Opposites Attract - Or Do They?:
Investigating Negated Verbs in Distributional Semantic Space

Master's Thesis

Presented to

The Faculty of the Graduate School of Arts and Sciences
Brandeis University
Graduate Program in Computational Linguistics
James Pustejovsky, Advisor

In Partial Fulfillment
of the Requirements for the Degree

Master of Science
in
Computational Linguistics

by
Jennifer Sarah Storozum

August 2018
Acknowledgements

Many thanks to my advisor James Pustejovsky, and the entire computational linguistics and linguistics faculty for their support and encouragement. Special thanks to Lotus Goldberg, for her wonderful teaching, patience and for being an all-around lovely human being. I couldn’t have done it without you, Lotus!

Thank you to my roommates, Chester and Meghan, who put up with my jigsaw puzzles taking over our coffee table and for enduring my late-night ramblings about distributional semantics and other hare-brained ideas. Thank you to all who helped with annotation: Nick, Chester, Justin, Nikhil, and Hannah. Double thank you to Nikhil, who also served as my second reader on short notice and gave me excellent comments and suggestions.

Thank you to my parents and my friends Jacob and Rey, for helping me retain my sanity and take care of life’s other necessities during the research and writing process, like food and transportation and human interaction.

Thank you to the MBTA commuter rail for making going to school fun every day. I love trains! Thank you to dogs for being good boys. Dogs should be allowed to vote. Lastly, thank you to the Facebook meme group Linguistics [ˈjɪt̚ˌpʰʊ̈ˌstɪn̩] for making negation a meme while I struggled through the last days of the semester and thank you Nick, Chester and Clay for putting up with me tagging you in every post related to negation with the comment "This is my thesis." Don’t forget: if p = true, ¬p = truen’t.

It’s just semantics.
ABSTRACT

Opposites Attract - Or Do They?: Investigating Negated Verbs in Distributional Semantic Space

A thesis presented to the Graduate Program in Computational Linguistics

Graduate School of Arts and Sciences
Brandeis University
Waltham, Massachusetts

By Jennifer Sarah Storozum

Negation is ubiquitous in natural language and a fundamental logical operator in formal semantics, yet it remains difficult to capture with distributional semantic models. In this thesis, we examine negation of verbal predicates by creating distributional semantic models of a medium-sized text corpus. We show that the effect negation has on vectors of negated verbal predicates is similar to that of dimensionality reduction, suggesting that negation in context, or conversational negation, functions as a kind of pragmatic salience signaling. We also introduce an annotated data set of alternative plausibility ratings for negated verbal predicates and find that antonyms and distributionally similar words are plausible alternatives for negated verbs.
Contents

ABSTRACT

List of Tables

List of Figures

Chapter 1. Introduction
  1.1 Motivations

Chapter 2. Formal and Distributional Semantics

Chapter 3. Mathematical Approaches to Compositionality
  3.1 The Additive Model
  3.2 Multiplicative Models
  3.3 Issues with Additive and Multiplicative Models
  3.4 Composition by Function Application

Chapter 4. Theories of Negation
  4.1 Logical Negation
  4.2 Negation as Alternatives
  4.3 Opposition
  4.4 Event Semantics

Chapter 5. Experiments
  5.1 Corpus
  5.2 Negated Verbs in Distributional Semantic Space
  5.3 Comparing Negated and Non-Negated Verbs in the Semantic Space
  5.4 Negated Verbs and Verbs with Opposition
  5.5 Negated Verbs with Opposition and Their Alternatives
5.6. Negated Verbs without Opposition

Chapter 6. Discussion and Conclusions

6.1 Future Work

Appendix

Bibliography
List of Tables

Table 1  Mean cosine similarity of verb pairs..............................................................27
Table 2  Mean cosine similarity of verb and antonym pairs........................................29
Table 3  Comparison of alternative plausibility scores by ........................................31
type of relation to the target verb
### List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>A hypothetical semantic space for <em>practical</em> and <em>difficulty</em></td>
<td>8</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Hypothetical vectors <em>practical</em> and <em>difficulty</em></td>
<td>8</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Sample vectors <em>practical</em> and <em>difficulty</em> and their composition (additive model)</td>
<td>9</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Hypothetical vectors <em>v</em> and <em>u</em> and their composition</td>
<td>10</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Sample vectors <em>practical</em> and <em>difficulty</em> and their composition (pairwise multiplication model)</td>
<td>11</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Hypothetical vectors <em>v</em> and <em>u</em> and their composition</td>
<td>11</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Composition by function application</td>
<td>13</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Comparing mixture-based composition and composition by function application</td>
<td>14</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Negated verbs in the counted verbal semantic space</td>
<td>25</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Negated verbs in the composed verbal semantic space</td>
<td>26</td>
</tr>
</tbody>
</table>
CHAPTER 1.

Introduction

Since the time of Plato and Aristotle, negation has been a topic of considerable thought. Despite its seeming simplicity, negation is actually a complex and interesting linguistic phenomenon worthy of closer study for many reasons. The ability to communicate abstractly about what isn’t and not just concretely about what is is a unique trait of human language (Horn 1972).

The concept of negation is deceptively simple in propositional logic, where a proposition p is negated simply with a logical operator to become ¬p. In a conversational dialogue or narrative text, however, the reality of negation and its meaning is much more nuanced and complicated. Negation in natural language can take several forms, including lexical items no and not, clitic n’t, and negative polarity items such as never and nobody. In addition, defining the correct scope of negation is crucial to capturing the correct meaning. Negation also has a variety of pragmatic uses as a discourse strategy.

In this thesis, we examine distributional semantic models of negation, specifically as applied to verbs, with the goal of answering some questions about event semantics and negation. What is the meaning of a negated event? Do all negations license alternatives? Can we model negation of verbal predicates with distributional semantic methods?

In Section 1.1, we will go over the motivations for this type of research. In Chapter 2 we will give a brief overview of distributional semantics, and mathematical methods for
composition in distributional semantic models will be discussed in Chapter 3. Then in Chapter 4, we will go more in depth about different formal semantic and compositional distributional semantic treatments of negation. In Chapter 5 we present experiments examining negated verbs and their results. Chapter 6 concludes with discussion and interpretation of our experimental results, as well as directions for future work.

1.1 Motivations

One of the key challenges in distributional semantics (DS) is effectively capturing the meaning of grammatical function words, such as *no* and *not*, and using them to compose longer phrases. The principle of compositionality, first introduced by Frege, is a cornerstone of formal semantic theory: "The meaning of a whole is a function of the meanings of its parts and their mode of syntactic combination" (Portner and Partee 2002). Compositionality is something that model-theoretic and truth conditional semantics capture quite well but DS struggles with, which makes achieving a compositional distributional semantics difficult.

On the more practical side, DS plays a large role in natural language processing (NLP) applications. Distributional word vectors (often called word embeddings in NLP literature) are frequently used for major NLP tasks such as machine translation, sentiment analysis, question answering and syntactic parsing. We hope that compositional models of distributional information will not only help provide a richer account of lexical semantics and inform linguistic theory, but also help make DSMs significantly more robust for increasing performance on NLP tasks.

Boleda and Herbelot argue that a "fully-fledged" semantics must be a *complete* theory of meaning, which includes accurately capturing compositionality. Although
distributional approaches have been useful in NLP for a variety of tasks, if it is necessary to account for compositionality in order to be considered a complete theory of meaning, then both DS as a whole and efforts in compositional DS still fall woefully short. Compositional distributional semantic models (DSMs) face many other challenges at the syntax-semantics interface and semantics-pragmatics interface, but negation offers a compelling starting point as it is the simplest logical operator and pervasive in natural language.
CHAPTER 2.

Formal and Distributional Semantics

Formal semantics refers to the study of meaning in the tradition of the seminal works by Frege on sense and reference, Montagovian model-theoretic semantics and Tarskian truth conditional semantics (Boleda and Herbelot 2017; Frege 1892; Montague 1974; Tarski 1944). In contrast to formal semantics, DS aims to represent meaning "distributionally," that is, deriving meaning from the distributions of words in context over large linguistic corpora.

DS and vector space models of meaning are rooted in the distributional hypothesis - that words with similar meanings occur in similar contexts - and the general notion that the meaning of a given word is determined by its context (Harris 1954; Firth 1957). Therefore, DSMs are based on the idea that semantic similarity of words can be predicted based on their contextual similarity and the meaning of a lexical item can be inferred from its distribution in large corpora. From this perspective, lexical ambiguities can be largely resolved by relying on context as an adequate gauge of meaning. As lexical meanings in DSMs are sensitive to context, DSMs are therefore adept at capturing subtly graded aspects of meaning that are often difficult to tease apart in formal lexical semantics (Baroni et al. 2014; Boleda and Herbelot 2017).

In order to quantify lexical meaning, "traditional" DS uses the concepts of vectors and tensors from linear algebra, and more recently, function application to compose two
items. Scalars are ordinary numbers, like 42 or .003. Vectors are ordered lists of scalars, similar to a $1 \times n$ dimension matrix, and the length of a vector is referred to as its dimensionality$^1$. A high-dimensional vector is simply a long vector, and each element in the vector indicates a dimension in the high-dimensional semantic space, like the $x$ or $y$ coordinates of a point on a two-dimensional graph. A tensor is an array of numbers indexed by $n$ number of indices. The number of indices is called a tensor's order. A vector is a first-order tensor, and a matrix is a second-order tensor, and tensors with three or more indices are referred to as higher-order tensors.

Meanings of words in DS are represented as high-dimensional vectors in a vector space. In standard practice, each element in a word vector is derived from co-occurrence counts indicating how frequently the target word appears in a certain context in a given corpus. These raw counts are then transformed into association weights such as pointwise mutual information$^2$ that have a normalizing effect on the data, accounting for the fact that some words are more likely to occur than others overall (see Jurafsky and Martin 2017). The numbers in each vector indicate a specific location in the vector space, and a difference in meaning is reflected in the difference in values of the elements that make up the word vector. Different values indicate different locations in the vector space. The difference in location (and by extension, meaning) between two word vectors (and thus, words) can be measured by taking the cosine of the angle formed by the vectors, a similarity metric called

---

$^1$ Vectors are denoted in boldface (ex: $\mathbf{v}$ denotes vector $\mathbf{v}$).

$^2$ Pointwise Mutual Information is a measure of how much information is shared between two entities, or how likely two events are to occur together measured on a scale from negative to positive infinity. Positive PMI indicates that two events are more likely than expected by chance to co-occur. In practice, negative PMI values are normalized to 0 for ease of calculation; this is called Positive Pointwise Mutual Information (PPMI).
**cosine similarity.**³ (For a detailed discussion of vector semantics, formal mathematical descriptions of how word vectors are calculated, and various normalization techniques other than PPMI, see Jurafsky and Martin 2017: Ch. 15.)

Two words that occur in similar contexts are semantically similar, and the cosine similarity score between them will be accordingly high. Because word vectors are high-dimensional, any slight difference in the value of two corresponding elements will impact the vector's location (and thus meaning), allowing graded degrees of meaning to be represented by vectors that are essentially similar except for a few elements.

It has been suggested many times (Erk 2013; Lewis and Steedman 2013; Grefenstette and Sadrzadeh 2011; Coecke et al. 2010) that there ought to be a way to combine formal and distributional approaches, given that they work towards the same goal and appear to be complementary in their strengths and weaknesses. An issue in formal semantics is that words are assumed to be disambiguated before composition (see discussion in Baroni et al 2014). This is made all the more challenging by the fact that a given word rarely has a static meaning across all contexts. The meaning of content words in the lexicon varies greatly based on context. DSMs supposedly capture these graded lexical meanings more accurately, as representations are derived from large amounts of data available in text corpora, however it is hard to believe that one representation of meaning should suffice across all possible contexts.

As briefly mentioned above, DS offers no framework for modeling compositionality, especially in situations where content words must combine with function or grammatical

---
³ The distributional representation of a word \( u \) as a vector and the cosine similarity of two word vectors \( u \) and \( v \) is the cosine of the angle between them, calculated as normalized dot product of the vectors:  

\[
\cos(u, v) = \frac{\sum_i u_i v_i}{\|u\| \|v\|}
\]
words in ways that adhere to syntactic rules for well-formed sentences. Intuitively, it makes sense that if we are willing to accept that words have distributional representations (i.e. that word meanings correlate with their typical context), longer phrases and sentences should have distributional representations as well (Baroni et al. 2014). The major issue is that as phrases become longer, their frequency decreases. Even with large text corpora scraped from the web, the data sparsity of longer phrases and sentences makes calculating accurate distributions above the lexical level difficult, not to mention that simply calculating existing distributions is not a particularly robust method for handling novelty or recursion in natural language.

Modeling negation with DS is particularly difficult in large part because words and their negated counterparts often occur in similar contexts, which in turn makes their vectors similar (Mohammad et al. 2013; Kruszewski et al. 2017). Negation doesn’t (and shouldn’t) change the general domain associated with a word, but rather the specific values in the word vector (Hermann et al. 2013). For example, blue and not blue both belong to the domain of colors. A common domain makes the values in their respective word vectors similar enough that they appear in the same region of distributional space, but in reality, blue and not blue refer to (at a minimum) two different hues. Consequently, current approaches to compositionality in DS have trouble capturing negation because it is not reflected in the word vector of the target word in a way that can capture these fine-grained distinctions.
CHAPTER 3.

Mathematical Approaches to Compositionality

In this chapter, we will briefly summarize the work of Mitchell and Lapata (2010), Guevara (2011) and Baroni et al. (2014) to outline mathematical techniques for vector composition. Because word meanings are represented as vectors in DSMs, composition in DS naturally involves linear algebraic methods to combine word vectors. Composition continues to be an active area of research in DS because there are a variety of ways to combine vectors, all of which have their own strengths and weaknesses and none of which are perfect analogs to composition.

<table>
<thead>
<tr>
<th></th>
<th>music</th>
<th>solution</th>
<th>economy</th>
<th>craft</th>
<th>reasonable</th>
</tr>
</thead>
<tbody>
<tr>
<td>practical</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>difficulty</td>
<td>1</td>
<td>8</td>
<td>4</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 1: A hypothetical semantic space for practical and difficulty

Using the data in Figure 1 (reproduced here from Mitchell and Lapata 2010: p. 1401, Figure 3), where each scalar represents a co-occurrence count found in a corpus, let us define a word vectors for practical and difficulty as:

\[
\text{practical} = [0 \ 6 \ 2 \ 10 \ 4] \\
\text{difficulty} = [1 \ 8 \ 4 \ 4 \ 0]
\]

Figure 2: Hypothetical vectors practical and difficulty
3.1. The Additive Model

A natural first idea for combining vectors would be to simply add them, and indeed this is what the additive model does:

\[ \text{Vector addition: } \mathbf{p} = \mathbf{u} + \mathbf{v} \]

Where vector \( \mathbf{p} \) is the sum of each element of vectors \( \mathbf{u} \) and \( \mathbf{v} \)

\[
\text{practical difficulty} = \text{practical} + \text{difficulty} = \\
\begin{bmatrix}
0+1 & 6+8 & 2+4 & 10+4 & 4+0 \\
\end{bmatrix} = \\
\begin{bmatrix}
1 & 1 & 4 & 6 & 14 & 4 \\
\end{bmatrix}
\]

Figure 3: Sample vectors \text{practical} and \text{difficulty} and their composition via the additive model, \text{practical} + \text{difficulty}

Modeling composition this way makes sense at first glance, since composition itself can be described as a sort of "addition" of the meanings of smaller parts to create the meaning of the whole. But even Mitchell and Lapata admit that a serious flaw in this approach is that it assumes that the order of the constituents doesn't matter. Obviously, syntax is critically important to meaning and must be taken into account. The same words in different combinations can have radically different meanings, such as the phrases \textit{cold head} and \textit{head cold}. The additive model would find these two phrases semantically equivalent because there is no mechanism for capturing the significance of word order.

Assuming that composition is a symmetric operation flies in the face of other linguistic intuitions as well. For example, the meaning of \textit{good dog} is not equivalent to an object or concept that is fifty percent \textit{good} and fifty percent \textit{dog}. Rather, \textit{good} is a function over the set of all dogs that restricts the set to only those dogs which are good. Extending the example to \textit{some dogs} clarifies this distinction. Clearly the meaning of \textit{some dogs} is not
fifty percent *some* mixed with fifty percent *dogs - some* is a quantifier that restricts the set of all dogs.

Another disadvantage of the additive model is that there is no way to normalize the results of vector addition. For example, if an element in one of the vectors has an unusually high value, that anomalous value gets carried over into the composed vector:

\[
\mathbf{v} = [0, 6, 2, 10, 4] \\
\mathbf{u} = [200, 8, 4, 4, 2] \\
\mathbf{v} + \mathbf{u} = [200, 14, 6, 14, 6]
\]

Figure 4: Hypothetical vectors \(\mathbf{v}\) and \(\mathbf{u}\) and their composition, \(\mathbf{v} + \mathbf{u}\)

In Figure 4, the value of the element at index 0 in \(\mathbf{v}\) is 0. This indicates that co-occurrence count for the word represented by \(\mathbf{v}\) and the feature represented at index 0 is 0, and thus the feature represented at index 0 is not salient to the meaning of the word. However, the value of index 0 in \(\mathbf{u}\) is 200, indicating this feature is highly salient to meaning of the word represented by \(\mathbf{u}\). In this model, there is no way to tell whether the feature at index 0 is actually salient to the meaning of the composed vector or not, yet it is included regardless. In this way, addition can change both the direction and magnitude of the original vectors, which in turn can radically affect the composed vector’s location in the semantic vector space and, by extension, its meaning.

3.2. Multiplicative Models

Mitchell and Lapata (2010) introduce several types of multiplicative models for vector composition. The simplest model multiplies two vectors pairwise:

*Pairwise vector multiplication:* \( \mathbf{p} = \mathbf{u} \times \mathbf{v} \)

Where each component in vector \(\mathbf{p}\) is the pointwise multiplication of each component of vectors \(\mathbf{u}\) and \(\mathbf{v}\)*
practical difficulty = practical $\times$ difficulty =

$$\begin{bmatrix}
0 & 1 & 6 & 8 & 2 & 4 & 10 & 4 & 0
\end{bmatrix} = \begin{bmatrix}
0 & 48 & 8 & 40 & 0
\end{bmatrix}$$

Figure 5: Sample vectors practical and difficulty and their composition under the pairwise vector multiplication model, practical $\times$ difficulty

Mitchell and Lapata (2010) argue that multiplicative methods capture the meaning of a composed phrase more accurately than additive measures because multiplication implicitly normalizes the resulting vector to some degree.

$$v = \begin{bmatrix} 0 & 6 & 2 & 10 & 4 \end{bmatrix}$$

$$u = \begin{bmatrix} 200 & 8 & 4 & 4 & 2 \end{bmatrix}$$

$$v \times u = \begin{bmatrix} 0 & 48 & 8 & 40 & 8 \end{bmatrix}$$

Figure 6: Hypothetical vectors v and u and their composition, $v \times u$

In contrast to the additive model, in Figure 6 the first element of the composed vector does not reflect the anomalously high value of the first element of u. The corresponding element of the composed vector is also zero, reflecting the low importance of the feature represented by that element. These zero values later are removed from the vector for computational efficiency in a process called dimensionality reduction. However, this model still does tell us whether the feature at index 1 is actually salient to the meaning of the composed vector or not.

3.3. Issues with Additive and Multiplicative Models

Mitchell and Lapata also describe other types of multiplicative approaches, but in the interest of brevity and clarity we will not discuss them here. An important issue to note

---

4 See Jurafsky and Martin (2017), ch. 16 for a treatment of dimensionality reduction strategies, such as singular value decomposition (SVD), principle components analysis (PCA), and latent semantic analysis (LSA), among others.
at this juncture is that although it seems intuitive that adding or multiplying word vectors could constitute composition, it remains unclear whether or not applying these mathematical operations to word vectors actually has any relation to semantics despite the fact that these models generally perform well in practice (Guevara 2011).

Both multiplicative and additive models fail to capture key linguistic insights about the nature of composition. The first issue with these models is the assumption that composition is a symmetric function. This means that because they contain the same words, phrases such as *dogs chase cats* and *cats chase dogs* end up with the same distributional representation in vector space despite having nearly opposite meanings. In fact, composition can be considered an inherently asymmetric function because syntax and semantics are so closely interrelated.

Second, composing different combinations of syntactic categories of words seems like it ought to require different approaches. Composing adjective-noun phrases (ANs) is semantically a very different task than composing verb-object phrases. Third, these models completely ignore the role of function words in language, such as determiners and quantifiers, and negation.

The problem is that both additive and multiplicative models are ways to “mix” word vectors, which is not the same as true composition (Baroni et al. 2014). Mixture approaches to vector composition are unable to properly capture the meaning of composed phrases, such as in the *good dog* example above, because even the most basic composed phrases such as ANs are not an equal mixture of the meaning of their components.

3.4. Composition by Function Application
Despite the theoretical issues with additive and multiplicative models of composition in DS, they continue to be used widely because they generally yield good results (Mitchell and Lapata 2010; Grefenstette and Sadrzadeh 2011). This is likely due to the fact that experiments in DS tend to ignore key aspects of meaning, like word order and function words.

Baroni et al. (2014) ultimately reject both additive and multiplicative methods of composition as they are do not truly capture compositionality. Instead, they propose that the distributional meaning of words is encoded in distributional functions, not word vectors themselves. These distributional functions take word vectors as arguments and return composed vectors as output, and grammatical words represent functions over word vectors. In short, composition is akin to function application.

In this model, a one-argument distributional function (defined by Baroni et al. as adjectives or intransitive verbs, similar to one-place predicates in first order logic) is represented with matrix-by-vector multiplication:

\[ \text{Function application: } \mathbf{F} \times \mathbf{a} = \mathbf{b} \]

Where \( \mathbf{F} \) is a matrix that represents the function, \( \mathbf{a} \) is a vector that represents the argument to the function, and \( \mathbf{b} \) is the resulting vector of the function application.

![Figure 7](image) Composition by function application, reproduced here from Baroni et al. (2014: p.273, table 3)

Adjective \textit{old} is a function applied to noun \textit{dog} via matrix-by-vector multiplication.
This model appears to be quite similar to the pointwise multiplicative model of Mitchell and Lapata discussed in Section 5.2 above. But actually, the model allows for the same matrix to act differently with different arguments by incorporating a richer and more flexible definition of features, as matrices encode more information than vectors. In additive and multiplicative models of composition in DS, the values that make up word vectors come from manipulating co-occurrence counts. But where do the values in the matrix $F$ come from? In a nutshell, the idea is to extract from a corpus vectors $a$ and $b$, and then use machine learning techniques to find the ideal weights (matrix $F$) that would produce the expected target output for a given input.

Figure 8, below, reproduced here from Baroni et al. (2014: p.268, Figure 2), illustrates visually the differences between function application and the additive models, showing that an additive model produces a new vector with a different direction and magnitude, while function application only adjusts the existing word vector.

Figure 8: Mixture-based composition (left) takes two vectors and "mixes" them through addition or multiplication to make a vector of the phrase. Composition by function application (right) moves an existing vector by applying an argument over the vector to shift it to a new position in the semantic space.
Only recently have attempts been made to model true composition in distributional semantics instead of simple addition or multiplicative models. In addition, few studies focus on verbs as nouns and adjectives and their combination (ANs) are much simpler and more well understood (Baroni et al. 2014; Asher et al. 2017, inter alia). As we will see in Section 4.4, the intricacies of event semantics make analyzing negated verbs difficult.
CHAPTER 4.

Theories of Negation

4.1. Logical Negation

Logical negation acts as a truth-functional operator that switches the truth value of a sentence (Horn 1989). If the truth value of a proposition $p$ is True, the truth value of $\neg p$ is False. However, logical negation is not enough to describe the complexities of negation in natural language. Horn rejects the Fregean view of negation that "all forms of negation are reducible to a suitably placed 'It is not the case that..." because logical negation can capture the semantics of negation but is inadequate at capturing pragmatic ambiguities. The function of negation in discourse, or conversational negation, is not quite as straightforward as simply causing truth values to switch. Of course some negated utterances do directly correspond to their logical interpretation, but in many cases negation in discourse serves an additional and crucial pragmatic function rather than purely a semantic one (Horn 1989; Giora 2006).

4.2. Negation as Alternatives

Given the logical definition of negation, $\neg p$ is equivalent the complement to the set $p$. If $p$ is the set of dogs, then the complement of the set, $\neg$dog, includes anything that isn't a dog, from cat to table. Thus, the utterance not dog implies that the reference is a member of
the set \( \neg \text{dog} \). Taking this into consideration, it becomes natural to think about conversational negation as a kind of alternative-licensing function.

Thinking of negation as generating some kind of alternative set has its roots in Horn’s principle of alternate implicatures, and Rooth’s alternative semantics theory of focus (Horn 1972; Rooth 1985). Alternate implicatures predict that the types of predicates in negated clauses are limited by conversational (i.e., pragmatic) factors, including the belief state of the interlocuters and their shared common ground. The general idea of Rooth’s alternative semantics is that the purpose of focus in a sentence is to elicit an alternative set in addition to the typical semantic interpretation. For example:

\[ \text{(1)} \]
Meghan drank water.
\[ \llbracket \text{Meghan drank water.} \rrbracket^o = \text{drink}(m, w) \]
Where \( o \) is ordinary semantic value

\[ \text{(2)} \]
Meghan drank water.
\[ \llbracket \text{Meghan drank [water]} \rrbracket^f = \{ \text{drink}(m, y \mid y \text{ is a liquid}) \} \]
Where \( f \) is focus semantic value that licenses a set of alternative values.

Depending on the context of the situation focus in a statement such as "It's not a good movie" could license a set alternative implicatures equally:

\[ \text{(3) a.} \quad \text{It's not a good movie.} \implies \text{It's just okay.} \]

\[ \text{b.} \quad \text{It's not a good movie.} \implies \text{It's terrible.} \]

\[ \text{c.} \quad \text{It's not (just) a good movie.} \implies \text{It's the best movie.} \]

In a conversational context, negation can license an implied, unspoken, and constrained set of alternatives the negated element is referring to (Kruszewski et al. 2017). Within this alternative set, some items are naturally more likely to be true than others. For
example, if one asks, "Is it raining?" and the response is, "No, it's...", before the speaker has even finished the utterance, it is likely that an alternative set of weather terms has already come to mind - sunny, cold, windy, etc. In a typical situation, one would expect "sunny" to be a more likely response than "cold," because the conversational context refers to precipitation and not temperature.

This framing of negation lends itself well to DS, as it clarifies the link from negation to the notion of contextual similarity, the core of the distributional hypothesis. Since conversational negation is arguably more common in natural language than purely logical negation, it makes sense to attempt to model negation in a way that can capture more of the semantics of natural language.

Some important questions to ask (and that we try to answer empirically in Chapter 5) include:

- Do all negations license alternatives?
- If so, can these alternatives be reliably determined by cosine similarity, or some other metric?
- What bounds are placed on the domain of the alternatives, if any? (i.e., hypernyms, antonyms, hyponyms, etc.)

Previous work shows that it seems to be the case that negated nouns and adjectives license alternatives (Aina, 2017; Kruszewski et al., 2017). But whether this behavior extends to verbs (and also other syntactic categories of content words, although outside the scope of this thesis) is unclear.

Consider this example from Kruzewski et al:
(4) a. I do not want to watch a movie this afternoon...

b. ... I want to study.

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

(4b) is much more plausible than (4c). The predicates "watch a movie" and "study" seem to be only broadly related by the fact that they are things that can be accomplished in an afternoon, whereas "become an architect" cannot. Even without the overt time expression, "I don't want to watch a movie, I want to study" is a much more plausible utterance than "I don't want to watch a movie, I want to become an architect." This seems to be because there is an implicit expectation of how much time watching a movie will take (probably about two or three hours), and any alternative to watching a movie must roughly match that implicit expectation. Consider the following examples:

(5) a. I don't want to watch a movie, I want to go ice skating.

b. I don't want to watch a movie, I want to play a game.

c. ? I don't want to watch a movie, I want to send my friend a text message.

The alternative in (5a) *ice skating* is atelic and the alternatives in (5b) *play a game* and (5c) *send my friend a text message* are telic, yet only (5a) and (5b) are plausible. This supports what causes (im)plausibility isn't related to an action's telicity, but rather some kind of sense of how long the action may take. Ice skating and playing a game could take roughly as much time as watching a movie but sending a text message will certainly not.

Predicate negation gives a clue to the hearer that some kind of surprising or unexpected information about the event is going to be shared. The negation in these sentences reveals the unspoken but shared presuppositions, beliefs, and expectations held by the speaker and hearer about the activity or situation being discussed - in this case, the
temporal aspect of the event, which is what renders (5c) implausible. As we will discuss in Section 4.4, an event minimally consists of an action done at a time and in a place.

4.3. Opposition

Opposites surround us in the natural world - night and day, hot and cold - and opposition is a concept intuitively understood by ordinary speakers. Looking at binary pairs of opposites, it becomes apparent that opposition is closely linked with negation: "not true" is naturally equivalent to "false," and "not dead" is equivalent to "alive." These binary opposition relations are often associated with adjectives. However, some verbs also have opposites (Cruse 1986). (Importantly, although many verbs with opposition relations are stative verbs, many are not.)

One type of opposition in verbs is *complementary opposition*. Complementaries divide a concept into two mutually exclusive categories, such as succeed and fail. These complementaries are often expressed with a single lexical item, in this case a single verb. Interestingly, different from the adjectival binary opposites mentioned above, verbal complementaries often come in lexical triples. These triples are composed of a pair of complementaries and a second pair of opposites that share a lexical item with the first pair, such as start : continue : stop or arrive : stay : leave (Cruse 1986).

Linking back to DS, Cruse points out that the members of an opposite pair (or triple) have almost identical distributions, that is, that they occur in similar contexts. If opposites are semantically close, as having identical distributions indicates, they are more likely to be suitable alternatives for each other if suitability is measured by cosine similarity.

How opposition is linked to negation for items in an opposite triple of verbs is less clear. According to Cruse, speakers said that *obey* is the opposite of *command*, yet *disobey* is
the opposite of obey. He speculates that this is because obey : disobey are somehow "better" opposites but does not expand on what exactly "better" means. We suspect that being "better" opposites actually means a higher cosine similarity, which we investigate in Section 5.4.

4.4. Event Semantics

As any elementary schooler can tell you, verbs are "action words." If we consider an action a type of event, verbs are central to the description of that event. This is the main argument of Neo-Davidsonian event semantics, where the logical form of an action refers to some kind of underlying event (Davidson 1967). A verb is considered to be a predicate of events where every event is bound by an existential quantifier and each argument is linked to a semantic role.

(6) ⟦Meghan loves Will.⟧
    First order logic: love(m, w)
    Neo-Davidsonian event semantics: ∃e[loving(e) ∧ agent(e) = m ∧ theme(e) = w]

If negated verbs are roughly equivalent to negated events, what is the meaning of a negated event? Negation is sometimes regarded as challenging for event semantics, due to the fact that the broader context of the negated expression can alter its truth value (Krifka 1989). Krifka explains negation in event semantics with what he calls maximal events, or the fusion of all events at a specific time, and that negation takes scope under the event quantifier. The example Krifka gives of a negated event, where MXT refers to the maximal event at time t and AG refers to the agent, is:

(7) ⟦John didn’t laugh.⟧ =
    ∃e[MXT(e, t_r) ∧ ¬∃e'[laugh'(e') ∧ AG(e', John') ∧ e' ≤ e] ∧ τ(e) ≤ τt_r]
For all times \( \tau \) within the given timeframe \( t \), for all laughing subevents \( e' \) of the laughing event \( e \), there is no time at which a laughing subevent occurred.

(8) \[ \text{⟦John laughed.⟧} = \exists e [\text{laugh'}(e) \land AG(e, \text{John'}) \land \tau(e) \leq \tau t] \]

Here we must ask why negated past events are fusional, while non-negated past events are not. Given the interpretation of \[ \text{⟦John didn't laugh.⟧} \] above, one would expect the interpretation of \[ \text{⟦John laughed.⟧} \] to be equivalent, except that there is a laughing subevent that occurred during the given timeframe. Minimally, an event (negated or not) involves a temporal or spatiotemporal interval, and consequently all of the subintervals within that interval. But negated events additionally implicate some kind of information that contrasts or conflicts with the speaker or hearer's expectation, belief state or common ground assumptions.

It is unclear if fusion is something that occurs at all. More recent proposals argue against fusional event semantics and challenge the idea that Neo-Davidsonian compositional semantics and event semantics are incompatible. By including existential quantification in the meaning of the verb, negation and other quantifiers take scope above the event quantifier, which solves the incompatibility of Neo-Davidsonian event semantics with negation (Champollion 2010; Champollion 2015).
CHAPTER 5

Experiments

5.1. Corpus

We used the Open American National Corpus (OANC), a subset of the ANC Second Release that is a freely available corpus of over 15 million words of spoken and written American English. We used only the written portion of the corpus for consistency. The corpus has been automatically tagged with Penn Treebank style part of speech tags by GATE’s ANNIE system. We preprocessed the text by lemmatizing the corpus with NLTK and normalizing all 'n't' tokens to 'not'. For training purposes, we take instances of "not" followed by a verb and concatenate them into one unit (i.e., not run -> not_run). In general, antonymy is considered to be a lexical relation and therefore apply only to discrete lexical items - this is also the case in WordNet where "words" comprised of more than one lexical item are joined with an underscore. For the purposes of our experiments, we removed all lemmas in WordNet longer than one word. The unedited corpus has of 17,420,804 tokens, of which 1,591,188 are verbs, with 6,081 types. The edited corpus has 13,626,884 tokens, of which 32,776 are NOT_verbs with 2,535 types.

We created two distributional semantic models using gensim to make a Word2Vec CBOW model with standard parameters of dimensions d = 300 and a sliding window of 5

---

5 http://www.anc.org/data/oanc/
6 http://www.gate.ac.uk/
7 http://www.nltk.org
8 https://radimrehurek.com/gensim/
words in context (Mikolov et al. 2013). One model is of vectors calculated the unedited lemmatized corpus, and the other is of the edited NOT_verb corpus.

5.2. Negated Verbs in Distributional Semantic Space

We used k-means clustering (where $k = 2$) to see if verbs from the counted NOT_verb corpus could be automatically classified into the correct cluster (negative/non-negative) based on their vectors alone. Verbs from the counted NOT_verb corpus were correctly classified as negative or not negative with 90% accuracy, which indicates that there is a substantial difference in the vectors of negated verbs and those of non-negated verbs.

Interestingly, the clustering algorithm only miscategorized non-negated verbs as negated verbs. Misclassified verbs included many with a negative prefix, such as disappoint, disappear, displace, and discourage, and some that are associated with generally negative contexts such as depress, regret, die, impair, neglect, prohibit, harm, mistake, and infringe. Unsurprisingly, the misclassified verbs have a substantially lower mean and median frequency count than average for the corpus, following the intuition that verbs that occur less frequently are more likely to be miscategorized by the clustering algorithm. The mean and median frequency counts for the misclassified verbs were 308 and 202, respectively, and the mean and median frequency counts for the corpus over all were 16,911 and 5,997 respectively (see Appendix for a complete list).

5.3 Comparing Negated and Non-Negated Verbs in the Semantic Space
Using the DISSECT toolkit, we trained a lexical function composition model using least squares regression to learn a NOT function, represented as a 300 x 300 matrix\(^9\) (Georgiana Dinu and Marco Baroni 2013).

Recall the equation for the function application model of composition: \(F \times a = b\), or for our purposes, \(\text{NOT} \times \text{verb} = \text{NOT\_verb}\). \text{NOT\_verb} vectors were extracted directly from the \text{NOT\_verb} corpus to use as the gold standard. The corresponding NOT function matrix was estimated by optimizing the mapping from the \text{verb} vectors onto the corpus-extracted \text{NOT\_verb} phrase vectors. We used this NOT function to create a third DSM, a composed semantic space of vectors of verbs multiplied by learned NOT function.

![Figure 9: Negated verbs (red) in the counted verbal semantic space (blue).](image)

To show the difference between negated verbs and non-negated verbs in the semantic space, we plotted the counted NOT\_verbs (922 samples) and their corresponding non-negated verbs (430 samples) in the space constructed from the counted NOT\_verb corpus. The dimensionality of the vectors was reduced with PCA. Counted NOT\_verbs (red

---

\(^9\) Train/Test split: 70/30  
Number of training pairs: 21,630 (1313 pairs ignored because one of them is not found in the semantic space)  
Number of test pairs: 9,834
dots) occupy a narrow band of space within the counted non-negated verbs (blue dots), suggesting that negated verbs occupy a different place in the semantic space and thus are semantically distinct from non-negated verbs.

In order to examine the general effectiveness of the lexical function composition model, we then plotted composed NOT_verbs from the composed semantic space (605 samples) and compared them to the counted non-negated verbs that correspond to the NOT_verbs from the peripheral space (334 samples). The dimensionality of the vectors was reduced with PCA. Composed NOT_verbs (red dots) occupy a narrow band of space within the counted non-negative verbs (blue dots). The composed NOT_verbs appear to follow the same general pattern as the counted ones, indicating that the learned NOT function performs the transformation that we would expect to see.

![Diagram](image)

**Figure 10**: Negated verbs (red) in the composed verbal semantic space (blue).

To further test the effectiveness of our learned NOT function, we selected 100 random pairs of verbs with a corresponding negated verb in the composed space. We then calculated cosine similarity for each random pair.
Table 1: Mean cosine similarity of verb pairs

<table>
<thead>
<tr>
<th>Example</th>
<th>Verb, Random verb</th>
<th>NOT_verb, Random NOT_verb</th>
<th>Verb, NOT_verb</th>
</tr>
</thead>
<tbody>
<tr>
<td>distort, advocate</td>
<td>.13</td>
<td>.18</td>
<td>.25</td>
</tr>
</tbody>
</table>

An average cosine similarity score of .13 indicates that random verb pairs have low similarity. The average cosine similarity of pairs of random negated verbs was slightly higher, indicating that negation 'pushes' verbs closer to one another in the semantic space. This 'squeezing' phenomenon can be seen in the 2-D representations of the distributional semantic space in Figure 9 and Figure 10 above.

Lastly, we calculated the average similarity of a word and its negated counterpart in the composed space, obtaining an average cosine similarity of .25, indicating that (verb, NOT_verb) pairs are significantly more similar on average than random pairs of verbs.

5.4. Negated Verbs and Verbs with Opposition

Contrary to Cruse's restriction that antonyms must be gradable adjectives, verbs, nouns, adjectives and other words pairs in WordNet are linked with "antonymic" relations (which Cruse would term "opposition," see Section 4.3) (Cruse 1986). Using WordNet, we collected verbs with antonyms. 967 out of 6,081 verb types had an antonym listed in WordNet, or about 16% of the total types. 962,912 instances out of 1,591,188 instances of verbs were those with antonyms, or about 60% of the total verb tokens. Although there are relatively few verbs with antonyms overall, verbs with antonyms occur significantly more frequently than verbs without antonyms.
Our hypothesis is that for verbs with opposition, a negated verb elicits an alternative that is somehow constrained by the semantics of the original verb. Consider the following pairs:

(9) a. not stay : leave
b. not walk : ?run, ?sit
c. not look : ?gaze, ?find

Despite *stay*, *walk* and *look* all having antonyms in WordNet, it is unclear and largely context dependent what lexical item should correspond to the negated verb. In contrast, it seems much more natural for negated gradable adjectives to elicit a lexical antonym as an alternative:

(10) a. not cold : hot
b. not good : bad
c. not old : young

For adjectives, previous research has suggested that *not good* is semantically closer (in terms of cosine similarity) to *good* than *bad* (Aina 2017; Hermann et al. 2013). But does this hold true for verbs as well? Is *not stay* closer to *stay* than *leave*? To answer this question, we calculated the mean cosine similarity between verbs and their negated counterparts, the mean similarity between verbs and their antonyms, and the mean similarity between negated verbs and the antonyms associated with their non-negated forms.
Table 2: Mean cosine similarity of verb and antonym pairs

<table>
<thead>
<tr>
<th></th>
<th>verb, NOT_verb</th>
<th>NOT_verb, antonym(s) of verb</th>
<th>verb, antonym(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean cosine similarity</td>
<td>0.22</td>
<td>0.17</td>
<td>0.27</td>
</tr>
</tbody>
</table>

We found that in all cases, the mean cosine similarity was quite low (< .30) and that the highest similarity was between verbs and their antonyms. This is unsurprising given that a word and its antonym are often used in the same context according to the distributional hypothesis of antonyms proposed by (Mohammad et al. 2008). What is notable is that verbs and their negated counterparts have a significantly higher mean similarity than negated verbs and antonyms.

5.5. Negated Verbs with Opposition and Their Alternatives

To explore the kinds of alternatives generated by negation of verbs with opposition, we created a human-annotated data set of alternative plausibility ratings to see if DS is an appropriate method of choosing alternatives to negated verbs. Humans rated the plausibility of sentences containing negated verbs and potential alternatives.

We selected the fifty most common verbs with antonyms (according to WordNet) from our corpus to use in the data set. Verbs are content words, and as such are easier to model in DS than function words. But polysemy is a much more potent challenge with verbs as the meanings of some verbs change more dramatically in different contexts. To combat this ambiguity, so-called 'light' or delexicalized verbs were excluded as they have little semantic meaning outside of a complete phrasal context. Even so, the most common non-delexicalized verbs also happen to be some of the most notoriously polysemous, such as *run*. We did not attempt to solve this problematic aspect of verbal polysemy, since the vector representation of a word supposedly represents all contexts equally.
On one hand, it is difficult to assess the meaning of polysemous words isolated from context. On the other hand, providing too much context could lead to unintended or biased readings. To give our target verbs a bit more context for human annotators, but not so much context that it forces one reading over another, we embedded the verbs in a minimal sentential context:

\[(11) \quad \text{They didn't V1 (it), they V2+past (it).}\]

Past tense was chosen to make a more natural sounding sentence. *They* was chosen as a subject that would agree with an uninflected target verb and also provide ambiguity of animacy, as some verbs were more suited to accept animate arguments and some were more suited to accept inanimate arguments. *It* was added as needed to satisfy the necessary arguments for transitive verbs, with minimal semantic content that could interfere with meaning. Other minimal alterations to sentences were made for fluency and coherency.

Potential alternatives were collected from WordNet. These included the first ten (or all, if there were less than ten) antonyms listed, the ten closest neighbors from our distributional model, the ten first hypernyms listed, the ten first hyponyms listed, and ten random verbs from the corpus. We then placed each verb and its pair into a minimal sentential context, varying and adding pronouns and verb inflections as necessary for grammaticality. A total of 1,364 sentences were generated.

Annotators rated each sentence on a plausibility scale of 0 to 3, with 0 meaning the
sentence contained an unknown word and 3 meaning a highly plausible sentence (see Appendix for more details and examples).  

### Table 3: Comparison of alternative plausibility scores by type of relation to the target verb

Differences in scores are tested for statistical significance with Welch’s two sample t-test ($t$) where *** = $p < .001$ and ** = $p < .05$

<table>
<thead>
<tr>
<th>Relation</th>
<th>Mean Scores</th>
<th>Median Mean Scores</th>
<th>Number of Pairs</th>
<th>Mean Variance</th>
<th>Number of Pairs with Score &gt; 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Antonym</td>
<td>2.34 ***$t$</td>
<td>2.66</td>
<td>66</td>
<td>0.27</td>
<td>43 (65%)</td>
</tr>
<tr>
<td>Neighbor</td>
<td>1.49 ***$t$</td>
<td>1.33</td>
<td>450</td>
<td>0.24</td>
<td>64 (14%)</td>
</tr>
<tr>
<td>Hyponym</td>
<td>1.30 **$t$</td>
<td>1.0</td>
<td>79</td>
<td>0.18</td>
<td>3 (3.7%)</td>
</tr>
<tr>
<td>Hyponym</td>
<td>1.19</td>
<td>1.0</td>
<td>263</td>
<td>0.11</td>
<td>11 (4.1%)</td>
</tr>
<tr>
<td>Random</td>
<td>1.19</td>
<td>1.0</td>
<td>448</td>
<td>0.11</td>
<td>13 (2.9%)</td>
</tr>
<tr>
<td>Totals/Avg</td>
<td>1.5</td>
<td>1.0</td>
<td>1306</td>
<td>0.18</td>
<td>134</td>
</tr>
</tbody>
</table>

In line with our expectations, the sentences that contained negated verb-antonym pairs were rated by human annotators as having the highest mean and median plausibility scores by a wide margin, along with the highest percentage of pairs that had a mean score higher than 2. Distributional neighbors received the next highest mean scores and percentage of pairs with a mean plausibility rating above 2. Unlike similar experiments with alternatives of negated nouns and adjectives, distributionally similar alternatives of verbs did not receive particularly high plausibility ratings overall, although they did receive statistically significantly higher ratings on average than random alternatives.

5.6. Negated Verbs without Opposition

The results of our previous experiment confirm our hypothesis that antonyms make the best alternatives for verbs with opposition. However, there is a sizable minority

---

10 Kruszewski et al. mention that their 1-5 scale was too broad and ambiguous and that ratings in the middle of the scale tended to be less meaningful because of their high variance. We reduced their scale to 1-3, and added 0 as an N/A marker because some of the words generated by WordNet were quite strange.
(roughly 40%) of instances of verbs without antonyms in WordNet in our corpus and it is unclear whether negated verbs without antonyms also elicit alternatives.

To determine this, we designed an experiment aiming to elicit alternatives to negated verbs without antonyms. We chose the fifty most common and least polysemous verbs in our corpus and placed them in a minimal sentential context. The polysemy score of each word was calculated by counting the number of verb synsets of that word. If a word's polysemy score was less than or equal to the average number of synsets per word, that word was included in our pool of verbs.

Since we wanted to first determine if negated verbs without antonyms licensed alternatives at all (and not which alternatives out of a set were "best" as in the previous experiment), we left the second half of the utterance empty for annotators to provide their answer. The instructions given to the annotators were simply to complete the sentence using as few words as possible with the first thing that came to their mind. The sentences were then read aloud with neutral intonation and the annotators responded verbally. As in the previous experiment, we used the past tense and added additional pronouns where needed to create a natural sounding, semantically neutral sentence:

(12) \[ I \text{ didn't } V1 \text{ it, I } \underline{\phantom{\text{ }}}. \]

Almost all the sentences did elicit some kind of alternative from the annotators. This could be for a variety of non-linguistic reasons. Most importantly, annotators seemed anxious or apologetic when they could not quickly come up with a good way to finish the sentence and felt pressured to give a response, despite being assured that it was okay if they couldn't think of anything. Perhaps surprisingly, about twenty percent of the alternatives elicited were similar if not exactly the same for each sentence across
annotators. One interesting case is the instance of a set phrase, such as when all three annotators responded to *I didn't mean it, I ___* with *am sorry*.

An important caveat to mention is that WordNet is not an exhaustive database of all opposition relations of all verbs. Some of the verbs that do not have antonyms according to WordNet clearly have some kind of opposition relation, such as *live*, which would explain why all three annotators agreed that its opposite would be *die*. Other sentences that were completed similarly by multiple annotators include verbs that occur in relatively high frequency collocations, such as *compare* and *contrast*.

Failure to elicit an alternative seemed to be linked to some kind of semantic anomaly or pragmatic infelicity of the minimal sentential context, and not necessarily related specifically to the semantics of the verb. Requiring annotators to provide a verbal response quickly may have also impacted annotators’ ability to think of a response.
CHAPTER 6.

Discussion and Conclusions

Our investigations of negated verbs in distributional semantic space showed that negated verbs occupy an area in the space that is distinct from verbs in general. Verbs in general seem to be in roughly normal distribution across the semantic space, while negated verbs occupy a tight band within the verb space. These observations are comparable to previous work done on negated adjectives in distributional semantic space, where negated adjectives were shown to cluster tightly in an area of space distinct from adjectives in general (Aina 2017). This clustering effect appears to empirically validate the theoretical framework proposed by Hermann et al (2013), where although the domain of the vectors remains the same (verbs) the values of negated verb vectors are systematically different than non-negated vectors. That the negated verbs cluster within the larger cloud of verb space seems to indicate that, similar to the results of work done on nouns and adjectives, a large amount of the information in the predicate remains encoded in the vector regardless of whether or not it is negated. This in turn suggests that the effect of negation is largely related to pragmatic aspects of meaning rather than semantic.

Incorporating negation into verb vectors, whether by function application or counting, brings verbs closer together in the semantic space. We posit that this 'squeezing' effect is due to negation acting as a kind of dimensionality reduction function where negation focuses the many data points of a vector to its principal components, that is,
distills the vector down to its most salient features. Under this interpretation, predicate negation reveals conversational salience given context.

The results of our experiments exploring negated verbs and their alternatives show that antonyms or opposites are the best alternative to negated verbs. Going forward, this may be a problem as DS is well suited to find synonyms, but antonym detection is a bit more of a challenge despite ongoing research. The second-best alternatives being distributionally similar verbs implies that distributionally similar verbs will probably be the most plausible alternatives for verbs without antonyms. Although we did not collect extensive data on verbs without antonyms for this thesis, if our hypothesis proves correct, DS is better suited to finding the alternatives of verbs without antonyms. However, just because distributionally similar verbs were rated as significantly more plausible alternatives than random verbs, it doesn't necessarily mean that they will be good alternatives as their overall plausibility scores were not high. For verbs without antonyms, what makes an alternative more plausible than another remains to be investigated.

Conducting an experiment to elicit alternatives to negated verbs without antonyms showed that there are indeed some contexts where negated verbs do not license alternatives. A major problem is that although native speakers do get some feeling of opposition from certain verb pairs, the concept of opposition isn’t as clearly defined in comparison to other lexical semantic relations (Cruse 1986).
6.1 Future Work

The logical progression of these experiments is to carry out a similar annotation task as described in Section 5.5, but using verbs without opposition to see if alternatives generated in the same fashion as in our experiment would produce similar results or not. It would also be worthwhile to replicate the alternative elicitation task described in Section 5.6, but with verbs with opposition in order to see if the most commonly elicited verb matches expectations based on our annotated data.

Another goal for future work is to use a larger and more balanced corpus to generate our distributional semantic model. Although useful for our purposes here as an exploratory study, more data would give us clearer insight into the way verbs and their alternatives behave in semantic space.

It would also be interesting to see if this approach might also be applicable cross-linguistically under different negation paradigms, since unfortunately much of the work on DS is focused on English. Languages with multiple methods of verbal negation, such as Korean, and languages with multi-word verbal negation constructions, such as French, would be very interesting to analyze further with distributional methods.
Appendix

Supplemental Data from Experiments in Chapter 5

Section 5.2: List of Verbs Misclassified by k-Means Clustering

invent protest sleep abide solicit fly foresee die marry pretend regret jump substitute duplicate disappoint let sing bode wait smoke love shoot impair reproduce intervene figure hesitate feature reside disappear interview stock fare spot cease evoke resemble depress neglect tempt trade shake equip mandate possess name preclude promulgate explode affiliate rock prohibit harm adhere print shut forgive abate bite diminish alienate elaborate circulate exhibit dictate hospitalize coincide react constrain merit stain dispute watch cry steal conform please label sweat wear spoil last consume overlook replicate entitle waste fire deter sound warrant originate tolerate cast kick mistake cross drink hang kid ratify rest touch displace dare like subscribe rush heed fulfil pronounce advertise mitigate oblige behave invade quit burn confer incline swear obstruct apologize war heal discourage eat file prevail spell infringe bother cooperate scare

Section 5.5: Verbs with Antonyms

Plausibility Scale Guidelines:
0: I don't know what one or more of the words in this sentence means.
1: This sentence seems implausible.
2: This sentence could be plausible given the right set of circumstances.
3: This sentence is highly plausible.

Examples Given to Annotators:
They didn't eat, they fasted - 3
They didn't eat, they talked - 2
They didn't eat, they stabilized - 1
They didn't eat, they auspicated - 0

Sample Annotated Data from Annotator #1:
They didn't add it, they supplemented it. 1
They didn't suggest it, they rectified it. 1
They didn't indicate it, they specified it. 1
They didn't bring it, they came. 1
They didn't change it, they violated it. 2
They didn't establish it, they developed it. 2
They didn't ask, they moved. 1
They didn't leave it, they buried it. 1

Section 5.6: Verbs without Antonyms

I didn't provide it, I _____ got it. bought it. supplied it.
I didn't report it, I _____ ignored it. called it in. concealed it.
I didn't contain it, I ______. let it go. put it away. released it.
I didn’t describe it, I ______ showed you a picture. drew a picture. obfuscated it.
I didn’t identify it, I ______. didn’t know what it was. hid it.
I didn’t mean it, I ______. ‘m sorry. ‘m sorry. ‘m sorry.
I didn’t want it, I ______. ate it. didn’t like it. needed it.
I didn’t compare it, I ______. contrasted it. just took it. contrasted it.
I didn’t base it, I ______. just took it. contrasted it. just took it. contrasted it.
I didn’t express it, I ______. held my tongue. repressed it. (contained all of my emotions in a sad box.)
I didn’t perform it, I ______. acted it. wrote it down. failed it.
I didn’t produce it, I ______. fixed it. hid it. borrowed it.
I didn’t obtain it, I ______. stole it. decided against it. left it.
I didn’t live, I ______. died. died. died.
I didn’t create it, I ______. destroyed it. copied it. found it.
I didn’t result it, I ______. began it. never happened.
I didn’t expect it, I ______. demanded it. I was surprised. I was surprised.
It didn’t occur, it ______. slipped my mind. happened. never happened.
I didn’t explain it, I ______. denied it. was confused. lied.
I didn’t affect it, I ______. allowed it. let it happen. didn’t touch it.
I didn’t detect it, I ______. smelled it. didn’t notice. missed it.
I didn’t publish it, I ______. threw it away. printed it. distributed it.
I didn’t define it, I ______. ignored it. looked it up. ???.
I didn’t cause it, I ______. affected it. wasn’t there. made a mistake.
I didn’t examine it, I ______. let it go. just took it. ignored it.
I didn’t measure it, I ______. estimated it. just guessed. guessed.
I didn’t generate it, I ______. plagiarized it. bought it. found it.
I didn’t test it, I ______. fixed it. tried it. guessed.
I didn’t estimate it, I ______. knew it exactly. just guessed. guessed.
I didn’t predict it, I ______. it happened. knew what happened. guessed.
I didn’t reveal it, I ______. covered it. kept it hidden. concealed it.
I didn’t propose it, I ______. somebody else did. went along with it. kept it to myself.
I didn’t understand it, I ______. didn’t know it. asked a question. failed.
I didn’t spend it, I ______. earned it. saved it. saved it.
I didn’t choose it, I ______. wanted it. just got it. selected it.
I didn’t select it, I ______. got it anyway. took what was offered. took whatever there was.
I didn’t induce it, I ______. was natural. it just happened. watched it.
I didn’t argue, I ______. agreed. just agreed. reconciled.
I didn’t learn it, I ______. memorized it. faked it. acquired it.
I didn’t hear it, I ______. saw it. had to look it up. smelled it. (sense)
I didn’t design it, I ______. copied it. bought it. relaxed.
I didn’t reflect it, I ______. ???. ???.
I didn’t achieve it, I ______. earned it. tried but failed. relaxed.
I didn’t calculate it, I ______. estimated it. gave up early. guessed.
I didn’t speak, I ______. listened. whispered. listened.
I didn’t seek it, I ______. found it. came across it. waited for it.
I didn’t study, I ______. slept. went to bed. relaxed.
I didn’t discuss it, I ______ ignored it. I didn’t care. just did it.
I didn’t share it, I ____ kept it. kept it to myself. kept it.
I didn’t assess it, I _____ made it. took it. ???.

**Verbs where 2+ annotators agree on the alternative**

share  keep keep keep
compare contrast take contrast
live die die die
measure estimate guess guess
estimate know guess guess
spend earn save save
select get take take
argue agree agree reconcile
speak listen whisper listen
study sleep sleep relax

**Verbs where annotators failed to provide an alternative**

base ???
express ?? held my tongue repress
result ? began ?
define ignore look_up ??
reflect ??
assess make take ???
Bibliography


Donald Davidson. 1967. The Logical Form of Action Sentences.


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


