIT	*adjective*	0.549	.0001
	noun	0.564	.0001
VLENGTH	*add*	0.566	.0001
	mult	0.521	.0001
	dil	0.565	.0001
	lf	0.509	.0001
COSINE	*add*	0.566	.0001
	mult	0.554	.0001
	dil	0.567	.0001
	lf	0.570	.0001
-DENSITY	*add*	0.512	.0001
	mult	0.498	.0001
	dil	0.535	.0001
	lf	0.516	.0001

Notes. Results of the accuracy analysis described in the text. Measures considered: MAJORITY BASELINE, thematic fit (THEMEFIT) for both *adjective* and *noun*, length of the composed phrase vector (VLENGTH), similarity between the noun component and the phrase vector (COSINE), *negative* density of the neighborhood of the phrase vector (-DENSITY). Compositional models considered: additive (*add*), multiplicative (*mult*), dilation (*dil*), lexical function (*lf*). Reported *p* values pertain to test of two-tailed difference from majority baseline level.

The results, in Table 3, replicate the general trends discussed in Section 5.2. In particular, we confirm that cosine with head noun is the most consistent measure, and that the density measure behaves in the opposite way from what we expected (the table reports results for $-1 \times$ density, which is more accurate than the positive measure).

5.4. Qualitative analysis of nearest neighbors of composed phrases

We have already stressed that a major advantage of the composition-based approach to semantic acceptability is that it does not only provide a measure to quantify the phenomenon, but it constructs full-fledged (distributional) semantic representations of the phrases of interest (unlike, e.g., the language modeling or thematic fit methods). These representations can be used to capture other semantic phenomena (e.g., measuring phrase similarity), but they can also be qualitatively assessed by looking at their nearest neighbors in semantic space. In Table 4, we provide examples of the top three nearest neighbors for a set of ANs in our test set.

The nearest neighbors of the *mult* function are quite odd for both acceptable and deviant ANs. The *add* model was able to model the acceptability judgments quite well, but we find that the nearest neighbors it predicts are strongly related to the component noun in all ANs, with no trace from the adjective. Both *dil* and *lf*, on the other hand, give more importance to the modifier. The meaning of the adjective seems to take over for deviant ANs when using the *lf* model, however we can see that in acceptable cases the nearest neighbors do represent the intuitive, functional combination of the meanings of the modifier and the head noun. This is the only composition model capable of capturing this. Thus, while from the quantitative results we should conclude with Blacoe and Lapata (2012) that there is no reason to adopt more complex methods of composition, the qualitative evidence supports the more sophisticated and linguistically motivated *lf* approach.

6. Conclusion

The ability to produce and understand linguistic expressions we never encountered before is one of the most powerful and fascinating aspects of the cognitive faculty of language. While linguists have worked for many decades on the syntactic aspects of linguistic productivity, the semantic factors that make a new phrase acceptable have been somewhat overlooked. Traditional methods from linguistics might be poorly equipped to handle semantic acceptability, since the latter is a graded phenomenon requiring large-scale, flexible commonsense knowledge about possible combinations. On the other hand, compositional distributional semantic models possess exactly these characteristics, and they might thus be well suited to account for what makes a new phrase semantically acceptable or deviant.

To face the problem empirically, we collected a large database of human semantic judgments about adjective-noun phrases that never occur in a very large corpus. We then

Table 4
Examples of the nearest neighbors of composed AN vectors

	add	*mult*	*dil*	*lf*
**empty fungus*	fungus	other cell	empty	empty tin
	other fungus	only cell	empty one	empty packet
	nematode	original cell	little space	empty container
**mathematical biscuit*	biscuit	complex one	mathematical	mathematical
	hot chocolate	new shape	mathematical tool	mathematical result
	sticky bun	effective idea	mathematical approach	mathematical system
**mental sunlight*	sunlight	secondary effect	mental	mental factor
	direct sunlight	considerable distress	mental health	mental experience
	natural sunlight	bipolar	mental promotion	mental fatigue
**moral protein*	protein	being	moral	moral
	new protein	potential movement	moral system	moral conscience
	basic protein	habitation	morality	moral sense
**written oak*	oak	late century	written	written
	beech	early century	oral	written form
	English oak	fifteenth	written exercise	written work
continuous uprising	uprising	period	continuous	constant warfare
	armed resistance	more period	uprising	constant conflict
	national uprising	probationary	armed resistance	continuous war
diverse farmland	farmland	rich habitat	farmland	diverse wildlife
	agricultural field	good habitat	agricultural field	varied habitat
	upland	diverse habitat	upland	rare flora
important coordinator	coordinator	pivotal	coordinator	instrumental role
	new coordinator	crucial	junior coach	integral role
	secondary coach	role	new coordinator	significant role
legendary province	province	several king	legendary	legendary city
	autonomous community	great king	province	famous
	prefecture	king	famed	great city
systematic likelihood	likelihood	risk	likelihood	systematic effect
	increased	acceptable risk	increased	systematic bias
	overall exposure	actual risk	adverse outcome	systematic approach

Notes. We report the top three nearest neighbors of the AN vectors—generated using *add*ittive, *mult*iplicative, *dil*ation, and *l*exical *f*unction model—in the semantic space. The asterisk (*) implies that the general acceptability score of the AN in the CF experiment (i.e., the number of times it was chosen as the more acceptable AN with respect to the number of times it was seen by participants) is less than 0.2. The other ANs reported here have a general acceptability score greater than 0.8.

proceeded to model these judgments with phrase representations derived by compositional distributional semantic models, showing that, even when other factors are taken into account, inherent characteristics of composed vector representations and their location in space explain a significant portion of variance in semantic acceptability. We also gained more specific insights into how composition affects interpretability, mainly through the potentially disrupting effect that an adjective can have on the overall meaning of the phrase. We achieved all this without the need to postulate special concept combination constraints, semantic types, or explicit selectional restrictions. (Lack of) acceptability in our system naturally emerges as a by-product of the composition process. We also showed that the linguistically inspired lexical function composition model provides qualitatively plausible semantic representations of unattested but meaningful phrases, further strengthening our view of compositional distributional semantics as a useful addition to a full-fledged theory of meaning.

In this spirit of seriously taking compositional distributional semantics as a linguistic theory, our next move will be to look at more specific patterns uncovered by our models, studying for example subclasses of adjectives and nouns, and how A-N relations such as redundancy (i.e., *wooden tree*) or oxymorons (i.e., *dry liquid*) affect acceptability. We want moreover to delve further into the issue of how polysemy and shifts in meaning interact with deviance. We intend in particular to study how compositional distributional semantics can capture the human ability to repair deviance by creative interpretation, for example coming up with figurative meanings. On a more psychological side, we would like to explore how our acceptability measures could be integrated in a realistic architecture for language processing. On the modeling side, we plan to develop richer measures of acceptability (in particular, supervised measures that can take the inner structure of phrase vectors into account), to attain further insights on what are the factors at work in determining the semantic success of a novel phrase.

Acknowledgments

We thank Roger Levy, our Cognitive Science editor, and the reviewers for constructive criticism. We thank Katrin Erk, Gemma Boleda, Angeliki Lazaridou, Yao-Zhong Zhang, Stephen Clark, and the members of the Composes group for various forms of feedback and advice. Special thanks to Jiming Li for helping with the Affinity Propagation method. This work was conducted while Eva Maria Vecchi was at the Center for Mind/Brain Sciences of the University of Trento, and it was funded by ERC 2011 Starting Independent Research Grant no. 283554 (COMPOSES).

Notes

1. We take the extensive success of distributional semantic methods at modeling a variety of meaning-related phenomena as evidence that such methods, while relying on purely distributional patterns, are indeed building vectors that are good proxies to a word *meaning*. We will thus refer to these models as "semantic" throughout the paper.

2. We adopt the (semantically) *acceptable* versus *deviant* terminology, but the same phenomenon might be termed semantic *plausibility*, *well-formedness*, or, with opposite polarity, semantic *anomaly*, *nonsensicality*, *meaninglessness*, etc. We emphasize the "deviant" end of the scale because we think the conceptual core of the problem lies in understanding what makes a phrase nonsensical, rather than in routine successful composition. Also on the terminological side, we believe that the terms *selectional preferences/restrictions* and *thematic fit* refer to the same phenomenon when studied in the specific context of predicate–argument relations.
3. The system of Erk and colleagues outperform earlier computational approaches to thematic fit, such as Resnik's (1996) WordNet-based method, and we do not consider these earlier approaches here.
4. The Spearman correlation between adjective family size and raw frequency is 0.67, and the Spearman correlation between noun family size and raw frequency is 0.71.
5. Kiela and Clark (2013) implicitly use vector length as one cue of (literal) semantic anomaly that they rely upon to identify idiomatic constructions. Since their phrases are composed by component-wise multiplication (see Section 4.3), they motivate the length heuristic as resulting from the dimensions of incompatible constituents canceling out when multiplicatively combined. We found VLENGTH to work well also with other composition methods (Section 5), supporting the more general interpretation we suggest in the text.
6. A reviewer points out that, since in general vectors of more frequent words or phrases will be longer, our reasoning suggests that frequency will correlate with semantic acceptability. This is an interesting prediction that we find quite credible: As a general tendency, frequent words and phrases might be easier to make sense of than rare ones, as we have more evidence about the topics they are about (one could even argue that "I live in Dayton, Ohio" is indeed less meaningful than "I live in New York" to someone very familiar with New York but who does not have the faintest idea about what Ohio is like). Still, we will show below that VLENGTH has a significant impact on acceptability even when frequency-based variables such as *family size* and LANGMOD are taken into account, so VLENGTH cannot be reduced to a frequency effect.
7. Spearman correlations between neighborhood isolation and neighborhood density for each composition model to be introduced below: *add*: 0.875; *mult*: 0.697; *dil*: 0.851; *lf*: 0.885.
8. So as not to bias the structure of the lower-dimensionality space toward the small subset of possible ANs we selected, only adjective and noun vectors were used to compute the SVD projection.
9. We used R lme4 package: http://CRAN.R-project.org/package=lme4.
10. Note that we are entering separate random factors for adjectives and nouns in first and second position. This is partly due to limitations of the package we are using to implement the analyses. If adjectives and nouns have the same effect when they occur in either position, given the large scale of the test set, the fitting procedure should naturally discover similar (opposite-sign) coefficients for the

adjective and noun intercepts. If, on the other hand, there is an interaction between lexical items and position, our structure can capture this. An analysis of the estimated random intercepts indicates that the former is the case. In the baseline model (see below) the pairwise sums of random intercepts associated to a given element in either first or second position is not different from zero, suggesting the the model tends to capture opposite effects for the same element in opposite positions. This applies to both nouns ($t(3717) = 0.0855$; $p = .9318$) and adjectives ($t(654) = 0.0842$; $p = .9329$). In Appendix D, we present an alternative analysis of the data with single intercepts for adjectives and nouns.

11. We checked for correlations between variables in each predictor set and there are no collinearity issues.
12. Obviously, the specific fit values we obtained are quite different from the ones previously reported, due to the different statistical model implemented and the different dependent variable (binary vs. aggregated continuous measure) considered.

References

Almuhareb, A., & Poesio, M. (2004). Attribute-based and value-based clustering: An evaluation. In D. Lin & D. Wu (Eds.), *Proceedings of the 2004 Conference on empirical methods in natural language processing* (pp. 158–165). Barcelona, Spain: Association for Computational Linguistics.

Andrews, S., Miller, B., & Rayner, K. (2004). Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap. *European Journal of Cognitive Psychology*, *16*(1–2), 285–311.

Asher, N. (2011). *Lexical meaning in context: A web of words*. Cambridge, UK: Cambridge University Press.

Baayen, R., Feldman, L., & Schreuder, R. (2006). Morphological influences on the recognition of monosyllabic monomorphemic words. *Journal of Memory and Language*, *55*(2), 290–313. doi:10.1016/j.jml.2006.03.008

Baayen, R., Davidson, D., & Bates, D. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Baldwin, T., Bannard, C., Tanaka, T., & Widdows, D. (2003). An empirical model of multiword expression decomposability. In E. Laporte, P. Nakov, C. Ramisch & A. Villavicencio (Eds.), *Proceedings of the workshop on multiword expressions* (pp. 89–96). Sapporo, Japan: Association for Computational Linguistics.

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

Baroni, M., & Zamparelli, R. (2010). Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In H. Li & L. Marquez (Eds.), *Proceedings of the 2010 conference on empirical methods in natural language processing (EMNLP 2010)* (pp. 1183–1193). Boston, MA: Association for Computational Linguistics.

Bertram, R., & Hyönä, J. (2003). The length of a complex word modifies the role of morphological structure: Evidence from eye movements when reading short and long finnish compounds. *Journal of Memory and Language*, *48*(3), 615–634.

Blacoe, W., & Lapata, M. (2012). A comparison of vector-based representations for semantic composition. In J. Tsujii, J. Henderson & M. Payca (Eds.), *Proceedings of the 2012 joint conference on EMNLP and CoNLL* (pp. 546–556). Jeju Island, South Korea: Association for Computational Linguistics.

Blei, D., Ng, A., & Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boleda, G., Vecchi, E. M., Cornudella, M., & McNally, L. (2012). First-order vs. higher-order modification in distributional semantics. In J. Tsujii, J. Henderson & M. Payca (Eds.), *Proceedings of the 2012 joint conference on EMNLP and CoNLL* (pp. 1223–1233). Jeju Island, South Korea: Association for Computational Linguistics.

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

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

Callison-Burch, C., & Dredze, M. (2010). Creating speech and language data with Amazon's Mechanical Turk. In C. Callison-Burch & M. Dredze (Eds.), *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk* (pp. 1–12). Los Angeles, CA: Association for Computational Linguistics.

Chomsky, N. (1957). *Syntactic structures*. Berlin: Mouton.

Coecke, B., Sadrzadeh, M., & Clark, S. (2010). Mathematical foundations for distributed compositional model of meaning. *Linguistic Analysis*, *36*, 345–384.

De Jong, N., Feldman, L., Schreuder, R., Pastizzo, M., & Baayen, R. (2002). The processing and representation of dutch and english compounds: Peripheral morphological and central orthographic effects. *Brain and Language*, *81*(1), 555–567.

Denhiére, G., & Lemaire, B. (2004). A computational model of children's semantic memory. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 297–302). Mahway, NJ: Lawrence Erlbaum.

Devereux, B., & Costello, F. (2006). Modelling the interpretation and interpretation ease of noun-noun compounds using a relation space approach to compound meaning. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Meeting of the Cognitive Science Society* (pp.184–189). Vancouver, Canada: Cognitive Science Society.

Dinu, G., Pham, N. T., & Baroni, M. (2013). General estimation and evaluation of compositional distributional semantic models. In P. Fung & M. Poesio (Eds.), *Proceedings of the ACL 2013 workshop on continuous vector space models and their compositionality* (pp. 50–58). Sofia, Bulgaria: Association for Cognitive Linguistics.

Erk, K. (2007). A simple, similarity-based model for selectional preferences. In A. Zaenen & van den Bosch A. (Eds.), *Proceedings of the association for computational linguistics (ACL 2007)* (pp. 216–223). Prague, Czech Republic: Association for Computational Linguistics.

Erk, K., Padó, S., & Padó, U. (2010). A flexible, corpus-driven model of regular and inverse selectional preferences. *Computational Linguistics*, *36*(4), 723–763.

Evert, S. (2005). The statistics of word cooccurrences. PhD dissertation, Stuttgart University.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Frege, G. (1892). Über sinn und bedeutung. *Zeitschrift fuer Philosophie und philosophische Kritik*, *100*, 25–50.

Frey, B. J., & Dueck, D. (2007). Clustering by passing messages between data points. *Science*, *315*(5814), 972–976.

Gagné, C. (2002). Lexical and relational influences on the processing of novel compounds. *Brain and Language*, *81*(1), 723–735.

Gagné, C., & Shoben, E. (2002). Priming relations in ambiguous noun-noun combinations. *Memory & Cognition*, *30*(4), 637–646.

Gardner, M., Rothkopf, E., Lapan, R., & Lafferty, T. (1987). The word frequency effect in lexical decision: Finding a frequency-based component. *Memory & Cognition*, *15*(1), 24–28.

Golub, G. H., Heath, M., & Wahba, G. (1979). Generalized cross-validation as a method for choosing a good Ridge parameter. *Technometrics*, *21*(2), 215–223.

Gordon, B. (1983). Lexical access and lexical decision: Mechanisms of frequency sensitivity. *Journal of Verbal Learning and Verbal Behavior*, *22*(1), 24–44.

Graves, W. W., Binder, J. R., & Seidenberg, M. S. (2013). Noun-noun combination: Meaningfulness ratings and lexical statistics for 2,160 word pairs. *Behavior Research Methods*, *45*(2), 463–469.

Grefenstette, G. (1994). *Explorations in automatic thesaurus discovery*. Boston, MA: Kluwer.

Grefenstette, E., & Sadrzadeh, M. (2011). Experimental support for a categorical compositional distributional model of meaning. In R. Barzilay & M. Johnson (Eds.), *Proceedings of the 2011 conference on empirical methods in natural language processing* (pp. 1394–1404). Edinburgh, UK: Association for Computational Linguistics.

Griffiths, T., Steyvers, M., & Tenenbaum, J. (2007). Topics in semantic representation. *Psychological Review*, *114*, 211–244.

Guevara, E. (2010). A regression model of adjective-noun compositionality in distributional semantics. In R. Basili & M. Pennacchiotti (Eds.), *Proceedings of the Association for Computational Linguistics GEMS workshop* (pp. 33–37). Uppsala, Sweden: Association for Computational Linguistics.

Harris, Z. S. (1968). *Mathematical structures of language*. New York: Wiley.

Hasher, L., & Zacks, R. (1984). Automatic processing of fundamental information: The case of frequency of occurrence. *American Psychologist*, *39*(12), 1372.

Jaeger, T. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*(4), 434–446.

Jones, M., & Mewhort, D. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*, 1-37.

Juhasz, B., Starr, M., Inhoff, A., & Placke, L. (2003). The effects of morphology on the processing of compound words: Evidence from naming, lexical decisions and eye fixations. *British Journal of Psychology*, *94*(2), 223–244.

Jurafsky, D., & Martin, J. (2008). *Speech and language processing* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Karypis, G. (2003). CLUTO: A clustering toolkit (Tech. Rep. No. 02-017). Minneapolis, MN: University of Minnesota Department of Computer Science.

Katz, G., & Giesbrecht, E. (2006). Automatic identification of non-compositional multi-word expressions using latent semantic analysis. In B. Villada Moiron et al. (Eds.), *Proceedings of the workshop on multiword expressions* (pp. 12–19). Sydney, Australia: Association for Computational Linguistics.

Keller, F., & Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, *29*(3), 459–484.

Kiela, D., & Clark, S. (2013, October). Detecting compositionality of multi-word expressions using nearest neighbours in vector space models. In D. Yarowsky et al. (Eds.), *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1427–1432). Seattle, WA: Association for Computational Linguistics.

Kuperman, V., Schreuder, R., Bertram, R., & Baayen, R. (2009). Reading polymorphemic Dutch compounds: Toward a multiple route model of lexical processing. *Human Perception and Performance*, *35*(3), 876.

Laham, R. D. (2000). Automated content assessment of text using latent semantic analysis to simulate human cognition. Dissertation, University of Colorado at Boulder.

Landauer, T., & Dumais, S. (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.

Lapata, M., McDonald, S., & Keller, F. (1999). Determinants of adjective-noun plausibility. In H. S. Thompson & A. Lascarides (Eds.), *Proceedings of the 9th conference on EACL* (pp. 30–36). Bergen, Norway: Association for Computational Linguistics.

Lapata, M., Keller, F., & McDonald, S. (2001). Evaluating smoothing algorithms against plausibility judgements. In B. Webber et al. (Eds.), *Proceedings of the 39th annual meeting on Association for Computational Linguistics (ACL 2001)* (pp. 354–361). Toulouse, France: Association for Computational Linguistics.

Li, J., Baroni, M., & Dinu, G. (2014). Improving the lexical function composition model with pathwise optimized elastic-net regression. In K. Toutanova & H. Wu (Eds.), *Proceedings of the Association for*

Computational Linguistics (ACL 2014) (pp. 434–442). Gothenburg, Sweden: Association for Computational Linguistics.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods*, *28*, 203–208.

McDonald, S. (2000). Environmental determinants of lexical processing effort. PhD dissertation, University of Edinburgh.

McRae, K., Spivey-Knowlton, M., & Tanenhaus, M. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*, 283–312.

Mitchell, J., & Lapata, M. (2008). Vector-based models of semantic composition. In J. D. Moore et al. (Eds.), *Proceedings of the Association for Computational Linguistics (ACL 2008)* (pp. 236–244). Columbus, OH: Association for Computational Linguistics.

Mitchell, J., & Lapata, M. (2009). Language models based on semantic composition. In P. Koehn & R. Mihalcea (Eds.), *Proceedings of the 2009 conference on empirical methods in natural language processing (EMNLP 2009)* (pp. 430–439). Singapore: Association for Computational Linguistics.

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science*, *34* (8), 1388–1429.

Montague, R. (1974). *Formal philosophy: Selected papers of richard montague*. New Haven, CT: Yale University Press.

Mullaly, A., Gagné, C., Spalding, T., & Marchak, K. (2010). Examining ambiguous adjectives in adjective-noun phrases: Evidence for representation as a shared core-meaning with sense specialization. *The Mental Lexicon*, *5*(1), 87–114.

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Tily, H. (2010). Crowdsourcing and language studies: the new generation of linguistic data. In C. Callison-Burch & M. Dredze (Eds.), *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk* (pp. 122–130). Los Angeles, CA: Association for Computational Linguistics.

Murphy, G. L. (1990). Noun phrase interpretation and conceptual combination. *Journal of Memory and Language*, *29*(3), 259–288.

New, B., Ferrand, L., Pallier, C., & Brysbaert, M. (2006). Reexamining the word length effect in visual word recognition: New evidence from the English lexicon project. *Psychonomic Bulletin & Review*, *13*(1), 45–52.

Nunberg, G., Sag, I., & Wasow, T. (1994). Idioms. *Language*, *70*, 491–538.

Padó, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2), 161–199.

Padó, S., Padó, U., & Erk, K. (2007). Flexible, corpus-based modelling of human plausibility judgements. In J. Eisner (Ed.), *Proceedings of the 2007 joint conference on EMNLP and CoNLL* (pp. 400–409). Prague, Czech Republic: Association for Computational Linguistics.

Partee, B. (2004). *Compositionality in formal semantics*. Malden, MA: Blackwell.

Pollatsek, A., Hyönä, J., & Bertram, R. (2000). The role of morphological constituents in reading Finnish compound words. *Journal of Experimental Psychology: Human Perception and Performance*, *26*(2), 820.

Rapp, R. (2003). Word sense discovery based on sense descriptor dissimilarity. In E. Hovy & E. Macklovitch (Eds.), *Proceedings of the 9th MT summit* (pp. 315–322). New Orleans, LA: Association for Machine Translation.

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, *61*, 127–159.

Rubenstein, H., & Goodenough, J. (1965). Contextual correlates of synonymy. *Communications of the ACM*, *8*(10), 627–633.

Sahlgren, M. (2006). The word-space model. PhD dissertation, Stockholm University.

Schmidt, L., Kemp, C., & Tenenbaum, J. (2006). Nonsense and sensibility: Inferring unseen possibilities. In R. Sun & N. Miyake (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 744–749). Austin, TX: Cognitive Science Society.

Schone, P., & Jurafsky, D. (2001). Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In L. Lee & D. Harman (Eds.), *Proceedings of the 2001 conference on empirical methods in natural language processing (EMNLP 2001)* (pp. 100–108). Pittsburgh, PA: Association for Computational Linguistics.

Schütze, H. (1997). *Ambiguity resolution in natural language learning*. Stanford, CA: CSLI.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In J. Tsujii, J. Henderson & M. Paayc (Eds.), *Proceedings of the 2012 joint conference on EMNLP and CoNLL* (pp. 1201–1211). Edinburgh, UK: Association for Computational Linguistics.

Sommers, F. (1971). Structural ontology. *Philosophia*, *1*(1), 21–42.

Trueswell, J., Tanenhaus, M., & Garnsey, S. (1994). Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, *33*, 285–318.

Turney, P., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

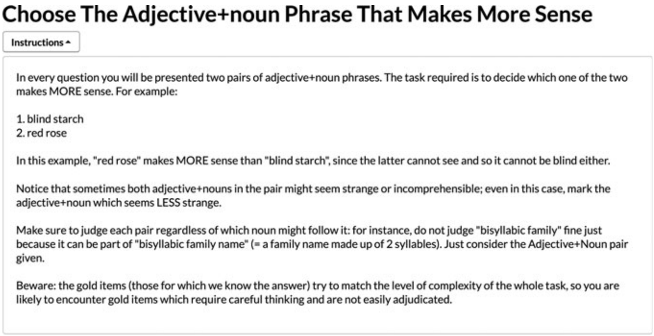
Vecchi, E. M., Baroni, M., & Zamparelli, R. (2011). (Linear) maps of the impossible: Capturing semantic anomalies in distributional space. In C. Biemann & E. Giesbrecht (Eds.), *Proceedings of the ACL 2011 workshop on distributional semantics and compositionality* (pp. 1–9). Portland, OR: Association for Computational Linguistics.

Vecchi, E. M., Zamparelli, R., & Baroni, M. (2013). Studying the recursive behaviour of adjectival modification with compositional distributional semantics. In D. Yarowsky et al. (Eds.), *Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP 2013)* (pp. 141–151). Seattle, WA: Association for Computational Linguistics.

Zanzotto, F., Korkontzelos, I., Falucchi, F., & Manandhar, S. (2010). Estimating linear models for compositional distributional semantics. In C.-R. Huang & D. Jurafsky (Eds.), *Proceedings of COLING* (pp. 1263–1271). Beijing, China: Coling 2010 Organizing Committee.

Appendix A: Evaluation materials

In the figures below, we show the instructions for the CF experiment as presented to the contributors (Fig. A1), as well as an example of the judgment task for a set of AN pairs (Fig. A2).



Choose The Adjective+noun Phrase That Makes More Sense

In every question you will be presented two pairs of adjective+noun phrases. The task required is to decide which one of the two makes MORE sense. For example:

1. blind starch
2. red rose

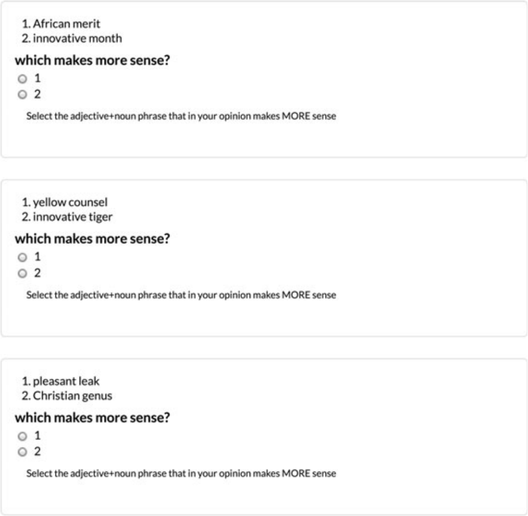
In this example, "red rose" makes MORE sense than "blind starch", since the latter cannot see and so it cannot be blind either.

Notice that sometimes both adjective+nouns in the pair might seem strange or incomprehensible; even in this case, mark the adjective+noun which seems LESS strange.

Make sure to judge each pair regardless of which noun might follow it: for instance, do not judge "bisyllabic family" fine just because it can be part of "bisyllabic family name" (= a family name made up of 2 syllables). Just consider the Adjective+Noun pair given.

Beware: the gold items (those for which we know the answer) try to match the level of complexity of the whole task, so you are likely to encounter gold items which require careful thinking and are not easily adjudicated.

Fig. A1. Screenshot of the instructions presented to the contributors to the CF task.



1. African merit
2. innovative month

which makes more sense?

○ 1
○ 2

Select the adjective+noun phrase that in your opinion makes MORE sense

1. yellow counsel
2. innovative tiger

which makes more sense?

○ 1
○ 2

Select the adjective+noun phrase that in your opinion makes MORE sense

1. pleasant leak
2. Christian genus

which makes more sense?

○ 1
○ 2

Select the adjective+noun phrase that in your opinion makes MORE sense

Fig. A2. Screenshot of a set of AN-AN pairs as presented to the contributors to be judged in the CF task.

Appendix B: Distributional semantic space evaluation

In order to evaluate the distributional semantic space used for our experiments, we validated it on three benchmarks. We first consider the correlation between the distance of noun vectors in the semantic space (described by their cosine distance) and human similarity judgments, based on the RG dataset provided in Rubenstein and Goodenough (1965) and consisting of 65 noun pairs rated by 51 subjects on a 0–4 similarity scale. For example, the nouns *food* and *rooster* resulted in a low similarity rating and should therefore be further from each other in the semantic space than, say, *gem* and *jewel*. Next, we consider a similar evaluation based on the correlation between distance in the semantic space and human similarity ratings of AN phrases, using the ML benchmark presented by Mitchell and Lapata (2010) in which 72 AN phrases were judged on a 1–7 similarity

Table B1
Semantic space quality evaluation

	RG	ML	AAMP
SoA	0.82	0.43	0.76
Full semantic space	0.82	0.42	0.67
Reduced semantic space	0.77	0.42	0.65

Notes. The first row reports the state of the art for each benchmark as reported in Baroni and Lenci (2010), for RG and AAMP, and in Mitchell and Lapata (2010), for ML. The second row reports the performance obtained with our distributional semantic space when no SVD is applied, and the third with SVD reduction to 600 dimensions. Figures of merit are Spearman's ρ for RG and ML and clustering purity for AAMP.

scale. Again, phrases like *national government* and *cold air* obtained low similarity scores from the participants, and thus their AN vectors should have a lower cosine score than the vectors for the phrases *certain circumstance* and *particular case* (since we are not evaluating composition models but the underlying semantic space, we use phrase vectors directly extracted from the corpus). Finally, we consider AAMP, the categorization dataset presented in Almuhareb and Poesio (2004), in which we evaluate the capacity to cluster a set of 402 concepts from WordNet (Fellbaum, 1998), balanced in terms of frequency and ambiguity—such as *acacia*, *ceremony* and *league*—into 21 categories, each selected from one of 21 unique WordNet beginners and represented by between 13 and 21 nouns—such as TREE, OCCASION, and SOCIALUNIT, respectively. Following standard practice for AAMP, we cluster distributional vectors with the CLUTO toolkit (Karypis, 2003), using repeated bisections with global optimization and CLUTO's default settings otherwise.

The results in Table B1 confirm that we are using a high-quality distributional semantic space, with non-tuned performance well within reach of the state of the art for all three benchmarks. Note that for most simulations we used the SVD-reduced space. The table also reports the (slightly better) results obtained without SVD reduction because for the multiplicative model, as discussed in Appendix C, we were constrained to use the latter space.

Appendix C: Composition method parameter tuning and evaluation

We optimized general and model-specific parameters using the unsupervised method suggested by Guevara (2010) and Baroni and Zamparelli (2010): The parameters were chosen to maximize the similarity of the model-generated phrase vectors to the corresponding corpus-extracted phrase vectors (about 3% of the ANs in our semantic space were set apart as tuning data for these purposes, and not used to train the *lf* matrices).

We considered whether to perform composition in full (10K dimensions) or SVD-reduced space (600 dimensions). The *mult* method was only tested on the full semantic space since SVD reduction introduces negative values, which are problematic for pointwise multiplication. The *lf* method was only tested in reduced space for efficiency reasons, and the tuning procedure picked the reduced space for *add* and *dil*. Moreover, based on the tuning results we chose to pre-normalize vectors before composition for all models.

For *add*, we explored 50:50, 40:60, and 30:70 weight ratios between adjective and nouns in either direction and picked 30:70 in favor of the noun.

For *dil* we tune the λ parameter determining the amount of stretching of one vector in the other's direction (see Eq. (5)), and the vector to stretch. We explore λ values {2.2, 4, 6, 8, 10, 12, 14, 16.7, 18, 20} (a range including the values which performed best in Mitchell & Lapata, 2010). We found that stretching the adjective in the direction of the noun by a factor of $\lambda = 20$ yielded the best performance in our parameter tuning experiments.

Table C1
Composition methods quality evaluation

Model	Our Implementation	Mitchell & Lapata
add	0.44	0.37
w.add	0.45	0.44
mult	0.40	0.46
dil	0.41	0.44
lf	0.37	–

Notes. Correlation scores (Spearman's ρ) between cosine distance of model-generated AN vectors and phrase similarity ratings from Mitchell and Lapata (2010), compared to the best reported results for their implementations.

Finally, for the *lf* model we must estimate a weight matrix for each adjective of interest. Following Guevara (2010) and Baroni and Zamparelli (2010), each matrix is estimated by solving the unsupervised least-squared error problem of approximating a set of corpus-extracted AN phrase vectors by a linear combination of the dimensions of the corresponding noun vectors (the number of N-AN vector pairs used for training ranged from 100 to over 1K items across the 663 adjectives). We set the matrix weights using Ridge regression with generalized cross-validation to automatically choose the optimal regularization parameter (Golub, Heath, & Wahba, 1979).

As a quality control, we verified that the tuned composition models obtained results comparable to those of Mitchell and Lapata (2010) for their AN phrase similarity benchmark (see Appendix B). This is confirmed by the results in Table C1.

Appendix D: Modeling proportional AN preference

As pointed out by the editor, one issue with the mixed-models statistical analysis reported in the article is that, due to constraints in the employed statistical package, we must artificially assume different intercept distributions for nouns and adjectives depending on whether they occur in the first or second AN of the tested pairs. We consider here an alternative approach in which we model the *proportion* of times an AN was chosen over all the times it occurred in a pair. In this way, each of the 26,137 distinct ANs was associated to a single value in the dependent variable, and we could enter in the analysis non-duplicated random intercepts for adjectives and nouns. In turn, a drawback of the analysis presented here is that it ignores the fact that the proportions of times different ANs are picked are not independent (every trial in which an AN is picked also counts as a trial in which another AN is discarded).

The alternative analysis takes the form of a set of multiple linear regressions, where, like in the main analysis, we add in turn each semantic measure to the set of baseline word- and language-model-based measures. Results are presented in Table D1 and their pattern is by and large consistent with those reported in the main article (Section 5.2 and Table 2).[12] In particular, we observe again that, with the exception of density, all

Table D1
Performance of the semantic measures in predicting the proportion of times an AN was chosen

Measure		*df*	AIC	logLik	Chisq	*p*
BASELINE		9	−6009.0	3013.5		
THEMEFIT	*adjective*	10	−6203.5	3111.8	196.51	.0001
	noun	10	−6355.1	3187.5	348.05	.0001
VLENGTH	*add*	10	−6507.2	3263.6	500.18	.0001
	mult	10	−6394.0	3207.0	387.00	.0001
	dil	10	−6296.4	3158.2	289.41	.0001
	lf	10	−6240.2	3130.1	233.14	.0001
COSINE	*add*	10	−6507.2	3263.6	500.18	.0001
	mult	10	−6454.8	3237.4	447.77	.0001
	dil	10	−6966.0	3493.0	959.00	.0001
	lf	10	−6738.1	3379.0	731.06	.0001
DENSITY	*add*	10	−6007.1	3013.6	0.0993	.7526
	mult	10	−6008.3	3014.2	1.3158	.2514
	dil	10	−6177.0	3098.5	169.98	.0001
	lf	10	−6009.2	3014.6	2.1479	.1428

Notes. Results for regression mixed effects models including both baseline and semantic measures as opposed to baseline measures only. Goodness-of-fit is expressed in terms of AIC and log-likelihood; *p* values based on chi-square tests. Measures considered: thematic fit (THEMEFIT) for both *adjective* and *noun*, length of the composed phrase vector (VLENGTH), similarity between the noun component and the phrase vector (COSINE), density of the neighborhood of the phrase vector (DENSITY). Compositional models considered: additive (*add*), multiplicative (*mult*), dilation (*dil*), lexical function (*lf*). *df*, degrees of freedom.

measures based on distributional semantics lead to significant improvements over the baseline. Both the length and cosine-to-head-noun cues are very effective, the second being particularly consistent across composition methods. In this analysis, the problems with density emerge more clearly. This cue only reaches significance when combined with dilation, and even in that case the effect is in the opposite direction than expected, just like in the main analysis.