Not unreasonable: Why two negatives don't make a positive

Michael Henry Tessler[1,3] & Michael Franke[2]

[1] Massachusetts Institute of Technology, Department of Brain and Cognitive Sciences

[2] University of Osnabrück, Department of Cognitive Science

[3] Stanford University, Department of Psychology

Author Note

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

Correspondence concerning this article should be addressed to Michael Henry Tessler, 43 Vassar St, Cambridge, MA 02139. E-mail: tessler@mit.edu

Abstract

Logic tells us that two negatives make a positive, but in language, things are not so black and white: A person *not unhappy* may not be entirely *happy*. We hypothesize that non-logical uses of double negatives like *not unhappy* stem from listeners entertaining flexible meanings for negation markers like *not* and *un*-, which context can then help disambiguate. We formalize this hypothesis in a computational model of language understanding, which predicts that *not unhappy* means something different than *happy* while also entertaining that single negations (*unhappy* and *not happy*) can be interpreted identically when context does not suggest otherwise. Across three experiments ($n = 995$), we confirm these predictions experimentally and further find that double negations that flagrantly use the same negation marker twice (e.g., *not not happy*) can also be interpreted in subtle ways. These findings suggest that even one of the most logical elements of language—negation—can mean many things at once.

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

Word count: 7228

Not unreasonable: Why two negatives don't make a positive

> *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."* (Orwell, 1946) (p. 357)

Introduction

For the most part, language reliably conveys our thoughts. When subtle feelings arise that are more difficult to express in common speech, speakers may resort to creative language (e.g., metaphors; Lakoff, 2008) or uncommon, perhaps novel terms like being "plateaued" (Bardwick, 1986) or residing in a "zone of indifference" (Sapir, 1944). There are systematic ways of expressing subtle gradations as well, such as with constructions involving double negations: A person *not unhappy* is probably not entirely happy. The interpretation of double negatives, in particular, is important from a logical perspective (Horn, 1989; Krifka, 2007) as well as a legal one, where such constructions are surprisingly common (Tiersma, 1999). More broadly, pinning down the meaning of negation—one of the most logical elements of language—is crucial for much of psychological science that depends upon linguistic information to convey task instructions and content (e.g., in the psychology of reasoning; Geurts, 2003; Lassiter & Goodman, 2015)

A common intuition about the meaning of double negatives is expressed by Jespersen (1924):

> [T]wo negatives do not exactly cancel one another [. . .]; the longer expression is always weaker: "this is not unknown to me" [. . .] means "I am to some extent aware of it," etc. (Jespersen, 1924) (p. 332)

In other words, *not unhappy* (a *negated antonym*) should indicate a slightly positive state, below that of *happy* but perhaps more positive than neutral (i.e., above the zone of indifference). These intuitions are subtle, not universally agreed upon, and further

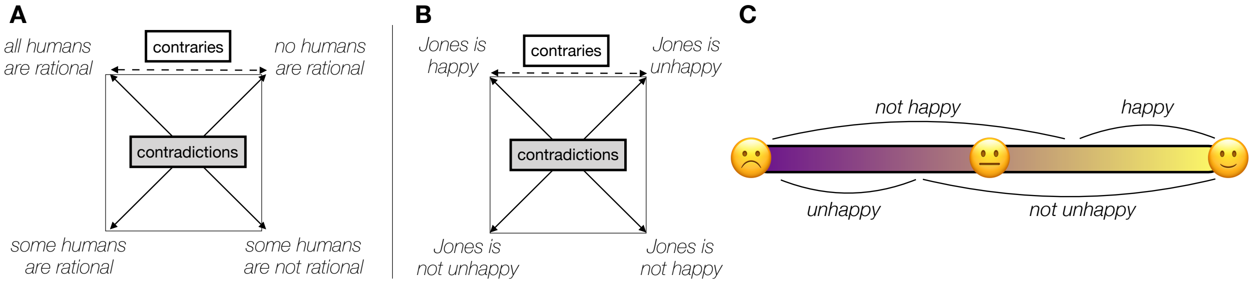


Figure 1. Contraries and contradictions are two different kinds of negation. A: Aristotle's classic Square of Opposition applied to quantifiers (all, some, none). B: An Aristotelean analysis of negated antonyms (happy, unhappy, not unhappy). C: Interpretations of negated antonyms in terms of *degrees of happiness* under an Aristotelean analysis.

complicated by the fact that the meaning of double negations (*not unhappy*) invariably depends upon the meaning of the component, single negations (*not*, *un-*), which are also difficult to pin down. Some argue that single negations have the same meaning (e.g., *not happy* = *unhappy*; Jespersen, 1917; Blutner, 2004), while others disagree (e.g., Krifka, 2007), citing examples like:

> 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.

Theorizing about the meaning of negation goes back to Aristotle, who noted that there are multiple ways of conveying an opposite meaning. *Contradictory* opposites must have opposite truth values, e.g., "No humans are rational" and "Some humans are rational". *Contrary* opposites cannot both be true, but can both be false, e.g., "All humans are rational" and "No humans are rational" (Figure 1A). An *Aristotelean analysis* consequently identifies one negation marker (*not*) as a contradictory opposite and another (*un-*) as contrary (Figure 1B); a double negative (*not unhappy*) then obtains a meaning distinct from that of the positive adjective (*happy*) by stipulating a difference between the two single negations (*unhappy* < *not happy*; Figure 1C; Horn, 1989, 1991; Krifka, 2007). An *Orwellian analysis* (*a*

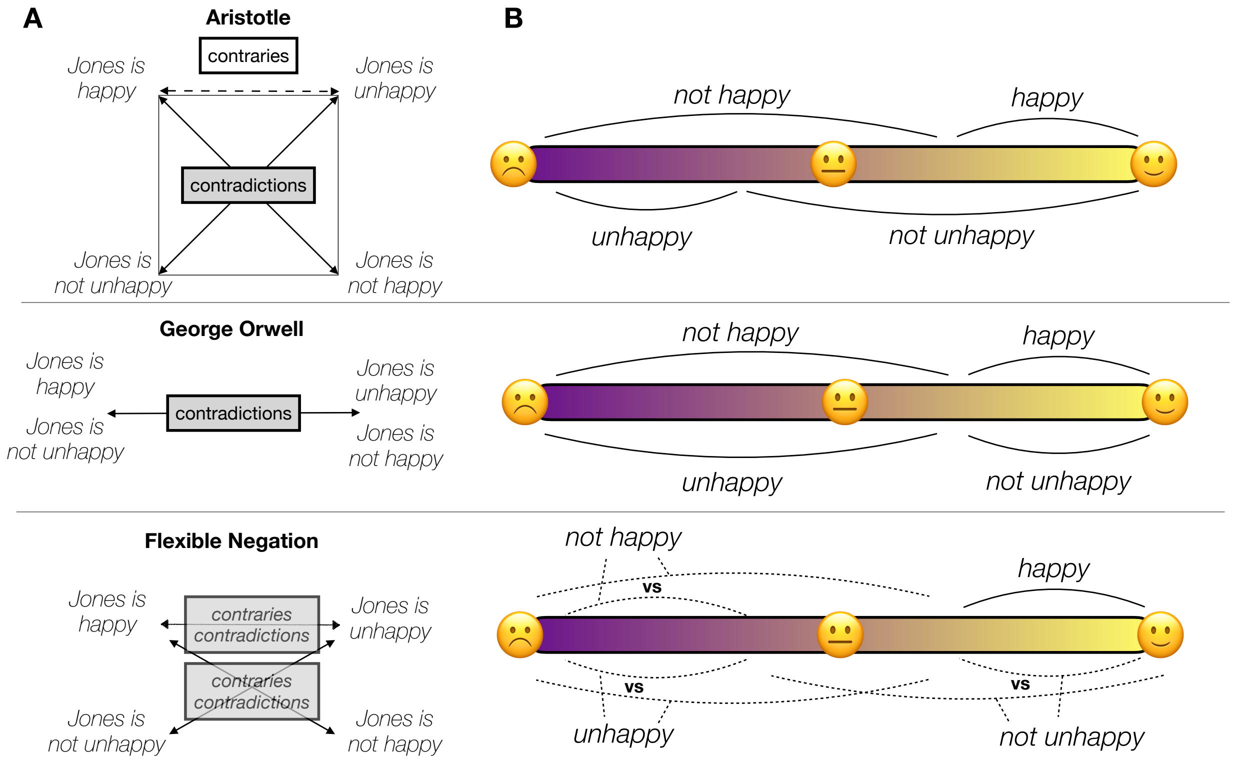


Figure 2. Alternative semantic hypotheses. A: Logical relations among negation markers under three different models of meaning. B: Literal meanings of antonym quartets on a happiness scale under three different models.

la the quote in the Preamble), in contrast, only acknowledges contradictory opposition (*unhappy* = *not happy*) and two negatives would be entirely redundant (*not unhappy* = *happy*; Figure 2; Orwell, 1946).[1] A third option, the Flexible Negation hypothesis developed here, is that the logical distinction between contradictory vs. contrary opposition creates an ambiguity in the meanings of negation markers such that either "not" or "un-" could be used to form either a contrary or contradiction (e.g., Russell, 1905; Martin, 1982; Wessel, 1993).

Ambiguity is ubiquitous in natural language, but seldom harms communication. Interlocutors rely on context to reason about each others' beliefs and goals to retrieve the intended meaning of underspecified words (Grice, 1975; Clark, 1996; Levinson, 2000).

[1] Models are named mnemonically. Aristotle and Orwell did not defend such theories themselves.

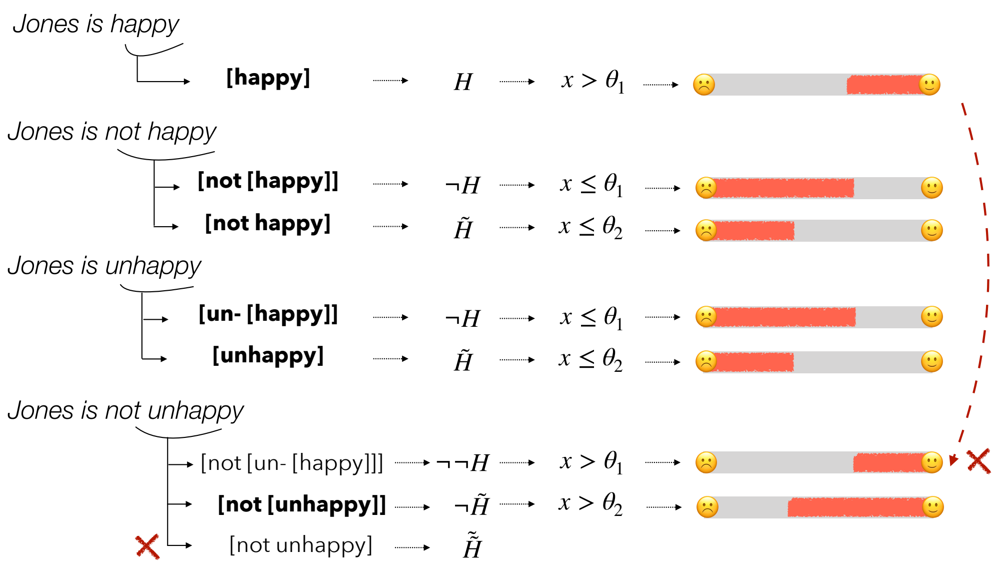


Figure 3. The Flexible Negation model reasons over a hypothesis space of meanings for antonym pairs and their negations. Both *not happy* and *unhappy* could signal either contradictory $\neg H$ or contrary negation $\tilde{H}$. *Not unhappy* can signal a double contradiction $\neg\neg H$ or a contradiction of a contrary $\neg\tilde{H}$. Contrary negation ˜ cannot take wide scope over other negation operators (see SI). A double contradiction $\neg\neg H$ is pragmatically unlikely, because the same meaning is expressed by just the simple positive H. Red bars denote the range of happiness values that are literally compatible with the adjective.

Theorizing about the meanings of negation must consider how the communicative context supports and potentially alters meanings, but heretofore no formal models have been developed that can adequately address the puzzles of double negatives while also accounting for pragmatic reasoning. We therefore investigate the question of the meaning of natural language negation by formalizing the semantic proposals—Aristotelean, Orwellian, Flexible Negation—in models of pragmatic communication that assume listeners interpret utterances as rational social actions. Our models combine previous, independently established modeling proposals on pragmatic reasoning (Franke & Jäger, 2016; Goodman & Frank, 2016; Scontras, Tessler, & Franke, 2018), the interpretation of gradable adjectives (e.g., *tall*, *happy*; Kennedy, 2007; Lassiter & Goodman, 2017), and ambiguity in the meanings of words (Bergen, Levy, &

Goodman, 2016). Our models are of listeners who hear utterances (e.g., *not unhappy*) and compute a posterior distribution over degrees of happiness, representing predictions about the interpretation of an utterance (Figure 4; see SI for mathematical details of each model), which we test in the experiments that follow.

The Flexible Negation model predicts a pattern of data that is unique from that of the Aristotelean and Orwellian models (Figure 4). When utterances involving a single negation marker are heard in isolation (e.g., *not happy* or *unhappy*), they are interpreted identically because of the ambiguity in whether each maps onto a contrary or contradiction (Figure 3). If the listener encounters a double negation (e.g., *not unhappy*), however, pragmatic reasoning helps disambiguate that the speaker likely intended a contradiction of a contrary (*a la* the Aristotelean account) because it would otherwise be a very costly manner of expressing the same meaning as the positive adjective (*happy*), which results from a double contradiction. The disambiguating power of hearing multiple negations (*[not] [un-]happy*) can also be observed for single negations (*not happy*, *unhappy*) if the listener hears multiple distinct utterances in the same context (e.g., Krifka, 2007's example, or simply: "She's not happy. He's unhappy."; Figure 4, *multiple utterances*). These indirectly contrastive inferences result from the fact that the listener has more evidence that the speaker associates different meanings with the different negation markers. The Flexible Negation model interprets the morphological antonym (*unhappy*) as more strongly negative than the utterance involving a negation particle (*not happy*) as a result of a cost difference between morphological and particle negation (e.g., with a word-based cost function: morphological markers do not add a new word to the utterance whereas negation particles do; see SI for further discussion of modeling assumptions).

We test these predictions in Experiments 1 & 2. Moreover, in Experiment 3, we investigate the dependence on linguistic form by having participants interpret expressions that flagrantly use the same negation marker twice (e.g., *not not happy*). In order to make sense of the utterance in a principled way, listeners would have to ascribe different meanings to the two different instances of *not*, suggesting an even richer picture of flexibility in negation.

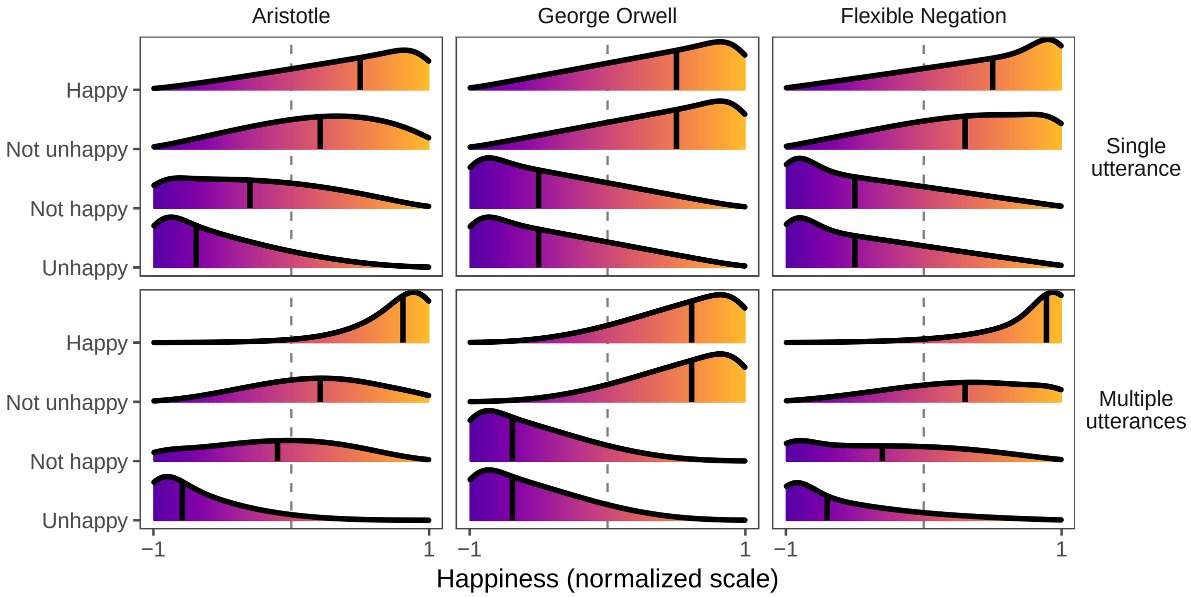


Figure 4. Model predictions for interpretations of antonym pairs and their negations under the three hypotheses. Black line shows the median of the distributions, in order to facilitate qualitative comparisons. *Aristotle* draws a distinction between all adjective types both when the adjective is heard in isolation (single utterance) and when adjectives are heard in the same context (multiple utterances). *George Orwell* never draws a meaning difference between the adjectives, even when they are heard in the same context. The *Flexible Negation* model generates a unique pattern of predictions: When adjectives are heard in isolation, the model draws no difference in meaning between *not happy* and *unhappy* but does distinguish *not unhappy* from *happy*; when the adjectives are heard in the same context, the model distinguishes among all the adjectives. Dashed line denotes the mid-point of the scale. Model predictions use minimally assumptive model parameters described in the SI.

Experiment 1: Negation in Isolation

Methods

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. All experiments reported here required participants with U.S. IP addresses, at least 95% work approval rating and English as a self-reported native language. The experiment took on average 3 minutes and participants were compensated $0.40.

Materials. We used adjectives that described properties of people. All of our adjectives were context-dependent, relative adjectives consistent with the definitions of Kennedy, 2007 and Kennedy & McNally, 2005. We consider *adjective sets* consisting of four related *adjective types* (see Table 2): positives (e.g., *happy*, *tall*), antonyms (e.g., *short*, *unhappy*), and their respective negations (*not* X).

In addition to analyzing morphological antonyms, we test a control set of items consisting of *lexical antonyms*, whose opposites are associated with distinct words (or, unique lexical items; e.g., *tall* and *short*). According to most theoretical proposals, these lexical antonyms should behave like bonafide contraries *a la* the Aristotelean account, and we

Adjective type	Definition	Examples
Positive	Positive-form scalar adjective	happy, mature
Negated positive	"not" + positive	not happy, not mature
Morphological antonym	Antonym created by morphology	unhappy, immature
Lexical antonym	Antonym with a unique lexical item	sad, childish
Negated morphological antonym	"not" + morphological antonym	not unhappy, not immature
Negated lexical antonym	"not" + lexical antonym	not sad, not childish
Negated negated positive (Expt. 3)	"not" + "not" + positive	not not happy, not not mature

Table 1

Informal definitions and examples of adjective types investigated.

include these items to verify that our experimental methods are able to adequately detect the behavioral signature associated with uncontroversial contraries. In total, twenty adjective sets were constructed, ten for lexical antonyms (*short*) and ten for morphological antonyms (*unhappy*).

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

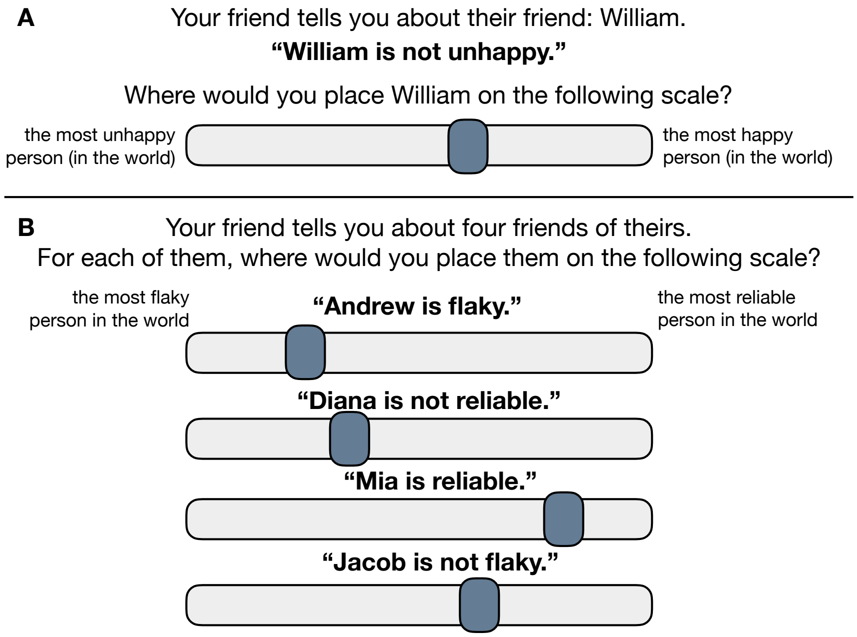
Table 2

Items in Experiment 1.

Procedure. On each trial, participants read a statement introducing a person using a gradable adjective of one of four *adjective types* from one of the two sets of *antonym types* (lexical vs. morphological) described in Materials. Participants rated the character on a scale from “the most *positive* person” to “the most *antonym* person” (item-dependent) using a slider bar (Fig. 5A). 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. No participant saw two instances from the same adjective set.

Results

Six participants were excluded for self-reporting a native language other than English, leaving a remainder of 114 participants for these analyses, which resulted in an average of 23


A
the most unhappy person (in the world)
the most happy person (in the world)
B
the most flaky person in the world
the most reliable person in the world

Your friend tells you about their friend: William.

"William is not unhappy."

Where would you place William on the following scale?

Your friend tells you about four friends of theirs.

For each of them, where would you place them on the following scale?

"Andrew is flaky."

"Diana is not reliable."

"Mia is reliable."

"Jacob is not flaky."

Figure 5. Example experimental trials for (A) single utterance (Expts. 1, 2) and (B) multiple utterances (Expts. 2, 3) conditions. "in the world" wording for endpoints was used in Expts. 2 & 3. (A) shows a trial from a morphological antonym set while (B) shows a lexical antonym set.

ratings for each unique adjective in our stimulus set. The qualitative predictions of our models concern the ordering within a set of alternatives for different antonym types (morphological vs. lexical). We expect the lexical antonyms to behave according the Aristotelean model, and thus show a total ordering: *short* < *not tall* < *not short* < *tall*. According to the Flexible Negation model, morphological antonyms should show a partial ordering: *unhappy* ≈ *not happy* < *not unhappy* < *happy*. We use the lexical antonyms as a control representing how antonyms should behave under the Aristotelean account. Thus, the key prediction is an interaction to see whether the antonym vs. negated positive contrast (*unhappy* vs. *not happy*) is different for morphological antonyms than it is for lexical antonyms, with morphological antonyms predicted to show no difference while lexical antonyms are predicted to show a difference.

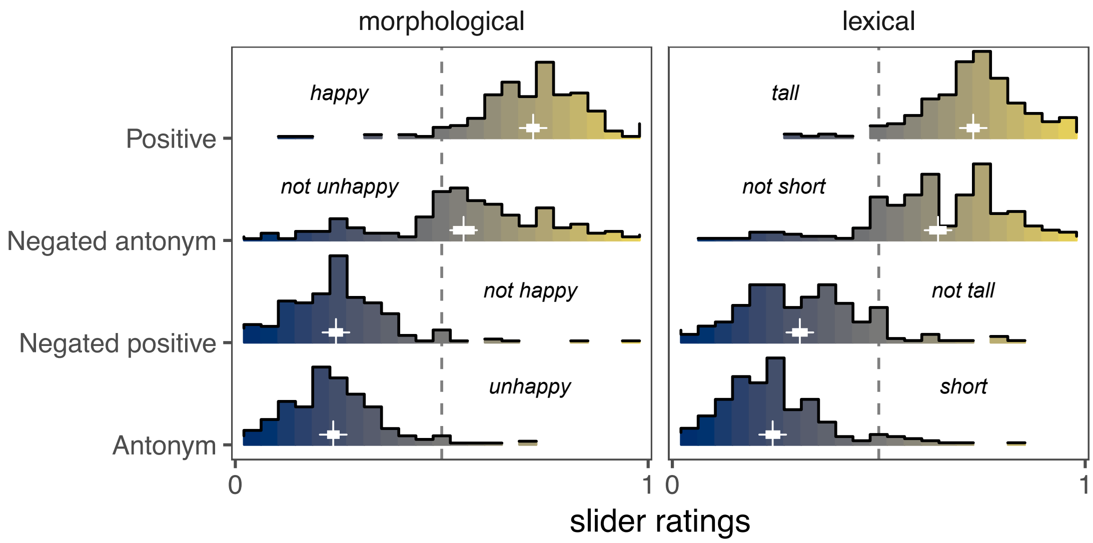


Figure 6. Experiment 1 results. Empirical histograms of responses for adjective sets with morphological antonyms (e.g., *happy/unhappy*) and lexical antonyms (e.g., *tall/short*). Dashed line indicates the midpoint of the scale. White bars denote bootstrapped 95% confidence intervals for the means.

Figure 6 shows the empirical distributions for each of the four adjective types for both morphological and lexical antonyms adjective sets. Critically, as predicted by the Flexible Negation model, adjective sets with morphological antonyms show only a partial ordering, with negated positives (e.g., *not happy*) and morphological antonyms (e.g., *unhappy*) receiving the same ratings; at the same time, the adjective sets with lexical antonyms show a full ordering (i.e., all adjective types receive distinct interpretations). To confirm these observations, we built a Bayesian generalized linear mixed model predicting the raw ratings in terms of fixed effects of *antonym type* (morphological vs. lexical), *adjective type*, and their interaction; the model included a maximal random-effects structure with random intercepts, slopes of *adjective type*, *antonym type*, and their interaction, by-participant and by-item.[2] Consistent with our predictions, lexical antonyms (short) were interpreted more negatively

[2] All regression models use a "zero-, one- inflated Beta" linking function which models data on the [0, 1] interval, using the `brms` package in R (Bürkner, 2017).

than their associated negated positive (not tall) in comparison to the morphological antonyms (happy) and their negated positives (not happy) the difference between the *antonym* vs. *negated positive* (i.e., an adjective type X morphological vs. lexical antonym type interaction; posterior mean and 95% Bayesian credible interval: $\beta = -0.283$ $[-0.556, -0.011]$; Cohen's d effect size: $d = -0.51$ $[-0.98, -0.05]$).[3] The negated positive vs. antonyms contrast (e.g., *unhappy* vs. *not happy*) was not appreciably different for the morphological antonyms ($\beta = -0.060$ $[-0.222, 0.100]$; $d = -0.06$ $[-0.32, 0.20]$), but with lexical antonyms (e.g., *short* vs. *not tall*), this difference was non-zero ($\beta = -0.343$ $[-0.568, -0.119]$; $d = -0.57$ $[-0.97, -0.17]$).

We observe further that negated morphological antonyms (e.g., *not unhappy*) were rated differently and lower than positive adjectives ($\beta = -0.343$ $[-0.568, -0.119]$; $d = -0.57$ $[-0.97, -0.17]$). We additionally observe that the distance between positive adjectives and negated morphological antonyms (e.g., *not unhappy*) was higher than positive adjectives and negated lexical antonyms (e.g., *not tall*; Figure 6), which manifested as an interaction in the negated antonym vs. positive contrast with the lexical vs. morphological items ($\beta = 0.332$ $[0.109, 0.548]$; $d = 0.70$ $[0.30, 1.10]$). Investigating the distribution of responses more closely, we see that negated antonyms received a distinct bimodal distribution wherein most ratings were slightly positive but a 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 (see SI for item-wise plots). This interpretation may be the result of participants attributing politeness to the speaker; when a speaker cares about a listener's self-image, they tend to endorse utterances with negation: *Not dishonest* may be an indirect way of saying that a person is not honest (Yoon, Tessler, Goodman, & Frank, 2017).

[3] We compute effect sizes by following the Bayesian technique described in (Kruschke, 2014) Ch. 16. See SI for details.

While we note the differences between lexical and morphological antonyms in this experiment, the direct comparison of these two kinds of adjectives is difficult. We have coded the antonyms in the morphological sets as positive and negative (or, antonym) by appealing to the morphology (e.g., *happy* is the positive adjective, while *unhappy* is the antonym); a similar assignment of the lexical antonyms to positive and negative is not so straightforward (Horn, 1989). Some pairs have a clear unmarked form: *tall* is the positive adjective because when describing the height of a person, we say *six feet tall* and not *six feet short.* For items that did not have a clear unmarked form (e.g., *fat* and *skinny*), we assigned the adjective that conveyed a greater amount to be the positive (i.e., fat conveys more weight than skinny); thus, the positive adjective is not necessarily the socially more desirable feature.[4] Because of these differences between lexical and morphological antonym sets, we treated this experiment as exploratory and curated a more tightly controlled set of materials for Experiments 2 & 3.

Experiment 2: Negation with Implicit Contrasts

We aim to replicate the previous findings using adjectives that describe the same semantic scales (e.g., *happy* vs. *unhappy* vs. *sad*). Also, we test our second prediction that hearing multiple utterances in the same context will produce the full ordering for morphological antonym sets (Figure 4).

[4] In addition, in the empirical data, we see a bimodal distribution of responses for the negated lexical antonyms and negated positives; thus, one may be concerned that this bimodal distribution is the result of an improper assignment of lexical antonyms to either the positive or antonym form (e.g., *skinny* is actually the positive adjective and *fat* is the antonym). If this were so, we would expect this bi-modality to occur across the items but not within items; however, when we look at the item-specific distributions of responses, we do not see clear evidence for this, but rather we see the bi-modality occurring within several items from the lexical sets (see SI).

Methods

Participants. We recruited 750 participants from MTurk. The experiment comprised four between-subjects experimental conditions arranged in a 2x2 design: *antonym type* (morphological vs. lexical) X *context* (single vs. multiple utterances). Three-hundred 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* → *unhappy*, *sad*; Table 4). Lexical antonyms were selected from a set of possibilities produced from a small survey (n=18) on MTurk eliciting "opposites" for a list of thirty positive-form adjectives which had morphological antonyms. In this antonym elicitation, participants saw the same material as in the main experiment (e.g., "Your friend tells you about their friend: William. *William is forgiving.*") and asked "What is the opposite of *[adjective]*?" ("What is the opposite of forgiving?"). From the list of freely-produced opposites, 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* → *resentful*; opposite of *kind* → *cruel*; opposite of *friendly* → *mean*). Ten out of the original thirty 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* → *cold*, but also *affectionate* → *cold*).

Procedure. In the *multiple utterances* conditions, participants rated all four adjective types simultaneously, each referring to a different person (Figure 5B), for a total of 12 trials. The *single utterances* conditions were similar to that of Experiment 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 Experiment 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 as “the most {*positive*, *negative*} person (of these four)” in the multiple utterances conditions. We note that by labelling the endpoints in this more explicit manner, we potentially decrease the size of the effects we observe

Results

Thirty-five participants were excluded for self-reporting a native language other than English, leaving 715 participants for these analyses. Results for each adjective type in each condition are shown in Fig. 7.

The Flexible Negation model predicts that morphological antonyms (e.g., *unhappy*) when heard in isolation should not be distinguished from negated positives (e.g., *not happy*; Figure 4). This lack of interpretative difference stands in contrast to (1) lexical antonyms, which we predict should behave according to the Aristotelean model and show an interpretative difference between antonyms and negated positives (adjective type by antonym type interaction) and (2) hearing the adjectives in the same context, which is predicted by all models to exaggerate differences between adjectives, but specific to morphological antonyms, result in an interpretative difference between morphological antonyms and negated positives where there was none before (adjective type by context interaction). To test this, we built a Bayesian generalized linear mixed-effects model predicting the raw ratings in terms of fixed

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 3

Items used in Experiment 2.

effects of *adjective type*, *antonym type* (morphological vs. lexical), and *presentational context* (single vs. multiple utterances), and their pairwise two-way and three-way interactions; the model included a maximal random-effects structure with random intercepts, slopes of *adjective type* by-participant, and random intercepts, slopes of *adjective type*, *antonym type*, *presentational context*, and their pairwise two-way and three-way interactions by-item.

We predict that when heard in isolation, morphological antonyms will not show an interpretative difference between *antonyms* and *negated positives* (consistent with the Flexible Negation model) while lexical antonyms will (consistent with the Aristotelean model). Consistent with this hypothesis, in the single utterance conditions, the *antonyms*

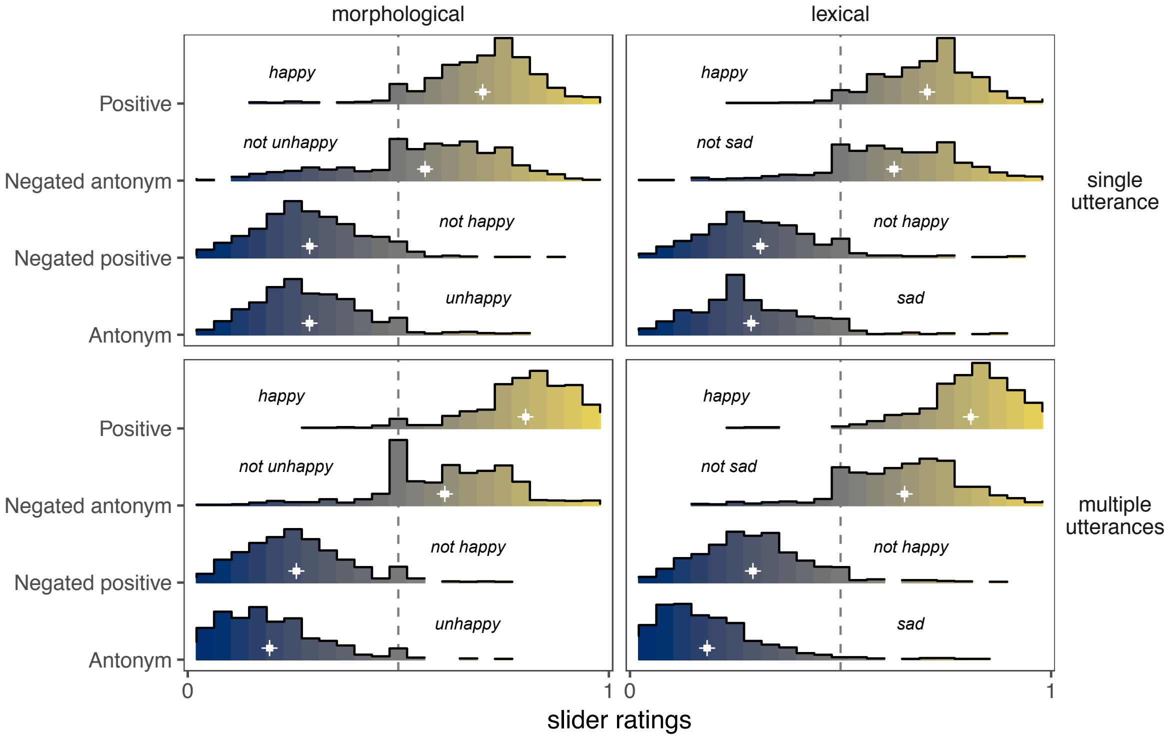


Figure 7. Experiment 2 results. Empirical histograms of responses for adjective sets with morphological antonyms (e.g., *happy/unhappy*) and lexical antonyms (e.g., *happy/sad*) for the single utterance and multiple utterances conditions. Dashed line indicates the midpoint of the scale. Width of white rectangles denotes bootstrapped 95% confidence intervals for the means.

were interpreted more negatively than *negated positive* for lexical antonyms than for morphological antonyms (i.e., an adjective type by antonym type interaction; $\beta = -0.109$ $[-0.199, -0.020]$; $d = -0.22$ $[-0.39, -0.05]$). At the same time, the *antonym* vs. *negated positive* difference for morphological antonyms (e.g., *unhappy* vs. *not happy*) was not different from zero ($\beta = -0.008$ $[-0.073, 0.057]$; $d = -0.00$ $[-0.13, 0.12]$), while the lexical antonyms (*sad*) were interpreted more negatively than the negated positives (*not happy*; $\beta = -0.117$ $[-0.187, -0.046]$; $d = -0.22$ $[-0.36, -0.09]$).

Our second prediction is that morphological antonyms will be interpreted differently than negated positive adjectives when they are heard in the same context. Specifically, we predict that morphological antonyms will be interpreted more negatively than negated positives in a context with multiple adjectival utterances. Consistent with this prediction, morphological antonyms were interpreted much more negatively than negated positives in the multiple utterances than in the single utterance condition (i.e., an adjective type by context interaction; $\beta = -0.366$ [-0.479, -0.254]; $d = -0.65$ [-0.86, -0.44]). Morphological antonyms were not interpreted differently than negated positives when they were heard in isolation (single utterance condition; reported above), but they were interpreted more negatively when they were heard in the same context (multiple utterance condition; $\beta = -0.374$ [-0.472, -0.274]; $d = -0.65$ [-0.84, -0.47]).

The observation that negated positives are interpreted similarly to morphological antonyms (in the single utterance condition) could result from the ambiguity in the meaning of negation markers as posited by the Flexible Negation account. It could also potentially result from an Aristotelean account with some kind of pragmatic strengthening of the meaning of a contradiction ("not happy") into a polar contrary ("unhappy"), a phenomenon referred to as *negative strengthening* or *inference towards the antonym* (Horn, 1989; Ruytenbeek, Verheyen, & Spector, 2017; Gotzner, Solt, & Benz, 2018). If that were occurring, one might expect that the presence of the morphological antonym in the same context (i.e., the *multiple utterances* condition) would weaken the interpretation of the negated positive (e.g., "not happy" *but not unhappy*). We observe a different pattern, however: When presented with multiple utterances, it is the morphological antonym that tends to be strengthened into a more negative interpretation ($\beta = -0.536$ [-0.681, -0.386]; $d = -0.98$ [-1.28, -0.68]); the interpretation of the negated positive also becomes more negative, though with a substantially smaller magnitude of difference ($\beta = -0.170$ [-0.303, -0.034]; $d = -0.33$ [-0.60, -0.06]; the relevant interaction was reported above). This pattern suggests that in our paradigm, in the absence of further context, morphological antonyms like *unhappy* are interpreted more like contradictions than negated positives like *not happy* are pragmatically strengthened to contraries.

The predictions of the Flexible Negation model (aimed at explaining morphological antonyms) and the Aristotelean model (aimed at explaining lexical antonyms) are ambiguous about the relevant three-way interaction (*antonym* vs. *negated positive* by lexical vs. morphological adjective type by context). On the one hand, for the *antonym* vs. *negated positive* contrast, we predict an interpretative difference for morphological antonyms only when the alternatives are presented together, whereas the difference is expected to occur for lexical antonyms in both context conditions. On the other hand, pragmatics operates in both models (Flexible Negation and Aristotle) to further differentiate the likely meaning of all of the adjectives as a result of being presented in the same context (i.e., all adjectives get more specific interpretations). Thus, it is not clear *a priori* what the prediction should be for the existence of a three-way interaction nor the direction of the interaction. As an exploratory analysis, we examined the three-way interaction in our the regression model and found the relevant three-way interaction was in the direction of lexical antonyms showing a larger *antonym* vs. *negated positive* difference in the multiple utterance condition ($\beta = -0.164$ $[-0.323, -0.005]$; $d = -0.25$ $[-0.55, 0.05]$).

We note the sizes of some of these effects are relatively small. In particular, the contrasts within the single utterance condition for morphological vs. lexical antonyms dance around a Cohen's *d* of 0.2, traditionally considered a "small effect". We suspect the smallness of these effect sizes are due to the subtlety of the phenomena under consideration and the fact that these judgments are either made by participants on different trials (thus, reflecting a small tendency to report in different regions of the negative portion of a continuous slider bar) or by different participants (lexical vs. morphological comparison). The effects are much larger in the multiple utterance conditions, where participants can deliberately select different positions of the slider bar with precision.

Experiment 3: Flagrant Double Negatives

Is the inferential cognitive mechanism that produces meanings for negated antonyms specific to the usage of distinct negation markers (e.g., *not* + *un-*) or is it a more general mechanism that is triggered when two negatives are encountered? We investigate this question using flagrant double negatives like *not not happy* in the *multiple utterances* context from Experiment 2, with two different sets of alternative utterances.

Methods Experiment 3a

Experiment 3a investigated the interpretation of a flagrantly double negative (e.g., *not not happy*) when it appears in the same context as the other utterances from Experiment 2b: the positive adjective (e.g., *happy*) and two negatives (e.g., *not happy*, *unhappy*).

Participants. We recruited 75 participants from MTurk to match the sample size of the same condition in Experiment 2. These numbers follow from the intention of getting approximately 45 ratings for each unique adjective in the experiment. The experiment took on average 5 minutes and participants were compensated $0.90. Exclusion criterion, sample size, procedure, and the analysis described below were preregistered: `osf.io/vjhak`.

Materials and procedure. The materials and procedure were identical to that of the *multiple utterances* condition of Experiment 2. The main difference in this experiment is that participants are presented with the following alternatives: positives, negated positives, morphological antonyms, and negated negated positives (e.g., *happy*, *not happy*, *unhappy*, *not not happy*).

Results Experiment 3a

All participants self-reported only English as their native language. Five participants were excluded for failing to respond correctly to at least 7 of the 10 memory check items, leaving 40 participants for these analyses.[5] Figure 8 shows results for each adjective type.

Our main hypothesis concerns the interpretation of flagrant double negative statements that use the same negation marker twice (*not not happy*). We built a Bayesian mixed-effects regression model with random by-participant and by-item intercepts and slopes (where the slopes refer to the effect of the adjective type). The two negatives did not simply cancel to make a positive: *not not happy* received a substantially more negative interpretation than *happy* ($\beta = -1.432$ $[-1.730, -1.138]$; $d = -3.48$ $[-4.13, -2.82]$). This suggests that participants actively try to make sense of seemingly redundant linguistic material in a way that would be informative for a speaker to produce. In addition, we replicate the result from Experiment 2 (multiple utterances condition) for the morphological antonyms vs. negated positives difference: *not happy* and *unhappy* were differentiated in meaning ($\beta = -0.262$ $[-0.409, -0.116]$; $d = -0.39$ $[-0.64, -0.14]$).

Contrary to our hypothesis, however, we did not find evidence that the flagrant double negative *not not happy* received on-average a positive interpretation ($\beta = 0.087$ $[-0.204, 0.365]$); inspection of the fitted random-effects suggested this was due to large participant-wise variation in the interpretation of these flagrant double negatives. Further, the interpretation of the flagrant double negative appears to be different from that of the negated antonym of Experiment 2 (multiple utterances condition), which did receive on average a positive interpretation ($\beta = 0.466$ $[0.348, 0.585]$). In the next experiment, we directly compare the interpretations of these two ways of expressing double negatives.

[5] This experiment was conducted eighteen months after Experiment 2 and, due to concerns about declining data quality on MTurk, included an additional memory check wherein participants had to select from a list of 10 items (5 real and 5 distractor) all of the items they could recall seeing in the experiment. We pre-registered the exclusion criterion of removing participants who failed to respond correctly to at least 7 of the 10 items.

Methods Experiment 3b

Experiment 3b directly compared the interpretation of two double negatives when observed in the same context (e.g., *not not happy* vs. *not unhappy*).

Participants. We recruited 50 participants from MTurk. This number was arrived at with the intention of getting approximately 20 ratings for each unique item in the experiment. The experiment took on average 5 minutes and participants were compensated $1.00. Exclusion criterion, sample size, procedure, and the analysis described below were preregistered: `https://osf.io/5zqd7`.

Materials and procedure. The materials and procedure were almost identical to that of the *multiple utterances* condition of Experiment 2 and Experiment 3a. In this experiment, participants were presented with five alternatives: positive adjectives, negated positives, morphological antonyms, negated morphological antonyms, and negated negated positives (e.g., *happy*, *not happy*, *unhappy*, *not unhappy*, *not not happy*). Since we observed substantially greater participant-wise variability than item-wise variability in Experiment 3a, we decided to shorten the experiment so participants completed 8 trials instead of the 12 used in Experiment 3a.

Results Experiment 3b

1 participant self-reported a language other than English as their native language. An additional 12 participants were excluded for failing to respond correctly to at least 7 of the 10 memory check items (see Experiment 3a), leaving 37 participants for these analyses. Figure 8b shows results for each adjective type.

In this experiment, we are concerned with whether or not participants will draw an interpretative difference between negated morphological antonyms (e.g., *not unhappy*) and doubly negated positives (e.g., *not not happy*). We built a Bayesian mixed-effects regression model with random by-participant and by-item intercepts and slopes (where the slopes refer to the effect of the adjective type). As suggested by the Experiment 2 (multiple utterance

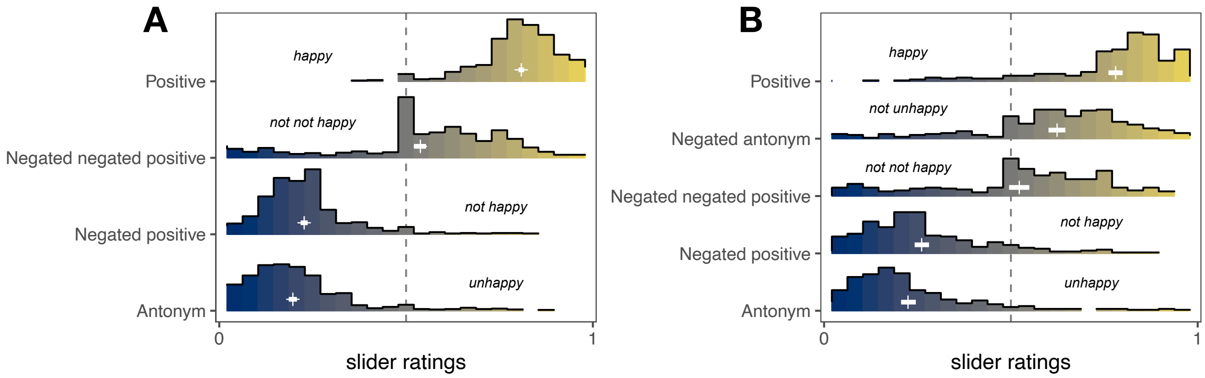


Figure 8. Experiment 3 results. A: Experiment 3a in which the doubly-negated adjective is presented in the context of the positive adjective, the morphological antonym, and the negated adjective. B: Experiment 3b in which the doubly-negated adjective is presented in the context of the same alternatives and the negated morphological antonym. Dashed line indicates scale midpoint. White bars denote bootstrapped 95% confidence intervals for the means.

condition) vs. Experiment 3a comparison that we reported above, in which the two double negatives are not presented in the same context, the doubly-negated positive adjective receives a lower overall rating than the negated morphological antonym ($\beta = -0.452$ $[-0.802, -0.102]$; $d = -0.84$ $[-1.44, -0.24]$). As in Experiment 3a, doubly-negated positive adjectives receive on average a rating not distinguished from the midpoint ($\beta = 0.065$ $[-0.243, 0.375]$; $d = 0.34$ $[-0.20, 0.90]$). Again, we replicate the results from Experiment 2 (multiple utterances condition) and Experiment 3a in which *not happy* and *unhappy* are subtly differentiated in meaning ($\beta = -0.180$ $[-0.316, -0.045]$; $d = -0.30$ $[-0.50, -0.10]$).

Examining the distribution of responses for *not not happy* vs. *not unhappy*, we observe that the negated negated adjective (*not not happy*) receives about twice as many negative responses (i.e., responses lower than the mid-point of the scale, assuming a 5-pt buffer on the 101-pt scale; in other words, responses < 0.45 on the scale; 27% vs. 14%) and about twice as many mid-point responses (assuming a 5-pt buffer; $0.45 <$ response < 0.55; 20% vs 11%)

than the negated morphological antonym (*not unhappy*). This variability in responses suggests that listeners will try to rationalize seemingly redundant linguistic material using a variety of strategies (see General Discussion).

General Discussion

Understanding language is a holistic process in which many factors—the meanings of words, uncertainty, context—come together to produce interpretations. Pragmatics in particular is a complicated, elusive, and often-ignored factor that is necessary to consider to understand how the meanings of words are computed in context. Across three experiments, we find that the interpretations of natural language negation can change based on subtle contextual circumstances like the presence of other uses of negation. These findings highlight just how deep the flexibility and context-sensitivity of language run and underscore the importance of discarding simplistic pictures of language understanding that assume transparent and static meanings of words—even logical words.

In Experiments 1 & 2, we discovered and confirmed a surprising empirical result predicted by the Flexible Negation hypothesis: *unhappy* (morphological antonyms) and *not happy* (negated positives) are interpreted identically, except when heard in the same context. The equivalent interpretations of morphological antonyms and negated positives could potentially arise via a different mechanism than the Flexible Negation model. If a speaker says they are “not good” when they could also have said they are “fine”, an implication may be that they are rather bad (so-called “negative strengthening”; Horn, 1989; Ruytenbeek et al., 2017; Gotzner et al., 2018). This alternative pragmatic hypothesis has never been articulated formally, however, and extending the Aristotelean model with a midpoint-denoting neutral utterance (e.g., “fine”) introduces new problems: The interval semantics for the neutral utterance leads to an interpretation for “not happy” that includes highly positive statements (i.e., not extremely happy but not just fine; see SI). Articulating the other pragmatic components necessary to derive the inferences described by Horn (1989) is an important area for future research.

We presented a computational solution to an age-old problem in natural language understanding: How to interpret double negatives (e.g., *not unhappy*; Horn, 1991; Krifka, 2007; Rett, 2014). We predicted and observed empirically the ordering hypothesized by Krifka (2007) for morphological antonyms (*unhappy* < *not happy* < *not unhappy* < *happy*), when a listener hears multiple adjectival utterances in the same context. Other accounts derive similar predictions based on different kinds of assumptions (Krifka, 2007; Rett, 2014; Cable, 2018), but only our model makes predictions about the context-dependence of these inferences. In Experiment 3, we saw that this inference goes beyond using distinct negation markers: Expressions that use the same negation marker twice (e.g., *not not happy*) are interpreted in a manner distinct from that of positive adjectives (e.g., *happy*) and that of negated antonyms (e.g., *not unhappy*). This result points to an underlying cognitive mechanism that flexibly allows for different meanings for the same negation marker within a single utterance.

Our work builds upon previous studies on negated adjectives (e.g., Giora, Balaban, Fein, & Alkabets, 2005) with our straight-forward response measurements, comparison of morphological and lexical antonyms, and situation within different presentational contexts. Considerable empirical work on negation has focused on *negative strengthening* (Ruytenbeek et al., 2017; Gotzner et al., 2018) to which we provide a new way of measuring the equivalence between negated positives and antonyms. Our results suggest, however, that in the absence of further context, morphological antonyms (*unhappy*) are interpreted more like contradictions than negated positives (*not happy*) as contraries. Our finding that negated antonyms are interpreted more negatively than positive adjectives is also consistent with negated antonyms (e.g., *not impossible*) being processed more easily then positive adjectives (e.g., *possible*) when the listener expects the degree to be low (something very likely to difficulty; Schiller et al., 2017).

Our model explains the interpretations of antonym pairs and their negations by assuming the speaker's utterances are chosen to convey information about the underlying degree (e.g., *how happy is John?*). This assumption about the speaker's communicative goal defines an implicit *Question Under Discussion* (or, *QUD*), which can easily change and complicate the picture presented here (Roberts, 2012; Beaver, Roberts, Simons, & Tonhauser, 2017). For example, utterances involving particle negation (e.g., *not happy*) are naturally produced to address a polar question (e.g., *Is John happy?*), which could result in different interpretations than the degree-based question. In the SI, we articulate a mechanism by which reasoning about the QUD could change interpretations of utterances with negation. The simple observation that negation particles ("not") are more likely under polar QUDs (*is John happy?*), however, is insufficient to explain our data: It does lead to "not happy" being interpreted similarly to "unhappy", but at the cost of washing out the difference between "not unhappy" and "happy" (see SI). QUDs are likely to play a role in understanding negation and the modeling approach we present here provides a framework for elaborating more complex hypotheses about the relationship between the QUD and utterances involving negation.

Our models aim to explain the modal interpretations for antonym pairs and their negations, but the empirical data is more nuanced and flexible than even our Flexible Negation model can account for. The interpretations we observe empirically for negated antonyms indicate a slightly positive state on average, but with consistently negative interpretations as well (e.g., "not UNreliable *[but kind of flaky]*"), potentially the result of politeness (Yoon et al., 2017). This flexibility in interpretation might be elicited by our text-based experiment, which leaves open exactly how our stimuli should sound to the ear; prosodic focus can more strongly constrain interpretations, which will be important to clarify using speech-based experiments. We also see evidence for strongly negative interpretations (e.g., very unhappy) from the flagrant double negatives of Experiment 3 (*not not happy*): The additivity of negations, or *negative concord*, is not often associated with standard

English, though it is relatively common cross-linguistically (e.g., in Italian: *non capisco niente*, literal translation: *I don't understand nothing;* Zeijlstra, 2004) including in African American Vernacular English (e.g., Mohammad Ali: "Ain't never been another fighter like me"; Labov, 1972; Howe, 2005), which suggests that negative concord could be a logical possibility that listeners entertain in their hypothesis space of meanings.

Our findings and modeling suggest that even seemingly rigid linguistic elements like negation have some kind of ambiguity in their meaning. Such an ambiguity could be exploited by speakers in subtle ways (e.g., as a kind of "dog whistle", in which a speaker wishes to say two things at once). Formalizing the structure of this ambiguity and how listeners reason about it is an important step to understanding the complexities that reside deep in human language.

References

Bardwick, J. M. (1986). *The plateauing trap: How to avoid it in your career... and your life.* American Management Association.

Beaver, D. I., Roberts, C., Simons, M., & Tonhauser, J. (2017). Questions under discussion: Where information structure meets projective content. *Annual Review of Linguistics*, *3*, 265–284.

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*, 488–514.

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of Statistical Software*(1), 1-28. doi: doi.org/10.18637/jss.v080.i01

Cable, S. (2018). The good, the 'not good', and the 'not pretty': Negation in the negative predicates of tlingit. *Natural Language Semantics*, *26*(3-4), 281–335.

Clark, H. H. (1996). *Using language.* Cambridge university press.

Franke, M., & Jäger, G. (2016). Probabilistic pragmatics, or why Bayes' rule is probably important for pragmatics. In *Zeitschrift für sprachwissenschaft* (pp. 3–44).

Geurts, B. (2003). Reasoning with quantifiers. *Cognition*, *86*(3), 223–251.

Giora, R., Balaban, N., Fein, O., & Alkabets, I. (2005). Negation as positivity in disguise. *Figurative language comprehension: Social and cultural influences*, 233–258.

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

Gotzner, N., Solt, S., & Benz, A. (2018). Scalar diversity, negative strengthening, and adjectival semantics. *Frontiers in Psychology*, *9*(SEP), 1–13. doi: 10.3389/fpsyg.2018.01659

Grice, H. P. (1975). Logic and Conversation. In *Syntax and semantics 3: Speech acts* (pp. 41–58).

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.

Howe, D. (2005). Negation in African American Vernacular English. *Aspects of English negation*, *132*, 173.

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.

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 345–381.

Krifka, M. (2007). Negated antonyms: Creating and filling the gap. *Presupposition and Implicature in Compositional Semantics*, 163–177.

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan.* Academic Press.

Labov, W. (1972). Negative attraction and negative concord in English grammar. *Language*, 773–818.

Lakoff, G. (2008). *Women, fire, and dangerous things.* University of Chicago press.

Lassiter, D., & Goodman, N. D. (2015). How many kinds of reasoning? inference, probability, and natural language semantics. *Cognition*, *136*, 123–134.

Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a Bayesian model of interpretation. *Synthese.*

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature.* MIT press.

Martin, J. N. (1982). Negation, ambiguity, and the identity test. *Journal of Semantics*, *1*(3-4), 251–274.

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

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

Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*, 1-69.

Russell, B. (1905). On denoting. *Mind*, *14*(56), 479–493.

Ruytenbeek, N., Verheyen, S., & Spector, B. (2017). Asymmetric inference towards the antonym: Experiments into the polarity and morphology of negated adjectives. *Glossa: a journal of general linguistics*, *2*(1), 92. doi: 10.5334/gjgl.151

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

Schiller, N. O., van Lenteren, L., Witteman, J., Ouwehand, K., Band, G. P., & Verhagen, A. (2017). Solving the problem of double negation is not impossible: electrophysiological evidence for the cohesive function of sentential negation. *Language, Cognition and Neuroscience*, *32*(2), 147–157.

Scontras, G., Tessler, M. H., & Franke, M. (2018). *Probabilistic language understanding: An introduction to the rational speech act framework.* Retrieved 2020-1-9 from https://problang.org.

Tiersma, P. M. (1999). *Legal language.* University of Chicago Press.

Wessel, H. (1993). Zur lösung einiger paradoxien. In W. Stelzner (Ed.), *Philosophie und logik: Frege-kolloquien, jena, 1989/1991* (Vol. 3). Walter de Gruyter.

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.*

Zeijlstra, H. (2004). *Sentential negation and negative concord.* Netherlands Graduate School of Linguistics.