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) suggested 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 adjectives* ("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 ($\neg$) of standard logic. Such a logically-transparent theory would predict that the two overt negation markers cancel each other out: *not unhappy* means $\neg\neg$*happy*, or just *happy*. Orwell (1946) voiced 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*. Two 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 (1924)'s intuition that *not unhappy* is a slightly positive state. Such an analysis would further depend on the meaning of *not happy* (Blutner, 2004), of which there is little agreement: Jespersen (1917) and Blutner (2004) argue that it means the same as *unhappy*, while Krifka (2007) cites examples like:

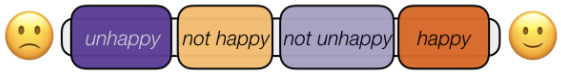

Figure 1: Possible ordering of antonyms and their negations.

> 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. *(taken from the internet)*

How does such a logical linguistic device—negation—give rise to a multiplicity of meanings? Both syntactic (Cable, 2017) or pragmatic (Rett, 2014) mechanisms could be at play, but heretofore there have been no computational accounts tested against human behavioral data. 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). We formalize the hypothesis that uncertainty about the meaning of overt negation markers (*un-*, *not*) interacts with pragmatic reasoning to give rise to fine-grained interpretations in the moment. These differences are shown to be sensitive to the presence of explicit alternative utterances, as suggestsed by Krifkas'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*). 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: $[\![happy]\!] = \lambda x.\,\text{happiness}(x) > \theta$, whose threshold variable θ 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 θ (Lassiter & Goodman, 2015). Formally, if Hx expresses that x is happy, contradictory opposition is standard, proposition-level negation $\neg Hx$. Contrary opposition, on the other hand, forms a new predicate $\tilde{H}x$ which introduces its own threshold $\theta_{\tilde{H}}$. Contradictory opposition is iterable ($\neg\neg Hx$) but contrary opposition is not (Horn, 1989). As a result, *not happy* and *unhappy* can mean either $\neg Hx$ or $\tilde{H}x$, while *not unhappy* may mean $\neg\neg Hx$ or $\neg\tilde{H}x$ (Fig. 2).

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

$$L_1(x,\theta,\mathcal{L} \mid u) \propto S_1(u \mid x,\theta,\mathcal{L}) \cdot P(x) \cdot P(\theta) \cdot P(\mathcal{L}) \quad (1)$$

$$S_1(u \mid x,\theta,\mathcal{L}) \propto \exp\left(\alpha \cdot \ln L_0(x \mid u,\theta,\mathcal{L}) - \text{cost}(u)\right) \quad (2)$$

$$L_0(x \mid u,\theta,\mathcal{L}) \propto \mathcal{L}(u,x,\theta) \cdot P(x) \quad (3)$$

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 α) trying to inform a naive 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 $\mathcal{L}$, where $\mathcal{L}(u,x,\theta)$ gives the truth-value of u in lexicon $\mathcal{L}$ applied to state x under threshold θ. The pragmatic listener has uncertainty about θ, which comes from an uninformed prior and is resolved by jointly reasoning about the likely degree $P(x)$, the likely lexicon $P(\mathcal{L})$, and the likelihood $S_1(u \mid x,\theta,\mathcal{L})$ that a cooperative information-maximizing speaker would utter the adjective given a degree x, threshold θ, and lexicon $\mathcal{L}$.

Any predicted qualitative differences between interpretations of antonym pairs and their negations will depend on the space of meanings considered in the lexicon prior $P(\mathcal{L})$. There are three natural possibilities. One is that negation markers map only onto contradictory meanings (the *logical negation* or *George Orwell* class of meanings; Fig. 2 blue dashed lines). Alternatively, morphological negation (*un-*) could express a bonafide contrary meaning (a la *sad*; Fig. 2 purple fuzzy lines). Finally, both overt negation markers (*un-*, *not*) could be mapped to either oppositional meaning—contrary or contradiction—up to the constraints of compositionality described above (full *uncertain negation* model; all meanings shown in Fig. 2). For maximal similarity to the full

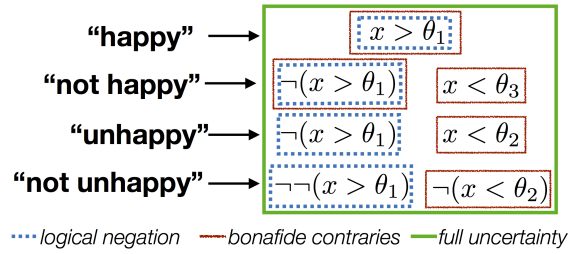

Figure 2: Space of possible meanings in the lexicon prior for the *logical negation*, *bonafide contraries*, and the full *uncertain negation* models.

uncertain negation model, we allow the *bonafide antonyms* model to maintain uncertainty about particle negation (*not*).[1]

For a given lexicon prior, listeners reason about which lexicon best explains a speaker's single utterance (Fig. 3, *single utterance*). If a speaker uses multiple utterances in the same context, the listener may have even more information about the speaker's lexicon. We thus also generate predictions for our models by conditioning L_1 on the observation of a speaker using all four adjective alternatives to describe four different individuals in the same context (Fig. 3, *multiple utterances*). Model predictions use the following minimally assumptive model parameters: $P(x) = \mathcal{N}(0,1); \alpha = 1; \text{cost}(un) = 1 < \text{cost}(not) = 2$.[2]

Upon hearing *not unhappy*, our *uncertain negation* model reasons that a truly compositional $\neg\neg$*happy* 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 *bonafide 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.

Behavioral experiments

The *uncertain negation model* predicts a partial ordering for morphological antonyms and their negations when heard in isolation (with antonyms $\approx$ 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

[1] Instead, in the *bonafide antonyms* model, one could fix the meaning of particle negation (*not*) to be logical negation. This turns out to not produce any qualitative difference from maintaining uncertainty about *not*'s meaning.

[2] Predictions are qualitatively similar when $\text{cost}(un) = \text{cost}(not)$.

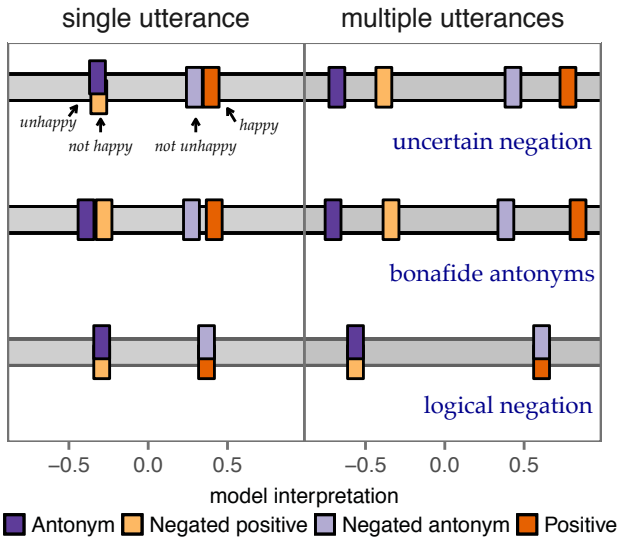

Figure 3: *Uncertain negation* listener model (Eq.1) posterior expectations on a normalized scale (x-axis) for different adjective types (color). The space of possible meanings is restricted for *bonafide antonyms* and *logical negation* simulations (Fig.2). Paddles are vertically-offset when overlapping.

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; in addition, participants who self-reported a native language other than English were excluded. The experiment took on average 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., *unhappy*) 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 1).[3] 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 $\{happy, unhappy, \dots\}$ set).

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

Table 1: Items in Experiment 1.

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 $r'_{ij} = \frac{r_{ij} - mean_j}{sd_j}$ 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.

To confirm these observations, we 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)[4], and their interaction; the model also included random intercepts and random slopes of *adjective type* by-participant and by-item.[5] Consistent with our observations, the difference between the *antonym* vs. *negated positive* levels of adjective type interacted significantly with antonym type (morphological vs. lexical; $\beta = 0.029$, $t(16) = 2.4$, $p = 0.029$).

[3] `http://www.enchantedlearning.com/wordlist/opposites.shtml`

[4] Throughout, we code adjective type using Helmert 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.

[5] 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).

We also observe that negated morphological antonyms (e.g., *not unhappy*) were rated lower than negated lexical antonyms (e.g., *not tall*; Fig. 5A). Closer investigation of responses revealed that negated antonyms (and not other adjective types) received a bimodal distribution: Most ratings were slightly positive but 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 present 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).

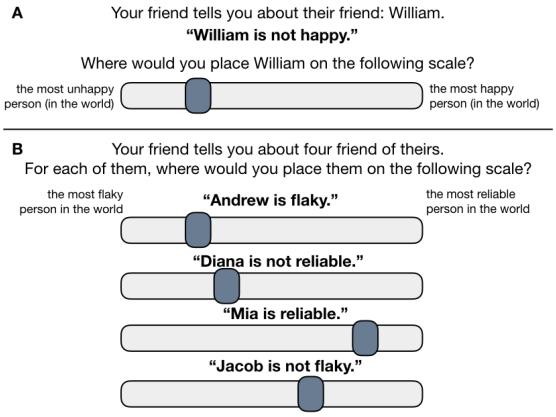

Figure 4: Example experimental trials for (A) single utterance (Expts. 1, 2) and (B) multiple utterances conditions (Expt. 2). "in the world" wording for endpoints was used in Expt. 2. (A) shows a trial from a morphological antonym set while (B) shows a lexical antonym set.

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 (Fig. 3).

Participants We recruited 750 participants from MTurk. The experiment comprised of four between-subjects experimental conditions arranged in a 2x2 design: *antonym type (morphological vs. lexical)* X *context (single vs. multiple utterances)*. 300 participants were assigned to each *antonym type* in the *single utterance* contexts, and 75 participants were assigned to each in 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 3 minutes and participants were compensated \$0.40; *multiple utterances* took on average 5 minutes and participants were compensated \$0.80. Exclusion criterion, sample size, procedure, and the analysis described below were preregistered: `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*; Table 2). 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-suited lexical antonym (e.g., *moral*) or for having a well-suited 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 (Fig. 4B), 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 Fig. 5B.

As we did in Expt. 1, we evaluate our hypothesis that morphological antonyms behave like the *uncertain negation* model (i.e., show a partial ordering) while lexical antonyms show a true ordering (like *bonafide contraries*). We considered data only from the *single utterances* conditions and built a linear mixed model predicting the unnormalized ratings in

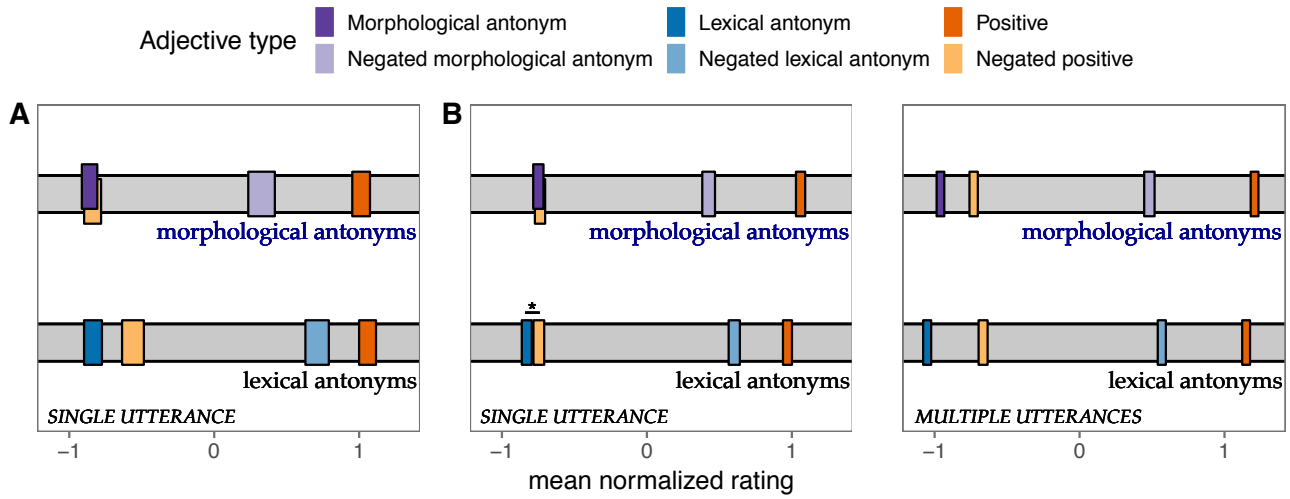

Figure 5: Empirical ratings for adjective sets with morphological antonyms (e.g., "unhappy") and lexical antonyms (e.g., "sad"). Width of paddle denotes a bootstrapped 95% confidence interval. Paddles are vertically-offset when overlapping. A: Expt. 1: Participants rated adjectives in isolation; a single participant saw both morphological and lexical antonym items. B: Expt. 2: Participants rated adjectives in isolation (left) or simultaneously (right); each ribbon denotes a between-participant condition.

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

Table 2: Items used in Experiment 2.

terms 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 ($\beta = 0.011$, $t(565) = 2.68, p = 0.0076$). We then analyzed the simple effects. Morphological antonyms were not significantly different than negated positives ($\beta = 5.3e-05$, $t(52) = 0.02, p = 0.98$), while lexical antonyms were interpreted more negatively than negated positives ($\beta = -0.011$, $t(280) = 3.66$, $p = 0.0003$).[6]

[6]The random effect structure for the simple effects models mirrored the full model. The only difference was that in analyzing the lexical antonyms, the random effect of adjective type by-item needed to be dropped in order for the model to converge.

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 ($\beta = 0.032$, $t(6457) = 6.73, p = 1.9e-11$), and in the correct direction (see Fig. 5B). As an exploratory analysis, we examined these effects in a full three-way interactive model. The relevant *antonym* vs. *negated positive* by adjective type (lexical, morphological) by context three-way interaction was in the direction of lexical antonyms showing a larger *antonym* vs. *negated positive* difference in the explicit context, but it was not significant ($\beta = 0.012$, $t(469) = 1.64, p = 0.1$).

Discussion

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 provides a computational solution to an outstanding puzzle in natural language understanding: How to

interpret double negatives (e.g., *not unhappy*; Krifka, 2007; Rett, 2014). We additionally 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 interpret overt negation markers (*un-*, *not*) predicts this very result, while alternative models that treat negation with a fixed meaning fall short. One limitation of this work is that we stipulate, rather than derive, differences in meaning for morphological vs. lexical antonym pairs (cf., Rett, 2014).

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 contradictory 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 priori*: 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.

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

Acknowledgements

This work was supported in part by NSF Graduate Research Fellowship DGE-114747 to MHT. We are grateful to our three reviewers and coordinator for helpful comments.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Bergen, L., Levy, R., & Goodman, N. D. (2016). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, *9*.

Blutner, R. (2004). Pragmatics and the lexicon. *Handbook of Pragmatics*, *488514*.

Bybee, J. L. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733.

Cable, S. (2017). The good, the 'not good', and the 'not pretty': Negation in the negative predicates of tlingit.

Franke, M., & Jäger, G. (2015). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. In *Zeitschrift fr sprachwissenschaft* (pp. 3–44).

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*(11), 818–829.

Horn, L. R. (1989). *A natural history of negation*. University of Chicago Press.

Horn, L. R. (1991). Duplex negatio affirmat.: The economy of double negation. *CLS 27-II: Papers from the Parasession on Negation*, 80–106.

Jespersen, O. (1917). *Negation in english and other languages*. Kobenhavn: Host.

Jespersen, O. (1924). *The philosophy of grammar*. London: Allen & Unwin.

Kennedy, C. (2007). Vagueness and grammar: the semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*, *30*, 1–35.

Krifka, M. (2007). Negated Antonyms: Creating and Filling the Gap. *Presupposition and Implicature in Compositional Semantics*, 163–177.

Lassiter, D., & Goodman, N. D. (2015). Adjectival vagueness in a Bayesian model of interpretation. *Synthese*.

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction*. MIT Press.

Morgan, E., & Levy, R. (2016). Abstract knowledge versus direct experience in processing of binomial expressions. *Cognition*, *157*, 384–402.

Orwell, G. (1946). Politics and the english language. *Horizon*.

O'Donnell, T. J. (2015). *Productivity and reuse in language: A theory of linguistic computation and storage*. MIT Press.

Qing, C., & Franke, M. (2014). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model. *Proceedings of SALT*, *24*, 23–41.

Rett, J. (2014). *The semantics of evaluativity*. Oxford University Press.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2017). "I won't lie, it wasn't amazing": Modeling polite indirect speech. In *Proceedings of the 39th annual meeting of the cognitive science society*.