A distributional study of negated adjectives and antonyms

Laura Aina*
Universitat Pompeu Fabra
Barcelona, Spain
laura.aina@upf.edu

Raffaella Bernardi
University of Trento
Trento, Italy
bernardi@disi.unitn.it

Raquel Fernández
University of Amsterdam
Amsterdam, The Netherlands
raquel.fernandez@uva.nl

Abstract

English. In this paper, we investigate the relation between negated adjectives and antonyms in English using Distributional Semantics methods. Results show that, on the basis of contexts of use, 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 (e.g., *happy - unhappy*).

Italiano. *In questo lavoro, analizziamo la relazione fra aggettivi negati e antonimi in inglese utilizzando metodi di Semantica Distribuzionale. I risultati mostrano che, sulla base dei contesti di uso, la negazione di un aggettivo (ad es. "not cold"; it.: "non freddo") è tipicamente più simile all'aggettivo stesso ("cold"; it.: "freddo") che al suo antonimo ("hot"; it.: "caldo"). Tale effetto è meno accentuato per antonimi derivati tramite affissi (ad es. "happy"-"unhappy"; it.: "felice"-"infelice").*

1 Introduction

Negation has long represented a challenge for theoretical and computational linguists (see Horn (1989) and Morante and Sporleder (2012) for overviews): in spite of the relative simplicity of logical negation ($\neg p$ is true $\leftrightarrow$ p is false), complexity arises when negation interacts with morphosyntax, semantics and pragmatics.

In this work, we focus on the negation of adjectives in English, expressed by the particle *not* modifying an adjective, as in *not cold*. A naïve account of these expressions would be to equate them to antonyms, and hence take them to convey the opposite of the adjective (e.g., *not cold* = *hot*). In fact, this simplifying assumption is sometimes made in computational approaches which model negation as a mapping from an adjective to its antonym (e.g., The Pham et al., (2015), Rimell et al., (2017)). However, a range of studies support what is known as *mitigation hypothesis* (Jespersen, 1965; Horn, 1972; Giora, 2006), according to which a negated adjective conveys an intermediate meaning between the adjective and its antonym (e.g., *not large* $\approx$ *medium-sized*). The meaning of the adjective is mitigated by negation, while some emphasis on it still persists in memory (Giora et al., 2005). This view is compatible with pragmatic theories predicting that the use of a more complex expression (*not large*) when a simpler one is available (*small*) triggers the implicature that a different meaning is intended (e.g., *medium-sized*) (Grice, 1975; Horn, 1984). Computational models predicting similar mitigating effects are those by Hermann et al., (2013) and Socher et al., (2012; 2013).

In this work, we investigate negated adjectives from the perspective of Distributional Semantics (Lenci, 2008; Turney and Pantel, 2010). We study antonymic adjectives and their negations in terms of their distribution across contexts of use: to this end, we employ an existing dataset of antonyms, whose annotation we further extend, and the distributional representations of these and their negated version, as derived with a standard distributional model. 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** (Joshi, 2012) (e.g., *small - large* and *happy - unhappy* respectively). More-

* Part of the work presented in this paper was carried out while the first author was at the University of Amsterdam.

over, we investigate the distinction between lexical antonyms that are **contrary or contradictory**, that is, those that have or do not have 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, negated adjectives are more similar to the adjective itself than to the antonym (e.g., *not cold* is closer to *cold* than to *hot*), regardless of the type of antonym or of negation. However, we find that morphological antonymy is closer to negation than lexical one is.

2 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), i.e., that are associated with opposite properties within the same domain (e.g., *hot* - *cold*). 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 (for each pair we obtain two triples). For example:

(1) $\langle$ *hot*, *cold*, *not* $\{$*hot*$|$*cold*$\}\rangle$

(2) $\langle$*happy*, *unhappy*, *not* $\{$*happy*$|$*unhappy*$\}\rangle$

As data, we make use of a subset of the Lexical Negation Dictionary by Van Son et al. (2016). This consists of antonym pairs in WordNet (Fellbaum, 1998) annotated for different types of lexical negation (Joshi, 2012). We consider adjective pairs that are either *lexical* antonyms, i.e., with distinct lexical roots (e.g., *cold* - *hot*), or *morphological* antonyms, i.e., derived by affixal negation (e.g., *happy* - *unhappy*).[1] In our analyses, we compare different subsets of the data: we explicate and motivate the distinctions in the following.

Lexical vs. morphological antonyms These two groups are usually taken to express the same lexical relation - i.e., 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*). For this reason, we keep triples derived from lexical and morphological antonyms distinct, and compare them in our analyses: in particular, we are interested in testing whether in a distributional space negation tends to be more similar to morphological antonymy than to lexical one. Besides this comparison, we apply other distinctions to the triples obtained with lexical and morphological antonyms respectively, in order to investigate further effects.

	adj.	not_adj.	# triples
Lexical antonyms	254715	1144	**198**
– contrary	336923	1057	68
– contradictory	298378	1031	28
Morphological antonyms	83232	1821	**185**
– simple negations	84744	2002	157
– double negations	122525	871	28

Table 1: Average frequency of adjectives and negated adjectives per class, and total number of triples $\langle a_1, a_2, not\ \{a_1|a_2\}\rangle$ considered.

Contrary vs. contradictory Lexical antonyms have been classified as either *contradictory* or *contrary* (Clark, 1974), depending on whether the negation of one entails the truth of the other, without the availability of a mid-value. Fraenkel and Shul (2008) provided psycholinguistic results showing that if an adjective is part of a contradictory pair, its negation is interpreted as closer to the antonym than if it is part of a contrary pair (e.g., *not dead* is interpreted as being closer to *alive* than *not small* to *large*). We aim to investigate this result in a distributional space, where we are able to quantify similarities between lexical items.

Since no data annotated with respect to this distinction is available, the three authors independently annotated the antonym pairs in the dataset as either contrary, contradictory or unclear, following the definition used by Fraenkel and Shul (2008).[2] Not surprisingly, the inter-annotator agreement is only moderate (Fleiss' $k = 0.37$): already Fraenkel and Shul (2008) noted that even for what they considered contradictory pairs it is possible to conceive a mid-value interpretation (e.g., *not dead* $\approx$ *half-dead*; Paradis and Willners (2006)). This suggests that the contrary

[1] In the dataset, the former are coded as *regular antonyms* and the latter as *direct affixal negations*.

[2] Annotation guidelines at `https://lauraina.github.io/data/notadj.pdf`

vs. contradictory distinction involves a continuum rather than a dichotomy. We leave this aspect to be further clarified by future research and, for the purpose of our analysis, only consider pairs classified with full agreement.

Simple vs. double negation In the case of morphological antonyms, one of the two adjectives is an affixal negation, and hence already contains a negating prefix (such as *un-* in *unhappy*): adding *not* thus gives rise to a double negation (e.g., *not unhappy*). These expressions have been widely studied in the literature due to their difference with double negation in logic (e.g., Bolinger (1972), Krifka (2007) and recently 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*) . Our goal is to test whether evidence for this effect is found in a distributional space: in particular, if two negations were to cancel each other out then the negation of an affixal negation (e.g., *not unhappy*) should be particularly close to the antonym (e.g., *happy*). We then test 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*).

3 Analyses

3.1 Methods

Previous studies about negation of adjectives described its effect as a meaning shift from the adjective towards the antonym, that can be measured in terms of semantic similarity (Fraenkel and Schul, 2008). Distributional Semantics offers us a data-driven method of quantifying this: we can represent expressions as vectors summarizing their large-scale patterns of usage and then interpret their proximity relations in terms of similarity.

To this aim, we build a distributional semantic model with standard techniques, but whose vocabulary includes, besides word units, also negated adjectives. In practice, each occurrence of a negated adjective (adjacent occurrence of *not* and an adjective without intervening words; e.g., we exclude cases like *not very cold*) is treated as a single and independent token (e.g., *not cold* $\rightsquigarrow$ *not_cold*). With this pre-processing, we train a word2vec CBOW model (Mikolov et al., 2013)[3] on the concatenation of UkWaC and Wackypedia-En corpora (2.7B tokens; Baroni et al., (2009)), setting parameters as in the best performing model by Baroni et al. (2014).[4] 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 that of Nguyen et al. (2016) or The Pham et al. (2015)). Indeed, we are interested in investigating characteristics of antonyms and negated adjectives in a standard distributional model, that is not fine-tuned to a particular task and where no assumptions about the structure of its space are incorporated. However, we assess the quality of the induced model through a similarity relatedness task, where we find that it achieves satisfying performances.[5]

For our analyses, we consider triples as those described in Section 2. Given a triple $\langle a_i, a_j, not\ a_i \rangle$ (e.g., *cold*, *hot*, *not cold*), we define the following score:

(3) *Shift* := *Sim*(not a_i, a_j) − *Sim*(not a_i, a_i)

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. This measures how much closer a negated adjective is to the antonym than to the adjective (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. Due to the well-known tendency of antonyms to be close in a distributional space (Mohammad et al., 2013), the absolute value of *Shift* is not expected to be high (a vector close to one is likely close to the other too). However, we can test whether a higher proximity is registered towards one of the two adjectives.

From the data introduced in Section 2, we only consider triples where each of the three elements occurs at least 100 times in the training corpus of the distributional model. Table 1 shows the number of triples considered for each class and the average frequency of adjectives and negated adjectives.[6] The number of contradictory triples is

[3] Gensim implementation.

[4] Vectors size: 400; window size: 5; minimum frequency: 20; sample: 0.005; negative samples: 1.

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

[6] Negated adjectives are overall less frequent than their non-negated counterparts, as shown in Table 1.

small due to the choice of keeping only antonyms for which we had full agreement in the annotation; double negations triples are few due to the limited frequency of these expressions in the corpus.[7]

3.2 Results and discussion

Table 2 shows the scores across the different categories mentioned in Section 2. Example triples for each category are given in Table 3, together with the nearest adjectives of each element in the triple.

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. On one hand, this could be seen 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, it shows that negated adjectives have a profile of use that is more similar to that of the adjective than to the antonym.

The two classes of antonyms differ significantly in the extent of this effect: negated adjectives are closer to a morphological antonym than a lexical one (e.g., *not perfect* vs. *imperfect*, *not wide* vs. *narrow*). Such similarity in distribution can be explained by the similarity in structure, and hence possibly in meaning, of negated adjectives and affixal negations. Yet, in spite of the higher similarity in use, affixal negation still does not seem equivalent to negation by *not*, due to the negative average *Shift* value.

Contrary vs. contradictory antonyms In contrast to the results from the linguistic literature (see Section 2), 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 adjectives tend to be more similar to the adjective itself than to the antonym.

This result points at the fact that distributional similarity is capturing a different type of similarity from that considered in the experiments of Fraenkel and Shul (2008). We cannot thus directly interpret our results as just a product of the mitigating aspect of negation. Distributional information may discriminate between the negation of an adjective and the antonym, even when the two seem intuitively equivalent (e.g., *not dead* is closer to *dead* than to *alive*): indeed, the use of one or the other may serve different functions (e.g., contradicting an expectation, politeness, etc.), leading them to appear in different contexts. Moreover, we find that, since continuous representations are able to capture nuanced differences, the alleged dichotomy between contrary and contradictory antonyms may become a continuum in distributional space: for example, one of the closest adjectives to *not dead* is *half-dead*. This further underscores the difficulty in distinguishing between contrary and contradictory antonyms which we had already encountered in the annotation.

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*).

As before, the result cannot be interpreted only in terms of mitigation (though, e.g., *not unhappy* is close to *unimpressed*, hence a mid-value between *happy* and *unhappy*). In general, 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)).

4 Conclusion

We have investigated negated adjectives using the tools of Distributional Semantics, 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 to the adjective itself. This can be seen as a result of the various functions of negation (e.g., mitigation, contradiction to an expectation, politeness) that may lead to different patterns of use for negated adjectives and antonyms. Further research may shed light on which type of contexts actually discriminate them, for example through a corpus study, and which other properties negated adjectives have in a distri-

[7] Full list of triples at `https://lauraina.github.io/data/notadj.pdf`

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

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).

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*

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

butional space, such as their interaction with scalar dimensions (e.g., *not hot* vs. *freezing, cold, lukewarm, hot* etc.; Wilkinson and Tim (2016)). Finally, while for the purpose of this study we opted for a standard word2vec model, one could test for the same effects with differently obtained distributional vectors.

Despite its current limitations in covering truth-related aspects of meaning, Distributional Semantics was shown by Kruszewski et al. (2017) to be apt to model at least some aspects of negation, especially if graded in nature, such as alternativehood. Our study provides supporting evidence for this line of research and in addition points at the utility of using Distributional Semantics to uncover nuanced differences in use between a negation and other expressions, even when logically equivalent. Moreover, we regard our results to be of general interest for the NLP community, since effects of negation like the ones we studied and how they are represented in a distributional space can be critical for tasks like 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

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

Marco Baroni, 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.

Marco Baroni, 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.

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

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

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

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

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

Rachel Giora, 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*, pages 233–258. Lawrence Erlbaum Associates.

Rachel Giora. 2006. Anything negatives can do affirmatives can do just as well, except for some metaphors. *Journal of Pragmatics*, 38(7):981–1014.

H. Paul Grice. 1975. Logic and conversation. *Syntax and Semantics*, pages 41–58.

Karl Moritz Hermann, 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.

Laurence R. Horn. 1972. *On the Semantic Properties of Logical Operators in English*. University of California, Los Angeles.

Laurence R. Horn. 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.

Laurence R. Horn. 1989. *A natural history of negation*. University of Chicago Press.

Otto Jespersen. 1965. *The philosophy of grammar*. University of Chicago Press.

Shrikant Joshi. 2012. Affixal negation: direct, indirect and their subtypes. *Syntaxe et semantique*, 13(1):49–63.

Manfred Krifka. 2007. Negated antonyms: Creating and filling the gap. In Uli Sauerland and Penka Stateva, editors, *Presupposition and implicature in compositional semantics*, pages 163–177. Palgrave McMillan.

Germán Kruszewski, 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*.

Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. *Italian journal of linguistics*, 20(1):1–31.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Proceedings of 2013 International Conference on Learning Representations (ILCR)*.

Saif M Mohammad, Bonnie J Dorr, Graeme Hirst, and Peter D Turney. 2013. Computing lexical contrast. *Computational Linguistics*, 39(3):555–590.

Roser Morante and Caroline Sporleder. 2012. Modality and negation: An introduction to the special issue. *Computational Linguistics*, 38(2):223–260.

Lynne Murphy. 2003. *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.

Kim Anh Nguyen, 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.

Carita Paradis and Caroline Willners. 2006. Antonymy and negation—the boundedness hypothesis. *Journal of pragmatics*, 38(7):1051–1080.

Laura Rimell, 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.

Richard Socher, 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.

Richard Socher, 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.

Michael Henry Tessler 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*.

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

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

Chantal van Son, 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*.

Michael Wiegand, 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.

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