Not logical: A distributional semantic account of negated adjectives

MSc Thesis *(Afstudeerscriptie)*

written by

Laura Aina
(born 30th November 1993 in Florence, Italy)

under the supervision of **Dr. Raquel Fernández**[1] and **Dr. Raffaella Bernardi**[2], and submitted to the Board of Examiners in partial fulfillment of the requirements for the degree of

MSc in Logic

at the *Universiteit van Amsterdam*.

Date of the public defense:
August 30th 2017

Members of the Thesis Committee:
Dr. Maria Aloni
Dr. Raffaella Bernardi
Dr. Tejaswini Deoskar
Dr. Raquel Fernández
Prof. Dr. Benedikt Loewe (chair)
Dr. Willem Zuidema

INSTITUTE FOR LOGIC, LANGUAGE AND COMPUTATION

[1] University of Amsterdam, Institute for Logic, Language and Computation
[2] University of Trento, Center for Mind/Brain Sciences

Abstract

The meaning of a negated adjective does not always correspond to the one of its antonym (e.g., *not small* $\neq$ *large*); indeed, linguistic theories and experimental data suggest that one of the functions of negation is to shift the meaning of the negated item but not necessarily flip it into the opposite (e.g. *not small* $\approx$ *medium-sized*). In this thesis, we study negated adjectives in English employing the perspective of Distributional Semantics. We first construct vectorial representations of these expressions based on their co-occurrences with contextual features in a large corpus. We then make use of these in a set of exploratory experiments aimed at clarifying their relationship with other expressions, such as antonyms (e.g. *not small* vs. *large*) and scale co-members (e.g., *not small* vs. *tiny*). In particular, we investigate negation in terms of pragmatic and "graded" notions which are apt to be studied in a distributional space: *alternativehood*, i.e., the degree of plausibility of alternatives to a negated item, and *mitigation*, i.e., the meaning shift from the original adjective. In addition, we design and evaluate a compositional method to model negation of adjectives as a function learnt directly from distributional data. Results suggest that negated adjectives have different profiles of use from other allegedly equivalent classes of expressions, and that, contrarily to what often is assumed, a data-driven modelling of negation is not entirely out of the scope of distributional methods. Overall, this thesis tackles research questions about the complex nature of negation and the open problem of modelling this phenomenon within Distributional Semantics.

Acknowledgements

I am deeply grateful to my thesis supervisors, Raquel Fernández and Raffaella Bernardi, for their precious advice and support: they have turned this demanding experience into a fruitful opportunity for learning and a very pleasant collaboration. I would also like to thank the members of the Thesis Committee for taking time to provide feedback to my work, the Dialogue Modelling Group at the ILLC for the useful discussions, and Malvina Nissim for her advice about the datasets to employ.

The Master of Logic, whose end I have with this thesis reached, has been a challenging experience. Unstable like Amsterdam's weather, it was often windy; however, when sunny, it lightened up all the view. I need to thank myself for having had the courage to start this adventurous journey and not to give up while at it. However, if I got here, it is because I had the luck to be guided by a group of people (and also things) that, in their own way, have been an inspiration for me to find my way.

I want to express my gratitude towards Raquel Fernández for being the best academic mentor I could wish for. Having had the possibility to rely on her help about my choices was invaluable and shaped my growth as a student, as a researcher and as a person. During the Master of Logic, I was lucky to have the opportunity to learn from excellent lecturers and researchers. I want to in particular thank my research project supervisors Arianna Betti and Marianna Bolognesi for having influenced me with their contagious enthusiasm. I also want to thank those that I had the chance to collaborate with for the quality of the Master and the environment at the ILLC: the members of the OC, MoL Room and Ex Falso committees, Maria Aloni and Tanja Kassenaar.

Thanks to all my friends at the ILLC for turning Amsterdam into a home and my worries into smiles. A few names, in particular: Bonan, Jo, Jonathan, Lisa, Natalia, Valentin, Zoi. Thanks to the friends that have been there for me despite the distance; in particular, my beautiful flowers in Pisa, Erica and Rosa. Thanks to the books by Italo Calvino for teaching me when to stop enforcing logic on reality and to rather contemplate its complexity. Thanks to my oven for turning my thoughts into cakes to share. Thanks to Amsterdam for having inspired so much poetry that I had never enough time to write it. Last but not least, I thank my family: my parents, my sister and my grandparents. While I was busy studying the "meaning of words", they taught me what *love* means in practice. Their constant support is a precious treasure. Grazie.

Contents

Chapter 1

Introduction

> Non domandarci la formula che mondi possa aprirti,
> sì qualche storta sillaba e secca come un ramo.
> Codesto solo oggi possiamo dirti,
> ciò che *non* siamo, ciò che *non* vogliamo.[1]
>
> — Eugenio Montale, Non chiederci la parola, *Ossi di seppia*, 1925

Negation is pervasive in natural language and yet more complex to produce and to process than affirmation (Horn, 1989; Wason, 1961). If the negation of a concept and the affirmation of its opposite are equivalent, why do we sometimes go through the bother of using the former rather than the latter? Perhaps because they are not, after all, equivalent and therefore used in the same way.

Researchers in Linguistics, Cognitive Science and Philosophy of Language have focused over time on studying the deeply complex role of negation in natural language, as well as its relationship with the concept of opposition (see Horn (1989) for an overview). At the same time, negation received a neat treatment in Logic, as that one-place connective in propositional logic ($\neg$) which flips the truth value of a proposition ($\neg p$ is true if and only if p is false) and participates in the Laws of Double Negation ($\neg\neg p \equiv p$) and Excluded Middle ($p \vee \neg p$). However, the simplicity of logical negation does not reflect the structure and use of negative statements in natural language (Horn and Kato, 2000). Linguistic negation is, in this sense, *not logical*.

In this thesis, we focus on the negation of adjectives in English (e.g., *not logical*, *not small*) and explore the type of meanings assigned to them by assuming a data-driven perspective. In particular, we carry out our investigation within the framework of Distributional Semantics (DS) (Lenci, 2008; Turney and Pantel, 2010), that is the family of approaches which construct semantic representations of expressions on the basis of their distributions across contexts of use.

[1] "Don't ask us for the phrase that can open worlds, / just a few gnarled syllables, dry like a branch. / This, today, is all that we can tell you: / what we are *not*, what we do *not* want."

But if negation is *not logical*, what is it that it is? We will use this example of negated adjective to introduce the various research questions that we will address in this thesis:

- If negation is *not logical*, what else could it alternatively be? Indeed, the negation of an item often suggests that another option might hold. A tradition in Formal Semantics and Psychology has for this reason taken negation to not only exclude the element it applies to but also to suggest other expressions as potential alternatives to it (among others, Horn (1972) and Oaksford and Stenning (1992)). Alternativehood was also studied as a graded notion, with the goal of determining the degree of plausibility of an alternative to a negated element (Wason, 1965; Clark, 1974). If negation is *not logical*, one plausible alternative might then be that it is *alternative-licensing*.

- If negation is *not logical*, is it then *absurd*? The negation of an adjective is sometimes taken to coincide precisely with the expression of the opposite meaning, i.e., a negated adjective denotes the same semantic content of the antonym (e.g., *not true* = *false*; *not small* = *large*). However, it was shown that one of the functions of negation is to act instead 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 the one of its antonym. As a consequence, it may express a *mitigated* sense of the original adjectives, in particular in those cases where a middle between the adjective and the antonym is not excluded (e.g., *not small* ≈ *medium-sized*) (Fraenkel and Schul, 2008). If negation is *not logical*, it does not necessarily have to be *absurd*: it might be, for example, *pragmatic*.

- If negation is *not logical*, is it *illogical*? Affixal negations are often taken to be synonymous with the negated adjectives (e.g., *illogical* = *not logical*). Are these expressions, however, used in the same contexts? Moreover, negations by affix and antonyms with a distinct lexical root (e.g., *illogical* and *absurd* respectively) have been taken to be different only in morphological terms (Joshi, 2012). But are they really part of a homogenous class? In particular, one may wonder whether they behave in the same way with respect to their similarities to relevant negated adjectives. If negation is *not logical*, it might be *illogical* or *absurd*, or perhaps be even different from those two.

- If negation was *logical*, would it be *not illogical*? The use of double negation in Logic has a nullifying effect: two negations cancel each other out (*duplex negatio affirmat*). However, in language use, double negations of the sort of *not illogical* are typically used in different contexts than the affirmative counterpart (*logical*), for example to attenuate the strength of a statement (Horn, 1989). It might then be that even in the case that negation was *not illogical*, it might still not be *logical*.

- If negation is *not logical*, how logical is it? Adjectives can be associated with scalar dimensions and express positive or negative degrees of a given property

(Kennedy and McNally, 2005), such as how much LOGIC there is or there is not in something *logical* or *fallacious*. When negation shifts the meaning of a scalar adjective it plausibly acts along this graded dimension and expresses a new degree in the scale. If negation is *not logical*, it still expresses some degree of LOGIC in it.

We here approach these research questions about the negation of adjectives assuming a DS perspective, and rely on distributional models to provide a descriptive account of this phenomenon. Many of the research questions about negated adjectives revolve around comparisons between expressions (e.g., negated adjectives vs. antonyms); for this reason, previous studies often resorted to the notion of semantic similarity (for example, in the work by Fraenkel and Schul (2008)). DS emerges as a very good methodology for analysing this phenomenon since it provides a data-driven way of comparing expressions. By constructing their vectorial representations on the basis of co-occurrences with contextual features, we can compare them in terms of geometric proximity in a high-dimensional space. On one side, this allows us to assume the desired empirical perspective; on the other, it gives us the possibility to investigate pragmatic differences between expressions, since representations are by construction sensitive to differences in use. Moreover, it was shown that the type of semantic similarity that is captured in a distributional model can be used as a predictor of alternativehood to a negated item (Kruszewski et al., 2017). Building on this finding, we investigate an alternative-licensing view on the negation of adjectives in the framework of DS.

Negation is, however, a big challenge for DS. Despite its success in accounting for lexical content, its development into a compositional DS (Baroni, 2013; Mitchell and Lapata, 2010) is now confronting researchers in this area with the difficulties of accounting for this and other linguistic phenomena involving function words, which are instead successfully modelled within Formal Semantics (Bernardi, 2014; Boleda and Herbelot, 2016). The approach to negation that is typically taken within DS is to design it as an operator on the basis of *a priori* assumptions about the behaviour that this is posited to have (among others, Nghia et al. (2015) and Rimell et al. (2017)). Negation is indeed mostly regarded to be a phenomenon which escapes the modelling potential of distributional methods. Kruszewski et al. (2017) point out that such a difficulty arises from the attempt to capture a negation that is essentially *logical* rather than *pragmatic*, or *conversational*. The latter has indeed a more “continuous” nature that distributional models may be apt to capture (for example, considering the graded aspect of alternativehood). Aligned with their purposes, we further investigate the potentialities of DS as a model of *pragmatic* negation.

In the first part of this thesis, we construct a distributional semantic model where negated adjectives are represented and treated as a lexical unit (e.g., *not-logical*): we describe in Chapter 3 the motivation and procedure employed, and some properties of the resulting space. We then employ this distributional model to carry out a set of exploratory experiments which address the above-mentioned research questions, and which we report in Chapter 4. By making use of an external dataset of affixal and regular antonyms (van Son et al., 2016), we compare negated adjectives with these classes of expressions, and explore the relationship between adjectives negated by means of *not* and a negative affix (e.g., *un-*) respectively, and between negation and antonymy. Moreover, through an annotation procedure we classify a set of antonymic pairs into contrary and contradictory pairs, depending on whether they admit a mid-value between the two or not (e.g., *small - large*, *present - absent*) respectively. We then proceed to test whether predictions put forward in the literature about the negation of adjectives from such classes are supported by distributional data. Finally, we study the relationship between the negation of an adjective and other adjectives from its scale, by making use of the adjectival scales collected by Wilkinson and Tim (2016).

Later on, in Chapter 5, we consider a different approach to the representation of negated adjectives: we exploit the observed vectors, that we previously analysed, to obtain a compositional function representing negation using machine learning techniques. In particular, we learn a linear transformation on the basis of distributional data such that when applied to the vector representing an adjective it yields a representation of its negation. We, therefore, investigate whether it is anyhow feasible to approach negation from an entirely data-driven perspective. We evaluate such a function on a specific phenomenon, namely on accounting for the differences between the presumably lexicalised meaning of relatively frequent negated adjectives (e.g., *not bad*) and their compositionally derived one.

Looking at the broader picture, this thesis contributes to linguistic research by presenting further empirical results about the nature of negated adjectives, in particular for what concerns their differences or similarities in use with other expressions. Moreover, we provide an exploration of the potentialities of distributional methods to account for negation, which is of general interest to the Computational Semantics community and challenges the idea that this phenomenon is outside the scope of the modelling potential of DS. Our results are also relevant to more applied Natural Language Processing tasks, and, in particular, to Sentiment Analysis, where the interpretation of attributes like *not good* is especially crucial (e.g., how negative should a review that describes a restaurant as *not good* be rated?).

Last but not least, in the process of exploring the behaviour of negated adjectives, we hope to shed some light on the general and complex issue of what negation in general is, or at least of what it is *not*.

Chapter 2

Previous research on negated adjectives

In this chapter, we give an overview of the research previously carried out on the topic of the negation of adjectives, in order to situate the present study in the context of the literature. We first consider theories and experimental results from the field of Linguistics (Section 2.1): we start with a recap about the semantics of adjectives and proceed at describing studies about their interaction with negation. Later, we focus on the work carried out on this topic within the framework of DS (Section 2.2): after a short introduction to the fundamentals of its approach, we give an overview of the models proposed to account for adjectives and, in particular, their negation.

2.1 Adjectives and their negation in Linguistics

2.1.1 Adjectival meaning

On a very general level, adjectives are expressions that modify the meaning contributions of nouns, allowing for conveying more fine-grained meanings that nouns alone would do (e.g., *shirt* vs. *blue shirt*) (Huddleston and Pullum, 2002). Syntactically, English adjectives can supply the predicate term for a copula (i.e., *be*) or epistemic verbs like *seem*, and compose recursively with nouns, giving rise to complex constituents. Thus, adjectives can appear in both *predicative* (e.g., *The tea is warm.*) or *attributive* (e.g., *warm tea*) positions. The semantic effect of their composition with other items in the sentence is, however, complex and variable, and crucially depends on the type of adjective and on the noun that they combine with (Kamp, 1975; Partee, 1995). For example, the composition of adjectives like *vegetarian* could be modelled as set intersection: a *vegetarian person* is someone that has the property of being vegetarian and the property of being a person. However, the same cannot be said about other members of this class, such as *skilful* or *former*: a *skilful poet* is a poet but is not necessarily skilful in general, and a *former student* is not even a student. Due to entailment patterns of this sort, adjectives have received various analyses in Formal Semantics both as properties (functions from entities to truth values) and high-order properties (functions from properties to properties) (see Kennedy (2012) for an overview).

One of the most important lexical relations between adjectives and fundamental for the organisation of the lexicon is that of *antonymy*: a pair of antonymic adjectives is such that the two share all relevant features except for one which causes their incompatibility (i.e., they cannot be both applied to the same noun phrase), namely that they are associated with opposite properties within the same domain (e.g., *hot - cold*, *present - absent*) (Murphy, 2003). Indeed, one cannot say that something is *hot* and *cold* at the same time, but to say that something is *hot* or *cold* is informative of the same property, namely perception of temperature. A crucial distinction, which dates back to Aristotle, is the classification of antonymic pairs between *contrary* and *contradictory* (Clark, 1974). Contrary adjectives are such that the negation of one does not entail the truth of the other: if something is *cold*, it is not necessarily *hot*, but may be *neither cold nor hot*; they thus admit a *tertium*, or what Jespersen (1965) calls a "*zone of indifference*". Conversely, contradictory antonyms are linked by a complementarity relation: the negation of one entails the truth of the other. For example, one is either *present* or *absent*, without the availability of a mid-value.

A group of adjectives of particular interest for this thesis is that of *scalar*, or *gradable*, adjectives, namely those whose encoded meaning is related to a particular value in a scalar dimension. For example, the adjectives *small* and *large* are taken to express particular measurements in the scale of SIZE (Figure 2.1). Because of their properties, this class has been analysed as expressing relations between entities and degrees, whereas degrees ordered with respect to a dimension are taken to constitute a *scale* (Kennedy and McNally, 2005). Adjectives which express positive and negative degrees of the same scale, like the antonymic pair *small* and *large*, are taken to be associated with inverse ordering on the shared domain (e.g., *X is larger than Y* ⇔ *Y is smaller than X*; we will expand this point in Chapter 4) (Kennedy, 1999).

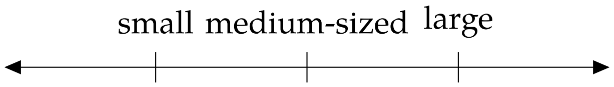

Figure 2.1: Examples of an adjectival scale of SIZE.

2.1.2 Negation of adjectives

Negation is a fundamental tool for natural language, which enriches it with the ability to express not only the truth but also the falsity of semantic contents. Such a *digital* property, however, encompasses the complex and various functions and forms that negation has in the actual use. For this reason, negation in natural language has historically represented a challenge for researchers in linguistics and philosophy (see Horn (1989) for an extensive overview). There is indeed a dramatic contrast between the simplicity of negation as it can be represented in a formal system and the complexity exhibited by instead linguistic negation, which emerges in interaction with principles of morphosyntax, semantics and pragmatics (Horn and Kato, 2000).

We here consider instances of the negation of adjectives to be the combination of a negative particle like *not* in English and an adjective, such as *not cold*.[1] Expressions of this kind happen to have a particular link with the notion of antonymy: indeed, one may be tempted to regard the negation of an adjective as equivalent to the assertion of its opposite (e.g., *not hot* = *cold*). However, negation is used in language not only to express denial and opposition (1), but also, among other functions, as a means of expressing contradiction to a common expectation (2), verbal politeness (3) and, last but not least, mitigation (4) (Giora, 2006).

(1) The student is *not present* (vs. *absent*).

(2) Despite the rumours, it turned out she was *not guilty* (vs. *innocent*).

(3) This painting is *not beautiful* (vs. *ugly*).

(4) The water is *not cold* (vs. *lukewarm*).

This diverse set of functions of negation can justify why speakers often opt for negative statements, despite these being typically more complex and harder to process than their affirmative counterparts (as shown by, among others, Wason (1961)): indeed, a negative statement may not always result in the same communicative import of an allegedly equivalent affirmative.

Negation as mitigation

We here focus in particular on the function of negation as mitigation. The *mitigation hypothesis* (see Jespersen (1965) and Horn (1972) for early formulations, and Giora (2006) for an overview) affirms that the negation of an adjective conveys a mitigated version of its meaning (e.g., *not large* ≈ *medium-sized*). In this sense, negation is described as a *modifier of degree*, such that it presupposes a bipolar dimension along which a meaning shift from an adjective towards its antonym occurs (Figure 2.2).

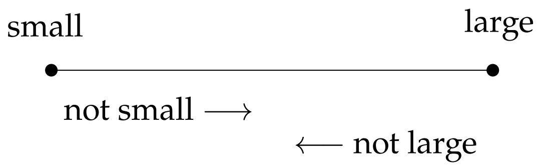

Figure 2.2: Example of an interaction between negation and a bipolar dimension defined by an antonymic pair, as predicted by the *mitigation hypothesis*.

Such an effect has been associated with two explanatory phenomena, possibly responsible in a complementary fashion. On one side, one could see the mitigation as

[1] At the syntactic level the occurrences of negative operators like *not* may be ambiguous between wide and narrow scope readings. For the purpose of this thesis, despite the potential simplification, we, however, align with most literature on negated adjectives which study *not* as a modifier of the adjective, and hence assume negation to take scope only over this constituent.

a result of the representational process: it arises as the product of the interaction between the negativity of the particle *not* and the meaning of the negated item, which is not suppressed but retained as accessible in memory (Giora et al., 2005). On the other, pragmatic inferences may be responsible for these non-literal interpretations: a non-parsimonious expression, such as a negated adjective, may be judged by the hearer to have been generated with a specific purpose (Grice, 1975; Horn, 1984). For example, the fact that one asserts that the water is *not cold* rather than saying that it is *hot* may suggest that she intends to convey an intermediate meaning between *hot* and *cold*.

Obviously, the interpretation of the negated adjective largely depends on its context of utterance. However, a stream of research focused on factors which impact on the amount of mitigation produced by the negation and are instead dependant on lexical properties of the adjective that is negated (Colston, 1999; Paradis and Willners, 2006; Fraenkel and Schul, 2008; Bianchi et al., 2011). In these studies, mitigation is typically operationalised in terms of semantic similarity of the negated adjective with the adjective itself or with the antonym.[2] We here in particular mention the work by Fraenkel and Schul (2008), which identify the feature of being part of a contrary (e.g., *hot - cold*) or contradictory antonymic pair (e.g., *dead - alive*) as a determining factor for the meaning shift applied by negation on an adjective.[3] They indeed show that if an adjective is part of an antonymic pair that bisects its domain in a dichotomous fashion, its negation is interpreted as closer to the antonym (e.g., *not dead* $\approx$ *alive*) than an adjective that is part of a contrary pair would (e.g., *not hot* $\neq$ *cold*) (Figure 2.3).

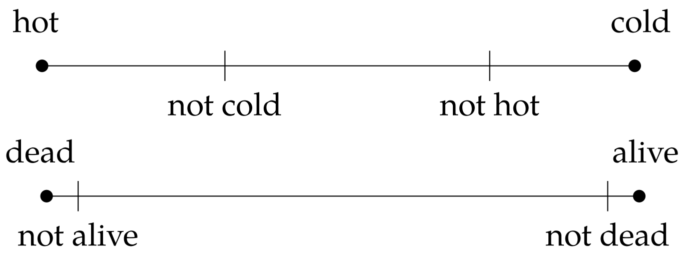

Figure 2.3: Example of mitigation as predicted by Fraenkel and Schul (2008) for contrary and contradictory pairs.

Intuitively, this corresponds to the idea that if no mid-value is available between two antonymic pairs (*tertium non datur*), as it is the case for contradictory ones, there is no

[2] Some see the mitigation as a process that weakens the meaning of the adjective that is negated (Giora et al., 2005); others instead regard it as an attenuation of the meaning of the antonym (Fraenkel and Schul, 2008). There is, however, agreement on the general idea of a meaning shift operated by negation which makes the meaning of the adjective closer to that of the antonym.

[3] Fraenkel and Schul (2008) also identify markedness as a determining feature for the meaning shift. However, in the present study we only focus on the results obtained for contrary and contradictory pairs, given the relatively more clear-cut definition of this class in comparison to the other predictors presented in the literature (e.g., markedness, negative or positive orientation, boundedness of the scale).

room for expressing a meaningful intermediate meaning: thus, a negated member of such a pair comes to express the same content as the antonym.

In these cases, negation shifts the meaning of an adjective towards the opposite in such a way that the property that this is associated to decrease (e.g., *not hot* approaches *cold* and hence expresses a smaller degree of heat than *hot* does). Interestingly, this is, however, not always the case: negated adjectives can also be used in sentences like (5) and (6), where the negation does not indicate a decrease of the property related to the adjective. For this reason, the negation of an adjective a was pointed out to be pragmatically ambiguous between a *less than a* and a *more than a* reading (Figure 2.4), whereas, however, the former still happens to be the default one (Horn, 1989).

(5) This is *not hot* - it is *scalding*!

(6) You are *not smart* - you are *brilliant*!

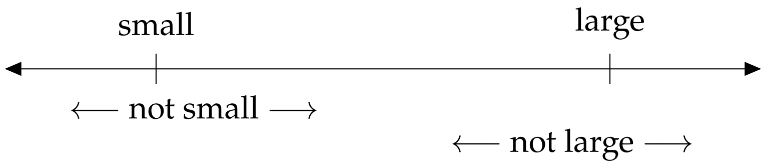

Figure 2.4: Example of an interaction between negation and a bipolar dimension defined by an antonymic pair, taking into account the pragmatic ambiguity of negation.

Negation as alternativehood

These interpretations of negated adjectives that we just saw are non-literal and pragmatic and can be seen an attempt to saturate a lack of sufficient informativity of negative utterances. Indeed, these are typically less informative than affirmative ones (Leech, 1981). For instance, while saying that something is *hot* is expressing a particular property that the object has, saying that it is *not cold* is instead only excluding a property that the object might have had. Such an attempt to reconstruct what an entity is on the basis of what it is *not* is accounted for by alternative-licensing views on negation. In this type of approaches, negation is taken to not only exclude the element that it applies to, but also to highlight a set of alternatives. In Formal Semantics, views of this sort have been presented in the *principle of alternate implicatures* by Horn (1972) and the theories of focus by Rooth (1992) and Krifka (1992) within Alternative Semantics and the structured meanings approach respectively.

An alternative set for a negative sentence is typically taken to be a set of semantic values which results from substituting the element that is negated with any value of the same semantic type (e.g., *not cold* ⇝ {*happy*, *hot*, *transparent*, *lukewarm*,...}). However, members of this set may differ in terms of their plausibility to constitute an alternative, not only depending on the context but also on the basis of the meaning of the negated item. A stream of research in Psychology focused indeed on studying the plausibility of alternatives to negated expressions, i.e., *alternativehood*, both in terms of constraints on an alternative set (Oaksford and Stenning, 1992; Oaksford, 2002) or as a graded notion (Wason, 1965; Clark, 1974). These studies emphasise a particular connection between alternativehood and semantic similarity. On one side, the alternatives primed by negation tend to be the most relevant and similar to the state of affairs that is negated; on the other, the interpretation is facilitated when the negated statement denies a possible presupposition, and hence something that may be believed to be true. It then seems reasonable that plausible alternatives would tend not to move away too much from the negated item. For example, alternative utterances to *The water is not cold* may likely substitute the negated adjective with *lukewarm* or *hot*, which are related to the same semantic domain, rather than the less relevant but typically true *transparent*, or the non-applicable *happy*.

Affixal negation

We here introduce a class of expressions which shares a substantial similarity with negated adjectives, namely *affixal negations*. These are morphologically complex expressions derived by the insertion of a negative affix (e.g., *un-*, *dis-*) to an adjective (e.g., *unhappy*, *dissimilar*). In particular, we focus on *direct* affixal negations, which Joshi (2012) defines as those which are linked to the original adjective by a relation of antonymy and that are arguably equivalent to the corresponding negated adjective (e.g., *unhappy* = *not happy*). Instead, *indirect* negations, like *infamous* and *subnormal*, despite of the negative connotation, encompass various types of semantic relations which cannot simply lead back to that of opposition.

The existence of a similarity between direct affixal negations and negated adjectives is not surprising: indeed, the two groups of expressions share a similar compositional structure, despite the fact that the latter exceeds the word boundaries. The negative affixes could indeed be seen as having the same function of the particle *not*. However, the incorporation of the negation into the adjective seems to bring in some differences. For example, a sentence with an affixal negation will count as an affirmative, unlike for negated adjectives, and hence possibly involve a different speech act (7); moreover, the compositional meaning of an affixal negation seems to be more subjected to a lexicalisation process, and hence more conventional: for instance, while negated adjectives licenses both *less than a* and *more than a* interpretations, negation by affix is always associated with a decrease of the property associated with the adjective (8, 9) (Horn, 1989).

(7) This is {*impossible*, *not possible*}.

(8) I am {*not happy*, *unhappy*} - I am sad.

(9) I am {*not happy*, # *unhappy*} - I am ecstatic.

Joshi (2012) considers antonymic pairs derived by affixation (e.g., *frequent* - *infrequent*) to be expressing the same lexical relation than antonyms with distinct lexical roots, namely *regular* antonyms (e.g., *wrong* - *right*), and hence to be different only at the morphological level. However, one problem with this classification may be encountered if considering in this picture mitigation effects: if affixal negations are equivalent to negated adjectives, then it is hard to see how they could be taken to express a relation of opposition exactly like regular antonyms, given that the negation of an adjective is not always interpreted as its antonym.

Negation and affixal negation come to interact in double negations constructions, such as *not uncommon*. These expressions have been studied in depth in the literature (see for example Horn (1984), Bolinger (1972) and Krifka (2007)) due to the fact that, unlike in logical negation ($\neg\neg p \equiv p$), the two negations do not seem to cancel each other out. Indeed, complex constructions of this kind tend to be instead associated to weaker meanings than the non-negated adjectives (e.g., *not uncommon* $\neq$ *common*), and to be used as a form of litotes or understatement (10), or in cases of hesitation or uncertainty (11).

(10) The damage was *not unproblematic* (vs. *problematic*).

(11) It is *not impossible* that it will rain tomorrow (vs. *possible*).

2.2 Adjectives and their negation in Distributional Semantics

2.2.1 Distributional semantics

Distributional Semantics (DS) is a computational framework for the representation of linguistic meaning; it consists of a family of data-driven methods which share a core assumption, known as *distributional hypothesis*, stating that similarity of semantic content correlates with similarity of contexts of use (Lenci, 2008). Following this idea, the distribution of an expression across contextual features is taken to be characterising of its meaning, whereas these are typically defined as the words that surround the occurrence of a lexical item within a certain span of text. Using DS methods, it is possible to summarise this information using the mathematical format of a vector, i.e., a set of numerical parameters identifying a point in a high-dimensional space.

The techniques that can be employed to construct such representations of expressions are, however, various, and can be clustered into two main types of resulting models (Baroni et al., 2014b). *Count* models make use of statistics of co-occurrences between target expressions and contextual features in a corpus: this information is collected in a set of weights dependent on the associativity between the former and the latter ones (Turney and Pantel, 2010). *Predict* models, instead, construct distributional representations using a neural network architecture trained on a corpus with the objective of predicting the context given a word, or vice-versa: by optimising the embeddings associated with the words to carry out this task, these are eventually transformed into representations of their distributions (Mikolov et al., 2013a).

Thanks to their vectorial format, representations of this sort can be compared to each other in terms of their geometric proximity in their high-dimensional space, known as *distributional* or *semantic* space. This allows to quantify the similarity between two expressions in a graded fashion by looking at distance measures between their vectors, such as *cosine similarity* (i.e., cosine of the angle between them). Because of how these representations are constructed, this methodology enables to capture fine-grained and nuanced differences between the distributions across contexts of the two expressions.

DS was shown to be successful at modelling many linguistic phenomena related to lexical meaning, such as semantic similarity prediction, synonymy detection, selectional preferences, concept categorisation and analogy (Baroni et al., 2014b). In addition, the framework was extended to account for the meaning of phrases and sentences in a compositional fashion. Various methodologies have been proposed in this respect (among others, Mitchell and Lapata (2010), Baroni and Zamparelli (2010), Socher et al. (2012) and Grefenstette et al. (2013)), ranging from simple operations on the distributional vectors to more complex operations making use of higher-order tensors estimated via machine learning techniques. These models were shown to be able to successfully carry out challenging tasks such as sentence similarity prediction (Marelli et al., 2014), and to account for some complex phenomena, especially involving the composition of content words (see the example of adjectival modification in the next subsection). However, many aspects of compositional meanings are still an open challenge for DS, in particular those related to the semantic contributions of function words (e.g., negation, quantification), which are instead more easily modelled employing formal approaches.

2.2.2 Modelling adjectival meaning

We here report some of the research carried out within DS on those aspects of adjectival meaning which we mentioned in Section 2.1, namely their semantic contribution, the lexical relation of antonymy, and the class of scalar adjectives.

As other content words, adjectives can be represented and compared in a meaningful way in the form of distributional vectors; in these cases, they are represented with the same format of the objects they typically modify in a sentence, i.e., nouns. While this is the standard approach when studying their lexical properties, this uniformity assumption may not be considered appropriate when, for example, using their representations for composing the ones of constituents above the word level. Indeed, adjectives have typically been studied in Formal Semantics as functions applied to nouns (Kamp, 1975). A class of approaches in DS (Grefenstette et al., 2013; Baroni et al., 2014a), inspired by Montague Grammar, proposed to model compositional operations as functional application, representing expressions with different semantic types as tensors of different orders. In particular, adjectives have been modelled as matrices (Baroni and Zamparelli, 2010), such that, when multiplied with a noun, they outputs a vector representing the adjective-noun phrase (i.e. *COLD* × *water* = *cold water*). Such matrices are estimated in a data-driven way from a training set of observed vectors of adjective-noun phrases. Because of the co-dependency of meaning between the adjective and the noun, this method was shown to better model the complex aspects of adjectival modification than other methods that instead treat adjectives as vectors (Boleda et al., 2012, 2013).

We here, however, make a step back and look at lexical properties of distributional representations of adjectives as corpus-derived vectors. In particular, we now focus on the semantic relation of antonymy. As pointed out by Mohammad et al. (2013), words with opposite meanings (e.g., *hot - cold*, *good - bad*) have the tendency to occur in similar contexts. This is aligned with the idea that, despite their incompatibility, they share many semantic properties, which will induce them to be used in similar contexts. However, due to this, distributional models have difficulties in distinguishing between antonyms and synonyms of a word, as both of them will typically be retrieved as its closest words. Many approaches have been proposed to overcome this issue, ranging from using *ad-hoc* measures for antonym detection to supervised algorithms that either make them distinguishable to a classifier or increase their distance in the semantic space (see for example Nguyen et al. (2016) for a brief overview of the methods proposed).

Despite not being able to clearly distinguish between synonyms and antonyms, distributional models have been, however, shown to capture scalar relationships between adjectives (e.g., *bad* < *okay* < *good* < *excellent*). Kim and de Marneffe (2013) devised a method to automatically construct adjectival scales exploiting simple spatial relationships between expressions. In particular, they assume intermediate points between two word vectors to represent intermediate meanings. Given a pair of antonymic adjectives, they are able to construct their adjectival scale by iteratively calculating mid-points between expressions. What this result seems to show is that, in spite of the proximity between antonyms, the intermediate space between them is typically populated in an ordered way by members of their scalar dimension. The gradability of adjectival scales seems to then have a counterpart in the continuous space of distributional models.

2.2.3 Modelling the negation of adjectives

Although DS traditionally focused on lexical meaning, its extension into a compositional DS emphasised the necessity to account for function words and the complex phenomena that involve them, such as negation, in order to provide a fully-fledged model of sentence meaning (Bernardi, 2014). As we saw in the previous subsection, compositional functions that involve content words, such as adjectival modification, have received a successful account by directly inducing these from distributional data. However, the same is usually not assumed to be feasible for function words.

On one hand, approaches like the one of Garrette et al. (2014) conceive the treatment of these expressions as entirely out of the scope of DS. They instead exploit the complementarity of distributional and formal approaches to meaning to account for compositionality, and propose to model relations between content words using DS and the contribution of function words using first-order logic. On the other hand, some have instead proposed to still model negation within the framework of DS, but, however, defining it as an operation in the semantic space on the basis of *a priori* assumptions about its behaviour.

For instance, Widdows and Peters (2003) expect a word meaning and its negated version not to share any feature, and hence model the latter as the orthogonal vector to the former. Coecke et al. (2010) and other related theoretical approaches, instead, consider the abstract scenario in which the truth value of a sentence is represented by the single vector $\vec{1}$ (*true*) or the origin $\vec{0}$ (*false*): negation is, in this context, treated as a matrix which swaps this, and hence entails the falsity of the sentence it is applied to. The approach proposed by Hermann et al. (2013) incorporates instead the idea that when *not* is applied to an adjective, the resulting phrase remains close to others from the same domain of the adjective (e.g., *blue* and *not blue* both belong to the domain of colours) but its value changes. In particular, they describe a framework where *domain* and *value* features are distinct in the representation of an adjective, and negation only modifies the latter. Rimell et al. (2017) implements a model of the negation of adjectives with a similar view: they introduce a neural network architecture to learn a mapping from an adjective to the negated version conditioned on the domain of the former, represented using the closest words to this in the semantic space. However, they train their model of negation by learning to map an adjective to its antonym, thus assuming a negated adjective and an antonym to be equivalent. A similar approach is taken by Nghia et al. (2015): they learn a matrix representing *not* as a mapping between the vectors of two antonyms, and to be multiplied with the adjective to yield the representation of its negation. Their choice of equating the meaning of a negated adjective and an antonym at training time is, however, a simplification: as we saw, negated adjectives do not always convey the same semantic content of an antonym.

Socher et al. (2012, 2013) propose instead a data-induced approach to modelling negation. They devise neural network models which learn representations of phrases and sentences with the objective of detecting the sentiment of a discourse (e.g., a movie review). Their approach to compositionality is then essentially task-driven. Interestingly, they evaluate these models with respect to their ability to capture the meaning of negated adjectives, which they expect to convey a mitigated version of the non-negated counterpart. They show that architectures of this sort are able to capture mitigation effects and correctly take them into account when assigning fine-grained sentiment labels. Such a result is obtained exploiting associativity patterns with not only contexts of use of expressions, but also sentiment labels of the discourses that they are used in.

As for affixal negation, this was attempted to be modelled by Marelli and Baroni (2015) in their work on morpheme combination at the word level. Similarly to the previously mentioned approach to adjectival modification, they treat affixes of different types, including negative ones like *un-*, as data-induced functions mappings lexical roots to derived forms (e.g., *acceptable* → *unacceptable*). Their model is able to correctly predict semantic intuitions about novel derived form. Although their focus is not on affixal negation, they show that it is possible to construct in a data-driven way compositional functions representing the semantic contribution of negative items.

Finally, we mention an approach to negation which focuses on noun phrases, but nevertheless largely inspired the study presented in this thesis. As we saw, the notion of alternativehood and semantic similarity appear to be connected: the plausible alternatives of a negated constituent are typically similar to this. Kruszewski et al. (2017) proposed to use similarity relations between expressions as captured by a distributional space to give an account of the alternative-licensing nature of negation. They show that the type of semantic similarity captured by a distributional model, i.e., proximity in the distributional space, provides an excellent fit to a dataset of alternative plausibility ratings. This consists of data collected in the following setting: subjects are presented with sentences in the form *This is not an X, it is a Y* and *There is not an X, but there is a Y* (e.g., *This is not a horse, it is a donkey*), and asked to provide a plausibility rating of the sentence. Very good results on this task are obtained using the cosine similarity between the negated constituent and the true alternative (in the example above, *horse - donkey*). Indeed, distributional similarity scores expressions as close when they tend to appear in the same contexts, and hence somehow measure their substitutability; this last notion is particularly aligned with the notion of alternativehood: one can indeed expect plausible alternatives to occur in similar contexts to the negated constituent. Crucially, the approach taken in their work opens up an interesting line of research where distributional semantics is employed to account for a pragmatic, or conversational, form of negation, which is arguably more "graded" in nature than logical negation, and thus more apt to be captured in a continuous space.

Conclusion

In this chapter, we reviewed studies on adjectives and their negation in Linguistics and Distributional Semantics. We focused, in particular, on those aspects which will become relevant in the course of this thesis, namely antonymy, adjectival scales, negation as mitigation, negation as alternativehood and affixal negation.

As we saw, the complexity of these phenomena as reported in Linguistics has a counterpart in the challenging task of modelling them within DS. In this thesis, we try to bridge between these two fields, and clarify some of the research questions that the each of them present, by making use of notions and methods from the other. We indeed believe that while Linguistics can benefit from the evidence provided by distributional methods, DS can be helped in its modelling purposes by a better awareness of the target linguistic phenomena.

Chapter 3

Distributional representations of negated adjectives

In order to give an empirical account of negated adjectives in English, we are interested in data-driven representations that reflect their large-scale use. Therefore, we construct a distributional semantic model using standard techniques in the field, but including as target items not only words but also phrases consisting of an adjacent occurrence of *not* and an adjective, such as *not logical*. First, we give the motivation for such an approach and describe the way we realise it in Sections 3.1 and 3.2 respectively; we then proceed to describe some aspects of the representations we obtain in Section 3.3.

3.1 Negated adjectives as a single unit

As mentioned in the Introduction, in this thesis we aim at statistically observe negated adjectives use in order to identify the kind of meanings which are typically assigned to them. DS is a natural choice for this goal: it allows us to build representations of expressions that approximate their semantic content and are by construction sensitive to differences in use. In the first part of our analyses (Chapter 4), we opt for representing negated adjectives by treating them as a single lexical unit rather than a multi-word phrase. We hence disregard, at least at this stage, their internal compositional structure and model them in practice as if they were a single word. We provide in this section the motivation for such an approach.

Building distributional representations of expressions larger than a word unit is not a standard approach in DS: typically, models are set up to build *observed*, i.e., corpus-derived, vectors only for unigrams of content words, such as nouns or adjectives. To obtain instead the representations of a multi-word phrase, one would then devise a compositional method which somehow merges the representations of its building blocks into a *composed* representation. However, we believe that, before designing a compositional function for negated adjectives, a better understanding of how negation modifies the representation of an adjective in a distributional space is required. Since this is precisely the object of our investigations, we first study negated adjectives as a single unit, and only later attempt at a compositional modelling in Chapter 5.

Nevertheless, the negation of an adjective does not result in an expression that behaves exactly like a single lexeme. A negated adjective tends to be a more complex construction than its non-negated counterpart (e.g., *not nice* vs. *nice*) both at the processing and linguistic level (Horn, 1989). Focusing on the latter, the negation of an adjective in English is syntactically marked (through the use of the particle *not*), results in a semantic content which is in most cases derived compositionally, and typically has a highly context-dependent interpretation due to complex semantic and pragmatic phenomena. In addition, the status of the phrase as a cohesive unit can be debated. On one side, the particle *not* may be seen as modifying a verb in the sentence rather than the adjective (e.g., *This (is not) good*) with complex implications for the scope of negation; on the other, the insertion of intervening words between *not* and the adjective is allowed (e.g., *This is not* {*that, very, too...*} *bad.*). Last but not least, even if treating a negated adjective as a single unit, its meaning would still be dependent on the noun phrase it is associated to, exactly like it happens for adjectives (e.g., *This man is not-tall* vs. *This building is not-tall.*). Nevertheless, we argue that treating negated adjectives as a unit is indeed a tenable approach with a purpose like ours.

We acknowledge that at the syntactic and formal semantic level this choice implies abstracting away from many of the complexities of these expressions. As we saw, adjectival meaning is better modelled by considering adjectives as functions applied to nouns. This is indeed the approach that is typically taken in Formal Semantics (Kamp, 1975) but also in compositional DS (Baroni and Zamparelli, 2010; Boleda et al., 2013). However, in this study, we apply a simplification and leave the aspect of the interaction with a noun to be accounted for in future research. Indeed, one of the fundamental tools that we can employ to study negated adjectives is to compare them to other expressions, and in particular, to adjectives. Eliciting similarity judgements is indeed the procedure that is often used in the literature to study their meaning (for example in the experiments by Fraenkel and Schul (2008)). To be able to easily model this in a distributional space, we are required to assume the same representation level, and somehow semantic type, for the types of expressions that we want to compare: for this reason, we model negated adjectives exactly like adjectives, in the form of observed vectors directly derived from their distributions.

But is it anyhow sensible to represent the meaning of multi-word phrases as if they were a unit? From the theoretical point of view, DS builds on the assumption that there is a correlation of some nature between the contexts of occurrence of an expression and its semantic and pragmatic content. There does not seem to be any limitation in this view that prevents it to be applied to expressions beyond the word boundaries, like negated adjectives, and study their meaning as a unit even when this has a compositional component. Indeed, although the internal interaction among the meanings of its building blocks, a multi-word expression still has an overall meaning which its use may reflect, and which we may be able to account for using DS. After all, even morphemes combination at the word level, i.e. affixation, such as *true* → *untrue*, or *think*

→ *rethink*, is a compositional process: yet this can be studied both considering functions that maps lexical roots (e.g., *true*, *think*) onto derived forms (e.g., *untrue*, *rethink*) (Marelli and Baroni, 2015), but also considering the latter expressions as unique and independent entities.

Applying the distributional methodology to phrases like negated adjectives, however, we encounter some practical limitations related to data sparsity. Typically, words have much more generic meaning than multi-word expressions and consequently occur in a wider range of contexts (e.g., *green* vs. *green apple*, *tall* vs. *not tall*), as well as substantially more often. Moreover, multi-word expressions lie in a continuum from semantic transparency and idiomaticity, whereas their meaning at the two poles is respectively entirely derived by looking at the meaning of its parts (e.g., *eat an apple*, *not vegetarian*), or instead be assigned in a conventional fashion to the expression as a whole (e.g., *kick the bucket*, *not bad*) (Fazly and Stevenson, 2008). The degree of lexicalisation of the phrase tends to impact on its frequency of occurrence in corpora, i.e., multi-word expressions with a fixed meaning tend to appear more often. As a result of these phenomena, phrases beyond the word level, and in particular compositional ones, are generally less frequent than words.

Since distributional representations are by construction sensitive to patterns of association in the data, their quality highly depends on the amount of relevant data that they had been trained on. As a consequence, except for frequent negated adjectives like *not bad*, we expect their distributional representations to be of lower quality and of less clear-cut content in comparison to, for example, the ones of adjectives. However, we consider as promising starting point the positive evaluation by Baroni and Zamparelli (2010) of corpus-derived vectors of adjective-noun pairs (e.g., *green apple*). In their methodology for learning compositional functions for adjectives, they are required as a first step to construct vectors of these bigrams: they found them to be meaningful representations, as well as an adequate benchmark to which compare the compositionally derived ones. On the other hand, we take into account both in the set-up and in the interpretation of our analyses the potential effects of low frequency on the vectors.

Finally, there is another main challenge for our approach. In our analyses, we make use of the distributional representations of negated adjective to, among other goals, study their link with antonymy. However, DS is known to struggle with this notion: adjectives with opposite meanings appear to be close in the semantic space (Mohammad et al., 2013). Although negation is not expected to always flip the meaning of an adjective into the antonym, its link with the notion of opposition is still crucial (the default interpretation seems to be a shift in meaning towards the opposite), but not marked in a discrete way in a distributional space. However, its continuous way of representing might as well be its advantage: the negation of adjectives, as we saw, can be seen as a graded phenomenon both in terms of mitigation and alternativehood. Moreover, although a distributional space might not be the ideal setting for an automatic identification of antonymic expressions, it seems to, however, capture their differences, as well as the gradability of intermediate meanings between them, when zooming into the region of the space where these are located (as the results of the experiments by Kim and de Marneffe (2013) show). For this reason, we believe that it is possible to study the relations of a negated adjective and its interaction with an antonymic pair in a distributional space, as we will do in Chapter 4.

3.2 Distributional semantic model

Given this motivation, we proceed to build a distributional semantic model where both words and negated adjectives are included as target items. To produce this, we make use of a large training corpus of English, namely the concatenation of the PoS-tagged versions of UkWaC (1.9B tokens) and Wackypedia-En (820M tokens) corpora (Baroni et al., 2009).

While we follow standard techniques at training time, we adapt the corpus data at pre-processing time for the purposes of our study. In particular, we process the corpus in order to merge adjacent occurrences of the particle *not* and an adjective as a single unit (e.g., `not nice` ⇝ `not_nice`).[1] Besides this procedure, we lemmatise the corpus, filter out stop-words, and keep part of speech labels for adjectives.

As we mentioned in Chapter 2, there exist various techniques to build a distributional semantic model given a training corpus. We opt for a Word2vec CBOW model (Mikolov et al., 2013a).[2] As the other models from the *predict* class, a model of this kind constructs distributional representations of expressions as a byproduct of optimising word embeddings in a prediction task: in particular, it learns to predict a term given a symmetric window of expressions at its left and right. Our choice of the model and its associated parameters relies on the extensive evaluation by Baroni et al. (2014b), which tested various combinations of techniques and parameters in a range of semantic tasks such as semantic relatedness prediction and synonymy detection. We set the parameters of our CBOW model as their best performing system across tasks (dimensionality of the vectors: 400; window of words: 5; minimum frequency threshold: 20; sample: 0.005; negative samples: 10). The resulting distributional model trained on the abovementioned corpus has a vocabulary of 719K items, among which 92K are adjectives and 1.8K are negated adjectives.

We evaluate the quality of the distributional space on a similarity relatedness task, in which the model is required to assign semantic similarity scores to a set of pairs of

[1] This procedure implies discarding some occurrences of negated adjectives. Requiring adjacency of the particle and the adjective, we discard all their occurrences with intervening words (e.g., *not too good*); these are however adjective modifiers such as *very*, *that*, *really* which alter the meaning of the adjective itself (in particular, most are modifiers of degree which would create a bias while studying negation itself as a modifier of degree). We also discard contracted occurrences of *not* such as *isn't*: while this reduces the occurrences of negated adjectives we can use to build our model, it is unclear whether these contracted forms bring in any semantic differences with the non-contracted or the auxiliary contracted ones (e.g., *That is not good*, *That's not good* vs. *That isn't good*), in particular at the level of the focus on the negative particle (see for example the overview by Pérez (2013)).

[2] Gensim implementation: `https://radimrehurek.com/gensim/models/word2vec`.

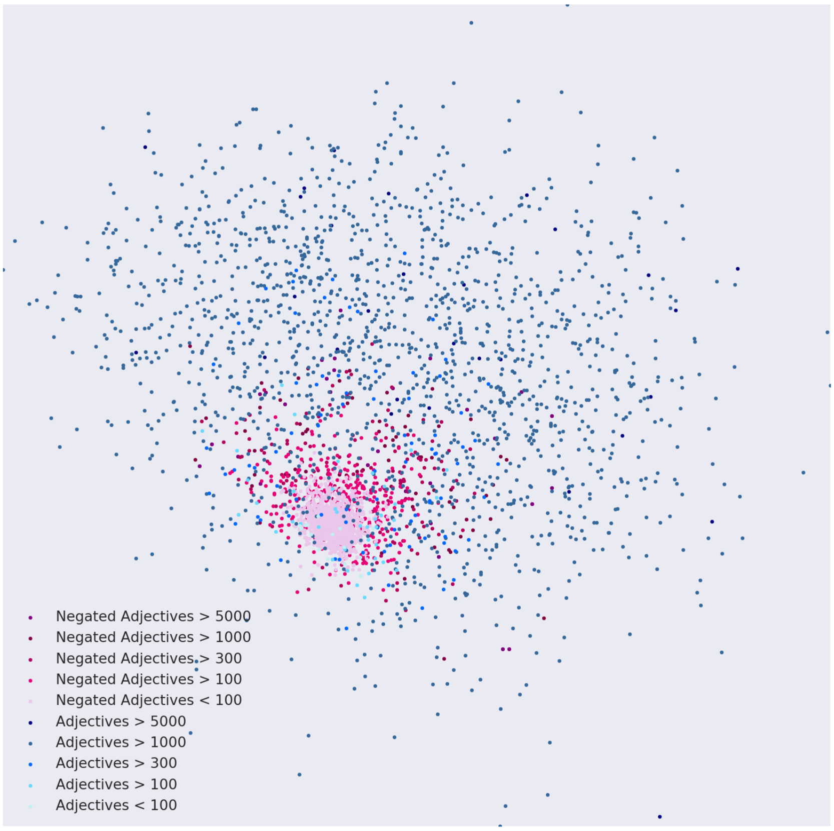

Figure 3.1: Negated adjectives and their corresponding adjectives in the two-dimensional semantic space (original space reduced using PCA).

content words. The results are then evaluated by looking at the correlation between the values and human-assigned similarity judgements. For this task, we use the MEN dataset (Bruni et al., 2014) (3K word pairs), and the cosine between vectors as a similarity score: the good performance in the task (Spearman's ρ: 0.75; $p = 0$; see results by Baroni et al. (2014b) for a comparison) makes us confident about the general quality of the distributional representations in the model.

3.3 Negated adjectives in the semantic space

3.3.1 Location in the semantic space

Once obtained our distributional representations of negated adjectives, we analyse some of their properties in the semantic space. An interesting feature we observe is that the vectors of negated adjectives tend to occupy a distinct region of the space from the one occupied by the adjectives, as it can be noticed in Figure 3.1.

In order to further assess the phenomenon, we apply a clustering algorithm, namely

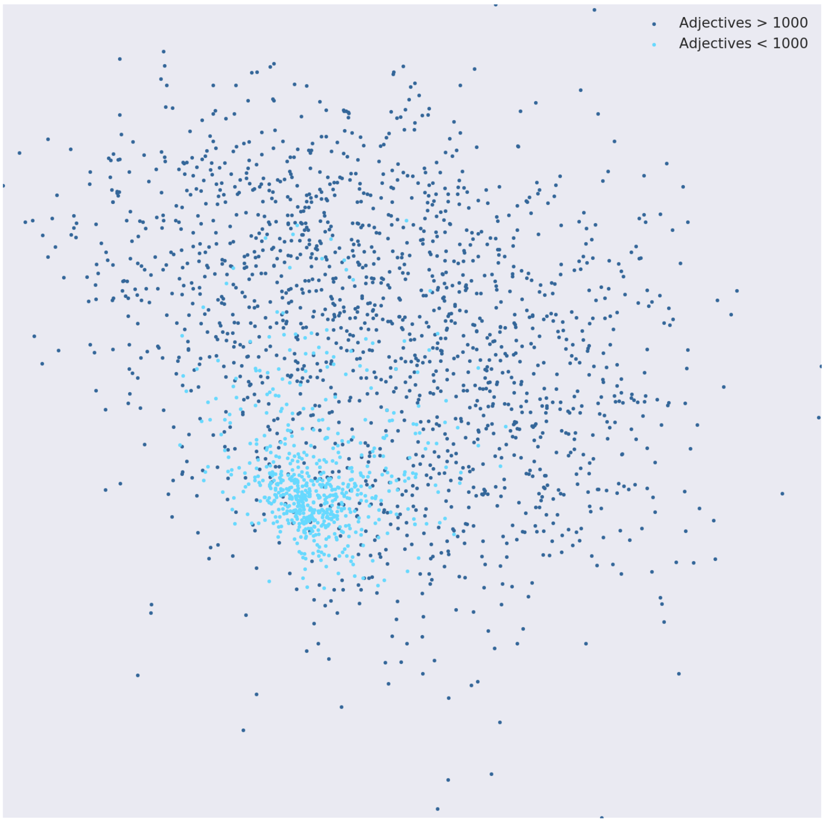

Figure 3.2: A sample of frequent and infrequent adjectives in the two-dimensional semantic space (original space reduced using PCA).

k-means (with $k = 2$), on the union of the set of all negated adjectives representations in our model and of their non-negated counterparts (e.g., $\{$*not big*, *not present*...$\} \cup \{$*big*, *present*...$\}$). Given vector representations of items, the algorithm partitions the group in such a way to maximise the similarity within each cluster. The results on internal evaluation show that the algorithm correctly classifies as non-negated or negated adjectives 74% of the data, confirming the observed clustering effect.

As we saw in Chapter 2, adjectives and negated adjectives are undoubtedly different classes of expressions, for which there are syntactic, semantic and pragmatic aspects which we can envisage pulling their placement in the semantic space apart. However, the effect observed here is rather drastic and induced us to consider the possibility that its major cause might be instead related to how the model is constructed, beyond the impact of linguistic features. In particular, as expected, there is a massive difference in the frequencies of negated and non-negated adjectives respectively in the training corpus: while the negated adjectives in our model occur on average around 400 times in the corpus, their related adjectives instead occur on average around 87K times. We then proceed to investigate the role of frequency in the clustering effect, and observe the following:

- Visualising the positions of negated adjectives and their related adjectives in the space, we can see that infrequent adjectives (less than 1K occurrences) tend to fall within the same region of negated adjectives (Figure 3.1). In addition, looking at a sample of frequent and infrequent adjectives only, the latter ones cluster in the space similarly to negated adjectives (Figure 3.2).

- The negated adjectives which are misclassified in the clustering algorithm, and which hence have less similar vectors than the rest of the group, have a much higher average frequency (around 5K) than the general mean. Moreover, applying the clustering algorithm to the dataset with increasing frequency thresholds leads to drops in the performances, suggesting that negated adjectives that occur relatively often in the corpus have distributional representations which are less distinguishable from the ones of adjectives. As it can be observed in Figure 3.1, they are indeed typically at the periphery of the cluster.[3]

- If looking at the same classification of the clustering algorithm and using it to predict whether an expression occurs more or less than 1K times in the training data, rather than whether it is negated or not, we obtain a very similar result (75% of correct classifications).

- There is a positive correlation (Spearman's ρ: 0.41; $p < 0.01$) between the frequency of a negated adjective and its cosine similarity with the original adjective (e.g., *good* to *not good*). We take this value to be generally indicative to how close the former is to the area where other words from the same semantic domain collocates (e.g., how close *not good* is to *good* but also *bad*, *decent*, *excellent* etc.).

- Negated adjectives are typically surrounded by infrequent expressions: the average frequency of an expression in the 20 closest ones to a negated adjective is 956, against the 2241 value obtained for a neighbour of an adjective.

Following these observations, although not excluding at all that also some linguistic features may have a role in this behaviour, we conclude that the major factor that causes the clustering of negated adjectives is actually their lower frequency of occurrence in the corpus data. As a result, not only do they tend to occupy a different region of the space from adjectives and be close to each other, but they also tend to be surrounded by other infrequent items.

This scenario is rather different from the one that is typically encountered studying the semantic relation of antonymy in the semantic space, which is similar, although not identical, to the relation of negation. Pairs with opposite meanings, e.g., *wide* and

[3] The higher frequency of these negated adjectives may be interpreted in terms of lexicalisation. Some of these are indeed almost fixed expressions like *not bad*, or *not familiar*, which arguably behave more like adjectives and have a less context-dependent and compositional meaning. We will come back to this aspect in Chapter 5 when looking at the compositional aspect of these expressions.

narrow, typically appear in similar regions of the space, since, despite their contrasting meaning, they share the same semantic domain and hence many of the contexts in which they occur (Mohammad et al., 2013). Instead, negated adjectives, even if pertaining to the same semantic domain of their original adjectives, tend to locate in a different region, due to a possibly insufficient amount of training data, in comparison to adjectives, to distribute them across the semantic space. However, adding more corpus data to the already large amount we are using would not eliminate this effect, since the disproportion would not be eliminated but only scaled.

Nevertheless, vectors of negated adjectives are far from being random. First of all, the fact that negated adjectives are close to each other is not counter-intuitive: their lower frequency can indeed be seen as an effect of their compositional nature (although this result comes with the cumbersome effect that they are in general close to other infrequent items). Interestingly, as some experiments reported in Chapter 4 show, the relations occurring between items of this class actually tend to replicate the ones holding among their non-negated counterparts, suggesting that the group has a meaningful internal structure. Moreover, despite the "distortion" introduced by the clustering effect, the vectors of negated adjectives tend to still be meaningful and similar to the ones of other words, both negated or not, that belongs to the relevant semantic domain. While we will come back to this topic again in Chapter 4, we present in the following subsection some statistics and examples to support this.

3.3.2 Semantic neighbours

We look here into the *semantic neighbours* of negated adjectives, that is expressions with highest geometric proximity as measured using, in this case, Cosine similarity. These are indicative of the type of meanings captured by negated adjectives, as they are indeed the words predicted to have the most similar semantic content.

Generally, as we saw earlier, negated adjectives tend to have more infrequent semantic neighbours than adjectives. In particular, they tend to have in their proximity more negated adjectives than their non-negated counterpart: the average number of negated adjectives in the 20 closest neighbours of a negated adjective is 3.5, in contrast with 0.4 for an adjective. However, despite this effect, their semantic representation is not "isolated" from the ones of other words in the same semantic domain: 60% of the negated adjectives have among their top 20 neighbours their related adjective and then possibly other similar words to it. On the other hand, the negated adjective is retrieved among the 20 neighbours of its related adjective 20% of the times. As an illustration, we report here the 10 closest neighbours to an adjective and its negation:

(12) COLD: *wet*, *chilly*, *warm*, *dry*, *freezing*, *hot*, *cold*, *frigid*, *not cold*, *icy*

(13) NOT COLD: *not warm*, *not hot*, *cold*, *warmish*, *chilly*, *frigid*, *muggy*, *warm*, *subzero*

Negated adjectives have a diverse behaviour in terms of the orientation exhibited by their semantic neighbours: while in some cases they suggest that the meaning of the adjective has been reversed (i.e., the neighbours are near-synonyms of the antonym), this is not always the case. Consider for example the closest neighbours of these expressions:

(14) NOT DIFFICULT: *not easy*, *not hard*, *difficult*, *impossible*, *easy*

(15) NOT EASY: *difficult*, *not difficult*, *hard*, *impossible*, *not hard*

In the case of *not difficult*, the list figures expressions that pertain to the scalar dimension of difficulty, although not pointing at a complete flip in meaning towards the opposite. On the other hand, the neighbours of the negation of the antonym, namely *not easy*, suggest a more substantial meaning shift operated by negation along the scalar dimension. Similar effects also occur with contradictory pairs, where the meaning flip is expected to be complete: for example, while the closest expression to *not present* is indeed the antonym *absent*, there does not seem to be the same reversal of meaning for *not absent* (its closest neighbour is still *absent*).

In general, negated adjectives have the tendency to have a strong similarity with the adjective that they were derived from (48% of the negated adjectives have it among the top 5 neighbours). As we will see later on in Section 4.2, this tendency is even stronger than the patterns registered instead with the antonym. This aspect may seem to contrast with the idea of the meaning shift operated by the negation on an adjective towards the antonym, especially for those cases where a stronger effect is expected (e.g., contradictory pairs). However, the phenomenon is actually aligned with the idea of Giora et al. (2005) that negation does not eliminate the negated concept, but instead retains a special relationship of accessibility with and emphasis on it. It should not then come as a surprise that the two are very similar, although it is interesting that distributional information often captures their non-trivial association.

We conclude from the qualitative analysis of a sample of semantic neighbours that the quality of the distributional representations of negated adjectives is generally adequate for the descriptive purposes of our analyses. They indeed reflect sensible expectations about their semantic content: they are similar to other expressions from the same semantic domain, both negated or not, and capture a particular connection with the adjective that they negate.

3.4 Semantic neighbours as alternatives

As shown by Kruszewski et al. (2017), DS can be employed to identify plausible alternatives introduced by a negative statement. We here take a similar approach and often interpret cosine similarity as a measure of the plausibility of an alternative to a negated adjective, and hence the semantic neighbours as the most plausible alternatives. The previous work focused on ranking the plausibility of alternatives to a noun introduced by negation (e.g., *There is not a dog here, there is a* {*cat, elephant, chair...*}.), and was mostly successful in the task by looking at the geometric proximity between the noun itself and the presented candidate. Simple semantic similarity scores between terms were then exploited, without resorting to modelling the negation of the term directly as an independent item.

In our case, focusing on adjectives, we can consider the following sentence pattern: *This is not X, it is Y*, whereas X and Y are adjectives and our goal is to rank candidate terms for Y. For example:

(16) This is *not fast*, it is...
 a. *slow*
 b. *medium-paced*
 c. *expensive*
 d. *blue*
 e. ...

Clearly, the type of alternatives, as well as the meaning, of an adjective are dependent on the noun phrase it modifies (e.g., *This car is not fast* vs. *This service is not fast*). However, our account summarises the diverse semantic behaviour of an adjective and a negated adjectives into unique representations that abstract way from the various noun phrases it might be associated to. While this is a simplification over the semantics of the adjective, it still allows us to look into its general meaning and typical alternatives independently of the particular context it is used in.

We consider two viable options for achieving this: on one side, in line with the work by Kruszewski et al. (2017), we can make use of the semantic similarity between X and Y (e.g., *hot - lukewarm*); on the other, having constructed distributional representations of negated adjectives, we can directly look at the semantic similarity between *not* X and Y (e.g., *not hot - lukewarm*). In order to obtain plausible alternatives candidates to a certain adjective a, we can then query model for the closest adjectives to either a or *not a*. While in some cases the two approaches may lead to similar results, in others either the terms retrieved or their ordering can instead be quite different. Considering the sentence in (16), for example, we obtain the following options for the top 5 alternative:

(17) FAST: *slow*, *quick*, *super-fast*, *high-speed*, *rapid*

(18) NOT FAST: *slowish*, *slow*, *manoeuvrable*, *medium-paced*, *super-fast*

The neighbours of an adjective, even if possibly consisting of items which similar orientation to the antonym, tend to be often synonyms, and hence fall into close points in its scalar dimension (i.e., express a similar degree of the property). Instead, the closest terms to the negated adjectives may reflect the meaning shift operated by negation and, as a result, express different degrees of the property. Due to the focus of our study on mitigation effects, we then investigate the potential of the second approach.

Conclusion

In this chapter, we gave the motivation for studying negated adjectives using their observed distributional representations as a unit, rather than as a compositional phrase. After having described how we construct such representations, we proceeded to give a first overview of their behaviour in terms of the global and local properties of their location in the semantic space. The vector representations of negated adjectives, although affected by low-frequency effects that gather them in a certain region of the space, are meaningful and reveal an interesting behaviour with respect to the type of semantic neighbours that they exhibit. Moreover, we presented an interpretation of semantic similarity relations as predicting the plausibility of alternatives introduced by the negation of an adjective.

The experiments that we will present in the following chapter aim at further clarifying these observations and ideas by exploring how the distributional model captures certain properties and phenomena of the negation of adjectives.

Chapter 4

Negation of adjectives in the semantic space

In this chapter, we present experiments aimed at investigating properties of negated adjectives looking at their large-scale use as captured by their observed vectors in a distributional semantic model. In our investigations, we use operations in the semantic space in order to carry out analyses that involve various aspects of negated adjectives, namely the meaning shift towards the opposite meaning (Section 4.2), and the interaction with a scalar dimension (Section 4.3). In order to explore these phenomena, we make use of external datasets of adjectives, which we describe in Section 4.1.

4.1 Datasets of adjectives

4.1.1 Antonyms

Some of our experiments require us to work on adjective pairs with opposite meanings (e.g., *open* - *closed*, *good* - *bad*), namely antonymic adjectives. For this purpose, we use adjectival pairs tagged as direct antonyms in WordNet (Fellbaum, 1998), which are individuated within the lexical resource as lexemes with clear opposite meanings, and which, differently from indirect antonyms, are psychologically salient and have a strong associative bond due to their frequent co-occurrence (e.g., *large* - *small* vs. *large* - *minuscule*). In particular, we make use of the Dictionary of Lexical Negation built by van Son et al. (2016), which is based on an annotated subset of the above-mentioned word pairs.

In this dataset, word pairs of direct antonyms are tagged following the categorisation by Joshi (2012), which we mentioned in Chapter 2. He takes lexical negation to include both affixal negations (e.g., *perfect* - *imperfect*), and regular antonyms (e.g., *wet* - *dry*), whereas the former group is in turn split up in direct and indirect negation, depending on whether the meaning of the affixed word is a direct antonym of the non-affixed counterpart (e.g., *clear* - *unclear* vs. *famous* - *infamous*). The Dictionary of Lexical Negation was built taking all pairs of direct antonyms from WordNet (nouns, adjectives and verbs) and then annotating these with information about whether they contain an affixal negation or they are regular antonyms, and, in the former case, whether it is a direct or an indirect one.

Given the focus of our experiments, we extract from the dictionary only adjective pairs, and, among these, those with a direct affixal negation on one side (which we refer to here as *affixal antonyms*), and those with a regular antonym on the other (*regular antonyms*). This gives rise to a dataset of 900 affixal and 620 regular antonymic pairs respectively, which will be further filtered on the basis of the coverage of our model for the adjectives and their negated counterpart.

As we saw in Chapter 2, this categorisation was proposed with the view that the differences between affixal direct negation and regular antonymy are merely morphological (the former is marked by a negative affix), and that they both are linked to a corresponding adjective by the relation of antonymy. However, in our experiments, we do not assume the sameness of the relation they express, but, instead, keep the two groups separate and check whether they exhibit the same behaviour with respect to negated adjectives across tasks.

4.1.2 Adjectival scales

In the case of scalar adjectives, i.e., those that express the degree of a property on a scale (e.g., *warm* expresses a value in the scale of TEMPERATURE) (Kennedy and McNally, 2005), negation can be seen as operating a meaning shift along the dimension of the scale itself. In order to analyse the behaviour of negation with respect to scalar dimensions, we make use of the golden dataset of adjective scales built by Wilkinson and Tim (2016).

This dataset consists of 12 adjective scales; its content was collected through elicitation tasks with the goal of producing a benchmark for tasks such as learning scalar relationships (for example, the work by Kim and de Marneffe (2013)). Through crowdsourcing, both the members and their ordering are collected for the scales of SIZE, DRYNESS, INTELLIGENCE, QUALITY, AGE, SPEED, DIFFICULTY, QUANTITY, BRIGHTNESS, SAMENESS, BEAUTY and TEMPERATURE. Each scale consists of 5 ordered adjectives on average. For instance, Figure 4.1 shows the scale constructed for the domain of SIZE:

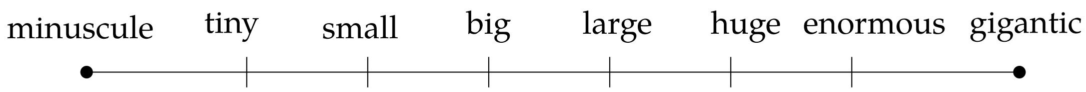

Figure 4.1: Scale of SIZE in the dataset by Wilkinson and Tim (2016)

In our experiments, we compare the distributional representation of a negated scalar adjectives with those of other members of the scale (e.g., *not small* in the scale of SIZE). It is, however, to be noted that, while the dataset gives us a golden standard for the ordering of adjectives, we do not have a benchmark, nor a clear-cut theoretical prediction, about where in the dimensional scale the negated adjectives should locate. One of our goals is to investigate this empirically with our analyses.

4.2 Antonyms and meaning shift

4.2.1 Motivation and set-up of the experiments

According to the mitigation hypothesis (Giora, 2006), when one member of a pair of antonymic adjectives is negated, it conveys a mitigated meaning which is intermediate between the ones of the two adjectives (e.g., *not small* $\approx$ *medium-sized*). The extent of the mitigation was shown to be dependent on, besides the context of utterance, a series of intrinsic properties of the adjective that is negated. We here present analyses aimed at assessing whether and how this aspect may be captured by a distributional model. In particular, we conceptualise the mitigation of sense operated by negation on an adjective as a shift in the meaning of the original adjective towards the one of the antonym.

Given a pair of antonymic adjectives $a_1 - a_2$, we look into the difference in meaning between both *not* a_1 and a_2, and *not* a_2 and a_1. For each antonymic pair, we then consider the triplet $(a_1, a_2, \textit{not } a_1)$ and $(a_2, a_1, \textit{not } a_2)$. We look at the differences in the meaning shift across different classes of these triplets, categorised as in Figure 4.2:

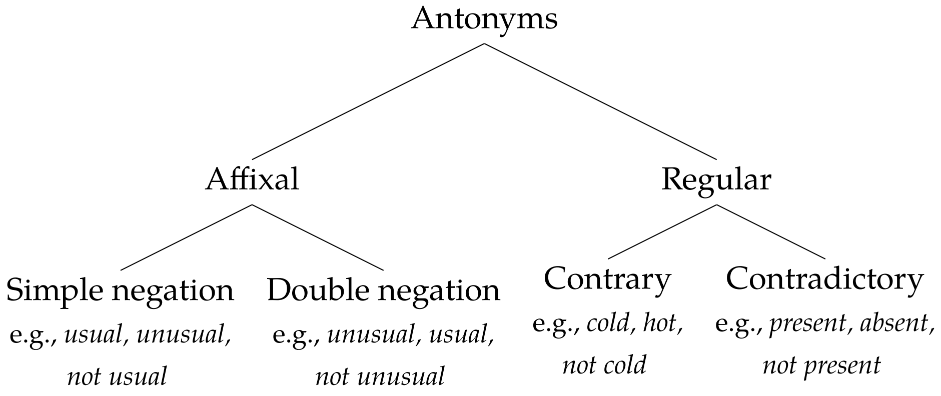
Antonyms
Affixal
Regular
Simple negation
e.g., usual, unusual, not usual
Double negation
e.g., unusual, usual, not unusual
Contrary
e.g., cold, hot, not cold
Contradictory
e.g., present, absent, not present

Figure 4.2: Categorisation of the triplets consisting of an antonymic pair and a negated adjective employed in our experiments.

Affixal and regular antonyms We first compare the behaviour of affixal and regular antonyms, by looking at the two differently annotated groups in the dataset of lexical negation that we described in Section 4.1.1. As we saw, the categorisation by Joshi (2012) assumed the two groups to be different only in morphological terms, and equating the meaning of the negation of an adjective to the one of the antonym, be it derived by affixation or a having a distinct lexical root. We here reconsider this assumption by comparing the two groups at the level of how close the negation of one of the members of the pair is to the one of the antonym, e.g., how close is *not hot* to *cold* or *not frequent* to *infrequent*. We are interested, on one side, in assessing whether negating by *not* or by a negative affix is in effect the same thing, and, on the other, whether the former can be equated to the relation of antonymy (i.e., whether negation flips the meaning of an adjective into its opposite).

Simple and double negation In the case of affixal negation, the procedure of looking at the meaning shift of the negated counterpart of both the adjectives leads us to also look into cases of double negations, such as *not unusual*: indeed, the meaning of *usual* is here modified first by the negative affix *un-* and then by the particle *not*. Whether the effect of this type of double negation is, as in logical negation ($\neg\neg p \rightarrow p$), a nullifying one (i.e., the meaning of the doubly negated adjective is the same as the adjective) depends on the extent of the meaning shifts induced by the two types of negations: only if both flip the meaning into the opposite one, it can be expected that they cancel each other out; otherwise there would be a sum of mitigation effects. In our analyses we separate the cases of *simple* (e.g., *not similar*) and *double* negation (e.g., *not dissimilar*) within the triplets containing an affixal negation, and look at their behaviour individually.

Contrary and contradictory antonyms As we saw, Fraenkel and Schul (2008) have provided experimental evidence indicating that whether an antonymic pair consists of contrary or contradictory adjectives is one of the factors that determine how much the meaning of a negated adjective is closer to the one of the antonym. Contrary pairs (e.g., *hot* - *cold*) are adjectives whose meaning lies in a continuum, whereas intermediate meanings along this dimension are possible (e.g., *lukewarm*); contradictory pairs (e.g., *present* - *absent*), instead, constitute a dichotomy such that the falsity of one implies the truth of the other. In line with the mitigation hypothesis, the extent of the meaning shift of an adjective towards the antonym was shown to be bigger when a *tertium* outside the antonymic pair is not available (e.g., one cannot be present and absent at the same time) than when there is a possible middle point.

In order to check whether distributional representations of negated adjectives replicate this experimental result, we partition regular antonyms into contrary or contradictory through an annotation procedure. The task was structured following the definition used by Fraenkel and Schul (2008): a pair of adjective $a_1 - a_2$ is tagged as contrary if the sentence pattern "*X is neither* a_1 *nor* a_2" is acceptable in a default context, and as contradictory otherwise (see Appendix A for more details). Indeed, while contrary adjectives can be used in two sentences that cannot be simultaneously true but can be simultaneously false (e.g., *The tea is neither hot nor cold*), in the case of contradictory pairs one of the two sentences must be true (e.g., *The student is either present or absent.*). In the annotation task, both the context and the subject of the sentence pattern are left unspecified, in order to obtain an intuitive judgement about the general relationship between the adjectives. For example:

(19) *hot* - *cold*: X is neither hot nor cold ⇝ contrary

(20) *present* - *absent*: ? X is neither present nor absent ⇝ contradictory

The annotation was carried out by three independent annotators on a list of 148 adjective pairs,[1] and achieved a moderate inter-rater agreement (Fleiss' $k = 0.37$). Given

[1] The annotation is applied to a subset of regular antonyms, taking into account the frequency threshold of 100 counts imposed on negated adjectives in the experiments.

these results, for the purpose of our experiments, we take into account only those antonymic pairs whose annotation the three raters had full agreement on (48 contrary and 20 contradictory pairs).

We find the low agreement level achieved not surprising given the complexity of the task and the simplification of the classification: rather than simply separable into two categories only, adjective pairs can be seen as lying in a continuum, and possibly be used as one or the other type depending on the context. Already Fraenkel and Schul (2008) noted that what they consider contradictory pairs may not be entirely dichotomous, as one may always find possible realistic interpretations of a negated member as a meaningful mid-value (e.g., *not dead* ≈ *half-dead*) (Paradis and Willners, 2006).

4.2.2 Methods

In our experiments we look into the meaning shift associated with negation from different angles and using different methodologies: while our main study consists of quantifying it in terms of similarity in the space, we also complement our observations with an investigation of its effects on the retrieved semantic neighbours of a negated adjective and its nature as a regularity in the space.

Similarity in the space The cornerstone of distributional semantic methods is the quantification of similarity between two expressions as geometric proximity between their vectors. It thus gives us a natural tool to measure the extent of the meaning shift of the negation of an adjective, by looking into the distance of its vector from the one of the antonym or the adjective itself.

In order to observe the phenomenon of meaning mitigation in the distributional space, we then define the following measures:

$$Sim(not\ a_1, a_2) := CosSim(not\ a_1, a_2) \tag{4.1}$$

$$Shift(not\ a_1, a_2) := CosSim(not\ a_1, a_2) - CosSim(not\ a_1, a_1) \tag{4.2}$$

$$Flip(not\ a_1, a_2) := [Shift(not\ a_1, a_2) > 0] \tag{4.3}$$

where $CosSim(a, b)$ stands for the cosine similarity between the vectors of a and b, a_1 and a_2 are antonymic adjectives. $Sim(not\ a_1, a_2)$ (4.1) measures how close the distributional vectors of a negated adjective is to the antonym (e.g., how similar *not hot* is to *hot*); being a cosine value, its range is [-1,1]. Instead, $Shift(not\ a_1, a_2)$ (4.2) looks at the difference between these values and measures how closer the negated adjective is to the antonym than to the adjective. The range of the value is [-1, 1]; it is positive when the similarity with the antonym is higher than the one with the adjective, and negative otherwise. Finally, $Flip$ (4.3) indicates whether the negated adjective is closer to the antonym than to the adjective, i.e., $Shift(not\ a_1, a_2) > 0$.

Neighbours of negated adjectives We look into the closest expressions of a negated adjective and compare these to the ones of the adjective itself and of the antonym. In particular, we consider a filtered set of neighbours: we query the model for the 20 closest adjectives which occur in the training corpus more than 100 times.[2] This choice allows us to look into the meaning of negated adjectives in terms of other adjectives: we find that while also other types of expressions may be appropriate semantic neighbours, this restriction allows us to focus on the paradigmatic class that the negation of adjectives applies to, and prevents us from circularly studying the meaning of negated adjectives in terms of other negated adjectives. Moreover, as we saw, this type of neighbours can be interpreted as a plausible alternatives set.

We analyse this filtered set in terms of intersection with the ones of the adjective and of the antonym, and of the relative frequency of the adjective and the antonym appearing in such a set.

Negation as regularity in the space While these measurements allow us to quantify the meaning shift and assess its extent in different categories, we are also interested in answering the following question: does the relation of negation and the meaning shift induced by it correspond to a regularity in the semantic space?

Distributional models have been shown to capture various types of syntactic and semantic relations, beyond the one of similarity, in the way word vectors are located in the space (Mikolov et al., 2013b): regularities are observed in terms of constant vector differences (*offsets*) between the representations of two words linked by a specific relation. For example, the plurality and the gender relations can respectively be captured by the facts that in the semantic space $cake - cakes \approx book - books$ and that $man - woman \approx king - queen$. Such a property can be exploited by setting up analogy tasks in the form of $a : b = c : d$ where d is to be retrieved exploiting the sameness of the relations occurring between a, b and c, d respectively.

We apply this methodology (Mikolov et al., 2013b; Levy et al., 2014) to check to which extent the relation of negation of an adjective (e.g., *beautiful* - *not beautiful*) is captured as a regularity in the semantic space. In particular, given a pair of antonymic adjectives a_1, a_2, we consider the analogy:

$$a_1 : not\ a_1 = a_2 : not\ a_2 \tag{4.4}$$

and aim at predicting a_2 given the other members of the analogy in the following way:

$$v_x = v_{a_2} - v_{a_1} + v_{not\ a_1} \tag{4.5}$$

$$x^* = argmax_w\ CosSim(v_x, w) \tag{4.6}$$

[2] The frequency threshold allows us to reduce the noise introduced by the data sparsity effects. We limit the search of adjectives at the 300 closest neighbours.

In (4.5), we compute the expected member of the analogy by taking the vector offset between a_1 and a_2, and add *not* a_1 to it; since no expression in the vocabulary might be in the exact position expressed by this vector, our prediction is then the expression in the model with the closest cosine similarity to it, as in (4.6).

4.2.3 Results

Similarity in the space

The results of the similarity analyses are presented in Table 4.1.[3] We focus in particular on $Shift$ and $Flip$ results, as they highlight the relationship of the negated adjectives with both the adjective and the antonym. We also report some examples of $Shift$ values in Table 4.2.

	Affixal	Regular		Simple	Double		Contraries	Contrad.
$Sim(not\ a_1, a_2)$ M	0.43	0.26	$***_t$	0.44	0.37	$*_t$	0.24	0.30
$Shift(not\ a_1, a_2)$ M	-0.04	-0.19	$***_t$	-0.03	-0.06		-0.18	-0.19
$Flip(not\ a_1, a_2)$	41%	10%	$***_\chi$	43%	25%		12%	7%
Datapoints	185	198		157	28		68	28

Table 4.1: Results of the similarity analyses on negated adjectives from different groups (M: mean value). Each negated adjective is compared to its adjective (a_1) and antonym (a_2); differences are tested for significance using Welch's t-test (t), Chi-squared test (χ) and Fisher's exact test (f) (*: $p < 0.05$: **: $p < 0.01$; ***: $p < 0.001$).

$Shift$	
happy, unhappy, not happy	0.31
unhappy, happy, not unhappy	-0.14
dead, alive, not dead	-0.09
small, large, not small	-0.09

Table 4.2: $Shift$ values of a sample of antonyms and negated adjectives triplets.

As can be seen, all the average values of $Shift$ are negative, meaning that in general negated adjectives tend to be closer to the adjective than the antonym. An account of negation which assumes it to flip the meaning of an adjective into its antonym would instead expect these values to be positive across all classes. This suggests that negated adjectives tend to be used more often in contexts in which the respective adjective would occur, rather than those that the antonym would.

[3] In all the analyses in this thesis, we use Welch's t-test for comparing populations, and Chi-squared test and Fisher's exact test for comparing sets of categorical data. In the second case, we use Fisher's exact test only when at least one value in the contingency table is smaller than 5.

We observe the following results with respect to the different categories that we considered (Figure 4.2):

- **Affixal and regular antonyms**:
  We observe a strong variation of behaviour across affixal and regular antonyms: negated adjectives are on average closer to an affixal antonym (e.g., *not perfect - imperfect*) than to a regular antonym (e.g., *not wide - narrow*), as reflected in the $Flip$ and $Shift$ values.

  This suggests that it might not be appropriate to consider these two categories as a unique and coherent one: there is, in fact, more similarity between the negation by *not* and by affix, than there is between the former and the relation of regular antonymy. This could be due to the similar compositional structure of negated adjectives and affixal negations: despite of the fact that one is a multi-word expression and the other is not, they both consists of a negative item (*not* and an affix such as *un-*, or *dis-*) and an adjective (e.g, *perfect* in *not perfect* and *imperfect*). Both may then preserve the emphasis on this, as well as express a mitigated version of its meaning. However, the two negative items, although similar, do not seem to be identical at the level of how they are used (the negated adjectives are still on average closer to the adjective than to the antonym).

- **Simple and double negation**:
  Although there is not a significant difference in terms of $Shift$ and $Flip$ for simple and double negations, the latter ones are significantly further from their antonym than simple ones (e.g., *not usual* is closer to *unusual* than *not unusual* is to *usual*).

  It is known that double negations are often used to attenuate assertions (Horn, 1989). These expressions, which are less natural and more complex than the non-negated counterpart (e.g., *not unlikely* vs. *likely*), tend to then appear in contexts where the use of the corresponding adjective would be too strong or too direct, such as cases of understatement, hesitation, or irony. This function of double negation may justify the here observed difference. What is clear, in general, is that the picture is not as simplistic as in classical logical negation: the two negative items do not doubly flip the meaning of the adjective into the opposite one, nullifying each other effect.

- **Contrary and contradictory antonyms**:
  In contrast with the difference in meaning shift expected from the literature, the behaviour of contrary (e.g., *small - large*)and contradictory pairs (e.g., *dead - alive*) is not different with respect to negated adjectives. For both classes, the distributional representations of the negated adjective are generally not very close to the one of the respective antonym, but intermediate between this and the one of the adjective they negate, and actually closer to the latter.

  One way to explain this effect could be to imagine that, although the conceptual dichotomy between the meanings of contradictory adjectives and the lack of a *tertium*, the contexts of use discriminate between negating a contradictory adjective or employing its antonym. Even if possibly pointing at the same semantic content, the contexts of use of *not dead* may be quite different from those of *alive* and actually more similar to those of *dead*, due to the various functions of negation (e.g., contradicting an expectation, understatement). Adding this result to the low agreement in the annotation task, it generally looks like these categories are generally more complex than predicted in the literature.

The results we obtained suggest a picture in which negated adjectives are not equivalent to antonyms, being them affixal or regular, and contrary or contradictory. As noticed by Horn (1989), effects of this kind can be put in relation with theories like the *Avoid Synonymy* principle (Kiparsky, 1982; Clark, 1992): at their core, these express the general tendencies in the lexicon not to allow a certain semantic slot to be filled by two expressions, and hence a "force to diversification" of meanings. In particular, Horn (1984) connects this phenomenon to, on one side, the countervailing tendency to simplification and minimisation of the lexicon (*Principle of Least Effort* by Zipf (1949)), and the theory of conversational implicatures of Grice (1975), on the other. He proposes that when a speaker opts for a more complex or less fully lexicalised expression over a simpler alternative, such as a negated adjective over an antonym, there is always a sufficient, although possibly different, reason (a phenomenon that he calls *division of pragmatic labour*). This could be for example the necessity of mitigating the meaning of an adjective rather than flipping it (21), but also the need of retaining the emphasis on a rejected concept (22), or attenuating the strength of a statement (23).

(21) This sandwich is *not bad* (vs. *good*) - it is decent.

(22) The plant that I forgot to water for a month is *not dead* (vs. *alive*).

(23) It is *not impossible* (vs. *possible*) that it will be sunny tomorrow .

Our results suggest that distributional representations of negated adjectives capture these differences in use, due to which they are different from both affixal and regular antonyms, and, in the case of double negations, from the original adjectives.

Neighbours of negated adjectives

	Affixal	Regular		Simple	Double	Contraries	Contrad.
a_1 in the 20 closest adj.	88%	89%		89 %	86%	81%	96%
a_1 as closest adj.	19%	19%		21%	11%	16%	32%
a_2 in the 20 closest adj.	70%	33%	$^{***}_{\chi}$	72 %	57%	25%	43%
a_2 as closest adj.	18%	7%	$^{**}_{\chi}$	18%	18%	7%	4%
Shared with a_1 M	4	4.1		4.1	4	3.8	4.8
Shared with a_2 M	4	1.8	$^{***}_{t}$	4.2	3	1.7	2.3
Datapoints	185	198		157	28	68	28

Table 4.3: Results of the analyses on the 20 closest adjectives to negated adjectives from different groups (M: mean value). Each negated adjective is compared to its adjective (a_1) and antonym (a_2); differences are tested for significance using Welch's t-test (t), Chi-squared test (χ) and Fisher's exact test (f) (*: $p < 0.05$: **: $p < 0.01$; ***: $p < 0.001$).

Table 4.3 shows the results of the neighbours analysis. The results for affixal and regular antonymy are aligned with those obtained during the quantification of the meaning shift. Negated adjectives have more often the associated affixal antonym in their proximate, and tend to share more neighbours with it than a regular antonym. Moreover, double negations are typically further away from their antonym. We also do not find any significant differences again between contraries and contradictory pairs with respect to their relation to the antonym. To illustrate and clarify the results, we report in Table 4.4 the 10 closest adjectives to a sample of negated adjectives.

As can be seen, the closest adjectives tend to be from the appropriate semantic domain; their meaning orientation ranges from the same (e.g., *not dead* - *lifeless*) to the opposite (e.g., *not small* - *large*), and, in particular, they often express an intermediate meaning between the one of the adjective that is negated and its reversal. For example, *not happy* is close to adjectives like *pleased*, *dissatisfied*, and *unimpressed*, which express a mitigated, but not opposite, sense of *happy*; similarly *not small* is close to many terms which expresses an intermediate size between *small* and *large* (e.g., *smallish, normal-sized*). An analogous behaviour can also be observed in contradictory regular antonyms, as can be seen in closest adjectives to *not dead* such as *half-dead* and *comatose*. What this suggests is that DS captures the fact that, at the level of the use, even dichotomies may actually be interpreted as continua (as pointed out already by Fraenkel and Schul (2008) when discussing the results of their comparison). Indeed, distributional methods, by representing expressions in a continuous fashion based on differences in use, are able to represent very nuanced meanings that arise due to subtle differences in distributions.

Adj.	Ant	Closest adjectives to *not* + adj.	Shared with adj.	Shared with ant.
happy	*unhappy*	*unhappy, unsatisfied, unsure, disappointed, dissatisfied, adamant, unimpressed, happy, annoyed, pleased*	*unhappy, pleased*	*disappointed, dissatisfied, unsatisfied*
unhappy	*happy*	*unhappy, adamant, disappointed, dismayed, unimpressed, relieved, pleased, gobsmacked, homesick, fed-up*	*disappointed*	*unhappy, pleased*
dead	*alive*	*dead, half-dead, alive, comatose, lifeless, asleep, drowned, unburied, reborn, grif-stricken*	*drowned, lifeless, half-dead, alive, unburied*	*dead, reborn*
small	*large*	*small, smallish, normal-sized, largish, middle-sized, large, medium-sized, big, minuscule, mega*	*large, smallish, minuscule*	*small, smallish*

Table 4.4: Closest adjectives to a sample of negated adjectives.

Negation as a regularity in the space

The results of the analogy task are reported in Table 4.5 both in terms of absolute precision, i.e., percentage of correct predictions of the negated adjective as the closest neighbour to the expected vector, and precision at k, i.e., percentage of correct predictions of the negated adjective as one of the top k neighbours of the expected vector. We only report values for affixal and regular antonyms, due to an insufficient number of datapoints for the other classes.[4]

	Affixal	Regular	
Precision	11%	20%	
Precision at 5	18%	57%	$^{**}_{\chi}$
Precision at 10	25%	66%	$^{**}_{\chi}$
Precision at 20	39%	71%	$^{**}_{\chi}$
Datapoints	44	100	

Table 4.5: Results of the analogy task $a_1 : \mathit{not}\ a_1 = a_2 : \mathit{not}\ a_2$ tuples for antonymic pairs a_1, a_2 with affixal and regular antonyms; differences are tested for significance using Chi-squared test (*: $p < 0.05$: **: $p < 0.01$).

The search space for this task is the entire vocabulary of the model (719K), meaning that a random baseline would have very low precision scores. Taking this into account, we can see that the model works quite effectively, suggesting that indeed negation is at least partially captured as a regularity in the space.

Regular antonyms retrieve the correct negated adjectives with this method substantially more often than affixal ones (more than half of the times when looking at the top 5 neighbours). This might be put in relation with the results of the meaning shift study: the fact that the negation of an adjective tends to be further from its regular antonym than from an affixal one suggests a different configuration of items in the space that perhaps allows for a better performance in the analogy task.

Moreover, it is important to notice that the analogy task that we carried out, namely $a_1 : \mathit{not}\ a_1 = a_2 : \mathit{not}\ a_2$ where $\mathit{not}\ a_2$ is unknown, is equivalent to trying to predict $\mathit{not}\ a_2$ with the analogy $a_1 : a_2 = \mathit{not}\ a_1 : \mathit{not}\ a_2$ (e.g., *good* : *bad* = *not good* : *not bad*).[5] This suggests that what could be captured as a regularity is not, or at least not only, the negation of an adjective, but the relation of antonymy, which would be captured as a constant offset both when linking adjectives (e.g., *good* - *bad*) and when linking negated adjectives (e.g., *not good* - *not bad*). This suggests that the relations occurring among adjectives have partially isomorphic counterparts in the way negated adjectives collocate in the space.

[4] We carry out the task only on antonyms pairs such that the negated adjectives of both the words occur more than 100 times.

[5] The formula that leads to the predicted vector for $\mathit{not}\ a_2$ is the same: $v_x = v_{a_1} - v_{a_1} + v_{\mathit{not}\ a_2}$.

4.3 Negated adjectives in the scale

4.3.1 Motivation

In the previous section, we assumed the meaning shift of negation to occur within a bipolar dimension defined by an antonymic pair (e.g., *good* - *bad*). We here look instead into the interaction of negated adjectives within a scalar dimension populated by more than two adjectives. In particular, we make use of the adjective scales collected by Wilkinson and Tim (2016) and presented in Section 4.1.2.

Consider the scale for QUALITY in Figure 4.3:

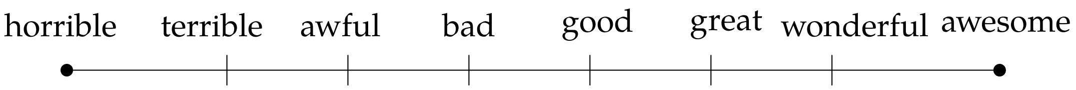

Figure 4.3: Scale of QUALITY in the dataset by Wilkinson and Tim (2016).

Where would *not good* or other negated members of the scale fall within such a scale? While the answer to this question is often context-dependant, we investigate whether distributional representations capture the main trends in this respect. Besides to its antonym, we compare here the negation of a scalar adjective to also other members of its scale, which lexicalise different degrees of the negated property (e.g., comparing *not good* to *terrible*).

Moreover, we were previously assuming the negation of an adjective to occur only in the direction of the antonym: in this case, *not good* could then only express the same or an intermediate meaning between *good* and *bad*. Nevertheless, a negated scalar adjective *a* is actually pragmatically ambiguous between the readings *less than a* or *more than a* (Horn, 1984), although the former interpretation is taken to be the default one:

(24) This cake is *not good*.

a. It is *okay* (less than good).

b. It is *extraordinary* (more than good).

(25) This cake is *not bad*.

a. It is *decent* (less than bad).

b. It is *awful* (more than bad).

While in (24a) and (25a) negation makes the meaning of the adjective shift towards the antonym (i.e., respectively, from *good* towards *bad*, and from *bad* towards *good*), in (24b) and (25b) the negated adjective expresses a degree of quality which is instead further away from the antonym. This pattern is however achieved asymmetrically: while *not good* proceeds towards left in the scale in Figure 4.3 to get close to the antonym, *not bad* proceeds towards right.

This phenomenon is due to the *polarity* of the adjective that is negated: some adjectives (typically, at the right side of the scale) express a positive value of the property (e.g., *good, great, wonderful, awesome* in the scale of QUALITY), while others express a negative one (e.g., *bad, awful, terrible, horrible*) (Kennedy, 1999; Sassoon, 2010). For this reason, *less than good* expresses a lower degree in the scale of quality, while *less than bad* expresses instead a higher degree; we obtain the opposite pattern if, instead, considering *more than good* and *more than bad*. This asymmetric effect has then direct consequences on the way that negated adjectives are disambiguated.

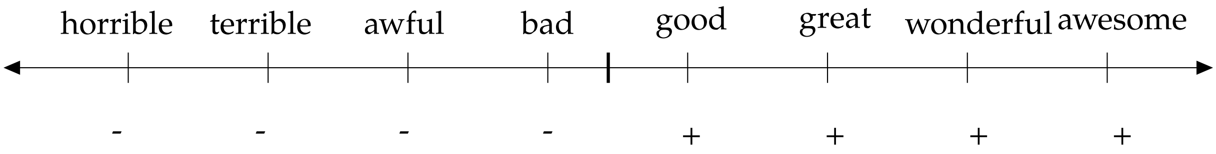

Figure 4.4: Scale of QUALITY in the dataset by Wilkinson and Tim (2016) represented with two orientations and polarity information.

The phenomenon of polarity is typically accounted for representing a scale as bidirectional: positive and negative adjectives make use of the same dimension, but differ in the direction of their ordering (Sapir, 1944; Kennedy, 1999) (see the example in Figure 4.4). However, the scales we work with here do not have information about the polarity of the adjectives and their consequent orientation, although we do take into account this aspect in the interpretation of our analyses.

In our experiments, we look at the scale members that the distributional representation of a negated adjective is close to: in particular, our research questions concern whether a particular tendency towards one of its possible readings (*less than a* or *more than a*) is captured, and whether we can provide an interpretation of its neighbours as plausible alternatives. We approach this aspect from different perspectives, evaluating the negated versions of the adjectives in the scales by Wilkinson and Tim (2016). We first design a method of producing an ordering relation over scalar expressions using cosine similarity and apply it to build scales with negated adjectives (Section 4.3.2). Secondly, we look into the semantic neighbours of negated scalar adjectives and check whether and which members of the scale are retrieved (Section 4.3.3).

4.3.2 Ordering the scales with negated adjectives

As a first attempt to account for the placement of negated adjectives in a scale, we design a general and simple procedure to order a set of adjectives that are members of a scalar dimension. Given a list of adjectives which constitutes a scale S and the highest pole p of such a scale (i.e., the right-most expression; e.g., *gorgeous* in the scale of BEAUTY in Figure 4.5), we order S in ascendant order of similarity from p; we then take such an ordered set to be our expected adjectival scale.[6]

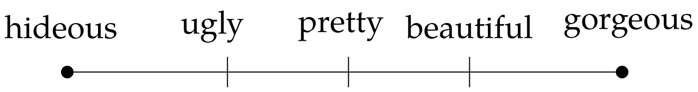

Figure 4.5: Scale of beauty in the dataset by Wilkinson and Tim (2016).

As a first step, we test this ordering method on the dataset of adjective scales: we order each scale by taking its list of members and the highest pole, and then compare the correct ordering with the derived one. For example, Figure 4.6 shows the fully correct results obtained for the scale of beauty given its exemplars and the similarity scores to the highest pole:

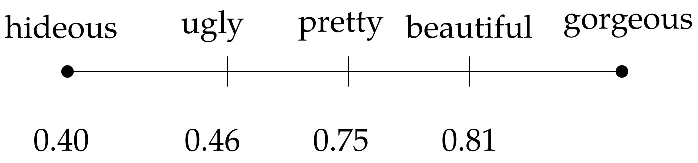

Figure 4.6: Scale of beauty as predicted by the cosine similarity ordering from the right-most adjective.

As a general evaluation of the method, we look at the overall correlation with the golden scales: we respectively concatenate the lists we build and the ones in the benchmark (excluding the highest poles in each scale, which are trivial cases for our method) and compare them in terms of Spearman's ρ. We obtain a value of correlation with the dataset of 0.64 ($p < 0.01$), which confirms the relatively good quality of this ordering procedure. This result is aligned with previous results, such as those by Kim and de Marneffe (2013), who have shown that distributional space capture relations of scalar membership and ordering of adjectives.

We then proceed to exploit this method to analyse negated adjectives in the scale. For each scale, we produce lists of members of a scalar dimensions by adding to the adjectives their negated counterparts.[7] We then apply the ordering procedure employed before and analyse the obtained scales. Recall that we do not have a benchmark for this kind of scalar dimensions, beyond intuitively expecting negated adjectives to fall between the two extremes of the scale (e.g., *not ugly* between *hideous* and *wonderful*). We here report two examples of scales obtained in this fashion in Figures 4.7 and 4.8.

Employing this method, negated expressions tend to appear in the left-most part of the scale: the one of temperature is an instance of this pervasive phenomenon. Because

[6] The method achieves comparable results using the lowest pole (i.e., the left-most expression; e.g., *hideous* in the scale of beauty) as p and the descendant order of similarity from p as an ordering relation. However, to simplify, we only report here the results obtained using one of the methods.

[7] We consider only adjectives that appear 100 times in the training corpus.

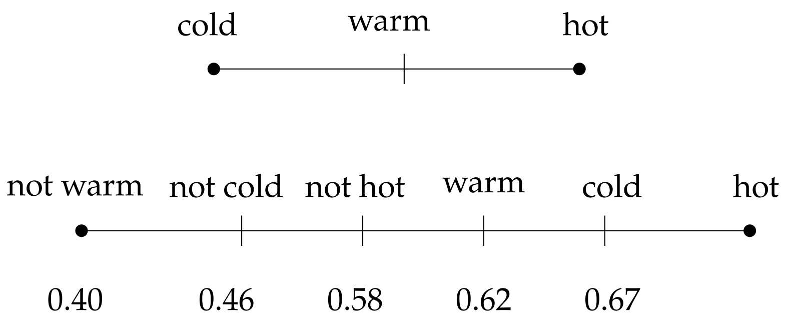

Figure 4.7: Scale of TEMPERATURE with negated adjectives as predicted by the cosine similarity ordering from the right-most adjective.

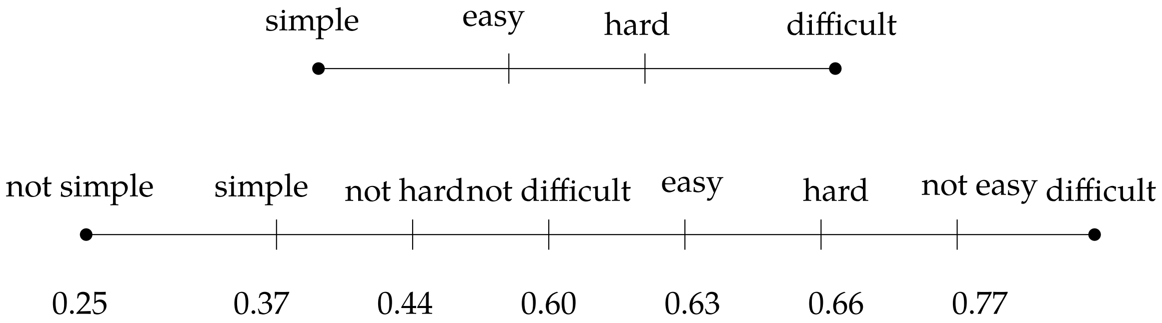

Figure 4.8: Scale of DIFFICULTY with negated adjectives as predicted by the cosine similarity ordering from the right-most adjective.

of this effect, only 22% of the adjectives actually falls within the two poles of the scale. This is not a surprising result if taking into account the scenario described in Chapter 4: negated adjective tend to collocate in a certain region of the space, although relatively close to their adjective and other words in the domain. As a consequence, when comparing expressions on the basis of their cosine similarity to a certain adjective, it is likely that related adjectives would have higher similarity scores than negated adjectives would: for this reason they tend to fall last in the ranking. Although we also obtain less biased results (for example in the scale of DIFFICULTY), we then consider this method to be not adequate for fruitfully studying this phenomenon in the given condition.[8] However, the analysis of the neighbours presented later will tackle our research questions with a different methodology.

Despite not being able to directly compare negated adjectives to adjectives in a unique scale, we can, though, employ this ordering method to compare, instead, the relations occurring among adjectives and negated adjectives respectively. The results of the analogy experiment suggested that the two groups have, to some extent, isomorphic internal structures. Indeed, the relation occurring between antonymic adjectives is captured in a similar way to the relation between two negated antonyms. This could be true also for other types of semantic relations such as scalar ones.

We then proceed to apply the ordering method on negated adjectives only: this amounts to negate the members in a golden scale and use the negated version of the highest pole to order the expressions in terms of the similarity to it. For example, we obtain the negated scale for DIFFICULTY depicted in Figure 4.9 using *not difficult* as reference pole.

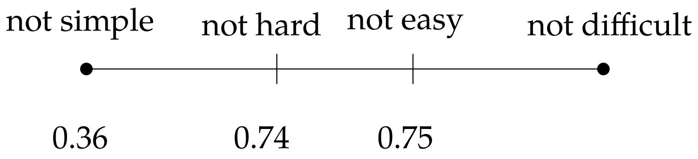

Figure 4.9: Negated scale of DIFFICULTY as predicted by the cosine similarity ordering from the right-most adjective.

Our goal is to check whether the same ordering relations are preserved between the adjectives and negated adjective (i.e., an ordering of *not a, not b, not c* corresponds to an ordering of *a, b, c*). We then look at the overall correlation with the positive scales: values of Spearman's ρ of around 0.51 are achieved with both the golden scales and, in particular, the one obtained with the ordering method ($p < 0.05$). This result, together

[8] For the same reasons, we did not attempt to use the automatic method to produce adjectival scales introduced by Kim and de Marneffe (2013): since this retrieves scale members in the proximate of antonymic pairs, the produced scales would most of the time not include negated adjectives, as these are typically located elsewhere in the space.

with the analogy task ones, confirms that distributional representations of negated adjectives tend to partially preserve the same internal relations of adjectives.

4.3.3 Alternatives in the scale

In this part of the experiments, we get back to the previous research question about where in the scale negated adjectives fall into. However, this time, we do not directly compare adjectives and negated adjectives to form a unique scale, aware of the limitations due to how the two groups collocate in the space. Instead, we look directly at the closest adjectives of negated scalar adjectives and check if and which other members of the same scale are retrieved.

As we saw in Chapter 2 and Section 3.3.2, semantic similarity and alternativehood with respect to a negated item are related notions, such that we may be able to estimate the former exploiting the latter. We here, in particular, follow the intuition that a negated adjective and its potential alternatives will tend to appear in similar contexts, and hence have close representations in a distributional model. Moreover, we then interpret the closest scale co-members to a negated scalar adjective as its most likely alternatives within the scale. We can use this to gain an insight about the default reading of negated adjective in ambiguous utterances like (24) and (25): whether for example *not bad* is closer to *decent* or to *awful* may tell us which one among the two is the typical interpretation.

For each scale in the golden dataset, we consider the negated members sufficiently represented in our distributional model (occurring more than 100 times). We then look at the members of the same scale that are retrieved among the 20 closest adjectives to the negated adjective, excluding the adjective itself. For example, the closest adjectives to *not small* which are part the scale of SIZE (4.1) are, in this order, *large, big, minuscule* and *tiny*.

We observe the following:

- The percentage of shared scale members which are retrieved by each negated adjective is 49%. This gives us an impression of how much the proximate of a negated adjective tend to include expressions from the same scale, and rank them high in terms of plausibility as alternatives. However, as noted by Wilkinson and Tim (2016), the scales in the dataset are incomplete, as they certainly do not exhaust the list of words belonging to the scalar dimension. Negated adjectives may be close to other relevant but not listed expressions: for example, other close adjectives to *not small* are *smallish*, *normal-sized*, *largish*, *middle-sized* and *medium-size*. For this reason, this percentage is to be read only as indicative of a general tendency of negated scalar adjectives to be close to co-members of their scalar dimension.

- 59% of the retrieved scale members are expressions to the left of the adjective that is negated.[9] Similarly, 65% of the times the closest member of the scale is at the

[9] We exclude the two extremes of the scale in the computation, as these can only retrieve scale members

left of the adjective. This means that there is a strong tendency for the negated adjectives in our data to be close to words which express a lower degree of the property than the original adjective (e.g., *not large* is close to *small*).

In order to clarify the last result, we resort to the above-mentioned effect of the polarity of the adjectives. As an exploratory study, we approximate information about the classification in positive and negative adjectives, by splitting up the ordered scale members in two halves, and taking the left one to represent the negative polarity and the right one the positive polarity: for example, in the scale of DIFFICULTY (Figure 4.8), *simple* and *easy* would be taken to express a positive value, while *hard* and *difficult* a negative one.[10]

We consider the joint frequency of negated adjectives to be classified as positive or negative and to retrieve a scale member in the closest adjectives to the right or to the left. In an analogous fashion, we investigate the relationship between the polarity of an adjective that is negated and having as closest scale member an adjective to the right or to the left. We observe the following results:

- The interaction between the polarity of the adjective that is negated and the location of the closest adjectives is significant (Chi-squared test; $p < 0.01$). The same result is obtained looking only at the location of the closest adjective (Fisher's exact test; $p < 0.05$).

- In particular, the negation of an adjective in the first half of the scale retrieves 67% of the times an adjective to the right as a close adjective: in particular, 71% of the times the closest scale member is indeed an adjective to the right. The predicted alternatives to the negation of a negative adjective suggest then most of the time a reading as *less than a*. For example, *not easy* retrieves as alternatives *difficult* and *hard*.

- The negation of an adjective in the second half of the scale retrieve 78% of the times an adjective to the left as a close adjective, and 90% of the times the closest one is indeed an adjective to the left. If interpreting the close adjectives as alternatives, this result suggests that most of the time the reading of the negation of a positive adjective is, as before, *less than a*. For example, *not difficult* retrieves as alternatives *easy*, *hard* and *simple*.

Indeed, the relationship between an adjective with negative polarity *a* and one to its right in an ordered scale, *b*, is such that *b* expresses an higher degree of the positive property associated with the scale, and hence a lower negative value than *a* (e.g., *easy* - *difficult*). If *not a* is interpreted as *b* (e.g., *not easy* as *difficult*), the reading of the negated adjective is then *less than a*. Conversely, if *b* is to the left of *a*, it expresses a lower degree

from either right or left.

[10] While this may not be always the accurate split for each scale, from a qualitative analysis it looks like this simplified classification often approximates the correct one.

of the positive property associated with the scale (e.g., *difficult*): if *not a* is interpreted as *b* (e.g., *not difficult* as *easy*), the reading of the negated adjective is then *more than a*. The opposite pattern is obtained if *a* is an adjective with positive polarity: if *not a* is interpreted as *b*, the reading will be *less than a* or *more than a* if respectively *b* is on the left or right of *a*.

The distributional vectors of negated adjectives seem then to represent more prominently the reading in which the meaning of the adjective shifts towards the opposite polarity. This is aligned with, on one side, the central role assigned to study the shift in meaning of negated adjective towards the antonym in theories like the mitigation hypothesis, and, on the other, to the idea that, even when an ambiguity is posited, *not a* is by default interpreted as *less than a*. Our analyses were based on a limited number of adjectival scales and negated adjectives, as well as an approximation of polarity information. However, they suggest that distributional data can be used to capture this pragmatic effect proposed in the literature, since what are predicted to be the most plausible alternatives are mostly associated with the default *less than a* reading while still sometimes pointing at the *more than a* reading.

Conclusion

In this chapter, we presented a series of experiments involving distributional representations of negated adjectives. We started by considering the interaction of these expressions with a bipolar dimension defined by a pair of antonyms. We found that negated adjectives emerge as not being equivalent in terms of their distribution from, on one side, affixal negations (e.g., *true* - *untrue*) and, on the other, regular antonyms (e.g., *not small* - *big*), even when the adjective that is negated is part of a contradictory pair (e.g., *not dead* - *alive*). We also observed that the relationship between an adjective and its negation can be retrieved as a regularity in the space, setting up an analogy task which makes use of antonymy relations (e.g., *small* : *not small* = *big* : *not big*).

Later, we studied negated adjectives in interaction with scalar dimensions populated by more than two adjectives (e.g., *not small* in the scale of SIZE). As a first finding, we observed that scalar relationships occurring between both adjectives and negated adjectives separately are captured using a simple ordering method which makes use of the proximity of scale members in a distributional space. In addition, we observed that using distributional similarity as an alternativehood measure, we most of the time retrieve as a plausible alternative one that corresponds to a shift in meaning towards the opposite (e.g. *bad* as an alternative to *not great*).

Overall, our results confirm previous findings about the nature of negated adjectives and suggest that these expressions should not be taken to be equivalent to antonyms: indeed, the former ones are subjected to mitigation and other effects which make their use diverge. Moreover, we have shown that the proposal of Kruszewski et al. (2017) to model alternativehood using distributional similarity yields interesting and meaningful results also when studying the negation of adjectives.

Chapter 5

A compositional approach to negated adjectives

In this chapter, we design a compositional method to model the meaning of negated adjectives. In the previous experiments, we represented them as a unit despite their typically compositional nature. While this was a useful tool for our exploratory analyses, we here exploit the distributional representations obtained in this fashion (observed vectors) for learning a general and data-driven function for the negation of adjectives.

In our evaluation, we focus on the compositional vs. lexicalised nature of relatively frequent negated adjectives (e.g., *not bad*). We first show that their observed vectors as a unit have different properties from the rest of the group, possibly due to their less compositional nature; we then proceed to compare these with the representations of the same expressions derived instead through the composition process.

5.1 A data-driven function for negation

5.1.1 Motivation

Function words, such as *not*, are a challenge for DS: despite the effectiveness of its methods on representing the meaning of content words and the promising results of its compositional methods, it is unclear how this toolbox should, or even could, be applied to model phenomena like negation or quantification. Nevertheless, if wanting to provide a complete model of compositionality distributional methods are required to account also for these. As we saw in Chapter 2, the available possibilities for modelling these range from presuming a division of labour between logical and distributional approaches to instead treat them as data-induced functions. Despite the latter approach to be typically excluded for negation, we here attempt to build a compositional function for the negation of adjectives by looking at the distributions of a training set of these expressions. Our focus here is indeed on exploring the general feasibility of an entirely data-driven approach to the modelling of this phenomenon.

Most of the previous compositional approaches to the negation of adjectives in DS based their modelling on *a priori* assumptions about what negation essentially does to the meaning of an adjective: in particular, Nghia et al. (2015) and Rimell et al. (2017) expect it to preserve the membership to the semantic domain but flip the meaning into the opposite. Although not resorting to logical approaches, they then model negation as a predefined operator, which is designed on the basis of theoretical expectations about its behaviour. However, such expectations may not always be correct: as experimental data about the interpretation of negated adjectives (among others, Fraenkel and Schul (2008)) and our results in Section 3.3 show, the way negated adjectives are understood and used do not always equate the ones of the antonym.

On top of that, one may also be simply curious to discover whether we can learn to negate adjectives merely by looking at how instances of this act are used, and forgetting everything one knows about negation, not only in terms of its effect on compositional meaning but also on a specific task. Approaches to negation like the one by Socher et al. (2013) learn indeed negation in a data-driven fashion; however, they optimise its representation for a specific task, namely detecting the sentiment of a discourse. In the present study, however, we do not take a task-driven approach and focus on how a representation of negation could arise using the mere product of co-occurrences. The idea would then be to set up a totally uninformed model that is required to generalise distributional properties of negated adjectives in order to be able to reproduce its effects of new data. This is precisely the approach that we take in this chapter.

Our research tackles then questions about the general "learnability" of negation using distributional data: is negation, or at least some aspects of it, something that can be learnt only on the basis of linguistic contexts of use? We believe that negation of adjectives can be a good starting point to investigate this question, given the relatively simple structure of these expressions (if compared to, for example, phrasal negation) and the similarity of these expressions with affixal negations, whose morpheme combination was already previously modelled in a data-driven way (Marelli and Baroni, 2015).

5.1.2 Compositional vs. lexicalised negation

How to evaluate a function for the negation of adjectives with an approach like ours is, however, not straightforward. Previous methods like the ones by Nghia et al. (2015) and Rimell et al. (2017) evaluated their functions coherently with the notion of negation they used for designing the function itself: they measured its quality in terms of accuracy in antonym detection. However, this does not correspond to the kind of behaviour we aim at modelling. In fact, we do not have any *a priori* assumptions about the behaviour of negated adjectives that we aim at modelling: a data-driven exploration of this phenomenon is indeed the main goal of our analyses.

Since we aim at understanding what type of operation our negation function is modelling, we decide to focus in our evaluation on a specific class of negated adjectives for which a compositional and non-compositional treatment may exhibit substantial differences, namely frequent ones. Indeed, expressions of this kind, such as *not bad* or *not* *happy* presumably have more lexicalised meanings, which, despite their internal structure, may have conventionalised in virtue of their frequent usage (e.g., *not bad* ≈ *fairly good*, *not happy* ≈ *dissatisfied*). For this reason, they might then have a less compositional nature, and act more similarly to independent lexical units.

Indeed, we can observe differences between the observed vectors of the most frequent ($n > 5000$) and the other negated adjectives ($100 < n < 5000$) in the distributional space. We already saw in Chapter 3 that the former ones do not exhibit the same clustering effect as the others. Moreover, we carry out further analyses on the two groups separately, making use of some of the measurements introduced in Section 4.2; the results are reported in Table 5.1.

	$100 < n < 5000$	$n > 5000$	
$Sim(not\ a_1, a_1)$ M	0.45	0.52	$*_t$
Shared 20 neighbours with a_1 M	2.7	4.4	$**_t$
Shared 20 closest adj. with a_1 M	4	5.4	$*_t$
AFFIXAL ANTONYMS			
$Sim(not\ a_1, a_2)$ M	0.42	0.60	$*_t$
$Shift(not\ a_1, a_2)$ M	-0.04	0.04	
$Flip(not\ a_1, a_2)$	39%	50%	
Shared 20 closest adj. with a_2 M	3.8	6.9	$*_t$
REGULAR ANTONYMS			
$Sim(not\ a_1, a_2)$ M	0.25	0.35	
$Shift(not\ a_1, a_2)$ M	-0.20	-0.08	
$Flip(not\ a_1, a_2)$	8%	42%	$***_\chi$
Shared 20 closest adj. with a_2 M	1.7	3.8	
Datapoints	567	29	

Table 5.1: Results of the comparisons between observed representations of negated adjectives with different frequencies (M: mean value). Each negated adjective is compared to its adjective (a_1) and antonym (a_2); differences are tested using Welch's t-test (t), Chi-squared test (χ) and Fisher's exact test (f) (*: $p < 0.05$: **: $p < 0.01$; ***: $p < 0.001$).

Given the limited sample size of frequent negated adjectives (29), the values and the results of the significance tests should be taken with a grain of salt, in particular those related to affixal and regular antonyms (there are 14 and 12 datapoints for the two categories respectively). Nevertheless, they suggest that frequent adjectives, as expected, are closer to their original adjective and the antonym ($Sim(not\ a_1, a_1)$, $Sim(not\ a_1, a_2)$ (4.1), i.e., cosine similarity of the negated adjective with the adjective and the antonym respectively), and consequently share more neighbours with them: this is to be expected given the lack of a low-frequency effects which isolate them in the space. Interestingly, they seem to be subjected to a larger meaning shift towards the antonym

($Shift(not\ a_1, a_2)$ (4.2), i.e., difference of the similarities between the negated adjective with the antonym and the adjective respectively, $Flip$ (4.3), i.e., whether a negated adjective is closer to the antonym than to the adjective). What this suggests is that the negation of an adjective may acquire during the process of lexicalisation a meaning which is closer to the one of the antonym. Indeed, it could be that, akin to affixal negations, their meaning is less pragmatically ambiguous and more rigidly associated with the orientation of the opposite meaning (e.g., *not easy* $\approx$ *difficult*).

In our evaluation, we then focus our attention on the most frequent negated adjectives in our corpus and compare their observed and composed representations. Our goal is to investigate in a semi-qualitative fashion whether they respectively capture the two faces of these expressions: on one side, their meaning as a combination of its building blocks, and, on the other, their default meaning conventionalised by the usage. At the same time, we use this group of adjectives and other relevant statistics to understand some of the general properties of the function of negation that we induce. While this may not be an exhaustive analysis of our compositional model, we regard this approach as the most appropriate and clear in this setting and postpone to future research other analyses on the resulting space.

5.1.3 Methods

Following our motivation, the type of technique that we employ for modelling compositional negation is one that does not introduce any *a priori* conjecture about how negation should behave. Instead, we propose to learn to negate adjectives by simply learning how to go from the distributional representations of adjectives to the ones of their observed vectors.

We follow the general framework to compositional DS introduced by Baroni and Zamparelli (2010) and Coecke et al. (2010) and further developed in Grefenstette et al. (2013) and Baroni et al. (2014a) (mentioned in Chapter 2). We model composition in terms of functional application, where a function consists of a linear transformation between two algebraic objects in a high-dimensional distributional space, and is directly learnt using observed vectors of expressions.

In particular, we are interested in the mapping from an adjective to a negated adjective, which we both represent as distributional n-dimensional vectors. Our function for negation consists then of a linear map between two vectors, namely a matrix (representing *not*) that, when multiplied with an adjective vector, yields a representation of its negation. Such a transformation is learnt from a set of training pairs consisting of the vector of an adjective and the one of its negation.

$$not\_adj = NOT \times adj \tag{5.1}$$

where NOT is the matrix associated to the particle *not* (e.g., *not* $\times$ *logical* $=$ *not logical*).

The distributional representations that we take into account to estimate the function are the observed vectors built as described in Chapter 3. In particular, our training data consist of pairs of adjectives and corresponding negated adjectives (e.g., *logical - not logical*). To derive the NOT function, we use machine learning techniques with the objective of maximising the quality of the mapping from the adjective vector to the negated adjective vector. Specifically, we estimate the weights in the NOT matrix using least squares regression, as implemented by Dinu et al. (2013) in the DISSECT toolkit.[1] In our setting, the independent and dependent variables for the regression equations are respectively the dimensions of the adjective and of the negated adjective vectors. Given the focus of our evaluation, we use as training data only pairs with a negated adjective that occurs $100 < n < 5000$ times, and as testing pairs those with a negated adjective that occurs more than 5000 times. This gives us training and testing datasets of respectively 567 and 29 pairs.

Our model shares similarities with the function for the negation of adjectives designed by Nghia et al. (2015). Except for the type of training data, we employ the same idea of modelling *not* as a matrix applied to an adjective, as well as the same learning algorithm. The crucial difference is here that while their negation function is learnt as a linear map from an adjective to its antonym (e.g., *wide* to *narrow*), ours is derived by the mapping from an adjective to its negated version (e.g., *wide* to *not wide*).

This methodology provides the data-driven approach that we require, since functions are estimated directly from the distributions of negated adjectives. Moreover, differently from other models of sentence compositionality, this methodology fits our simple setting: we do not aim here at providing a complete model of sentence compositionality and are able to only focus on the meaning of this type of constructions. We also find an interesting analogy between the idea of learning negation as a transformation in the space applied to the representation of the adjective, and the way it is often described in the literature as having a mitigation or shift effect on the meaning of the adjective.[2]

As a final remark, recall that, due to the low-frequency effects that we described in Chapter 3, negated adjectives tend to group in a region of the space with other infrequent items. Because our function of negation is learnt as a mapping from the adjective to the negated adjective using their observed representations, this aspect is going to be directly inherited in our compositional model. What the linear transformation is going to presumably amount to is an operation that "attracts" the adjective into this area of the space. This introduces the same complications that we encountered in our previous experiments. However, because our goal here is to just explore whether learning something about negation is possible in an entirely data-driven setting, we leave to future research to provide a methodology to cope with these effects.

[1] Dissect toolkit: `http://clic.cimec.unitn.it/composes/toolkit`. We, in particular, use the implementation of the Lexical Function with Least Squares Regression learner and intercept.

[2] As in the previous chapters, we make a simplification about the semantics of adjectives and the scope of *not*: we discard the role of the noun phrase that the adjective may interact with in the sentence (e.g., *not fast car*). A possible and natural extension of our compositional account could then be to include noun phrases in the picture and, for example, treat *not* as a function that applies to adjective-noun pairs modelled as by Baroni and Zamparelli (2010) (e.g., $not\_red\_car = not(red(car))$).

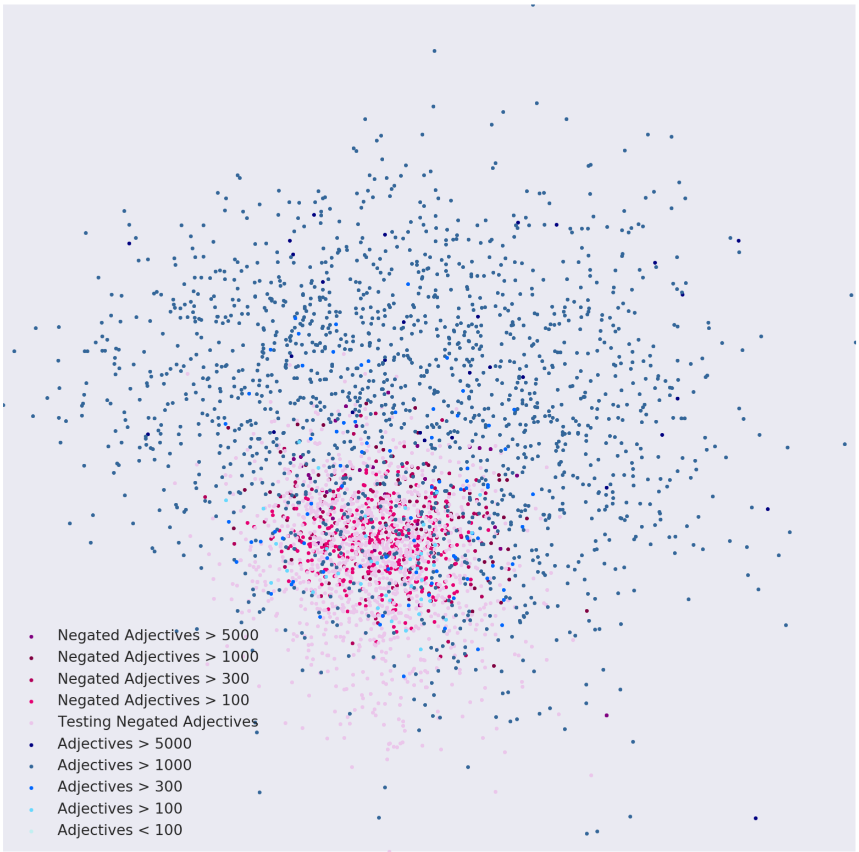

Figure 5.1: Negated adjectives and their corresponding adjectives in the two-dimensional semantic space with composed vectors (original space reduced using PCA).

5.2 Evaluation: lexicalised negated adjectives

Once obtained our function for negation as described in Section 5.1.3, we build a new semantic space where all negated adjective representations in the previous model, including training and testing ones, are substituted by the compositionally derived ones (by multiplication of the matrix representing NOT and the vector of the adjective).

5.2.1 Results

As can be seen in Figure 5.1, the distributional space with composed negated adjectives exhibits a similar phenomenon to the space with observed ones: negated adjectives cluster in a region of the space. The result is expected and inherited on the basis of the training data. If comparing Figures 3.1 and 5.1, we can see that in the compositional model very infrequent adjectives seem to be now more spread around the region where they typically cluster than their observed counterparts.

Figure 5.2 shows a heat map of our NOT function, where individual values con-

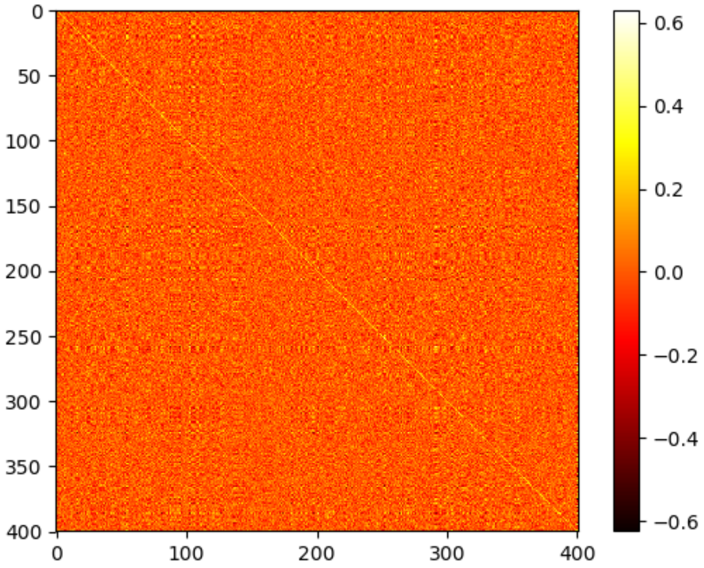

Figure 5.2: Heat map of the NOT function learnt as described in Section 5.1.3.

tained in the matrix are represented as colours in a scale. As can be seen, large positive values are concentrated along the diagonal, while the remaining ones are mostly close to 0 and heterogeneously distributed. What this shows is that the NOT function has similar properties to an identity matrix:[3] for this reason, applying it to an adjective would yield a vector which substantially resembles the adjective itself. This effect is also described by Nghia et al. (2015) for their NOT function, which they instead obtain learning to map antonyms in the semantic space: since these are typically very close in the space, the matrix is identity-like, with a substantial discrepancy between values in and out of the diagonal. However, this effect seems to be less marked in our negation function, as values of the same magnitude to the ones on the diagonal appear heterogeneously also in other cells of the matrix. Indeed, what our function seems to do is to shift the vector of an adjective in a different area of the space, although still preserving a certain similarity with it (see Table 5.4).

In the compositional space, vectors of negated adjectives that were part of the training data are very similar to the composed counterparts (the average cosine similarity is 0.9), since the compositional function was indeed learnt to optimise the map between these adjectives to their negation. Instead, when looking at the unseen testing data, the cosine similarity between observed and composed vectors is 0.38. One may see this as, on one side, a lack of generalisation of the function, but also, and more interestingly, as a modification of how these negated adjectives are now represented. Recall that our testing data consist of frequent negated adjectives, which exhibit different properties from the less frequent ones. What the negation function has likely learnt to do is to make these negated adjectives behave more like the latter ones, having been trained

[3] An identity matrix is a square matrix such that all elements of the diagonal are 1 and all others are 0; multiplying a matrix by an identity matrix gives, as a result, the matrix itself.

	Composed	Observed			
	$n > 5000$	$100 < n < 5000$		$n > 5000$	
$Sim(not\ a_1, a_1)$ M	0.36	0.45	***$_t$	0.52	***$_t$
Shared 20 neighbours with a_1 M	1	2.7	***$_t$	4.4	***$_t$
Shared 20 closest adj. with a_1 M	3.5	4		5.4	*$_t$
AFFIXAL ANTONYMS					
$Sim(not\ a_1, a_2)$ M	0.33	0.42	**$_t$	0.60	**$_t$
$Shift(not\ a_1, a_2)$ M	-0.07	-0.04		0.04	
$Flip(not\ a_1, a_2)$	29%	39%		50%	
Shared 20 closest adj. with a_2 M	3.6	3.7		6.9	
REGULAR ANTONYMS					
$Sim(not\ a_1, a_2)$ M	0.19	0.25		0.35	
$Shift(not\ a_1, a_2)$ M	-0.13	-0.20		-0.08	
$Flip(not\ a_1, a_2)$	17%	8%		42%	
Shared 20 closest adj. with a_2 M	1.2	1.7		3.8	*$_t$
Datapoints	29	567		29	

Table 5.2: Results of the comparisons between composed representations of testing negated adjectives and observed ones from different frequency thresholds (M: mean value). Each negated adjective is compared to its adjective (a_1) and antonym (a_2); differences are tested using Welch's t-test (t), Chi-squared test (χ) and Fisher's exact test (f) (*: $p < 0.05$: **: $p < 0.01$; ***: $p < 0.001$).

precisely on these.

We here present further results aimed at clarifying this aspect. In particular, we compare the statistics obtained for our testing adjectives with two classes from the non-compositional space. We look at their differences with the observed vectors of, on one side, their counterparts ($n > 5000$), and, on the other, less frequent negated adjectives ($100 < n < 5000$), which we compared in Section 5.1.2. This allows us to investigate whether frequent negated adjectives, which plausibly have more lexicalised meanings, when treated compositionally behave instead like those which appear less often and possibly have less conventionalised, and hence more compositional, meanings.

Table 5.2 reports the results of our comparison. Again these results are to be interpreted cautiously given the limited size of the testing data. Composed negated adjectives tend to be further away from their corresponding adjectives than both the observed frequent and infrequent ones, despite being closer to the behaviour of the latter ($Sim(not\ a_1, a_1)$). The average number of shared neighbours is consequently aligned with this result. However, the number of shared closest adjectives (more frequent than 100 times), both with the adjective and the antonym, does not exhibit instead the same scenario (this query can indeed filter out some of the clustering effects). Interestingly,

	Composed	Observed
not happy	*loath, unhappy, unwilling, reluctant, willing, glad, afraid, ashamed, unable, happy*	*unhappy, unsatisfied, unsure, disappointed, dissatisfied, adamant, unimpressed, happy, annoyed, pleased*
not bad	*daft, bad, clued-up, picky, rubbishy, stupid, crappy, decent*	*okay, bad, crappy, nice, lousy, decent, daft, mediocre, dodgy, stupid*
not possible	*preferable, unsuitable, impractical, inaccessible, low-level, possible, unavoidable, inadvisable, impracticable, lossy*	*possible, necessary, impossible, unable, difficult, able, impractical, insufficient, sufficient*
not uncommon	*uncommon, prevalent, rare, commonplace, endemic, common, widespread, worrying, symptomless, melanistic*	*uncommon, commonplace, most, many, prevalent, common, typical, prone, rare, unusual*
not true	*unscriptural, true, inexact, erroneous, fallacious, specious, untrue, spurious, imperfect, contrary*	*untrue, true, fallacious, absurd, disingenuous, ignorant, contrary, unfounded, self-evident, baseless*
not easy	*difficult, easy, straightforward, impossible, straigh-forward, preferable, simple, tricky, distateful, complicated*	*difficult, hard, impossible, easy, tricky, unable, frustrating, adept, able, incapable*

Table 5.3: Closest adjectives to a sample of negated adjectives in their composed and observed versions.

the results obtained on the composed representations resemble more the behaviour of the observed ones of the less frequent negated adjectives in terms of the meaning shift towards the antonym ($Sim(not\, a_1, a_2)$, $Shift(not\, a_1, a_2)$, $Flip$, shared closest adjectives). In particular, frequent negated adjectives, when treated compositionally, seem to register a smaller degree of shift in the direction of the antonym.

To clarify and exemplify the results, we complement our analyses with some qualitative comparisons between testing adjectives in their observed and composed version. In particular, we present here the closest adjectives and $Shift$ values of a sample of negated adjectives, respectively in Tables 5.3 and 5.4. This is also meant as an illustration of the type of meanings captured by the composed vectors, which, despite requiring a qualitative interpretation, are probably the most valuable source of evidence about the quality of the negation function in this setting.

	$Shift$	
	Composed	Observed
true, false, not true	-0.04	-0.17
bad, good, not bad	-0.06	-0.08
easy, difficult, not easy	0.01	0.23
happy, unhappy, not happy	0.08	0.31
possible, impossible, not possible	-0.12	-0.01
uncommon, common, not uncommon	-0.19	-0.08

Table 5.4: $Shift$ computed on a sample of antonyms and negated adjective triplets in their composed and observed versions.

As can be seen, the proximate adjectives of negated adjectives in the compositional model appear to be often meaningful related expressions, and in particular also presumably plausible alternatives to the negated item. There tends to be also a substantial overlap with the closest adjectives to the non-compositional counterpart. This itself is a very interesting result, given how these vectors were obtained. We saw that the transformation in the space tends to shift the adjective representation further away from where the observed vector would be, and, as a consequence, the composed vector is usually quite different from the latter. However, the closest adjectives to composed representations show that, despite this effect, the NOT function still seems to adequately map the adjective to a point in the space with analogous similarity relations to the observed vector. However, these do not necessarily need to be identical, given what we observed about the behaviour of frequent negated adjectives. For example, the closest adjectives of the non-compositional *not easy* seem to hint more at a reversal of meaning of *easy* into the opposite than the ones of the composed representation, which instead suggests the presence of mitigation effects.

Conclusion and discussion

In this chapter, we presented and evaluated a compositional model of negated adjectives. In particular, we estimated a matrix whose multiplication with the vector of an adjective represents the functional application of *not* to the adjective. Such a matrix was directly induced from distributional data through regression techniques, without incorporating in its design any *a priori* conjecture about the effects of negation. In our evaluation, we focused on negated adjectives whose associated meaning is presumably more lexicalised. Our results suggest that by treating them as a unit or as a compositional phrase we can model their conventionalised or compositional meaning respectively.

All in all, the results we presented, although not easily interpretable and not evaluating the compositional model on a large scale, suggests that it is actually possible to learn and generalise at least some aspects of negation just by looking at distributional representations of negated items. Our model was set up as an exploratory study; for this reason, we did not attempt any kind of optimisation in the learning phase in order to improve its performances. Moreover, the negation function was learnt in suboptimal conditions (i.e., trained on a relatively small number of vectors, which were in turn trained on relatively few occurrences of the expressions). Yet the results are promising, in particular when looking at the proximate of negated adjectives. We hence believe that there is space for these results to be largely improved with more complex training settings and techniques.

Negation does not then seem to be entirely out of the scope of entirely bottom-up DS approaches. Distributional vectors are indeed well-suited to account for its pragmatic and graded aspects of mitigation and alternativehood: by exploiting these representations, we can then generalise the modelling of these phenomena into a compositional function. Nevertheless, other characterising notions of negations like the ones of truth value or opposition have instead a discrete nature and are hence less easily model in the continuous space. More efforts are then still required to understand whether these could be integrated into distributional semantic models, or whether instead a division of labour between distributional and formal approaches is required in order to obtain a full model of linguistic negation.

Chapter 6

Conclusion

In this thesis, we reported exploratory analyses on properties of negated adjectives in English (e.g., *not logical*, *not small*) as represented in a distributional semantic model. We here provide an overview of the results obtained, as well as ideas for future research.

In the first part of the thesis (Chapters 3 and 4), we constructed and made use of a distributional model where negated adjectives are treated as a unique lexical item. This allowed us to provide a data-driven account of these expressions, as it emerges summarising their large-scale distributions across contexts of use in the form of vectorial representations. We have shown that the representations of negated adjectives as a unit tend to be meaningful, despite the compositional nature of these expressions and data sparsity effects. Moreover, we found that the spatial relation occurring between each of these and the representation of the corresponding adjective tends to be regular, and can, for this reason, be often successfully captured setting up an analogy task. We also provided evidence of the fact that some of the relations occurring between negated adjectives in the space (antonymy and scalar relationships) tend to replicate the ones occurring between their non-negated counterparts.

In our analyses, we focused on two aspects of the negation of adjectives, namely mitigation and alternativehood, which we modelled in terms of similarity relations between expressions in the distributional space. Interestingly, we found that the distributional vectors of negated adjectives tend to be quite different from the ones of other expressions which they are often equated to. In particular, we observed the following:

- Negated adjectives tend to be generally closer in a distributional space to their corresponding adjective than to the antonym (e.g., *not hot* is closer to *hot* than to *cold*; *not similar* is closer to *similar* than to *dissimilar*). Their semantic representation is then somehow intermediate between the adjective and the antonym.

- However, not all negated adjectives are expected to behave in this way: negations of members of antonymic pairs that do not admit a *tertium* (i.e., contradictory pairs) are expected to convey the same meaning as the antonym (e.g., *not true* = *false*). Nevertheless, this class did not emerge as different in our analyses.

- Affixal negations are distributionally more similar to negated adjectives than regular antonyms are and hence tend to have more similar patterns of use to these (e.g., *not similar - dissimilar*). For this reason, we found that treating affixal and regular antonyms as part of a coherent class may not be an appropriate assumption.

- Finally, we found that double negations of an adjective (e.g., *not dissimilar*) tend to have a more different representation from their antonym (e.g., *not dissimilar - similar*) than simple negations (e.g. *not similar - dissimilar*).

We put these results in relation to two main phenomena. Negated adjectives were shown not to suppress but retain the emphasis on the concept that is negated, and express a mitigated version of its meaning (Giora et al., 2005). It could then be for this reason that their distributional representations tend to still be very close to the adjective that is negated, and even closer to this than to the opposite meaning (i.e., antonym). Moreover, the negation of an adjective tends to have a quite peculiar profile of use that cannot be simply lead back to that of the antonym, even when the two constitute a contradictory pair. This suggests that a complex expression like a negated adjective is used in different contexts of use, which makes it different from other allegedly equivalent expressions, in particular at the pragmatic level.

Moreover, we found that the similarity relations of distributional vectors of negated scalar adjectives seem to capture what is usually taken to be their default interpretation. Indeed, the adjectives in the scale that are predicted to be the most plausible alternatives to the negated item express smaller degrees of the relevant property conveyed by it (e.g., *not big* expresses a lower degree of positive size than *big*; *not small* expresses a lower degree of negative size than *small*). At the same time, less typical, but still plausible, alternatives are not ruled out (e.g., *huge* as alternative of *not big*).

In the second part of the thesis (Chapter 5), we exploited the previously analysed observed vectors of negated adjectives to instead obtain their composed versions. We learnt through regression techniques a matrix, representing *not*, such that when multiplied with the vector of an adjective yields the vectorial representation of its negation. Such a linear transformation was learnt entirely from a set of training distributional data as a mapping from the vectors of an adjective and their negation. This setting contrasts with the typical approach to negation in DS, namely to design it as a function on the basis of *a priori* assumptions about its effects. We evaluated our compositional approach on a set of frequent negated adjectives, which appear to have a different behaviour from the others due to their possibly more lexicalised and less compositional meanings (e.g., *not bad*): we compared their observed and composed representations, and found the latter to behave more closely to the less frequent and possibly more compositional negated adjectives. This suggests that with a compositional treatment one can model their compositional meaning, while by treating them as a unit one can capture their conventionalised one. On a more general level, we found that composed representations of negated adjectives are sufficiently meaningful not to discard the feasibility of a data-driven approach to negation in DS, although many questions still remain open about how to obtain a complete model of negation within the DS framework.

All in all, our findings provide interesting results on two fronts. On one hand, they offer new empirical evidence for linguistic research about the negation of adjectives and, more in general, arguments for conceiving negation, and in particular pragmatic negation, as a more "graded" phenomenon that usually taken to be. On the other hand, we believe that our results provide useful caveats for the Natural Language Processing community. In particular, we regard certain assumptions about negated adjectives made for modelling purposes (such as in the work by Nghia et al. (2015) and Rimell et al. (2017)) counter-productive, since they do not reflect their actual interpretations and use. As existing experimental data and our results suggest, the negation of an adjective is not the same as its antonym. For this reason, we suggest to either take into account mitigation and pragmatic effects in the assumptions made about negation or, instead, to abandon all the assumptions and use entirely data-driven methods. Indeed, as noted by Wiegand et al. (2010), whether, for example, *not bad* is taken to be equivalent to *good* or not may be particularly crucial, for example, in Sentiment Analysis. In fact, Kamoen et al. (2015) show that the way people interpret negated adjectives in reviews is not the same as a mere reversal of meaning (e.g., *not bad* $\neq$ *good*) and this affects the perceived sentiment of the text.

However, there are still many aspects of the negation of adjectives that need to be clarified; in particular, even within our work, many phenomena could be further examined in future research. In the course of our experiments, we made several simplifications, ranging from discarding the dependency of meaning between an adjective and the noun it modifies to assumptions about the scope of negation. Both our investigations and compositional approach could thus be extended to account for these phenomena. Moreover, we found the range of experiments one could devise to study the distributional representations of negated adjectives to be particularly wide. For the sake of this first study, we restricted it to the set of analyses that we here presented. However, it would be interesting to, for example, carry out a corpus study to detect the contextual features which make the negation of an adjective different from the adjective itself or the antonym, or analyse the negation of adjectives whose scalar dimension is more complex than the ones we considered (e.g., colour terms like *not blue*). Another aspect one could investigate is if and how representations of negated adjectives like the ones we built can account for scalar inference patterns which involve this type of expressions, as the ones studied by Van Tiel et al. (2016) (e.g., *This is difficult.* $\rightsquigarrow$ *It is not impossible*). In addition, given the low-frequency effects on the vectors of negated adjectives, one could experiment with different techniques to construct these, both at pre-processing and training time. The same applies to our compositional approach: being our goals merely exploratory, there are still many technical variables which one could test to improve the quality of our model.

Finally, another interesting research direction could be to evaluate the ability of distributional methods to predict plausible alternatives to a negated adjective against human judgements. In this case, one would be required to construct a dataset similar to the one built by Kruszewski et al. (2017) for investigating the negation of nominal predicates: the idea would be to collect plausibility ratings for sentences containing a negated adjective and providing an alternative to it (*This is not X, it is Y.*; e.g., *This is not big, it is* {*small, medium-sized, blue, true*}.). In our experiments, we indeed used the notion of alternativehood as an interpretative tool; however, a natural and fundamental extension of this work would be to check how tight the connection between the plausibility of an alternative and its distributional similarity to it is at the empirical level. We believe that, given the results presented in this thesis, there are reasons to be optimistic about the success of a distributional model like ours in this task.

Bibliography

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

Marco Baroni and Roberto Zamparelli. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1183–1193, 2010.

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

Marco Baroni, Raffaela Bernardi, and Roberto Zamparelli. Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, 9, 2014a.

Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 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, 2014b.

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

Ivana Bianchi, Ugo Savardi, Roberto Burro, and Stefania Torquati. Negation and psychological dimensions. *Journal of Cognitive Psychology*, 23(3):275–301, 2011.

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

Gemma Boleda, Eva Maria Vecchi, Miquel Cornudella, and Louise McNally. First-order vs. higher-order modification in distributional semantics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 1223–1233, 2012.

Gemma Boleda, Marco Baroni, The Nghia Pham, and Louise McNally. Intensionality was only alleged: On adjective-noun composition in distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS)*, pages 35–46, 2013.

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

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

Eve V. Clark. 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*, page 171. Routledge, 1992.

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

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

Herbert L Colston. "Not good" is "bad," but "not bad" is not "good": An analysis of three accounts of negation asymmetry. *Discourse Processes*, 28(3):237–256, 1999.

Georgiana Dinu, Nghia The Pham, and Marco Baroni. Dissect - distributional semantics composition toolkit. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL): System Demonstrations*, pages 31–36, 2013.

Afsaneh Fazly and Suzanne Stevenson. A distributional account of the semantics of multiword expressions. *Italian Journal of Linguistics*, 1(20):157–179, 2008.

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

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

Dan Garrette, Katrin Erk, and Raymond Mooney. 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, pages 27–48. Springer, 2014.

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

Rachel Giora, Noga Balaban, Ofer Fein, and Inbar Alkabets. 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, 2005.

Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. Multi-step regression learning for compositional distributional semantics. *Proceedings of the 10th International Conference on Computational Semantics (IWCS)*, pages 131–142, 2013.

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

Karl Moritz Hermann, Edward Grefenstette, and Phil Blunsom. "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, 2013.

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

Laurence R. Horn. Toward a new taxonomy for pragmatic inference: Q-based and r-based implicature. *Meaning, form, and use in context: Linguistic applications*, pages 11–42, 1984.

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

Laurence R. Horn and Yasuhiko Kato. Introduction: Negation and polarity at the millennium. In Laurence R. Horn and Yasuhiko Kato, editors, *Negation and Polarity. Syntactic and Semantic Perspectives*, pages 1–19. Oxford University Press, 2000.

Rodney Huddleston and Geoffrey K Pullum. *The Cambridge Grammar of the English Language*. Cambridge University Press, 2002.

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

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

Naomi Kamoen, Maria B.J. Mos, and Willem F.S. Dekker. A hotel that is not bad isn't good. the effects of valence framing and expectation in online reviews on text, reviewer and product appreciation. *Journal of Pragmatics*, 75:28 – 43, 2015.

Hans Kamp. Two theories about adjectives. In Edward Keenan, editor, *Formal semantics of natural language*, pages 123–155. Cambridge University Press, 1975.

Christopher Kennedy. *Projecting the adjective: The syntax and semantics of gradability and comparison*. Routledge, 1999.

Christopher Kennedy. Adjectives. In Delia Graff Fara and Gillian Russell, editors, *Routledge Companion to Philosophy of Language*. Routledge, 2012.

Christopher Kennedy and Louise McNally. Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 81:345–381, 2005.

Joo-Kyung Kim and Marie-Catherine de Marneffe. 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, 2013.

Paul Kiparsky. Word-formation and the lexicon. In *Proceedings of the Mid-America Linguistics Conference*, 1982.

Manfred Krifka. A compositional semantics for multiple focus constructions. In *Informationsstruktur und Grammatik*, pages 17–53. VS Verlag für Sozialwissenschaften, 1992.

Manfred Krifka. 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, 2007.

Germán Kruszewski, Denis Paperno, Raffaella Bernardi, and Marco Baroni. There is no logical negation here, but there are alternatives: Modeling conversational negation with distributional semantics. *Computational Linguistics*, 2017.

Geoffrey N. Leech. Pragmatics and conversational rhetoric. In Herman Parret, Marina Sbisà, and Jef Verschueren, editors, *Possibilities and limitations of pragmatics*, pages 413–442. John Benjamins Publishing Company, 1981.

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

Omer Levy, Yoav Goldberg, and Israel Ramat-Gan. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Language Learning (CoNLL)*, pages 171–180, 2014.

Marco Marelli and Marco Baroni. Affixation in semantic space: Modeling morpheme meanings with compositional distributional semantics. *Psychological review*, 122(3): 485, 2015.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. Semeval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval)*, pages 1–8, 2014.

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

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of 2013 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), pages 746–751, 2013b.

Jeff Mitchell and Mirella Lapata. Composition in distributional models of semantics. *Cognitive Science*, 34(8):1388–1429, 2010.

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

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

The Pham Nghia, Angeliki Lazaridou, and Marco Baroni. 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, 2015.

Kim Anh Nguyen, Sabine Schulte im Walde, and Ngoc Thang Vu. 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, 2016.

Mike Oaksford. Contrast classes and matching bias as explanations of the effects of negation on conditional reasoning. *Thinking & Reasoning*, 8(2):135–151, 2002.

Mike Oaksford and Keith Stenning. Reasoning with conditionals containing negated constituents. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 18 (4):835, 1992.

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

Barbara Partee. Lexical semantics and compositionality. *An invitation to cognitive science: Language*, pages 311–360, 1995.

José Ramón Varela Pérez. Operator and negative contraction in spoken british english: a change in progress. In Bas Aarts, Joanne Close, Geoffrey Leech, and Sean Wallis, editors, *The Verb Phrase in English: Investigating Recent Language Change with Corpora*, page 256. Cambridge University Press, 2013.

Laura Rimell, Amandla Mabona, Luana Bulat, and Douwe Kiela. 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, 2017.

Mats Rooth. A theory of focus interpretation. *Natural language semantics*, 1(1):75–116, 1992.

Edward Sapir. Grading, a study in semantics. *Philosophy of science*, 11(2):93–116, 1944.

Galit Weidman Sassoon. The degree functions of negative adjectives. *Natural language semantics*, 18(2):141–181, 2010.

Richard Socher, Brody Huval, Christopher D. Manning, and Andrew Y. Ng. 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, 2012.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D Manning, Andrew Y. Ng, Christopher Potts, et al. 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, 2013.

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

CM van Son, CWJ van Miltenburg, R Morante Vallejo, et al. Building a dictionary of affixal negations. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, 2016.

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

Peter C. Wason. Response to affirmative and negative binary statements. *British Journal of Psychology*, 52(2):133–142, 1961.

Peter C. Wason. The contexts of plausible denial. *Journal of Verbal Learning and Verbal Behavior*, 4(1):7 – 11, 1965.

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

Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. 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, 2010.

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

George Kingsley Zipf. *Human Behaviour and the Principle of Least-Effort*. Addison-Wesley Press, 1949.

Appendix A

Annotation of regular antonyms

Data Each annotator is given a list of 148 adjectives pairs with opposite meanings (antonyms). The list is obtained according to the following criteria:

- The adjectives are tagged as direct antonyms in WordNet (Fellbaum, 1998) ;
- The adjectives are tagged as regular antonyms in the Dictionary of Lexical Negation (van Son et al., 2016);
- Both the adjectives occur more than 100 times in the training corpus of the distributional model (UkWac + Wacky);
- The negated adjective of at least one of the members of the pair occurs more than 100 times in the training corpus of the distributional model (UkWac + Wacky).

Examples of adjective pairs

beautiful, ugly

free, bound

actual, potential

alive, dead

Annotation guidelines Given a pair of words a_1 and a_2, the annotator considers the following sentence pattern:

(a) *X is neither* a_1 *nor* a_2

- The annotator tags the pair as `x` if she does not know the meaning of the word, or does not interpret the two words as having opposite meanings.
- The annotator tags the pair as `1` (contrary) if (a) is acceptable under a default context.
- The annotator tags the pair as `0` (contradictory) if (a) is not acceptable under a default context.

The annotator is asked to carry out the task following for each case her first intuitive judgement.

Agreement Interraters agreement is computed using Fleiss' k, which generalises Cohen's k to cases with more than two coders.

- Agreement on the three categories (`1, 0, x`): Fleiss' $k = 0.38$;
- Agreement on two categories (1, 0): Fleiss' $k = 0.37$.

Dataset For the purpose of the final categorisation used in the analyses, only pairs that have been categorised as either `1` or `0` are considered, hence discarding those for which at least one annotator tagged them as `x`, or for which there was not full agreement.

Examples of pairs tagged as contradictory

alive, dead

innocent, guilty

optional, obligatory

Examples of pairs tagged as contrary

beautiful, ugly

wide, narrow

full, empty