The lexical and grammatical sources of neg-raising inferences

Hannah Youngeun An
University of Rochester

Aaron Steven White
University of Rochester

Abstract

We investigate *neg(ation)-raising* inferences, wherein negation on a predicate can be interpreted as though in that predicate's subordinate clause. To do this, we collect a large-scale dataset of neg-raising judgments for effectively all English clause-embedding verbs and develop a model to jointly induce the semantic types of verbs and their subordinate clauses and the relationship of these types to neg-raising inferences. We find that some neg-raising inferences are attributable to properties of particular predicates, while others are attributable to subordinate clause structure.

1 Introduction

Inferences that are triggered (at least in part) by particular lexical items provide a rich test bed for distinguishing the relative semantic contribution of lexical items and functional structure. One class of such inferences that has garnered extended attention is *neg(ation)-raising*, wherein negation on a predicate can be interpreted as though in that predicate's subordinate clause (Fillmore, 1963; Bartsch, 1973; Horn, 1978; Gajewski, 2007). For example, a neg-raising inference is triggered by (1) while one is not triggered by (2).

(1) Jo doesn't think that Bo left.
$\rightsquigarrow$Jo thinks that Bo didn't leave.

(2) Jo doesn't know that Bo left.
$\not\rightsquigarrow$Jo knows that Bo didn't leave.

Though accounts vary with respect to whether neg-raising inferences are explained as a syntactic or a pragmatic phenomenon, all associate these inferences with particular predicates in some way or other—e.g. *think, believe, suppose, imagine, want*, and *expect* are often taken to be associated with neg-raising inferences as a matter of knowledge one has about those predicates, while *say, claim, regret*, and *realize* are not (Horn, 1971, 1978).

One challenge for such approaches is that whether a neg-raising inference is triggered varies with aspects of the context, such as the predicate's subject—e.g. (3a) triggers the inference that the speaker thinks Jo didn't leave—and tense—e.g. (3b) does not trigger the same inference as (3a).

(3) a. I don't know that Jo left.
b. I didn't know that Jo left.

While some kinds of variability can be captured by standing accounts, other kinds have yet to be discussed at all. For example, beyond a predicate's subject and tense, the syntactic structure of its clausal complement also appears to matter: (4a) and (5a) can both trigger neg-raising interpretations, while (4b) and (5b) cannot.

(4) a. Jo wasn't thought to be very intelligent.
b. Jo didn't think to get groceries.

(5) a. Jo wasn't known to be very intelligent.
b. Jo didn't know to get groceries.

Should these facts be chalked up to properties of the predicates in question? Or are they general to how these predicates compose with their complements? These questions are currently difficult to answer for two reasons: (i) there are no existing, lexicon-scale datasets that measure neg-raising across a variety of contexts—e.g. manipulating subject, tense and complement type; and (ii) even if there were, no models currently exist for answering these questions given such a dataset.

We fill this lacuna by (i) collecting a large-scale dataset of neg-raising judgments for effectively all English clause-embedding verbs with a variety of both finite and non-finite complement types; and (ii) extending White and Rawlins' (2016) model of s(emantic)-selection, which induces semantic type signatures from syntactic distribution, with a module that associates semantic types with the inferences they trigger. We use this model to jointly induce semantic types and their relationship to neg-raising inferences, showing that the best fitting model attributes some neg-raising inferences to properties of particular predicates and others to general properties of syntactic structures.[1]

We begin with background on theoretical approaches to neg-raising, contrasting the two main types of accounts: syntactic and pragmatic (§2). We then present our methodology for measuring neg-raising across a variety of predicates and syntactic contexts (§3) as well as our extension of White and Rawlins' s-selection model (§4). Finally, we discuss the results of fitting (§5) our model to our neg-raising dataset (§6).

2 Background

Two main types of approaches have been proposed to account for neg-raising interpretations: syntactic and pragmatic (see Zeijlstra 2018; Crowley 2019 for reviews). We do not attempt to adjudicate between the two here—rather aiming to establish the explanatory devices available to each for later interpretation relative to our modeling results.

Syntactic Approach In syntactic approaches, neg-raising interpretations arise from some syntactic relation between a matrix negation and an unpronounced embedded negation that is licensed by the neg-raising predicate. This is classically explained via a syntactic rule that "raises" the negation from the subordinate clause to the main clause, as in (6), though accounts using alternative syntactic relations exist (Fillmore 1963; Kiparsky 1970; Jackendoff 1971; Pollack 1976; Collins and Postal 2014, 2017, 2018; cf. Klima, 1964; Zeijlstra, 2018; see also Lasnik, 1972).

(6) Jo does not believe Bo did ___ leave.

Evidence for syntactic accounts comes from the distribution of negative polarity items, Horn-clauses, and island phenomena (Horn, 1971; Collins and Postal, 2014, 2017, 2018; cf. Zwarts, 1998; Gajewski, 2011; Chierchia, 2013; Horn, 2014; Romoli and Mandelkern, 2019).

Purely syntactic approaches to neg-raising have effectively one method for explaining variability in neg-raising inferences relative to subject, tense, and subordinate clause structure (as discussed in §1): if a certain lexical item—e.g. *know*—occurs in some sentence that licenses a neg-raising inference—e.g. (5a)—and another that doesn't—e.g. (5b)—one must say that the structure in the first differs from the second in such a way that the first allows the relevant syntactic relation while the second does not. This implies that, even in cases like (3a) v. (3b), where there is no apparent structural difference (beyond the subject), the structures differ on some neg-raising-relevant property. This can be implemented by saying that, e.g. the same verb can select for two different structural properties—one that licenses neg-raising and one that does not—or that the verb is somehow ambiguous and its variants differ with respect to some neg-raising-relevant, syntactic property.

Semantic/Pragmatic Approach In semantic/pragmatic approaches, neg-raising interpretations are logically derived from an *excluded middle* (EM or *opinionatedness*) inference (Bartsch, 1973; Gajewski, 2007); see Horn (1978), Horn and Bayer (1984), Romoli (2013), and Xiang (2013) for implicature-based approaches (see also Tovena, 2001; Homer, 2015). This approach posits that, anytime a neg-raising predicate v is used to relate person x with proposition p, the hearer assumes that either $x\ v\ p$ or $x\ v\ \neg p$. For example, in the case of *believe*, as in (7), the hearer would assume that Jo either believes that Bo left or that Bo didn't leave.

(7) Jo believes that Bo left.
   a. *truth conditions:* x BELIEVE p
   b. *inference:* x BELIEVE $p \vee x$ BELIEVE $\neg p$

The EM inference is impotent in the positive cases but drives further inferences in the negative, where the first EM disjunct is cancelled by the truth conditions: if Jo doesn't believe that Bo left and Jo believes that Bo left or that Bo didn't leave, then Jo must believe that Bo didn't leave.

(8) Jo doesn't believe that Bo left.
   a. *truth conditions:* $x \neg$ BELIEVE p
   b. *inference:* $\cancel{x \text{ BELIEVE } p \vee}$ x BELIEVE $\neg p$

To capture non-neg-raising predicates, one must then say that some predicates trigger the EM inference, while others don't (Horn, 1989). However, such lexical restrictions alone cannot exhaustively explain the variability in whether verbs trigger presuppositions with certain subjects, as noted for (2) and (3a). To explain this, Gajewski (2007) posits that neg-raising predicates are soft presupposition triggers. Effectively, the EM inferences are defeasible, and when they are cancelled, the neg-raising

[1] Data are available at megaattitude.io.

inference does not go through (Abusch, 2002). This is supported by cases of explicit cancellation of the EM inference—e.g. the neg-raising inference (9c) that would otherwise be triggered by (9b) does not go through in the context of (9a).

(9) a. Bill doesn't know who killed Caesar. He isn't even sure whether or not Brutus and Caesar lived at the same time. So...
  b. Bill doesn't believe Brutus killed Caesar.

  ---

  c. $\not\rightsquigarrow$ Bill believes Brutus didn't kill Caesar.

This sort of explanation relies heavily on semantic properties of particular verbs and naturally covers variability that correlates with subject and tense differences—e.g. (3a) v. (3b)—since facts about how one discusses their own belief or desire states, in contrast to others belief states, at different times plausibly matter to whether a hearer would take make the EM inference. The explanation for variation relative to subordinate clause structure is less clear but roughly two routes are possible: (i) some property of the subordinate clause licenses (or blocks) EM inferences; and/or (ii) predicate ambiguity correlates with which subordinate clause structure (or property thereof) a predicate selects.

Abstracting the Approaches Across both approaches, there are roughly three kinds of explanations for neg-raising inferences that can be mixed-and-matched: (i) lexical properties might directly or indirectly (e.g. via an EM inference) license a neg-raising inference; (ii) properties of a subordinate clause structure might directly or indirectly license a neg-raising inference; and/or (iii) lexical and structural properties might interact—e.g. via selection—to directly or indirectly license a neg-raising inference. We incorporate these three kinds of explanation into our models (§4), which we fit to the data described in the next section.

3 Data

We develop a method for measuring neg-raising analogous to White and Rawlins' (2018) and White et al.'s (2018) method for measuring veridicality inferences triggered by predicates. With the aim of capturing the range of variability in neg-raising inferences across the lexicon, we deploy this method to test effectively all English clause-embedding verbs in a variety of subordinate clause types—finite and nonfinite—as well as matrix tenses—*past* and *present*—and matrix subjects—*first* and *third person*.

Method Participants are asked to answer questions like (10) using a 0-1 slider, wherein the first italicized sentence has negation in the matrix clause and the second italicized sentences has negation in the subordinate.[2]

(10) If I were to say *I don't think that a particular thing happened*, how likely is it that I mean *I think that that thing didn't happen*?

Because some sentences, such the italicized in (11), sound odd with negation in the matrix clause, participants are asked to answer how easy it is to imagine someone actually saying the sentence—again, on a 0-1 slider. The idea here is that the harder it is for participants to imagine hearing a sentence, the less certain they probably are about the judgment to questions like (10).

(11) How easy is it for you to imagine someone saying *I don't announce that a particular thing happened*?

Acknowledging the abuse of terminology, we refer to responses to (11) as *acceptability responses*. We incorporate these responses into our model (§4) as weights determining how much to pay attention to the corresponding *neg-raising response*.

Materials We use the MegaAcceptability dataset of White and Rawlins (2016) as a basis on which to construct acceptable items for our experiment. MegaAcceptability contains ordinal acceptability judgments for 50,000 sentences, including 1,000 clause-embedding English verbs in 50 different syntactic frames. To avoid typicality effects, these frames are constructed to contain as little lexical content as possible besides the verb at hand—a method we follow here. This is done by ensuring that all NP arguments are indefinite pronouns *someone* or *something* and all verbs besides the one being tested are *do*, *have* or *happen*. We focus on the six frames in (12)–(17).

(12) [NP _ that S]
Someone knew that something happened.

(13) [NP _ to VP[+EV]]
Someone liked to do something.

(14) [NP _ to VP[-EV]]
Someone wanted to have something.

(15) [NP be _ that S]
Someone was told that something happened.

[2] The full task instructions are given in Appendix A.

(16) [NP be _ to VP[+EV]]
Someone was ordered to do something.

(17) [NP be _ to VP[-EV]]
Someone was believed to have something.

These frames were chosen so as to manipulate (i) the presence and absence of tense in the subordinate clause; (ii) the presence or absence of a direct object; and (iii) the lexical aspect of the complement. The frames with direct objects were presented in passivized form so that they were acceptable with both communicative predicates—e.g. *tell*—and emotive predicates—e.g. *sadden*—the latter of which tend to occur with expletive subjects. Lexical aspect was manipulated because some verbs—e.g. *believe*—are more acceptable with nonfinite subordinate clauses headed by a stative than ones headed by an eventive, while others—e.g. *order*—show the opposite pattern.

In light of the variability in neg-raising inferences across the same verb in different tenses—compare again (3a) and (3b)—we aim to manipulate the matrix tense of each clause-taking verb in our experiment. This is problematic, because the MegaAcceptability dataset only contains items in the past tense. We could simply manipulate the tense for any acceptable sentences based on such past tense items, but some verbs do not sound natural in the present tense with some subordinate clauses—compare the sentences in (18).

(18) a. Jo wasn't told that Mary left.
b. Jo isn't told that Mary left.

To remedy this, we extend MegaAcceptability with tense information by constructing new sentences in both present and past progressive tense and collecting acceptability judgments for the sentences.[3] Combined with MegaAcceptability, our extended dataset results in a total of 75,000 verb-tense-frame pairs: 50,000 from the MegaAcceptability dataset and 25,000 from our dataset. From this combined dataset, we take past and present tense items rated on average 4 out of 7 or better (after rating normalization), for our experiment. This yields 3,968 verb-tense-frame pairs and 925 unique verbs. With our subject manipulation (first v. third person), the number of items doubles, producing 7,936 items. Table 1 summarizes the distribution of verbs in each frame and tense.

To construct items, we follow the method of White et al. (2018) of "bleaching" all lexical category words in our sentences (besides the subordinate clause-taking verb) by realizing NPs as *a particular person* or *a particular thing*. Verbs are replaced with *do*, *have*, or *happen*. This method aims to avoid unwanted typicality effects that might be introduced by interactions between our predicates of interest and more contentful items elsewhere in the sentence.[4]

Matrix tense	Frame	# verbs
past	NP _ that S	556
	NP _ to VP[+EV]	400
	NP _ to VP[-EV]	359
	NP be _ that S	255
	NP be _ to VP[+EV]	461
	NP be _ to VP[-EV]	460
present	NP _ that S	413
	NP _ to VP[+EV]	219
	NP _ to VP[-EV]	155
	NP be _ that S	176
	NP be _ to VP[+EV]	268
	NP be _ to VP[-EV]	246

Table 1: # of verbs acceptable in each tense-frame pair based on our extension of MegaAcceptability.

We partition items into 248 lists of 32 items. Each list is constrained such that (i) 16 items had a first person subject, and 16 items had a third person subject; (ii) 16 items contain a *low frequency* verb and 16 items contain a *high frequency* verb, based on a median split of the frequencies in the SUBTLEX_US word frequency database (Brysbaert and New, 2009); (iii) 16 items are *low acceptability* and 16 items are *high acceptability*, based on a median split of the normalized acceptabilities for items selected from our extension of the MegaAcceptability dataset; (iv) no verb occurred more than once in the same list; (iv) items containing a particular combination of matrix tense and syntactic frame occur in rough proportion to the number of verbs that are acceptable with that tense-frame combination based on our extension of the MegaAcceptability dataset (see Table 1).

Participants 1,108 participants were recruited through Amazon Mechanical Turk to give 10 ratings per sentence in the 248 lists of 32—i.e. the end result contains 79,360 ratings for each of neg-raising and acceptability judgments. Participants were not allowed to respond to the same list more than once, though they were allowed to respond to

[3] See Appendix B for details.

[4] Because this method has not been previously validated for measuring neg-raising, we report two validation experiments in Appendix C, which demonstrate that the measure accords with judgments from prior work.

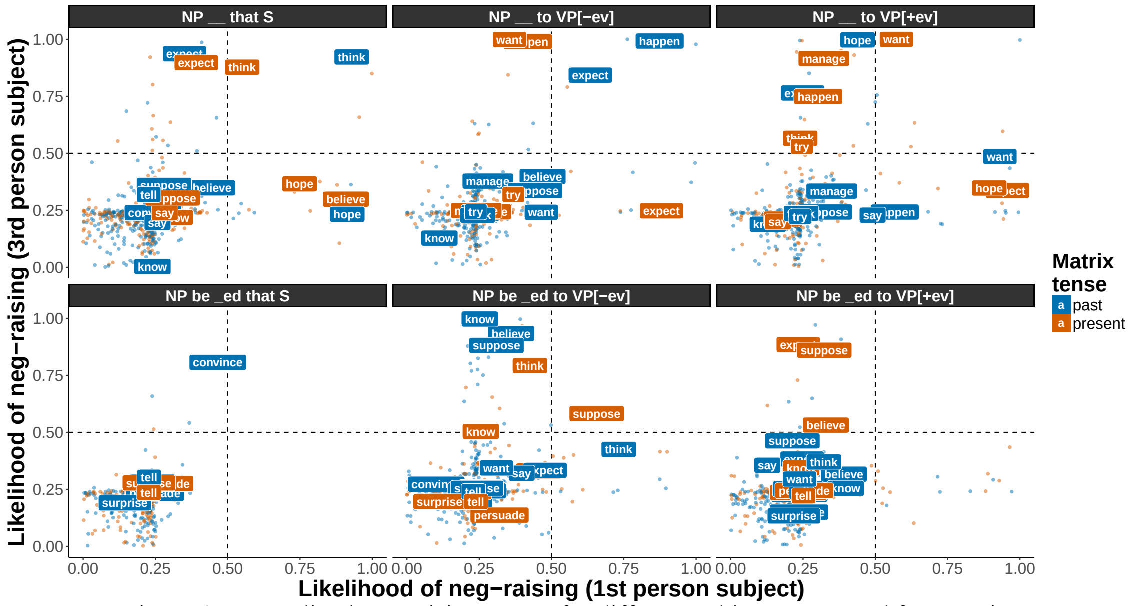

Figure 1: Normalized neg-raising scores for different subject, tense, and frame pairs.

as many lists as they liked. Each participants responded to 2.3 lists on average (min: 1, max: 16, median: 1). Of the 1,108 participants, 10 reported not speaking American English as their native language. Responses from these participants were filtered from the dataset prior to analysis. From this, responses for 27 lists were lost (∼1% of the responses). This filtering removed at most two judgments for any particular item.

Results Figure 1 plots the normalized neg-raising scores for verbs in different subject (axes)-tense (color)-frame (block) contexts.[5] A verb (in some tense) being toward the top-right corner means that it shows strong neg-raising inferences in both first person and third person subject, while a verb being towards the bottom-right corner means that it shows neg-raising behavior with a first person subject but not with a third person subject. The converse holds for the top-left corner: neg-raising behavior is seen with the third person subject but not the first. We see that our method correctly captures canonical neg-raising predicates—e.g. *think* and *believe* with a finite complements and *want* and *expect* with infinitival complements—as well as canonical non-neg-raising predicates—e.g. *know* and *say* with finite complements and *try* and *manage* with infinitivals.

4 Model

We aim to use our neg-raising dataset to assess which aspects of neg-raising inferences are due to properties of lexical items and which aspects are due to properties of the structures they compose with. To do this, we extend White and Rawlins' (2016) model of s(emantic)-selection, which induces semantic type signatures from syntactic distribution, with a module that associates semantic types with the *inference patterns* they trigger.

Our model has two hyperparameters that correspond to the theoretical constructs of interest: (i) the number of lexical properties relevant to neg-raising that it assumes; and (ii) the number of structural properties relevant to neg-raising that it assumes. In §5, we report on experiments aimed at finding the optimal setting these two hyperparameters, and we analyze the parameters of the model fit corresponding to these hyperparameters in §6.

S-selection Model White and Rawlins' (2016) model of s-selection aims to induce verbs' semantic type signatures—e.g. that *love* can denote a relation between two entities and *think* can denote a relation between an entity and a proposition—from their syntactic distribution—e.g. that *love* is acceptable in NP _ NP frames and that *think* is acceptable in NP _ S frames. They formalize this task as a boolean matrix factorization (BMF) problem: given a boolean matrix $\mathbf{D} \in \mathbb{B}^{|\mathcal{V}|\times|\mathcal{F}|} = \{0,1\}^{|\mathcal{V}|\times|\mathcal{F}|}$, wherein $d_{vf} = 1$ iff verb $v \in \mathcal{V}$ is acceptable in syntactic frame $f \in \mathcal{F}$, one must induce boolean matrices $\mathbf{S} \in \mathbb{B}^{|\mathcal{V}|\times|\mathcal{T}|}$ and $\mathbf{P} \in \mathbb{B}^{|\mathcal{T}|\times|\mathcal{F}|}$, wherein $s_{vt} = 1$ iff verb v can have semantic type signature $t \in \mathcal{T}$ and $p_{tf} = 1$ iff t can be mapped onto syntactic frame f, such that (19): verb v is acceptable in frame f iff v has some type t that can be mapped (or *projected*) onto f.

[5] See Appendix D for details on normalization.

(19) $d_{vf} \approx \bigvee_t s_{vt} \wedge p_{tf}$

As is standard in matrix factorization, the equivalence is approximate and is only guaranteed when there are as many semantic type signatures $\mathcal{T}$ as there are frames $\mathcal{F}$, in which case, the best solution is the one with $\mathbf{S} = \mathbf{D}$ and $\mathbf{P}$ as the identity matrix of dimension $|\mathcal{T}| = |\mathcal{F}|$. Because this solution is trivial, $|\mathcal{T}|$ is generally much smaller than $|\mathcal{F}|$ and determined by fit to the data—in BMF, the count of how often $d_{vf} = \bigvee_t s_{vt} \wedge p_{tf}$.

As an estimate of $\mathbf{D}$, White and Rawlins use the MegaAcceptability dataset, which we use in constructing our neg-raising dataset (§3). Instead of directly estimating the boolean matrices $\mathbf{S}$ and $\mathbf{P}$, they estimate a probability distribution over the two under the strong independence assumption that all values s_{vt} and p_{tf} are pairwise independent of all other values. This implies (20).[6]

(20) $\mathbb{P}(d_{vf}) = 1 - \prod_t 1 - \mathbb{P}(s_{vt})\mathbb{P}(p_{tf})$

White and Rawlins treat $\mathbb{P}(d_{vf})$ as a fixed effect in an ordinal mixed effects model, which provides the loss function against which $\mathbb{P}(s_{vt})$ and $\mathbb{P}(p_{tf})$ are optimized. They select the number of semantic type signatures to analyze by setting $|\mathcal{T}|$ such that an information criterion is optimized.

Neg-Raising Model We retain the main components of White and Rawlins' model but add a notion of *inference patterns* associated both with properties of verbs, on the one hand, and with semantic type signatures, on the other. In effect, this addition models inferences, such as neg-raising, as arising via a confluence of three factors: (i) properties of the relation a lexical item denotes—e.g. in a semantic/pragmatic approach, whatever property of a predicate triggers EM inferences; (ii) properties of the kinds of things that a predicate (or its denotation) relates—e.g. in a syntactic approach, whatever licenses "raising" of the negation; and (iii) whether a particular verb has a particular type signature. With respect to (ii) and (iii), it is important to note at the outset that, because we do not attempt to model acceptability, semantic type signatures play a somewhat different role in our model than in White and Rawlins': instead of determining which structures a verb is compatible with—i.e. (non)finite subordinate clauses, presence of a direct object, etc.—our model's type signatures control the inferences a particular verb can trigger when taking a particular structure. As such, our model's semantic type signatures might be more easily construed as properties of a structure that may or may not license neg-raising.[7]

Our extension requires the addition of three formal components to White and Rawlins' model: (i) a boolean matrix $\mathbf{\Psi} \in \mathbb{B}^{|\mathcal{V}|\times|\mathcal{I}|}$, wherein $\psi_{vi} = 1$ iff verb $v \in \mathcal{V}$ has property $i \in \mathcal{I}$; (ii) a boolean tensor $\mathbf{\Phi} \in \mathbb{B}^{|\mathcal{I}|\times|\mathcal{J}|\times|\mathcal{K}|}$, wherein $\phi_{ijk} = 1$ iff property i licenses a neg-raising inference with subject $j \in \mathcal{J}$ and tense $k \in \mathcal{K}$; and (iii) a boolean tensor $\mathbf{\Omega} \in \mathbb{B}^{|\mathcal{T}|\times|\mathcal{J}|\times|\mathcal{K}|}$, wherein $\omega_{tjk} = 1$ iff semantic type signature $t \in \mathcal{T}$ licenses a neg-raising inference with subject j and tense k.

Analogous to White and Rawlins, we treat our task as a problem of finding $\mathbf{S}, \mathbf{P}, \mathbf{\Psi}, \mathbf{\Phi}, \mathbf{\Omega}$ that best approximate the tensor $\mathbf{N}$, wherein $n_{vfjk} = 1$ iff verb v licenses neg-raising inferences in frame f with subject j and tense k. This is formalized in (21), which implies that $n_{vfjk} = 1$ iff there is some pairing of semantic type signature t and inference pattern i such that (i) verb v has semantic type signature t; (ii) verb v licenses inference pattern i; (iii) semantic type signature t can map onto frame f; and (iv) both t and i license a neg-raising inference with subject j and tense k.

(21) $n_{vfjk} \approx \bigvee_{t,i} s_{vt} \wedge \psi_{vi} \wedge \phi_{ijk} \wedge p_{tf} \wedge \omega_{tjk}$

Also analogous to White and Rawlins, we aim to estimate $\mathbb{P}(n_{vfjk})$ (rather than n_{vfjk} directly) under similarly strong independence assumptions: $\mathbb{P}(s_{vt}, \psi_{vi}, \phi_{ijk}, p_{tf}, \omega_{tjk}) = \mathbb{P}(s_{vt})\mathbb{P}(\psi_{vi})$ $\mathbb{P}(\phi_{ijk})\mathbb{P}(p_{tf})\mathbb{P}(\omega_{tjk})$. This implies (22), where $\mathbb{P}(\zeta_{vtifjk}) = \mathbb{P}(s_{vt})\mathbb{P}(\psi_{vi})\mathbb{P}(\phi_{ijk})\mathbb{P}(p_{tf})\mathbb{P}(\omega_{tjk})$.

(22) $\mathbb{P}(n_{vfjk}) = 1 - \prod_{t,i} 1 - \mathbb{P}(\zeta_{vtifjk})$

We design the loss function against which $\mathbb{P}(s_{vt})$, $\mathbb{P}(\psi_{vi})$, $\mathbb{P}(\phi_{ijk})$, $\mathbb{P}(p_{tf})$, and $\mathbb{P}(\omega_{tjk})$ are optimized such that (i) $\mathbb{P}(n_{vfjk})$ is monotonically related to the neg-raising response r_{vfjkl} given by participant l for an item containing verb v in frame f with subject j and tense k (if one exists); but (ii) participants may have different ways of using the response scale. For example, some participants may prefer to use only values close to 0 or 1, while others may prefer values near 0.5; or some participants may be prefer lower likelihood values while others may prefer higher values. To implement this, we incorporate (i) a fixed scaling term σ_0; (ii)

[6]See Appendix E for the derivation of (20).

[7]Alternatively, they might be construed as (potentially cross-cutting) classes of syntactic structures and/or semantic type signatures that could be further refined by jointly modeling acceptability (e.g. as measured by MegaAcceptability) alongside our measure of neg-raising inferences.

a fixed shifting term β_0; (iii) a random scaling term σ_l for each participant l; and (iv) a random shifting term β_l for each participant l. We define the expectation for a response r_{vfjkl} as in (23).

(23) $\hat{r}_{vfjkl} = \text{logit}^{-1}\left(m_a \nu_{vfjk} + \beta_0 + \beta_l\right)$
where $\nu_{vfjk} = \text{logit}\left(\mathbb{P}(n_{vfjk})\right)$
$m_l = \exp\left(\sigma_0 + \sigma_l\right)$

We optimize $\mathbb{P}(s_{vt})$, $\mathbb{P}(\psi_{vi})$, $\mathbb{P}(\phi_{ijk})$, $\mathbb{P}(p_{tf})$, and $\mathbb{P}(\omega_{tjk})$ against an L1 loss.[8] To take into account that it is harder to judge the neg-raising inferences for items that one cannot imagine hearing used, we additionally weight the above-mentioned L1 loss by a normalization of the acceptability responses for an item containing verb v in frame f with subject j and tense k. We infer this value from the acceptability responses for an item containing verb v in frame f with subject j and tense k given by participant l, assuming a form for the expected value of a_{vfjkl} as in (24)—analogous to (23). (Unlike ν_{vfjk} in (23), α_{vfjk} in (24) is directly optimized.)

(24) $\hat{a}_{vfjkl} = \text{logit}^{-1}\left(m'_l \alpha_{vfjk} + \beta'_0 + \beta'_l\right)$
where $m'_l = \exp\left(\sigma'_0 + \sigma'_l\right)$

The final loss against which $\mathbb{P}(s_{vt})$, $\mathbb{P}(\psi_{vi})$, $\mathbb{P}(\phi_{ijk})$, $\mathbb{P}(p_{tf})$, $\mathbb{P}(\omega_{tjk})$ are optimized is (25).[9]

(25) $\sum \alpha'_{vfjk}|r_{vfjkl} - \hat{r}_{vfjkl}| + |a_{vfjkl} - \hat{a}_{vfjkl}|$

where $\alpha'_{vfjk} = \text{logit}^{-1}(\alpha_{vfjk})$.

5 Experiment

We aim to find the optimal settings, relative to our neg-raising data, for (i) the number $|\mathcal{I}|$ of lexical properties relevant to neg-raising that it assumes; and (ii) the number $|\mathcal{T}|$ of structural properties relevant to neg-raising that it assumes.

Method We conduct a five-fold cross-validation for each possible pairing of $|\mathcal{I}| \in \{1, 2, 3, 4\}$ and $|\mathcal{T}| \in \{1, 2, 3, 4\}$, wherein we pseudorandomly partition our items into five sets (folds), fit the model with those hyperparameters to the responses to items in four of these sets (80% of the data), and then compute the loss on the held-out set. The same partition is used for every setting of $|\mathcal{I}|$ and $|\mathcal{T}|$. The assignment of items to folds is pseudorandom because we ensure that each fold contains at least one instance of a particular verb in a particular frame; otherwise, on some folds, the model might have no data upon which to predict that verb in that frame.

[8] We also considered a Beta likelihood loss but found that fits were not possible due to numerical instability.

[9] An additional term (not shown) is added to encode the standard assumption that the random effects terms are normally distributed with mean 0 and unknown variance.

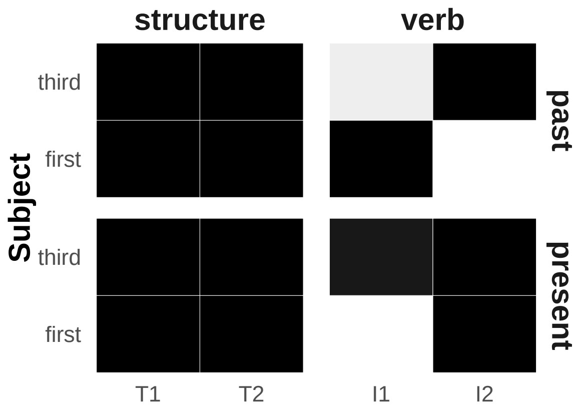

Figure 2: Relationship between structure properties and neg-raising inferences $\mathbb{P}(\mathbf{\Omega})$ (left) and verb properties and neg-raising inferences $\mathbb{P}(\mathbf{\Phi})$ (right) with different subjects (y-axis) and tenses (top v. bottom).

Implementation We implement our model in `tensorflow 1.12.0` (Abadi et al., 2016). We use the Adam optimizer (Kingma and Ba, 2015) with default hyperparameters.

Results We find that the model fit with the lowest average loss on held-out data across folds is the one where $|\mathcal{I}| = 2$ and $|\mathcal{T}| = 2$. This result may suggest that that neg-raising is not purely a product of lexical knowledge: properties of the subordinate clause that a predicate combines with also influence whether neg-raising inferences are triggered. This is a surprising finding from the perspective of prior work, since (to our knowledge) no existing proposals posit that the syntactic structure of a predicates subordinate clause can influence whether neg-raising inferences are triggered. We next turn to analysis of this model fit to understand how our model captures patterns in the data.

6 Analysis

Figure 2 gives our model's estimate of the relationship between structural properties and neg-raising inferences $\mathbb{P}(\mathbf{\Omega})$ (left) and lexical properties and neg-raising inferences $\mathbb{P}(\mathbf{\Phi})$ (right) with different subjects (y-axis) and tenses (top v. bottom). We see that the model infers that having either of the two structure properties (left) makes a structure compatible with neg-raising inferences neg-raising inferences, regardless of the subject and tense. In contrast, the two lexical properties (right) differ on whether they license neg-raising inferences: I1 correlates licensing of neg-raising in-

Frame	T1	T2
NP _+T that something happened.	0.00	1.00
NP be+T _ed that something happened.	0.93	0.00
NP _+T to do something.	0.00	1.00
NP be+T _ed to do something.	1.00	0.18
NP _+T to have something.	1.00	0.10
NP be+T _ed to have something.	1.00	0.02

Table 2: Relationship between structure properties and structures $\mathbb{P}(\mathbf{P})$.

ferences with tense—requiring a combination of a first person subject and past tense or a third person subject and present tense—while I2 blocks neg-raising only with the particular combination of a first person subject and past tense. This pattern is interesting because it suggests that the model captures the variability across different subjects and tenses observed in Figure 1 as a matter of lexical, not structural, properties. This result would make intuitive sense insofar as such variability arises due to pragmatic reasoning about how one tends to discuss their own belief or desire states, in contrast to others belief states, at different times.

The model does not explain all variability in neg-raising inferences via lexical properties, though; some variability is explained as a fact about which structural properties a structure can have. This can be seen in Table 2, which shows the relationship between structure properties and structures $\mathbb{P}(\mathbf{P})$. We see that the model effectively partitions the frames into two sets: ones that have property T1 and others that have property T2. This implies that, insofar as a verb selects only a single property and a structure that verb is acceptable in does not have that property, a sentence containing that verb in that structure will not be neg-raising.

To illustrate, Figure 3 gives the relationship between selected verbs—separated into neg-raising (top) and non-neg-raising (bottom) based on prior work (§2)—and structure properties $\mathbb{P}(\mathbf{S})$ (left) and verb properties $\mathbb{P}(\mathbf{\Psi})$ (right). The neg-raising verbs universally select both structure types and most select both lexical properties. In contrast, the non-neg-raising verbs show two distinct patterns: some, such as *love* and *hate*, are inferred to be non-neg-raising because the model does not assign any relevant lexical property to those verbs; others have the relevant lexical properties but are inferred to be non-neg-raising with particular structures because they do not select the relevant structural properties. For example, *know* has both I1 and I2, but it only takes a single structural property T1, to licence inferential patterns. As seen in

Figure 3: Relationship between selected verbs and the structure properties $\mathbb{P}(\mathbf{S})$ (left) and verb properties $\mathbb{P}(\mathbf{\Psi})$ (right). Verbs are separate into neg-raising (top) and non-neg-raising (bottom) based on prior work (§2).

Table 2, T1 is missing for the structure [NP _ that S], thus blocking the neg-raising inference in this structure—compatible with the data presented in Figure 1. But because [NP be _ to VP[-EV]] does have property T1, neg-raising is licensed for *know* in that structure—again, compatible with the data presented in Figure 1.

7 Conclusion

We presented a probabilistic model to induce the mappings from lexical sources and their grammatical sources to neg-raising inferences. We trained this model on a large-scale dataset of neg-raising judgments that we collected for 925 English clause-embedding verbs in six distinct syntactic frames as well as various matrix tenses and subjects. Our model fit the best when positing two lexical properties and two structural properties. This is a surprising finding from the perspective of prior work, since (to our knowledge) no existing proposals posit that the syntactic structure of a predicates subordinate clause can influence whether neg-raising inferences are triggered. Our findings suggest new directions for theoretical research attempting to explain the interaction between lexical and structural factors in neg-raising.

References

Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283.

Dorit Abusch. 2002. Lexical alternatives as a source of pragmatic presuppositions. *Semantics and Linguistic Theory*, 12:1–19.

Renate Bartsch. 1973. negative transportation gibt es nicht. *Linguistische Berichte*, 27(7).

Marc Brysbaert and Boris New. 2009. Moving beyond kucera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior Research Methods*, 41:977–990.

Gennaro Chierchia. 2013. *Logic in grammar: polarity, free choice, and intervention*, first edition. Oxford University Press, Oxford.

Chris Collins and Paul Martin Postal. 2014. *Classical NEG Raising: An Essay on the Syntax of Negation*. Book collections on Project MUSE. MIT Press.

Chris Collins and Paul Martin Postal. 2017. Interclausal neg raising and the scope of negation. *Glossa: A Journal of General Linguistics*, 2:1–29.

Chris Collins and Paul Martin Postal. 2018. Disentangling two distinct notions of neg raising. *Semantics and Pragmatics*, 11(5).

Paul Crowley. 2019. Neg-raising and neg movement. *Natural Language Semantics*, 27(1):1–17.

Mark Davies. 2017. Corpus of Contemporary American English (COCA).

Charles J. Fillmore. 1963. The position of embedding transformations in a grammar. *WORD*, 19(2):208–231.

Jon R. Gajewski. 2007. Neg-raising and polarity. *Linguistics and Philosophy*, 30(3):289–328.

Jon R. Gajewski. 2011. Licensing strong npis. *Natural Language Semantics*, 19(2):109–148.

Vincent Homer. 2015. Neg-raising and positive polarity: The view from modals. *Semantics and Pragmatics*, 8(4):1–88.

Laurence Robert Horn. 1971. Negative transportation: unsafe at any speed? *Papers from the seventh regional meeting, Chicago Linguistic Society*, pages 120–133.

Laurence Robert Horn. 1978. Remarks on neg-raising. *Syntax and Semantics*, 9:129–220.

Laurence Robert Horn. 1989. *A Natural History of Negation*. University of Chicago Press.

Laurence Robert Horn. 2014. The cloud of unknowing. In Jack Hoeksema and Dicky Gilbers, editors, *Black Book: A Festschrift for Frans Zwarts*, pages 178–196. University of Groningen, Groningen, the Netherlands.

Laurence Robert Horn and Samuel Bayer. 1984. Short-circuited implicature: A negative contribution. *Linguistics and Philosophy*, 7(4):397–414.

Ray S. Jackendoff. 1971. On some questionable arguments about quantifiers and negation. *Language*, 47(2):282–297.

Adam Kilgarriff, Vt Baisa, Jan Buta, Milo Jakubek, Vojtch Kov, Jan Michelfeit, Pavel Rychl, and Vt Suchomel. 2014. The sketch engine: ten years on. *Lexicography*, pages 7–36.

Diederik P. Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*.

Paul Kiparsky. 1970. Semantic rules in grammar. In Hreinn Benediktsson, editor, *The Nordic Languages and Modern Linguistics*, pages 262–285. Visindafelag Islendinga, Reykjavik.

Edward S. Klima. 1964. *Negation in English*. Englewood Cliffs, NJ: Prentice-Hall.

Howard Lasnik. 1972. *Analyses of negation in English*. Ph.D. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830.

Jay M. Pollack. 1976. A re-analysis of neg-raising in english. *Working Papers in Linguistics*, 21:189–239.

Jacopo Romoli. 2013. A scalar implicature-based approach to neg-raising. *Linguistics and Philosophy*, 36(4):291–353.

Jacopo Romoli and Matthew Mandelkern. 2019. Whats not to like. *Linguistic Inquiry*, 0(ja):1–21.

Nadine Theiler, Floris Roelofsen, and Maria Aloni. 2017. Whats wrong with believing whether? *Semantics and Linguistic Theory*, 27:248–265.

Lucia M. Tovena. 2001. Neg-raising: negation as failure. In Jack Hoeksema, Hotze Rullmann, Vctor Snchez-Valencia, and Ton van der Wouden, editors, *Perspectives on Negation and Polarity Items*, pages 331–356. John Benjamins, Amsterdam.

Aaron Steven White and Kyle Rawlins. 2016. A computational model of s-selection. *Semantics and Linguistic Theory*, 26:641–663.

Aaron Steven White and Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. *Proceedings of the 48th Meeting of the North East Linguistic Society*.

Aaron Steven White and Kyle Rawlins. 2019. Frequency, acceptability, and selection: A case study of clause-embedding. Under Review at *Glossa*.

Aaron Steven White, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2018. Lexicosyntactic inference in neural models. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4717–4724.

Yimei Xiang. 2013. Neg-raising: Focus and implicatures. *Proceedings of Sinn und Bedeutung*, 18:487–503.

Hedde Zeijlstra. 2018. Does neg-raising involve neg-raising? *Topoi*, 37(3):417–433.

Frans Zwarts. 1998. Three types of polarity. In Fritz Hamm and Erhard Hinrichs, editors, *Plurality and Quantification*, pages 177–238. Springer Netherland, Dordrecht.

A Instructions

In this experiment, you will be asked to answer questions about what a person is likely to mean if they say a particular sentence.

Your task will be to respond about the likelihood on the slider that will appear under each question, where the left side corresponds to *extremely unlikely* and the right side corresponds to *extremely likely*.

For instance, you might get the question *If I were to say John has three kids, how likely is it that I mean John has exactly three kids?* with a slider. In this case you would move the slider handle fairly far to the right (toward *extremely likely*), since if someone says "John has three kids", it's pretty likely that they mean that John has exactly three children.

If the question were *If I were to say some of the boys left, how likely is it that I mean all of the boys left?*, then you might move the slider pretty far to the left (toward *extremely unlikely*), since it would be odd if someone says "Some of the boys left and by that, I mean all of the boys left".

And if the question were *If I were to say Ann didnt greet everyone politely, how likely is it that I mean Ann was unwelcoming to every single person?*, you might leave the slider in the middle (which corresponds to *maybe or maybe not*), since quite often such sentence can be used to mean Ann greeted some people politely but not all, or to mean Ann was not polite to every single person.

Try to answer the questions as quickly and accurately as possible. Many of the sentences may not be sentences that you can imagine someone ever saying. Try your best to interpret what a speaker would mean in using them. After each question, you will be given a chance to tell us whether the sentence you just responded to isn't something you can imagine a native English speaker ever saying.

Not all questions have correct answers, but a subset in each HIT do. Prior to approval, we check the answers given for this subset. We will reject work containing a substantial number of answers that do not agree with the correct answer.

When the experiment is over, a screen will appear telling you that you are done, and a submission button will be revealed.

B Data

We extend White and Rawlins' (2016) MegaAcceptability by collecting acceptability judgments for sentences in present and past progressive tenses. Then, by running another experiment to link MegaAcceptability and the extended MegaAcceptability datasets, we combine the two datasets and create a new large-scale dataset. Both the extended MegaAcceptability and linking datasets are available at megaattitude.io.

Extended MegaAcceptability Our test items are selected and modified from the top 25% most acceptable verb-frame pairs from the MegaAcceptability dataset of White and Rawlins (2016). The selected dataset corresponds to 12,500 verb-frame pairs, with 1000 unique verbs and the same 50 subcategorization frames (35 in active voice and 15 in passive voice) that are used in MegaAcceptability. Given the 12,500 verb-frame pairs, we construct new sentences in both present and past progressive tense, resulting in a total of 25,000 items. Example sentences for present and past progressive tense are in (26) and (27).

(26) Someone trusts which thing to do.

(27) Someone was gloating about something.

The experiment consists of 500 lists of 50 sentences. Each list is constructed such that (i) each frame shows up once in a list, making each list contain 50 unique frames, if possible; (ii) otherwise, the distribution of frames are kept as similar as possible across lists; and (iii) no verbs appear more than once in a list. We gather 5 acceptability judgments per sentence, so there are 125,000 judgments in total for 25,000 items. The judgments are on a scale of 1 to 7. To avoid typicality effects, we construct the frames to contain as little lexical content as possible besides the verb at hand. For this, we instantiate all NP arguments as indefinite pronouns *someone* or *something* and all verbs besides the one being tested as *do* or *happen*. For this experiment, we recruited 565 participants, where 562 speak American English as their native language.

Linking experiment Our extended MegaAcceptability dataset contains present and past progressive tenses for the verb-frame pairs that were rated as acceptable in past tense from the existing MegaAcceptability dataset, which makes the two datasets built on two different scales. To combine the existing MegaAcceptability dataset and our extended dataset, which contain different tenses and annotators in their items, we run another experiment to help us link the two datasets on the same scale. We choose 50 items, each with a unique verb, by selecting 26 items from our dataset (14 in present tense and 12 in past progressive tense) and 24 items from MegaAcceptability (all in past tense). A half of the items chosen from each dataset are rated with the lowest acceptability scores while the other half are with the highest acceptability scores; the items with the lowest acceptability scores consist of 8 in the present, 6 in the past progressive, and 12 in the past tense and so do the items with the highest acceptability scores. (28) - (30) show example items with the lowest acceptability scores and (31) - (33) with the highest acceptability scores.

(28) Someone demands about whether something happened.

(29) Someone was judging to someone that something happened.

(30) Someone invited which thing to do.

(31) Someone is distracted.

(32) Someone was teaching.

(33) Someone dared to do something.

The linking experiment contains 35 unique frames, listed in (34).

(34) a. NP (be) _ {so, about NP, about whether S, S, that S, to NP that S, that S[-TENSE]}
  b. NP (be) _ NP {∅, to NP, VPing, to VP[+EV], which NP S}
  c. NP (be) _ {∅, VP, VPing, VP[+/-EV]}
  d. NP (be) _ which {NP to VP, NP S}
  e. NP (be) _ whether {to VP, S, S[FUTURE], that S[FUTURE]}
  f. S, NP _

The linking experiment is built in a very similar manner to our Extended MegaAcceptability experiment above. Ordinal acceptability judgments are collected on a scale of 1 to 7. We recruit 50 participants to rate a list of 50 items for our linking experiment. All of the 50 participants reportspeaking American English as their native language.

After running the linking experiment, we normalize the ratings using the procedure described in White and Rawlins (2019). Then by fitting the linear regression implemented in scikit-learn (Pedregosa et al., 2011) to the normalized ratings of our linking experiment, we make a new dataset combining MegaAcceptability and our extended dataset. To do so, we first fit a subset of the linking data to the normalized MegaAcceptability data. Next, we fit a subset of the normalized data of our extended data to the linking data, leading to two regression models: one mapping a subset of the linking data to a subset of the MegaAcceptability data, and the other mapping a subset of our extended data to a subset of the linking data. We then go about using these regression models to map items of our extended MegaAcceptability dataset to the linking data, as well as to map all items of the linking data to the MegaAcceptability dataset. In turn, this gives us a new and combined large-scale dataset of acceptability judgments for sentences in three different tenses (*past, present*, and *past progressive*) and 50 different syntactic frames.

C Validation Experiments

We conduct experiments aimed at validating our method for measuring neg-raising. In both experiments, we test the same set of 32 clause-embedding verbs, half of which we expect to show neg-raising behavior and the other half we do not (based on the literature discussed in §2). For neg-raising verbs, we refer to the neg-raising predicates listed in Gajewski (2007) and Collins and Postal (2018); and for non-neg-raising verbs, we choose factive verbs and those that Theiler et al. (2017) claim are not neg-raising. The experiments differ with respect to whether we employ "bleached" items (as in the data collection described in the main body of the paper) or "contentful" items, which are constructed based on sentences drawn from English corpora.

Materials We select neg-raising and non-neg-raising verbs such that half of each type takes infinitival subordinate clauses and half takes finite subordinate clauses. Table 3 shows the 32 verbs we choose for the pilot. Some verbs listed as taking one kind of subordinate clause can also take the other. In these cases, we only test that verb in the subordinate clause listed in Table 3.

The matrix subject (first v. third person) and matrix tense (present v. past) are manipulated for each predicate: (35) schematizes four items from our bleached experiment and (36) schematizes four items from our contentful experiment.

Subordinate clause	Neg-raising	Non-neg-raising
Finite	think, believe, feel, reckon, figure, guess, suppose, imagine	announce, claim, assert, report, know, realize, notice, find out
Infinitival	want, wish, happen, seem, plan, intend, mean, turn out	love, hate, need, continue, try, like, desire, decide

Table 3: Verbs used in validation experiments

(35) {I, A particular person} {don't/doesn't, didn't} want to do a particular thing.

(36) {I, Stephen} {don't/doesn't, didn't} want to introduce new rules.

Items for the bleached experiment were constructed automatically using the templates, which select *to have a particular thing* for *turn out* and *seem* as their subordinate clause, *to do a particular thing* for other verbs taking infinitival subordinate clauses, and *that something happened* for the verbs taking finite subordiante clauses. Items for the contentful experiment were constructed by replacing all bleached words (*a particular person, a particular thing, do, have*, and *happen*) from the bleached experiment items by contentful lexical words.

The high content sentences are constructed based on sentences sampled from the Corpus of Contemporary American English (Davies, 2017) and the Oxford English Corpus (Kilgarriff et al., 2014). The contentful items are modified so that third person subject is a proper name and sentences do not include any pauses or conjunctions. To allow possible item variability, we created five contentful items per each bleached item.

For the bleached experiment, four lists of 32 items each were constructed by partitioning the resulting 128 items under the constraints that (i) every list contains every verb with exactly one subject (*first*, *third*) and tense (*past*, *present*) and (ii) every subject-tense pair is seen an equal number of times across verbs. We ensure that the same level of a particular factor is never assigned to the same verb more than once in any list and that the items in a list are randomly shuffled. To construct items, we control four conditions: neg-raising, embedded complement, matrix subject, matrix tense. Neg-raising and embedded complements are pre-determined for each verb, while matrix subject and matrix tense are randomly selected for a verb in each task. The same constraints applied for the contentful experiment except that the test items were partitioned into 20 lists of 32 instead of four lists because the total number of sentences for the contentful experiment is five times bigger than the bleached experiment.

Participants For the bleached experiment, 100 participants were recruited such that each of the four lists was rated by 25 unique participants. For the contentful experiment, 100 participants were recruited as well, to have each of the 20 lists of 32 rated by five unique participants. No participant was allowed to rate more than one list. In each experiment, one participant out of 100 reported not speaking American English natively and this participant's responses were filtered prior to analysis.

Analysis We test whether our task correctly captures canonical (non-)neg-raising verbs using linear mixed effects models. For both validation experiments, we start with a model containing fixed effects for NEGRAISING (*true*, *false*; as in Table 3), random intercepts for PARTICIPANT, VERB, and (in the contentful validation) ITEM. Nested under both verb and participant, we also included random intercepts for MATRIX SUBJECT (*1st*, *3rd*) and MATRIX TENSE (*past*, *present*) and their interaction. We compare this against a model with the same random effects structure but no effect of NEGRAISING. We find a reliably positive effect of NEGRAISING for both the bleached experiment ($\chi^2(1) = 34.5$, $p < 10^{-3}$) and the contentful experiment ($\chi^2(1) = 19.8$, $p < 10^{-3}$). This suggests that participants responses are consistent with neg-raising inferences being more likely with verbs that have previously been claimed to give rise to such inferences.

D Normalization

For the purposes of visualization in §3, we present normalized neg-raising scores. These scores are derived using a mixed effects robust regression with loss the same loss (25) as for the model described in Section 4, except that, unlike for the model, where ν_{vfjk} is defined in terms of the model, for the purposes of normalization, both ν_{vfjk} in (23) and α_{vfjk} in (24) are directly optimized. Figure 1 plots $\text{logit}^{-1}(\exp(\sigma_0)\nu_{vfjk})+\beta_0$.

E Model Derivation

$$
\begin{aligned}
\mathbb{P}(d_{vf}) &= \mathbb{P}\left(\bigvee_t s_{vt} \wedge p_{tf}\right) \\
&= \mathbb{P}\left(\neg\neg \bigvee_t s_{vt} \wedge p_{tf}\right) \\
&= \mathbb{P}\left(\neg \bigwedge_t \neg(s_{vt} \wedge p_{tf})\right) \\
&= \mathbb{P}\left(\neg \bigwedge_t \neg(s_{vt} \wedge p_{tf})\right) \\
&= 1 - \mathbb{P}\left(\bigwedge_t \neg(s_{vt} \wedge p_{tf})\right) \\
&= 1 - \prod_t \mathbb{P}\left(\neg(s_{vt} \wedge p_{tf})\right) \\
&= 1 - \prod_t 1 - \mathbb{P}\left(s_{vt} \wedge p_{tf}\right) \\
&= 1 - \prod_t 1 - \mathbb{P}(s_{vt})\mathbb{P}(p_{tf})
\end{aligned}
$$