Not unreasonable: Carving vague dimensions with contraries and contradictions

Michael Henry Tessler (mhtessler@stanford.edu)
Department of Psychology, Stanford University

Michael Franke (mchfranke@gmail.com)
Department of Linguistics, University of Tübingen

Abstract

Language provides multiple ways of conveying the opposite: A person not happy can be unhappy, sad, or perhaps neither, just not happy. Rather than being redundant, we hypothesize that uncertainty about the meaning of negation markers allows listeners to derive fine-grained distinctions among these various alternatives. We formalize this hypothesis in a probabilistic model of gradable adjectives (e.g., happy), and use this to address an outstanding puzzle: how to interpret double negations (e.g., not unhappy). Our model makes surprising additional predictions about a putative difference between morphological antonyms (unhappy) and negated positives (not happy): Listeners should judge unhappy as more sad than not happy only when confronted with alternatives in context; when interpreted in isolation, we predict no difference in understanding. Two behavioral experiments confirm consistent orderings of interpretations that interact with the presentational context in the way predicted. These findings support the hypothesis that listeners represent uncertainty even about the most logical elements of language.

Keywords: semantics; pragmatics; negation; Bayesian cognitive model; Rational Speech Act

Introduction

If “Jones is not unhappy”, does that mean that Jones is happy? Jespersen (1924) believed not:

[T]wo negatives do not exactly cancel one another […] the longer expression is always weaker: “this is not unknown to me” or “I am not ignorant of this” means “I am to some extent aware of it,” etc. (p.332)

Negated antonyms (e.g., not unhappy) are thought to occupy a particular region of their associated scale (e.g., happiness), below positive antonyms (happy) but above negated positives (not happy) and antonyms (unhappy) (Fig. 1). Krifka, 2007). A straight-forward, compositional analysis, however, would map morphological negation via affixation (e.g., un- and negation particles (e.g., adverbial not) to proposition-level negation (¬) of modern standard logic. Such a logically transparent theory would predict that the two overt negation markers cancel each other out: not unhappy means ¬¬happy, or just happy. Orwell (1946) voices this opinion:

Banal statements are given an appearance of profundity by means of the “not un-” formation. […] It should be possible to laugh the “not un-” formation out of existence by memorizing this sentence: “A not unblack dog was chasing a not unsmall rabbit across a not ungreen field.” (p.357)

Alternatively, morphological antonyms (unhappy) could behave like lexical antonyms (sad) which seem clearly to express contrary opposition. Contraries cannot both be true but they can both be false (Horn, 1989): Jones may be neither happy nor sad, neither tall nor short. Not unhappy, then, would compete with happy as a pragmatic alternative and so would be contextually strengthened to a more specific interpretation, namely to a neutral or indifferent state (Horn, 1991), contra Jespersen’s intuition that not unhappy is a slightly positive state. It’s been noted that not happy should also be an alternative (Blutner, 2004), but how pragmatics contextually strengthens not happy has yet to be fully spelled out (Krifka, 2007). Furthermore, the meaning of not happy is at least as controversial as that of unhappy: Jespersen (1917) and Blutner (2004) argue that the two are basically identical in meaning, while Krifka (2007) cites examples like the following (taken from the internet):

It’s an absolutely horrible feeling to be unhappy, and I don’t even think I was unhappy, just not happy, if you know what I mean.

How does such a logical linguistic device—negation—give rise to a multiplicity of meanings? We posit that uncertainty about the meaning of overt negation markers (un-, not) interacts with communicative reasoning principles to give rise to fine-grained interpretations in the moment. We propose a probabilistic model of pragmatic reasoning in the Rational Speech Act tradition (Franke & Jäger, 2015; Goodman & Frank, 2016), combining previous work on gradable adjectives (e.g., tall; Lassiter & Goodman, 2015; Qing & Franke, 2014) with elements of lexical uncertainty (Bergen, Levy, & Goodman, 2016). Our model interprets not unhappy as a slightly positive state and predicts qualitative differences between morphological antonyms (e.g., {happy, unhappy}) and lexical antonyms which lack overt negation markers (e.g., {tall, short}). These differences are shown to be sensitive to the presence of explicit alternative utterances, as exists in Krifka’s example above. We compare model predictions to novel data from two experiments that measure interpretations for different kinds of adjectives in different contexts, uncovering subtle but reliable differences.

Computational model

Negation is the semantic operation of forming an opposite, but there are multiple kinds of semantic opposition. A contrary opposition is one where both predicates cannot be true at the same time, but both can be false (e.g., tall and short). Figure 1: Possible ordering of antonyms and their negations.
A contradictory opposition is one where the falsity of one predicate entails the truth of the other (e.g., even/odd positive integer). We posit that listeners have uncertainty about whether negation markers (not, un-) express contradictory or contrary opposition, and examine its effect on the interpretation of negated gradable adjectives (e.g., tall, happy).

Formal linguistic theories capture the meaning of gradable adjectives as a threshold function: \( \text{[happy]} = \lambda e . \text{happiness}(e) > \theta \), whose threshold variable \( \theta \) is supplied by the context (Kennedy, 2007). Here, we introduce contrary and contradictory negations into a pragmatic model that reasons about a speaker’s likely values for \( \theta \) (Lassiter & Goodman, 2015). Formally, if \( Hx \) expresses that \( x \) is happy, contradictory opposition is standard, bivalent, proposition-level negation \( \neg Hx \). Contrary opposition is a predicate-forming operation \( Hx \) which introduces its own free threshold \( \theta_H \) for gradable predicates. The latter is not iterable (Horn, 1989), so while it makes sense to iterate \( \neg \neg Hx \), it is impossible to iterate contrary negation. As a result, not happy and unhappy can be construed as either \( \neg Hx \) or \( \bar{H}x \), while not unhappy may be \( \neg \neg Hx \) or \( \neg \bar{H}x \) (Fig. 2).

Listener uncertainty about the interpretation of negation markers can be modeled as uncertainty about the speaker’s lexicon \( L \) (Bergen et al., 2016). We combine this technique with the model of Lassiter & Goodman (2015) that derives plausible thresholds \( \theta \) for gradable adjective interpretation:

\[
\begin{align*}
L_1(x, \theta, L \mid u) &\approx S_1(u \mid x, \theta, L) \cdot P(x) \cdot P(\theta) \cdot P(L) \\
S_1(u \mid x, \theta, L) &\approx \exp(\alpha \cdot \ln L_0(u \mid x, u, \theta, L) - \text{cost}(u)) \\
L_0(u \mid x, \theta, L) &\approx L(u, x, \theta) \cdot P(x)
\end{align*}
\]

Eqs. (1–3) are a Rational Speech Act (RSA) model, a recursive reasoning model wherein a pragmatic listener \( L_1 \) tries to resolve the intended meaning of an utterance \( u \) (e.g., “Jones is not unhappy”) by combining its prior beliefs about the degree of Jones’ happiness \( P(x) \), with the generative process of the utterance, a speaker model \( S_1 \). The speaker model \( S_1 \) describes an approximately rational agent (with degree of rationality \( \alpha \)) trying to inform a na"ive listener \( L_0 \) about the degree \( x \). The literal listener \( L_0 \) updates its prior beliefs \( P(x) \) via an utterance’s literal meaning in lexicon \( L \), where \( L(u, x, \theta) \) gives the truth-value of \( u \) in lexicon \( L \) applied to state \( x \) under threshold \( \theta \). The pragmatic listener has uncertainty about \( \theta \), which comes from an uninformed prior and is resolved by jointly reasoning about the likely degree \( P(x) \), the likely lexicon \( P(L) \), and the likelihood \( S_1(u \mid x, \theta, L) \) that a cooperative information-maximizing speaker would utter the adjective given a degree \( x \), threshold \( \theta \), and lexicon \( L \).

Does this model predict any qualitative differences between interpretations of positive adjectives (happy), antonyms (unhappy), negated positives (not happy) and negated antonyms (not unhappy)? Predictions depend on the space of possible meanings considered in the lexicon prior \( P(L) \). There are three natural possibilities for this space of meanings (Fig. 2). Our full uncertain negation model considers the full space of logically possible ways of mapping overt negation markers (un-, not) onto oppositional meanings (contraries and contradictions). Another possibility is for morphological negation (un-) to express a bona fide contrary meaning (a la sad), while maintaining uncertainty about particle negation (not; Fig. 2 purple dashes). A third class of meanings considers both negation markers to map only onto contradictory meanings (the logical negation or George Orwell class of meanings; Fig. 2 orange solid lines). We consider these different spaces of possible meanings as model variants for purposes of comparison.

For a given lexicon prior, listeners reason about which lexicon best explains a speaker’s utterance. Multiple utterances heard in the same context can help resolve the meaning of each other, and thus yield different interpretations than if the utterances were heard in isolation. We thus generate predictions for our models by conditioning \( L_1 \) on the observation of a speaker using all four adjective alternatives to talk about four different individuals in addition to a single utterance in isolation (Fig. 3 multiple utterances). Model predictions assume the following model parameters: \( P(x) = \mathcal{N}(0,1); \alpha = 1; \text{cost}(un) = 1; \text{cost}(not) = 2 \)

Upon hearing not unhappy, our uncertain negation model reasons that a truly compositional \( \neg\neg\neg Hx \) is implausible (intuitively because the speaker could have said the simpler happy) and interprets the utterance as signalling a slightly positive state (Fig. 3). When conditioning on a single utterance, uncertain negation does not differentiate antonyms (unhappy) from negated positives (not happy), as Jespersen (1917) and Blutner (2004) surmised. But when it hears multiple utterances in the same context, the model predicts that unhappy is more sad than not happy, producing the ordering hypothesized by Krifka (2007) in Fig 1. The bona fide contraries class of meanings also yields interpretations of negated antonyms as slightly positive, but predicts Krifka’s ordering for both single and multiple utterance conditioning. The logical negation class does not differentiate between negated antonyms and positives, nor between negated positives and antonyms. All models have more extreme interpretations when they condition on multiple utterances.

\[\text{1} \] Predictions are qualitatively similar when \( \text{cost}(un) = \text{cost}(not) \).
Behavioral experiments

The uncertain negation model predicts a partial ordering for morphological antonyms and their negations when heard in isolation (with antonyms ≈ negated positives), but a full ordering when present in the same context (Fig 3). As a control condition, we examine antonyms which do not have overt negation markers (e.g., short). These lexical antonyms should behave like bonafide antonyms, which predicts a full ordering regardless of context. Expt. 1 was exploratory and informed our computational modeling. Expt. 2 is a larger, more stringent, preregistered (osf.io/p7f25/) replication.

Experiment 1: Single utterances

Participants We recruited 120 participants from Amazon’s Mechanical Turk (MTurk). This number was arrived at with the intention of getting approximately 25 ratings for each unique item in the experiment. In all experiments, participants were restricted to those with U.S. IP addresses and at least a 95% work approval rating. The experiment took on average 3.3 minutes and participants were compensated $0.40.

Procedure On each trial, participants read a statement introducing a person using a gradable adjective of one of four adjective types: positives (e.g., happy, tall), antonyms (e.g., short, unhappy), and their respective negations (not X). Antonyms were one of two types: morphological (e.g., unhapp) and lexical (e.g., short). Participants rated the character on a scale from “the most positive person” to “the most antonym person” (item-dependent) using a slider bar (Fig 4A). Participants rated one sentence at a time and saw items from both antonym types throughout the experiment. Each participant completed a total of 16 trials, with exactly 2 repetitions of each adjective type for each antonym type.

Materials We used adjectives that described properties of people. We refer to a collection of the four associated adjective forms—positives, antonyms (morphological or lexical), and their negations using the particle “not”—that have the same positive adjective as an adjective set (e.g., one adjective set is happy, unhappy, not happy, not unhappy). 10 adjective sets were constructed for each antonym type (total 20) from an informal survey of the linguistics literature and taken from a list of “common opposites” available online (Table 2). Each trial of the experiment used an adjective from a distinct adjective set (e.g., if a participant rated unhappy, they rated no other adjective from the {happy, unhappy, ...} set).

Results 6 participants were excluded for self-reporting a native language other than English, leaving a remainder of 114 participants for these analyses.

The qualitative predictions of our models concern the ordering within a set of alternatives for different antonym types (morphological vs. lexical). To visualize the data, we compute normalized responses on a participant-wise basis (i.e., normalized response $r_{ij} = \frac{y_{ij} - \text{mean}_j}{\text{sd}_j}$ for trial i and participant j). Fig 5A shows the mean normalized responses and bootstrapped 95% confidence intervals for each of the four adjective types for morphological and lexical antonyms. Critically, as predicted by the uncertain negation model, adjective sets with morphological antonyms show only a partial ordering, while those with lexical antonyms show a full ordering.

Table 1: Items in Experiment 1.

<table>
<thead>
<tr>
<th>Morphological antonyms</th>
<th>Lexical antonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>attractive, unattractive</td>
<td>beautiful, ugly</td>
</tr>
<tr>
<td>educated, uneducated</td>
<td>brave, cowardly</td>
</tr>
<tr>
<td>friendly, unfriendly</td>
<td>fat, skinny</td>
</tr>
<tr>
<td>happy, unhappy</td>
<td>hard-working, cowardly</td>
</tr>
<tr>
<td>honest, dishonest</td>
<td>loud, quiet</td>
</tr>
<tr>
<td>intelligent, unintelligent</td>
<td>proud, humble</td>
</tr>
<tr>
<td>interesting, uninteresting</td>
<td>rich, poor</td>
</tr>
<tr>
<td>mature, immature</td>
<td>strong, weak</td>
</tr>
<tr>
<td>polite, impolite</td>
<td>tall, short</td>
</tr>
<tr>
<td>successful, unsuccessful</td>
<td>wise, foolish</td>
</tr>
</tbody>
</table>

http://www.enchantedlearning.com/wordlist/opposites.shtml

Throughout, we code adjective type using Helmer coding, which compares levels of a factor to the average of preceding levels, in order to compare antonym vs. negated positive levels of the adjective type factor.

This, and all subsequent regression models, were the maximal mixed-effects model that converged for the data set that additionally explained significantly more variance than models with simpler mixed-effects structures, using the lme4 package in R (Bates, Mächler, Bolker, & Walker, 2015).
vs. negated positive levels of adjective type interacted significantly with antonym type (morphological vs. lexical): $\beta = 0.029$, $SE = 0.012$, $t(16) = 2.4$, $p = 0.029$.

We also observe that negated morphological antonyms (e.g., not unhappy) were rated lower than negated lexical antonyms (e.g., not tall; Fig5A). Closer investigation of responses revealed that negated antonyms (and not other adjective types) received a bimodal distribution: The vast majority of ratings were slightly positive, while a clearly distinguishable minority distribution of ratings were slightly negative (e.g., not dishonest meaning not honest). This weakly negative interpretation for negated antonyms was seen at least somewhat in every item and in most participants. This interpretation may be the result of participants attributing politeness to the speaker: Not dishonest may be an indirect way of saying that a person is not honest (Yoon, Tessler, Goodman, & Frank, 2017).

![Figure 4](osf.io/p7f25/)

Figure 4: Example experimental trials for (A) single utterance (Expts. 1, 2) and (B) multiple utterances conditions (Expt. 2). “in the world” wording for endpoints was used in Expt. 2. (A) shows a trial from a morphological antonym set while (B) shows a lexical antonym set.

**Experiment 2: Single and multiple utterances**

Expt. 1 revealed an asymmetry: Lexical antonyms (e.g., short) were clearly distinguished from negated positives (e.g., not tall), whereas morphological antonyms were not (e.g., unhappy $\approx$ not happy). In Expt. 1, our adjective sets varied both in terms of their antonym type (morphological vs. lexical) as well as the actual degree scales being described (e.g., height for tall/short vs. happiness for happy/unhappy). Many adjective sets have both morphological and lexical antonyms (e.g., happy/unhappy/sad). Here, we aim to replicate the asymmetry findings using adjectives that describe the same semantic scales. Also, we test our second prediction that hearing multiple utterances in the same context will produce the full ordering for morphological antonym sets (Fig5).

**Participants** We recruited 750 participants from MTurk. The experiment comprised of four between-subjects experimental conditions arranged in a 2x2 Latin Square design: antonym type (morphological vs. lexical) X context (single vs. multiple utterances). 300 participants were assigned to each of the single utterance conditions, and 75 participants were assigned to each of the multiple utterances conditions. These numbers follow from the intention of getting approximately 45 ratings for each unique adjective in the experiment.

The single utterance task took on average 2.8 minutes and participants were compensated $0.40; multiple utterances took on average 5.5 minutes and participants were compensated $0.80. Participants who self-reported a native language other than English were excluded. This exclusion criterion, our planned sample size, the procedure, and the analysis described below were preregistered: [osf.io/p7f25/](osf.io/p7f25/).

**Materials** To best isolate the contribution of morphological vs. lexical antonyms, we curated adjective sets consisting of words for properties of people, such that both types of antonyms existed for the same positive adjective (e.g., happy $\rightarrow$ unhappy, sad; Table3). Lexical antonyms were selected from a set of possibilities produced from a small survey (n=18) on MTurk eliciting “opposites” for a list of 30 positive-form adjectives which had morphological antonyms (asking participants in the same experimental context as our interpretation studies, “What is the opposite of e.g., forgiving?”). From the list of freely-produced opposites (the vast majority of which were not morphological), the first author chose the one that intuitively best conveyed the same scalar dimension as the morphological antonym and which was not already used as a lexical antonym for another item (e.g., opposite of forgiving $\rightarrow$ resentful; opposite of kind $\rightarrow$ cruel, because opposite of friendly $\rightarrow$ mean). Ten out of the original 30 items were dropped for either not having such a well-suitied lexical antonym (e.g., moral) or for having a well-suitied lexical antonym that conflicted with another item (e.g., compassionate $\rightarrow$ cold, but also affectionate $\rightarrow$ cold).

**Procedure** In the multiple utterances conditions, participants rated all four adjective types simultaneously, each referring to a different person (Fig4), for a total of 12 trials. The single utterances conditions were similar to that of Expt. 1: Participants rated one sentence at a time (e.g., “Greg is not unhappy”), each from a unique adjective set (e.g., never rated both unhappy and not happy), completing a total of 12 trials, with exactly 3 repetitions of each adjective type (positive, antonym, and their negations). In contrast to Expt. 1, antonym type (morphological vs. lexical) was a between-participants factor. In addition, the slider bar endpoints were relabeled to “the most {positive, negative} person in the world”; without “in the world”, there is a salient interpretation of the endpoints indicating “the most {positive, negative} person (of these four)” in the multiple utterances conditions.

**Results** 35 participants were excluded for self-reporting a native language other than English, leaving 715 participants for these analyses. Mean normalized responses for each adjective type in each condition are shown in Fig5.
As we did in Expt. 1, we evaluate our hypothesis that morphological antonyms behave as predicted by the *uncertain negation* model (i.e., show a partial ordering) while lexical antonyms show a true ordering (like *bonafide contraries*). To do so, we considered data only from the *single utterances* conditions and built a linear mixed model predicting the raw, unnormalized ratings in terms of fixed effects of *antonym type* (morphological vs. lexical), *adjective type* (Helmert coded in order: antonym, negated positive, negated antonym, positive) and their interaction; the model also included random intercepts and random slopes of *adjective type* by-participant and by-item. Consistent with our hypothesis, the interaction between the *antonym* vs. *negated positive* levels of adjective type and antonym type (morphological vs. lexical) was significant: \( \beta = 0.011, SE = 0.004, t(565.1) = 2.68, p = 0.0076 \). We then analyzed the simple effects. Morphological antonyms were not significantly different than negated positives \( \beta = 5.3e-05, SE = 0.0027, t(52.5) = 0.02, p = 0.98 \), while lexical antonyms were interpreted more negatively than negated positives \( \beta = -0.011, SE = 0.003, p = 0.0003 \).

Our second main hypothesis is that context (single vs. multiple utterances) modulates the interpretive difference between morphological antonyms and negated positives. Specifically, we predict that morphological antonyms will be interpreted more negatively than negated positives in a context with multiple utterances. To evaluate this hypothesis, we considered data only from the morphological antonyms conditions and built a linear mixed model predicting the raw, unnormalized ratings in terms of adjective type, context (single vs. multiple utterances) and their interaction; the model also included random intercepts and random slopes of adjective type by-participant and by-item. This interaction was also significant: \( \beta = 0.032, SE = 0.0047, t(6457.1) = 6.73, p = 1.9e-11 \), and in the correct direction (see Fig. 5B).

### Table 2: Items used in Experiment 2.

<table>
<thead>
<tr>
<th>Positive adjective</th>
<th>Morphological antonym</th>
<th>Lexical antonym</th>
</tr>
</thead>
<tbody>
<tr>
<td>affectionate</td>
<td>unaffectionate</td>
<td>cold</td>
</tr>
<tr>
<td>ambitious</td>
<td>unambitious</td>
<td>lazy</td>
</tr>
<tr>
<td>attractive</td>
<td>unattractive</td>
<td>ugly</td>
</tr>
<tr>
<td>educated</td>
<td>uneducated</td>
<td>ignorant</td>
</tr>
<tr>
<td>forgiving</td>
<td>unforgiving</td>
<td>resentful</td>
</tr>
<tr>
<td>friendly</td>
<td>unfriendly</td>
<td>mean</td>
</tr>
<tr>
<td>generous</td>
<td>ungenerous</td>
<td>stingy</td>
</tr>
<tr>
<td>happy</td>
<td>unhappy</td>
<td>sad</td>
</tr>
<tr>
<td>honest</td>
<td>dishonest</td>
<td>deceitful</td>
</tr>
<tr>
<td>intelligent</td>
<td>unintelligent</td>
<td>stupid</td>
</tr>
<tr>
<td>interesting</td>
<td>uninteresting</td>
<td>boring</td>
</tr>
<tr>
<td>kind</td>
<td>unkind</td>
<td>cruel</td>
</tr>
<tr>
<td>mature</td>
<td>immature</td>
<td>childish</td>
</tr>
<tr>
<td>patriotic</td>
<td>unpatriotic</td>
<td>traitorous</td>
</tr>
<tr>
<td>polite</td>
<td>impolite</td>
<td>rude</td>
</tr>
<tr>
<td>rational</td>
<td>irrational</td>
<td>crazy</td>
</tr>
<tr>
<td>reliable</td>
<td>unreliable</td>
<td>flaky</td>
</tr>
<tr>
<td>resourceful</td>
<td>unresourceful</td>
<td>wasteful</td>
</tr>
<tr>
<td>sincere</td>
<td>insincere</td>
<td>fake</td>
</tr>
<tr>
<td>tolerant</td>
<td>intolerant</td>
<td>bigoted</td>
</tr>
</tbody>
</table>

Many dimensional scales lack units. Speakers cannot say they are 42 units *happy* like they can say they are 6’1” *tall*. Instead, speakers can use modifiers and morphemes to carve more precise meanings from otherwise vague dimensions. A person *not unhappy* is neither sad nor truly happy, but residing in some marginally positive state that is difficult to refer to because degrees of happiness lack precise units.

This work resolves an outstanding puzzle in natural language understanding: How to interpret double negatives (e.g.,
not unhappy). Krifka (2007) critiques previous pragmatic theories for either being underdetermined (Blutner, 2004) or making the wrong prediction (Horn, 1991). Here, we propose a theory of negated antonym interpretation by introducing negation into a pragmatic model of adjective interpretation. But morphological antonyms (unhappy) do not behave like bonafide antonyms (sad). We discovered and confirmed a surprising empirical result that challenges “established” intuitions in linguistics: unhappy and not happy are not immediately differentiated, except when both are present in the same context. Our model that represents uncertainty about how to parse overt negation markers (un-, not) predicts this very result, while alternative models that treat negation with a fixed meaning fall short. While the items used in our experiment were all evaluative statements about people, we think it is not unreasonable to expect the asymmetries observed in our data to generalize to other negated antonyms.

It is noteworthy that we are able to recover, both in our model and empirically, the ordering predicted by Krifka (2007) for morphological antonyms when a listener hears multiple adjectival utterances in the same context (multiple utterances condition). This work thus carries with it an account of a robust linguistic intuition: Potentially equivalent expressions receive differential interpretations when observed uttered by the same speaker in close proximity. Reasoning about lexical ambiguity, listeners conclude that a choice of different expressions may be most likely for a speaker who differentiates meanings. More generally, the inferences modeled here can be seen as an instance of mutual exclusivity (Markman, 1989), in which listeners resolve uncertainty about multiple elements of meaning simultaneously.

Our formalization of lexical uncertainty about the meaning of natural language negation builds on a growing movement to treat the combinatorial rules of grammar as not totally separable from the lexicon (e.g., Bybee, 2006; O’Donnell, 2015). Recent psycholinguistic evidence supports the idea that utterances which are heavily used will be processed as unique lexical entries while less frequent phrases will be understood compositionally (Morgan & Levy, 2016). The two types of negation meaning we considered—contrary and contradiction opposition—can be seen as a lexicalized form of opposition (with the adjective receiving its own threshold variable) and a compositional rule (logical negation), respectively. In our modeling, we assumed all lexica (all logically-possible interpretations of negation) were equally likely: A further test of our negation uncertainty model would be to see if frequency can serve as a proxy for this prior over lexica.

To negate is to make true false, but for statements that are truly vague, the behavior of negation is not so obvious. We present a computational explanation for why this is so, and provide empirical data that sheds new light on the age old question of meaning and opposition.

Acknowledgements

This work was supported in part by NSF Graduate Research Fellowship DGE-114747 to MHT.

References


Experimental paradigms, computational models, analysis scripts, and data for this paper can be found at [https://mhtess.github.io](https://mhtess.github.io)