Negated Adjectives and Antonyms in Distributional Semantics: *not similar*?

Laura Aina* **
Universitat Pompeu Fabra, Spain

Raffaella Bernardi †
University of Trento, Italy

Raquel Fernández ‡
University of Amsterdam,
The Netherlands

We investigate the relation between negated adjectives and antonyms pairs in English (e.g., not cold *vs.* hot - cold*) using Distributional Semantics. We build vector representations of a set of antonyms and their negations on the basis of their contexts of use, and compare the similarities of the negated adjectives to each of the adjective in their antonym pair. We find that in a distributional semantic model a negated adjective (e.g.,* not cold*) is typically more similar to the adjective itself (*cold*) than to its antonym (*hot*). The effect is less strong for antonyms that share their lexical root (morphological; e.g.,* happy - unhappy*). No difference is observed between simple and double negations (e.g.,* not happy, not unhappy*), and contrary and contradictory antonyms (e.g.,* hot - cold, dead - alive*). Our results provides insights on negated adjectives, and in general the type of similarity captured by Distributional Semantics.*

1. Introduction

Negation has long posed challenges to researchers in Theoretical and Computational Linguistics (see Horn (1989) and Morante and Sporleder (2012) for overviews). Similarly to logical negation ($\neg p$ is true $\leftrightarrow$ p is false), natural language negation allows to express the falsity of a content. However, when interacting with morphosyntax, semantics and pragmatics, it exhibits a diversity of functions and forms, which go beyond the simplicity of the logical connective (Horn and Kato 2000). We here focus on the **negation of adjectives in English**, as expressed by the particle *not* modifying an adjective – e.g., *not cold*. In particular, we study the relation between these expressions and the antonym pair constituted by the adjective that is negated and its opposite (e.g., *not cold* vs. *cold-hot*). We carry out our investigation using the methods of **Distributional Semantics**, hence representing expressions on the basis of their use in a corpus (Lenci 2008).

It has been noted that when an adjective is negated, addressees not only conclude that the property denoted by it does not apply, but also tend to infer that an alternative property from the same domain applies (e.g., *not cold* $\approx$ *lukewarm*, *hot*, etc.; Horn (1989), Fraenkel and Schul (2008)). For instance, one can take the negation of an adjective to directly convey the opposite of the adjective, that is its antonym (e.g., *not cold* =

* Department of Translation and Language Sciences; Email: laura.aina@upf.edu.
** Part of the work presented in this paper was carried out while at the University of Amsterdam.
† CiMeC, Center for Mind/Brain Sciences and DISI, Information Engineering and Computer Science; Email: raffaella.bernardi@unitn.it
‡ ILLC, Institute for Logic, Language and Computation; Email: raquel.fernandez@uva.nl

hot). However, when a speaker opts for a complex expression when a simpler one is available – choosing to use a negation instead of an antonym – this is typically to serve a particular purpose (Horn 1984). For instance, a range of studies support what is known as *mitigation hypothesis* (see Jesperseen (1965) and Horn (Horn 1972) for an early formulation, and Giora (2006) for an overview): a negated adjective tends to be understood as conveying an intermediate meaning between the adjective and its antonym (e.g., *not large* $\approx$ *medium-sized*).

In this work, we study antonymic adjectives and their negations in a distributional semantic model. To this end, we employ an existing dataset of antonyms, whose annotation we further extend; we then obtain distributional representations of these and their negated version. This allows us to conduct a data-driven study of negation and antonymy that covers a large set of instances. We compare pairs of antonyms with distinct lexical roots and those derived by affixation, i.e., **lexical and morphological antonyms** (e.g., *small - large* and *happy - unhappy*, respectively; Joshi (2012)). Moreover, we investigate the distinction between lexical antonyms that are **contrary or contradictory**, that is, those that do or do not allow an available intermediate value (Fraenkel and Schul 2008): e.g., something *not cold* is not necessarily *hot* - it could be *lukewarm* - but something *not present* is *absent*. As for negations of morphological antonyms, we compare instances of **simple and double negation**, where the latter occurs if the antonym that is negated is an affixal negation (e.g., *not unhappy*).

Our analyses show that, when considering distributional information, a negated adjective (e.g., *not cold*) is typically more similar to the adjective itself (*cold*) than to its antonym (*hot*). Such effect is less strong for antonyms derived by affixation and which then share the same lexical root (e.g., *happy - unhappy*). This suggests that there is a tendency for expressions sharing a lexeme to appear in similar contexts. No difference is observed between simple and double negations (e.g., *not happy*, *not unhappy*), and contrary and contradictory antonyms (e.g., *hot - cold*, *dead - alive*). The latter result contrasts with previous experimental evidence showing that inferential relations between expressions (e.g., *not dead* entails *alive*) affects how similar the negation is perceived to the antonym (Fraenkel and Schul 2008). We hypothesize that this is because distributional similarity is not tailored to capture inferential relations but rather differences in use between expressions. Therefore, even when these seem logically equivalent, they may still emerge as different to a distributional semantic model. Our results provide novel insights on negated adjectives, and in general the type of similarity captured by Distributional Semantics.

2. Related Work: Negation in Distributional Semantics

In Distributional Semantics (henceforth, DS), the distribution of an expression across context in a corpus is taken to be representative of its content, and summarized into a vectorial representation (Lenci 2008; Erk 2012; Turney and Pantel 2010; Baroni, Dinu, and Kruszewski 2014). A vast body of research in Computational Linguistics and Cognitive Science has shown that this methodology is very successful at modeling the meaning of content words (e.g., adjectives). However, these data-driven and bottom-up techniques are not as successful when modeling function words and the complex phenomena that they involve – for instance *not* and in general negation (Bernardi 2014; Boleda and Herbelot 2016). In the case of negation, this is primarily due to the fact that it acts as a truth-reversing operator, whereas in the first place DS lacks a built-in method to account for truth conditions. For this reason, some methods, like that by Garrette et al. (2014), consist of hybrid approaches between distributional and formal semantics,

while others aimed to find a counterpart of logical operations in the distributional space (Widdows and Peters 2003; Coecke, Sadrzadeh, and Clark 2010).

Some approaches have been proposed to specifically model or evaluate the negation of adjectives. Hermann et al. (2013) design a framework where domain and value features of an adjective are separately represented, and when *not* is applied to an adjective, the resulting phrase remains close to others from the same domain of the adjective but its value within the domain changes. Similarly, Rimell et al. (2017) introduce a neural network architecture to learn a mapping from an adjective to the negated version conditioned on the domain of the former. To train this model, they learn negation as a mapping from an adjective to its antonym; a similar idea was also employed by The Pham et al. (2015). As these approaches rest on the assumption that antonyms and negations are equivalent, they do not take into account mitigation effects, nor in general peculiarities of negation that makes it differ from antonymy. However, Socher et al. (2012, 2013) showed that a neural network learning general compositional operations as a byproduct of a sentiment analysis task can actually capture such effects, and correctly taking them into account when assigning fine-grained labels.

Finally, we mention an approach to negation which focuses on noun phrases, but is nevertheless very relevant to our study. Kruszewski et al. (2017) proposed to use similarity relations between expressions as captured by DS to account for the alternatives triggered by the use of a negation. They show that these provide an excellent fit to data of alternative plausibility ratings. For example, in the sentence *This is not a dog, it is a cat* the distributional similarity between *dog* and *cat* is used to measure how plausible the latter is as an alternative to the former. Crucially, this approach showed that, in spite of the difficulties in modeling truth-related aspects of negation, DS can still provide valid contributions to its study. We build on this research line and employ DS as an investigative tool to study the relation between negation and antonymy.

3. Motivation and Data

We are interested in how negation acts with respect to pairs of adjectives connected by the lexical relation of **antonymy** (Murphy 2003), that is, associated with opposite properties within the same domain (e.g., *hot - cold* with respect to temperature). In particular, we want to compare the negation of one of the antonymic adjectives with itself and its antonym respectively (e.g., *not cold* vs. *cold* and vs. *hot*). Our data of interest are then triples obtained starting from an antonymic pair and negating one of the two items, as in Ex. (1). In the following, we present how we construct such a dataset. Since we are interested in analyzing different types of antonyms and negations, we apply a classification to these data (Figure 1), which we describe and motivate.

Example 1
(a) ⟨ *hot*, *cold*, *not* {*hot*||*cold*}⟩
(b) ⟨*happy*, *unhappy*, *not* {*happy*||*unhappy*}⟩

3.1 Dataset

In order to build a dataset of triples as in Ex. 1, it is sufficient for us to have a dataset of antonymic pairs of adjectives. Indeed, we can automatically obtain the negation of each adjective by simply making it precede by *not*. Note that for each pair we can obtain two triples, by negating each of the adjectives in turn.

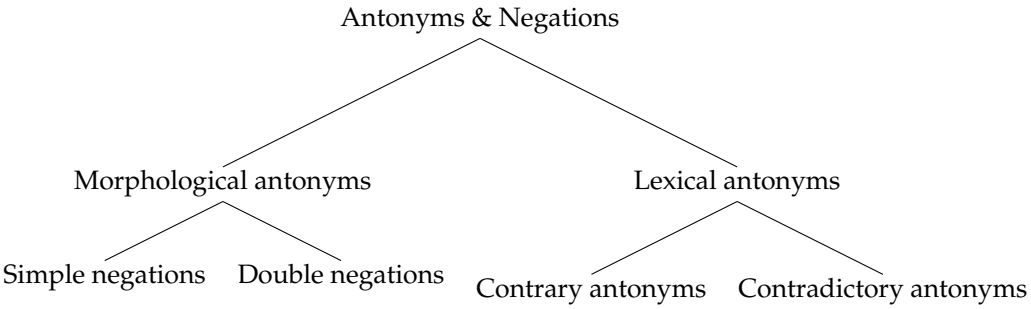


Figure 1
Categorisation of the triples consisting of an antonymic pair and a negated adjective employed in our experiments.

As a dataset of antonyms, we make use of a subset of the **Lexical Negation Dictionary** by Van Son et al. (2016). This consists of word pairs tagged as antonyms in WordNet (Fellbaum 1998) classified into different types of lexical negation, following the categorization of Joshi (2012). In particular, lexical negation is taken to include both *affixal negations* (e.g., *perfect - imperfect*), and *regular antonyms* (e.g., *hot - cold*). The former is in turn split into *direct* and *indirect* negation, depending on whether the meaning of the affixed word actually expresses an opposite property of the non-affixed counterpart (e.g., *imperfect* expresses a different degree than *perfect* in the scale of perfection, while the difference between *famous - infamous* is not about the degree of fame and they are not incompatible with each other). Therefore, we consider as antonymic adjectives only those pairs in the Lexical Negation dictionary that either involve a direct affixal negation on one side, and a relation of regular antonymy on the other. We refer to these as **lexical** antonyms, i.e., with distinct lexical roots (e.g., *cold - hot*), or **morphological** antonyms, i.e., sharing the lexical root (e.g., *happy - unhappy*). The motivation to use this dataset is that it leverages a large-scale resource such as WordNet, but also allows us to filter out pairs that do not actually correspond to the relation of antonymy in the sense of direct opposition. Finally, as it comes equipped with the classification between lexical and morphological antonyms, it enables us to investigate differences between these two groups (see following section).

From this list of antonyms, we build our dataset of triples as explained above. Our analyses methods leverage on occurrences of expressions in a corpus, which we use to obtain their distributional representations (see details in Section 4). To ensure that we only consider representations based on a relatively high number of occurrences, we enforce a frequency threshold: to be employed in the analyses, each of the elements of the triple (the two adjectives and the negation) needs to occur at least 100 times in the corpus. Table 1 shows the final number of triples for each class and the average frequency of their elements after this filtering. As it can be expected, negated adjectives are overall less frequent than adjectives.[1]

1 Full list of triples at `https://lauraina.github.io/data/notadj.csv`

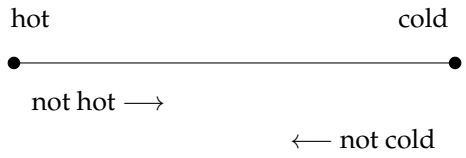


Figure 2
Interaction of negation with an antonym pair: negation shifts the meaning of an adjective towards its antonym.

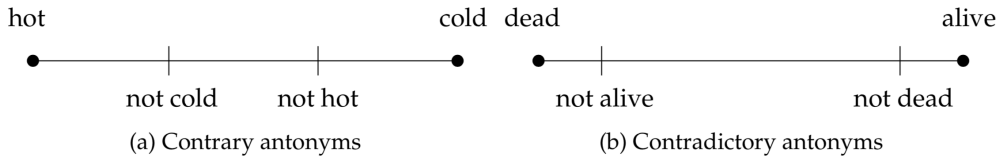
3.2 Antonyms: Lexical vs. Morphological

We are interested in investigating the relationship between negation by *not* and antonymy. A non-parsimonious expression – e.g., a negated adjective – tends to trigger the implicature that a different meaning from a simpler alternative – an antonym – is intended (Grice 1975; Horn 1984). For instance, it has been shown that one of the functions of negation is to act as a modifier of degree (Giora et al. 2005): it alters the meaning of the adjective it applies to and shifts it more or less close to its antonym (Figure 2). Such mitigation in meaning has been explained in pragmatic terms, but also as a result of the representational process: it arises due to the interaction between the negativity of the particle *not* and the meaning of the adjective, which is retained as accessible in memory (Giora et al. 2005).

We aim to investigate the semantic shift that results from applying negation to an adjective, and the extent that this makes it closer to the antonym. We analyze this using the measures of similarity yielded by a distributional semantic model – e.g., is *not hot* closer to *cold* than to *hot*? Differently from previous studies, we here compare negated adjectives and antonyms solely on the basis of their contexts of use. Moreover, we compare the behavior of lexical and morphological antonyms in this respect. As we mentioned earlier, these two classes of antonyms are usually taken to express the same lexical relation, namely opposition - and to be different only on morphological terms. However, such difference might affect their relation with negated adjectives: indeed, affixal negations have a morphological structure that resembles negated adjectives (e.g., *un-happy* vs. *not happy*). It may then be that antonyms with a negative affix are more similar to the negated adjective than antonyms with a distinct lexical root – e.g., is *not frequent* closer to *infrequent* than *not hot* to *cold*? To investigate this, we compare triples derived from lexical and morphological antonyms.

3.3 Lexical Antonyms: Contrary vs. Contradictory

An important distinction within the lexical antonyms group is that between contradictory and contrary pairs (Clark 1974). In the case of contradictory antonyms, the negation of one of the adjectives entails the truth of the other, without the availability of a mid-value (e.g., *not dead* implies *alive*). The opposite is true for contrary pairs (e.g., *not hot* does not imply *cold*, since a mid-value, such as *lukewarm*, exists). In other terms, contradictory pairs constitute a dichotomy, while contrary ones lie in a continuum. Fraenkel and Shul (2008) provided psycholinguistic results showing that if an adjective is part of a contradictory pair, its negation is interpreted as being closer to the related antonym than if it is instead part of a contrary pair (e.g., *not dead* is interpreted as being closer to *alive* than *not small* to *large*). This corresponds to the intuition that when



(a) Contrary antonyms (b) Contradictory antonyms

Figure 3
Example of the effects observed by Fraenkel and Shul (2008): a negated adjective that is part of a contradictory pair typically expresses a meaning that is closer to the adjective's antonym than a negated adjective that is part of a contrary pair.

negation shifts the meaning of the adjective towards its opposite, such shift is bigger if a mid-value between these is not available (see Figure 3). However, Fraenkel and Shul (2008) also noted that, in spite of this general result, some variation may occur: even for contradictory pairs it may be possible to conceive a context where a mid-value interpretation is available (e.g., *not dead* $\approx$ *half-dead*; Paradis and Willners (2006)).

We replicate the analysis of Fraenkel and Shul (2008) with DS, where similarities between expressions can be quantified in terms of the geometric distance between their distributional representations (details in Section 4). The antonyms pairs in the Lexical Negation Dictionary do not come with a classification into contrary and contradictory; in general, to the best of our knowledge, no large-scale dataset annotated with this information is available. Therefore, the three authors independently annotated the lexical antonym pairs extracted by the Lexical Negation Dictionary as contrary or contradictory, following the definitions reported by Fraenkel and Shul (2008).[2] In particular, we tag as contrary those pairs a, b such that it is acceptable to say that something is "neither *a* nor *b*", and viceversa for contradictory ("neither *hot* nor *cold*" vs. "neither *dead* nor *alive*"). Antonyms could also be tagged as *unclear*, if none of these options clearly fit.[3] Not surprisingly, the inter-annotator agreement is only moderate (Fleiss' $k = 0.37$). As we mentioned above, the distinction into contrary and contradictory antonyms presents some limitations: these are likely to cause difficulties when scaling the classification up to a large set of antonyms and attempting to reach agreement over it. Therefore, our low agreement in the annotation further underscores the possibility that the availability of an alternative between two antonyms may be a contextual matter, and that the contrary vs. contradictory distinction may rather be a graded one. We leave this aspect to be clarified by future research and, for the purpose of our analysis, only consider antonyms pairs classified with full agreement. Table 1 reports the values after such filtering: as it can be seen, we had to exclude almost 50% of the lexical antonyms triples. In particular, this filtering leaves us with a small number of triples involving contradictory antonyms.

3.4 Morphological Antonyms: Simple vs. Double Negations

In the case of morphological antonyms, one of the two adjectives is an affixal negation, and hence already contains a negating prefix (such as *im-* in *imperfect* or *un-* in *unhappy*):

2 We only annotated lexical antonyms pairs that we could analyze in our setup. That is, we first filtered all triples on the basis of their frequency of the corpus employed (see Section 4) and then annotated the lexical antonyms in these data.

3 Annotation guidelines at `https://lauraina.github.io/data/notadj.pdf`.

Table 1
Total number of triples $\langle a_1, a_2, not\ \{a_1 \| a_2\}\rangle$ used in the experiment, after filtering by frequency and annotation agreement , and average frequency of adjectives and negated adjectives in these triples per class.

	# triples	frequency	
		adj.	*not* adj.
Lexical antonyms	**198**	254K	1K
– contrary	68	337K	1K
– contradictory	28	298K	1K
Morphological antonyms	**185**	83K	1.8K
– simple negations	157	85K	2K
– double negations	28	122K	0.9K

adding *not* thus gives rise to a double negation (e.g., *not imperfect*; see Figure 2). Since our dataset of triples is obtained by negating both of the adjectives in the set of antonymic pairs, datapoints involving morphological antonyms encompass both simple and double negations (see Ex. 1). Double negations have been widely studied in the literature due to their difference with double negation in logic (e.g., Bolinger (1972), Krifka (2007), Tessler and Franke (2018)). While in logic two negations cancel each other out ($\neg\neg p \equiv p$), in natural language double negations are typically employed to weaken the meaning of the adjective that is negated twice (e.g., *not unhappy* $\neq$ *happy*). We here test whether evidence for this effect is found in a distributional semantic model: in particular, if two negations were equivalent to no negation at all we would expect that the negation of an affixal negation (e.g., *not unhappy*) is particularly close to the antonym (e.g., *happy*). Therefore, we check whether simple (e.g., *not happy*) and double (e.g., *not unhappy*) negations exhibit similar trends in relation to an antonym pair (*happy* vs. *unhappy*).

We classify our morphological antonyms data into simple and double negations, as follows. From the Lexical Negation Dictionary, we know which adjective among an antonym pair is the affixal negation; if a triple includes a negation of an affixal negation (e.g., *not imperfect*), it is classified into the double negations groups, and viceversa for simple negations. As Table 1 shows, we could only consider a small number of double negations; this is due to the fact that these expressions are rarely produced, and therefore few occur in our corpus more than the frequency threshold.

4. Methods

Previous studies about negation described its effect on an adjective as a meaning shift towards the antonym, that can be measured in terms of **semantic similarity** (Fraenkel and Schul 2008). DS offers us a data-driven method of quantifying this, leveraging the occurrences of antonyms and negations in a corpus. Within this framework, expressions - here, adjectives and their negated versions - are represented as vectors of continuous values, summarizing their distribution across contexts of use in a corpus. Crucially, we can interpret the proximity relations of the expressions in the resulting high-dimensional space (e.g., cosine between their vectors) in terms of similarity relations between them. By definition, this measures similarity of distribution: the more two expressions appear in the same contexts, the more they will be distributionally similar.

While representing words in this fashion is standard for adjectives and in general word units, it is more atypical as a way to represent multi-word phrases, such as a negations. These are more typically represented using compositional operations (Baroni

2013). In the case of negation, these often incorporate assumptions about its resulting behavior (e.g., Hermann et al. (2013), Rimell et al. (2017)). To avoid introducing any bias, we simply represent negated adjectives by treating them as a single unit (see details below). This approach allows us to study negation in a bottom-up fashion, moreover covering a large set of instances. An important caveat is in place: as in standard Distributional Semantics, each expression is assigned one vector abstracting away from all their uses. Therefore, we focus on the main regularities in the use of adjectives and negated adjectives – as captured in their distributional vectors – leaving an investigation of their context-sensitive behavior to future work.

4.1 Distributional Semantic Model

To build a distributional semantic model with negated adjectives, we employ an existing algorithm but apply a particular pre-processing of the training corpus. In particular, we want the vocabulary of our model to include, besides word units, also negated adjectives. We pre-process the training corpus as follows: adjacent occurrences of the particle *not* and an adjective are merged (e.g., *not cold* $\rightsquigarrow$ *not_cold*), therefore treating each negated adjective as a single type, independent of the related adjective. Moreover, we employ part of speech labels for adjectives, to distinguish occurrences of a form as a different part of speech (e.g., *poor* as adjective or noun, as in *the poor*). Finally, we remove function words, as to avoid syntactic differences between adjectives and negated adjectives (e.g., the former ones appear both in predicative and attributive position, while the latter ones almost exclusively in attributive position).

The corpus we employ is the concatenation of UkWaC and Wackypedia-En corpora (2.7B tokens; Baroni et al., (2009)). We train a word2vec CBOW model (Mikolov et al. 2013) on this corpus, setting its hyperparameters as in the best performing model by Baroni et al. (2014).[4] We are interested in investigating characteristics of antonyms and negated adjectives in a distributional semantic model that is not fine-tuned to a particular task and where no assumptions about the structure of its space are incorporated. Therefore, we do not carry out any hyperparameters search, nor we employ any ad hoc techniques aimed at, for example, amplifying the distances between antonyms in the semantic space (such as those by Nguyen et al. (2016) or The Pham et al. (2015)). We assess the quality of the induced model through a similarity relatedness task, where we find that it achieves satisfying performances.[5]

4.2 Quantitative Analysis

We consider triples as in Ex. 1, derived as described in Section 3. Given a triple $\langle a_i, a_j, not\ a_i \rangle$ (e.g., *cold*, *hot*, *not cold*), we define the following score:

$$\textit{Shift} := \textit{Sim}(\text{not } a_i, a_j) - \textit{Sim}(\text{not } a_i, a_i) \quad (1)$$

4 Vectors size: 400; window size: 5; minimum frequency: 20; sample: 0.005; negative samples: 1. We employ the Gensim implementation of CBOW from `https://radimrehurek.com/gensim/models/word2vec`

5 Spearman's ρ of 0.75 on the MEN dataset (Bruni, Tran, and Baroni 2014); see results by Baroni et al. (2009) for a comparison.

Table 2
Average *Shift* scores, with standard deviation, for each category. ***: significant difference between categories in the row ($p < 0.001$, Welch's t-test).

Lexical antonyms	$-.19$ ($\pm.16$)	Morphological antonyms	$-.04$ ($\pm.16$)	***
Contrary antonyms	$-.18$ ($\pm.15$)	Contradictory antonyms	$-.19$ ($\pm.16$)	
Simple negations	$-.03$ ($\pm.17$)	Double negations	$-.06$ ($\pm.11$)	

where $i \neq j$, and *Sim*(not a_i, a_j) and *Sim*(not a_i, a_i) are the cosine similarities of the negated adjective with the antonym and the adjective, respectively. *Shift* measures how much closer a negated adjective is to the antonym than to the adjective it self (i.e., how much closer *not cold* is to *hot* than to *cold*), and hence how much negation shifts the meaning of an adjective towards that of the antonym. When the *Shift* score positive, the negated adjective is closer to the antonym, and viceversa. Due to the well-known tendency of antonym pairs to be close in a distributional space (Mohammad et al. 2013), the absolute value of *Shift* is not expected to be high: since antonyms tend to be close to each other, a vector that is close to one is also likely to be close to the other. However, this is practically not a problem, as by comparing the similarities with the two antonyms, respectively, we can still assess whether a higher proximity is registered towards one of them. For instance, Kim and De Marneffe (2013) have shown that, in spite of the relative proximity of two antonyms, meaningful relationships among other members of their domain can still be captured: they were able to retrieve adjectival scales by looking at intermediate points between the two antonyms' vectors.

5. Results

Table 2 shows the scores across the different categories described in Section 3. Example triples for each category are given in Table 3, together with the nearest adjectives of each element in the triple. Figure 4 offers a visualization of the results for the different categories.[6]

5.1 Lexical vs. Morphological Antonyms

The average *Shift* scores of both classes are negative, showing that a negated adjective is typically closer to the adjective than to the antonym. Indeed, as shown in Table 3, the nearest neighbor of a negated adjective is often the related adjective.[7] At first glance, one could interpret this result as supporting the idea that negated adjectives express an intermediate meaning between that of the adjective and the antonym (e.g., *not small* is close to *normal-sized*). More in general, considering the setup of our experiments, it shows that negated adjectives have a profile of use that is more similar to that of the adjective than to the antonym.

6 The Figure serves as a visualization of the results for each category, and not for the particular triple that is reported as example. That is, we locate the elements of the example triples on the basis of the average *Shift* score of their category, and not the score of the specific triple.

7 Note that we treated negated adjectives and adjectives as completely separate types. In particular, the occurrence of a negated adjective does not count also as an occurrence of an adjective. Therefore, the result cannot be led back to introducing an overlap in distribution.

Table 3
Nearest adjectives in the semantic space for the three elements in some sample triples.

Contrary antonyms	**small**: *large, tiny, smallish, sizeable, largish*	**large**: *small, sizeable, huge, vast, smallish*	**not small**: *small, smallish, normal-sized, largish, middle-sized*
Contradictory antonyms	**dead**: *drowned, lifeless, half-dead, wounded, alive*	**alive**: *dead, awake, unharmed, beloved, tortured*	**not dead**: *dead, half-dead, alive, comatose, lifeless*
Simple negations	**similar**: *analogous, identical, comparable, dissimilar, same*	**dissimilar**: *similar, different, distinct, unrelated, identical*	**not similar**: *similar, dissimilar, identical, distinguishable, analogous*
Double negations	**happy**: *glad, pleased, contented, nice, kind*	**unhappy**: *disappointed, dissatisfied, unsatisfied, resentful, anxious*	**not unhappy**: *unhappy, adamant, disappointed, dismayed, unimpressed*

The two classes of antonyms differ significantly in the extent of this effect: the Shift score is higher for morphological antonyms than for lexical ones. We found that this is due to the fact that negated adjectives and morphological antonyms are significantly closer to each other than it is the case for lexical antonyms, whereas there is not a significant difference between how close the negation is to the adjective.[8] The higher similarity between negations and morphological antonyms can be justified by the fact that one of the two morphological antonyms is an affixal negation. Its structure would then be much more similar to that of negated adjectives than it is the case for lexical antonyms: both affixal negations and negations by *not* are formed by a negative particle and the adjective itself (e.g., *not* vs. *un-, im- etc.* + adjective). This might impact on the similarity between, i.e., *not happy* and *unhappy*, as well as that between *not unhappy* and *happy* (see Section 5.3 for a comparison between simple and double negations).

Overall, the picture that emerges is one where sharing a lexeme – in particular, as a separate word – impacts on the distributional similarity between expressions. Indeed, negated adjectives tend to be more similar to the original adjective – which they specifically include as a separate word – than to the antonym (e.g., *not cold - cold* vs. *hot*). However, if also the antonym shares the lexical root – which is the case for morphological antonyms (e.g., *not imperfect – perfect*), this effect is less strong, as the negated adjective is also quite close to the antonym. One way to explain this is by positing that the contexts of use of an expression may be affected by a lexeme that they include, due to the connotations that this carries. For instance, in a context where too-direct expressions are to be avoided for politeness, expressions like *incorrect* or *not correct* – sharing *correct* – may be preferred to *wrong*.

5.2 Contrary vs. Contradictory Antonyms

In contrast to results from the linguistic literature (see Section 3), the behavior of contrary and contradictory antonym pairs is not significantly different in our analysis. When we look into a distributional space, even for contradictory antonyms, the negated

8 The average Sim(not a_i, a_j) is 0.43 and 0.16 for morphological and lexical antonyms, respectively

adjectives tend to be more similar to the adjective itself than to the antonym (e.g., *not dead* is closer to *dead* than to *alive*).

While this result may seem counterintuitive at first, we posit the following explanation. The contrary vs. contradictory distinction taps into inferential relations between expressions - e.g., *not present* implies *absent*. These constraint the potential readings of negation (e.g., an intermediate meaning is not available, so it cannot be intended) and affect speakers when prompted to judge the similarity between negations and antonyms (Fraenkel and Schul 2008). However, Distributional Semantics organizes its lexicon in terms of general relatedness relations, with no built-in method to capture inferential relations between expressions (Bernardi 2014; Boleda and Herbelot 2016). Yet, this does not mean that distributional similarity is not informative regarding negation. It indeed captures a different and possibly complementary type of similarity from that tackled, for example, in the experiments of Fraenkel and Shul (2008). Even if the negation of an adjective and the antonym may seem intuitively equivalent, the use of one or the other may serve different functions (e.g., contradicting an expectation, politeness, emphasis etc.) leading them to appear in different contexts. Therefore, the negation of an adjective from a contradictory pair may not be so similar to its antonym when looking at their distribution. One could argue that this effect is a shortcoming of DS; on the contrary, we regard it be an interesting property that naturally arises as an artefact of representing expressions solely on the basis of their contexts of use.

Finally, we find that, since continuous representations are able to capture nuanced differences, the alleged dichotomy between contrary and contradictory antonyms tends to become a continuum in the distributional space. For example, one of the closest adjectives to *not dead* is *half-dead*: this suggests a gradability of the *dead-alive* scale, in spite of the pair being categorized as contradictory. This further underscores the complexity of the contrary and contradictory distinction which we had already encountered in the annotation procedure.

5.3 Simple vs. Double Negations

There is not a significant difference between negated adjectives that are instances of simple and double negations: crucially, it is not the case that double negations are very close to the antonym as a result of the two negations canceling each other out (e.g., *not unhappy* is closer to *unhappy* than to *happy*). The results could be interpreted in terms of mitigation: for instance, *not unhappy* is close to *unimpressed*, a mid-value between *happy* and *unhappy* (Table 3). More in general, following an analogous rationale to what noted earlier, it suggests that the contexts of use of double negations are more similar to the ones of the adjective that is negated than to those of its antonym. Indeed, double negations typically appear in contexts where the use of the "logically" equivalent alternative (i.e., the antonym) is to be avoided for pragmatic reasons, as possibly too strong or direct (e.g., *not unproblematic* vs. *problematic*; Horn, (1984)).

6. Discussion and Conclusion

We have investigated negated adjectives using the tools of DS, which allows us to quantify the similarities between expressions on the basis of how they are used. Our analyses show that, when considering contexts of occurrence, negating an adjective does not make it closer to the antonym than it is to the adjective itself. We hypothesized that this effect may partially be due to mitigation effects (Giora et al. 2005), but more

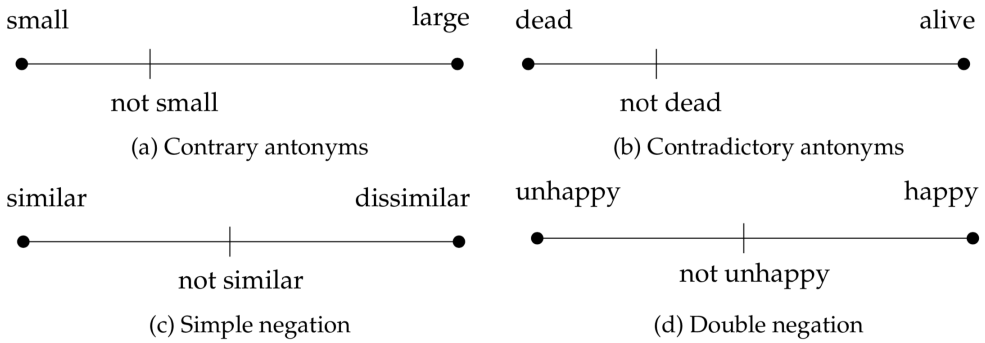


Figure 4
Visualization of the results; the negation is located as distant from each antonym following the average *Shift* score in Table 2 (e.g., center of the scale: *Shift* = 0; closer to the adjective: *Shift* < 0).

in general due to the different functions that these expressions are used in, which ultimately reflect their contexts of use. This follows from the type of methodology which we employ in our study, namely the distributional one.

Our results align with observations in the linguistic literature. Language has been noted to exhibit a force to diversification of meanings, such that full synonymic expressions tend to be avoided (Kiparsky 1982; Clark 1992). In particular, according to the *division of pragmatic labor* posited by Horn (1984), if a speaker opts for a more complex or less fully lexicalized expression over a simpler alternative (e.g, a negated adjective over an antonym) this is justified by a particular function. This could be for example that of expressing an intermediate meaning, but also retaining the emphasis on a rejected concept, or attenuating the strength of a statement. Our results suggest that distributional representations may be sensitive to such differences in use. In particular, we found that sharing a lexical root affects how similar the distributions are: this suggests that certain lexemes are associated to particular contexts. Moreover, we could not conclude that relations of distributional similarity align with inferential relations between expressions, that is *not dead* is not closer to *alive* than to *dead*. This suggests that: 1) again, the lexical root has an impact on the contexts of occurrence; 2) alternative expressions, such as antonyms or negated adjectives, are not fully interchangeable, and therefore used in the same contexts, even when logically equivalent (e.g., the case of contradictory antonyms and double negations). Further research may shed light on which type of contexts characterize the two types of expression, for example through a corpus study. Moreover, it would be interesting to assess which other properties negated adjectives have in a distributional space, such as their interaction with scalar dimensions (e.g., *not hot* vs. *freezing, cold, lukewarm, hot* etc.; Wilkinson and Tim (2016)) and implicatures (Van Tiel et al. 2016).

Our study is in line the proposal put forward by Kruszewski et al. (Kruszewski et al. 2017) that, even though Distributional Semantics may not be the right tool to represent truth-related aspects of meaning, it can be still very useful to study certain aspects of negation. In our case, we have showed that DS is probably not the right tool to capture inferential relations involving negations, but it can be used to quantify how similar the use of negations is to that of other expressions. For the purpose of this study we have focused on one distributional semantic model; however, it would be interesting to test

the robustness of our findings through an evaluation of various models. Moreover, one could apply an analogous methodology to analyze other types expressions which are taken to convey almost identical semantic content but may be used in different ways; for example, the quantifiers *most* and *more than a half* in English (Hackl 2009), or different forms of absolute superlatives in Italian (e.g., *molto bello - bellissimo*; Berlanda (2013)).

Our study primarily relates to research questions in Linguistics; however, we regard our results to also be of interest for Natural Language Processing. Aspects of negation as the ones we studied, as well as their effect on distributional semantic models, can be critical for tasks like stance detection or sentiment analysis (e.g., what does it imply that a costumer is *not happy* or *not unhappy* with a product?; Wiegand et al, (2010)).

Acknowledgments

We thank Gemma Boleda and Malvina Nissim for their useful suggestions. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 715154), and by the Catalan government (SGR 2017 1575). This paper reflects the authors' view only, and the EU is not responsible for any use that may be made of the information it contains.

References

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

Baroni, Marco, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language resources and evaluation*, 43(3):209–226.

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 238–247, Baltimore, Maryland, USA, June 22-27.

Berlanda, Sara. 2013. Constructional intensifying adjectives in italian. In *Proceedings of the 9th Workshop on Multiword Expressions*, pages 132–137, Atlanta, Georgia, 13-14 June.

Bernardi, Raffaella. 2014. Distributional semantics: A montagovian view. In Claudia Casadio, Bob Coecke, Michael Moortgat, and Philip Scott, editors, *Categories and Types in Logic, Language, and Physics*. Springer, pages 63–89.

Boleda, Gemma and Aurélie Herbelot. 2016. Formal distributional semantics: Introduction to the special issue. *Computational Linguistics*, 42(4):619–635.

Bolinger, Dwight. 1972. *Degree words*. Walter de Gruyter.

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

Clark, Eve V. 1992. Conventionality and contrast: pragmatic principles with lexical consequences. In Eva Kittay and Adrienne Lehrer, editors, *Frames, Fields, and Contrasts: New Essays in Semantic and Lexical Organization*. Routledge, pages 171–188.

Clark, Herbert H. 1974. Semantics and comprehension. In Thomas A. Sebeok, editor, *Current trends in linguistics: Linguistics and adjacent arts and sciences*, volume 12. Mouton, pages 1291–1428.

Coecke, Bob, Mehrnoosh Sadrzadeh, and Stephen Clark. 2010. Mathematical foundations for a compositional distributional model of meaning. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING): Tutorial Abstracts*, pages 1–4, Beijing, China, August 23-27.

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

Fellbaum, Christiane. 1998. *WordNet: An Electronic Lexical Database*. MIT press.

Fraenkel, Tamar and Yaacov Schul. 2008. The meaning of negated adjectives. *Intercultural Pragmatics*, 5(4):517–540.

Garrette, Dan, Katrin Erk, and Raymond Mooney. 2014. A formal approach to linking logical form and vector-space lexical semantics. In Harry Bunt, Johan Bos, and Stephen Pulman, editors, *Computing meaning*, volume 4. Springer, pages 27–48.
Giora, Rachel. 2006. Anything negatives can do affirmatives can do just as well, except for some metaphors. *Journal of Pragmatics*, 38(7):981–1014.
Giora, Rachel, Noga Balaban, Ofer Fein, and Inbar Alkabets. 2005. Negation as positivity in disguise. In Albert N. Katz and Herbert L. Colston, editors, *Figurative language comprehension: Social and cultural influences*. Lawrence Erlbaum Associates, pages 233–258.
Grice, H. Paul. 1975. Logic and conversation. *Syntax and Semantics*, pages 41–58.
Hackl, Martin. 2009. On the grammar and processing of proportional quantifiers: most versus more than half. *Natural Language Semantics*, 17(1):63–98.
Hermann, Karl Moritz, Edward Grefenstette, and Phil Blunsom. 2013. "Not not bad" is not "bad": A distributional account of negation. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)*, pages 74–82, Sofia, Bulgaria, August, 9.
Horn, Laurence R. 1972. *On the Semantic Properties of Logical Operators in English*. University of California, Los Angeles.
Horn, Laurence R. 1984. Toward a new taxonomy for pragmatic inference: Q-based and R-based implicature. *Meaning, form, and use in context: Linguistic applications*, pages 11–42.
Horn, Laurence R. 1989. *A natural history of negation*. University of Chicago Press.
Horn, Laurence R. and Yasuhiko Kato. 2000. Introduction: Negation and polarity at the millennium. In Laurence R. Horn and Yasuhiko Kato, editors, *Negation and Polarity. Syntactic and Semantic Perspectives*. Oxford University Press, pages 1–19.
Jespersen, Otto. 1965. *The philosophy of grammar*. University of Chicago Press.
Joshi, Shrikant. 2012. Affixal negation: direct, indirect and their subtypes. *Syntaxe et semantique*, 13(1):49–63.
Kim, Joo-Kyung and Marie-Catherine de Marneffe. 2013. Deriving adjectival scales from continuous space word representations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1625–1630, Seattle, Washington, USA, October, 18-21.
Kiparsky, Paul. 1982. Word-formation and the lexicon. In *Proceedings of the Mid-America Linguistics Conference*.
Krifka, Manfred. 2007. Negated antonyms: Creating and filling the gap. In Uli Sauerland and Penka Stateva, editors, *Presupposition and implicature in compositional semantics*. Palgrave McMillan, pages 163–177.
Kruszewski, Germán, Denis Paperno, Raffaella Bernardi, and Marco Baroni. 2017. There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics. *Computational Linguistics*, 42(4).
Lenci, Alessandro. 2008. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.
Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of 2013 International Conference on Learning Representations (ICLR)*, Scottsdale, Arizona, USA, May, 2-4.
Mohammad, Saif M., Bonnie J. Dorr, Graeme Hirst, and Peter D. Turney. 2013. Computing lexical contrast. *Computational Linguistics*, 39(3):555–590.
Morante, Roser and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.
Murphy, Lynne. 2003. *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.
Nguyen, Kim Anh, Sabine Schulte im Walde, and Ngoc Thang Vu. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 454–459, Berlin, Germany, August 7-12.
Paradis, Carita and Caroline Willners. 2006. Antonymy and negation—the boundedness hypothesis. *Journal of pragmatics*, 38(7):1051–1080.
Rimell, Laura, Amandla Mabona, Luana Bulat, and Douwe Kiela. 2017. Learning to negate adjectives with bilinear models. In *Proceedings of the 15th Annual Meeting of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 71–78, Valencia, Spain, April 3-7.
Socher, Richard, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 Joint*

Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 1201–1211, Jeju Island, Korea, July, 12-14.

Socher, Richard, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D Manning, Andrew Y. Ng, Christopher Potts, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642, Seattle, USA, October, 18-21.

Tessler, Michael Henry and Michael Franke. 2018. Not unreasonable: Carving vague dimensions with contraries and contradictions. In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*, Madison, Wisconsin, USA, July, 25-28.

The Pham, Nghia, Angeliki Lazaridou, and Marco Baroni. 2015. A multitask objective to inject lexical contrast into distributional semantics. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and The 7th International Joint Conference of the Asian Federation of Natural Language Processing (ACL-IJCNLP)*, pages 21–26, Beijing, China, July 26-31.

Turney, Peter D. and Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.

van Son, Chantal, Emiel van Miltenburg, and Roser Morante Vallejo. 2016. Building a dictionary of affixal negations. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, Osaka, Japan, December, 12.

Van Tiel, Bob, Emiel Van Miltenburg, Natalia Zevakhina, and Bart Geurts. 2016. Scalar diversity. *Journal of Semantics*, 33(1):137–175.

Widdows, Dominic and Stanley Peters. 2003. Word vectors and quantum logic: Experiments with negation and disjunction. *Mathematics of language*, 8(141-154).

Wiegand, Michael, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. 2010. A survey on the role of negation in sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing (NeSP-NLP)*, pages 60–68, Uppsala, Sweden, July 10.

Wilkinson, Bryan and Oates Tim. 2016. A gold standard for scalar adjectives. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*, Portorož, Slovenia, May, 23-28.