Proceedings of SALT 31: 000–000, 2021

Intensional Gaps: Relating veridicality, factivity, doxasticity, bouleticity, and neg-raising*

Benjamin Kane
University of Rochester

William Gantt
University of Rochester

Aaron Steven White
University of Rochester

Abstract We investigate which patterns of lexically triggered doxastic, bouletic, neg(ation)-raising, and veridicality inferences are (un)attested across clause-embedding verbs in English, as well as how those categories map onto morphosyntactic distribution. To carry out this investigation, we use a multiview mixed effects mixture model to discover the inference patterns captured in three lexicon-scale inference judgment datasets: two existing datasets, MegaVeridicality and MegaNegRaising, which capture veridicality and neg-raising inferences across a wide swath of the English clause-embedding lexicon, and a new dataset, MegaIntensionality, which similarly captures doxastic and bouletic inferences. We focus in particular on inference patterns that are correlated with morphosyntactic distribution, as determined by how well those patterns predict the acceptability judgments in the MegaAcceptability dataset. We find that there are 15 such patterns attested. Similarities among these patterns suggest the possibility of underlying lexical semantic components that give rise to them. We use principal component analysis to discover these components and suggest generalizations regarding the (in)compatibility of these components and the possibility of explaining lexical gaps.

Keywords: inference patterns, veridicality, factivity, neg-raising, doxasticity, bouleticity, acceptability

1 Introduction

Gaps in logically possible patterns of lexically triggered inferences have long played an important role in semantic theory because they suggest potentially deep constraints on lexicalization (Horn 1972; Barwise & Cooper 1981; Levin & Rappaport Hovav 1991: a.o.). Recently, there has been substantial progress on explaining such gaps in the domain of propositional attitude predicates. Much of this work focuses on lexically triggered inferences that are associated with predicates' intensional properties—in particular, lexically triggered *veridicality inferences* (1a), *neg(ation)-raising inferences* (1b), *doxastic inferences* (1c), and *bouletic inferences* (1d).

(1) A predicate v triggers...
   a. ...{*veridicality*, *antiveridicality*} *inferences* in sentence np (*not*) v s iff the use of np (*not*) v s consistently triggers the inference that {s, *not* s} across contexts.

* This work was supported by NSF-BCS-1748969 (*The MegaAttitude Project: Investigating selection and polysemy at the scale of the lexicon*). All data and code are available at megaattitude.io

b. ...*neg(ation)-raising inferences* in a sentence np *not* v s iff the use of np *not* v s can trigger the inference that np v *not* s in some contexts.
c. ...{*doxastic*, *antidoxastic*} *inferences* in a sentence np (*not*) v s iff the use of np (*not*) v s consistently triggers the inference that {np *believes* s, np *believes not* s} across contexts.
d. ...{*bouletic*, *antibouletic*} *inferences* in a sentence np (*not*) v s iff the use of np (*not*) v s consistently triggers the inference that {np *wants* s, np *wants not* s} across contexts.

These inferences are of interest for at least two reasons. First, they display apparent correlations with each other across lexical items—potentially suggesting some core set of lexical properties that interact to give rise to them. For instance, Anand & Hacquard (2014) suggest that, while there are predicates for which doxastic inferences are "foregrounded" (as diagnosed by sensitivity to semantic operators like negation)—e.g. (2a) ⇝ (4a), (3a) ⇝ (5a)—and predicates for which bouletic inferences are entailments and doxastic inferences are "backgrounded" (as diagnosed by *in*sensitivity to semantic operators like negation)—e.g. (2b) ⇝ (4a), (3b) ⇝ (4a), (2b) ⇝ (4b), (3b) ⇝ (5b)—there are no predicates for which doxastic inferences are "foregrounded" and bouletic inferences are "backgrounded" (see also Hooper 1975; Heim 1992; Anand & Hacquard 2013).

(2) a. Jo knew that Bo left.
b. Jo liked that Bo left.

(3) a. Jo didn't know that Bo left.
b. Jo didn't like that Bo left.

(4) a. Jo believed that Bo left.
b. Jo wanted Bo to have left.

(5) a. Jo didn't believe that Bo left.
b. Jo didn't want Bo to have left.

Second, these inferences appear to correlate with morphosyntactic distribution—potentially suggesting that said lexical properties may be formally represented, rather than solely a byproduct of how conceptual representations interact with pragmatic reasoning. For example, veridicality and neg-raising inferences have been claimed to correlate with interrogative selection (Hintikka 1975; Karttunen 1977; Zuber 1983; Berman 1991; Ginzburg 1995; Lahiri 2002; Egré 2008; George 2011; Uegaki 2015; Theiler, Roelofsen & Aloni 2017, 2019; Elliott, Klinedinst, Sudo & Uegaki 2017; Uegaki & Sudo 2019; Roberts 2019; cf. White & Rawlins 2018) and mood/finiteness selection (see Giannakidou & Mari 2021; and references therein), and doxastic and bouletic inferences have been claimed to correlate with mood/finiteness selection (Bolinger 1968; Hooper 1975; Farkas 1985; Portner 1992; Giorgi & Pianesi 1997; Giannakidou 1997; Quer 1998; Villalta 2000, 2008).

A major remaining challenge in this domain is that, not only is the space of possible inference patterns vast—computed naïvely and considering only the effect of matrix polarity, there are 1,458 possible inference patterns—even the cleanest measurements of at least veridicality and neg-raising using inference judgment tasks display substantial gradience (White & Rawlins 2018; An & White 2020). This gradience makes it difficult to ascertain which inference patterns are attested because it makes it difficult to determine whether a particular sentence should be considered to trigger a particular inference. Such a difficulty may be unavoidable—e.g. because the gradience is endemic to the inferences themselves (Tonahuser & Degen under review). But it could also be that gradience is partly or wholly a product of task effects and that there are discrete, formally represented lexical properties that are active in triggering these inferences—a strong possibility given what we know about gradience in acceptability judgment tasks (see Schütze & Sprouse 2014).

Our aim in this paper is to derive a full taxonomy of lexically triggered doxastic, bouletic, neg-raising, and veridicality inference patterns associated with predicates that take finite clausal complements. To carry out this derivation while addressing the challenges posed by gradience, we apply a soft clustering model to all three lexicon-scale inference judgment datasets.

Two of these datasets already exist: one focused on veridicality inferences (the MegaVeridicality dataset; White & Rawlins 2018) and the other focused on neg-raising inferences (the MegaNegRaising dataset; An & White 2020). We review these datasets in §2. No similar, lexicon-scale dataset capturing doxastic and bouletic inferences currently exists. To address this gap, we extend the methodology used to collect MegaVeridicality and MegaNegRaising to collect doxastic and bouletic inferences for 725 finite clause-taking predicates, covering a wide variety of semantic classes. These include cognitive predicates—e.g. *think*, *know*, *remember*, *forget*—emotive predicates—e.g. *hope*, *fear*, *love*, *hate*—and communicative predicates—e.g. *say*, *tell*, *notify*, *convince*—among others. A unique challenge that arises in collecting judgments for these inferences is that they are anchored to a particular role—e.g. in a convincing, we know that the convincee believes the content of the convincing afterward, but we can draw no similar inferences about the beliefs of the convincer. In §3, we report on an experiment validating a novel method for capturing role-anchored inferences while avoiding typicality effects on the judgments.

After describing our lexicon-scale data collection in §4, we describe the model we use to discover inference patterns and the verbs that trigger them in §5. To improve the chances that these clusters correspond to some formally represented lexical property, we select the number of clusters to assume by selecting the clustering that best predicts the acceptability judgments in the MegaAcceptability dataset (White & Rawlins 2016, 2020) with the minimum number of clusters. We then use principal component analysis to extract generalizations about the nature of the inference patterns we find before concluding in §6.

2 Existing lexicon-scale datasets

Our work builds on two previous lexicon-scale datasets capturing veridicality (the MegaVeridicality dataset; White & Rawlins 2018) and neg-raising inferences (the MegaNegRaising dataset; An & White 2020). The MegaVeridicality dataset contains veridicality judgments for 517 finite clause-embedding English verbs in a variety of syntactic frames, along with manipulations for voice and matrix polarity. These verbs were selected on the basis of their acceptability in those frames according to the MegaAcceptability dataset (White & Rawlins 2016, 2020), which contains acceptability judgments for 1,000 English clause-embedding verbs in 50 syntactic frames.

A major challenge to collecting inference judgments lies in disentangling the lexical effects of the predicate on the inference from the potentially confounding effects of world knowledge or sentential context. For instance, if a respondent were to indicate that (6b) follows from (6a), it would be difficult as the experimenter to tell whether their judgment was due to the semantics of *know* (as desired) or to the respondent's knowledge of world history.

(6) a. Jo knew that Napoleon was defeated at Waterloo.
b. Napoleon was defeated at Waterloo.

Thus, to isolate predicate-specific effects on veridicality inferences, White & Rawlins build on a method they developed in White & Rawlins 2016 for constructing low-content sentences. Specifically, they solicit judgments by presenting participants with a *bleached* sentence, as in (7a), and then ask (7b), where the possible responses are *yes*, *no*, and *maybe or maybe not*.

(7) a. Someone {knew, didn't know} that a particular thing happened.
    b. Did that thing happen?

This bleaching is carried out by instantiating all DP arguments in a frame with indefinite pronouns and by replacing all verbs except for the target verb with *do*, *have*, or *happen*, as appropriate. Note that White & Rawlins are further able to capture factivity by manipulating the negation on the verb and seeing whether participants judge that the inferences goes through in both conditions. The results of their experiments broadly confirm the validity of this approach, as the judgments they collect are able to distinguish a large set of canonically veridical, non-veridical, factive, and non-factive predicates.

An & White (2020) take a similar approach to investigating neg-raising inferences. Participants are presented with questions like (8) and respond using a bounded slider ranging from 0 to 1, with 0 indicating *not likely at all* and 1 indicating *very likely*.

(8) 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*?

As in White & Rawlins 2018, the authors select predicates based on their acceptability in different frames and with different tenses based on data from MegaAcceptability. The resulting MegaNegRaising dataset comprises judgments for 925 clause-embedding verbs in six frames in both past and present tenses, and with both first and third person subjects—e.g. the past tense analogue of (8) with a third person subject would be (9).

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

In this paper, we focus only on the items containing finite embedded clauses.

3 Validating a templatic approach

We aim to capture doxastic and bouletic inferences across the entire lexicon. As for veridicality and neg-raising inferences, a major obstacle to collecting such data at scale is that, using standard item construction methods, it can be difficult to ensure that one is isolating inferences triggered by the predicate of interest (in some syntactic context) rather than surrounding lexical material (in conjunction with world knowledge). For instance, *boast* tends to trigger an inference that the boaster believes the content of the boast, but this inference can fail to be triggered in cases where the boaster is widely assumed to be a willful liar, as in (10).

(10) Trump boasted that he won in 2020.

(11) Trump doubts that he won in 2020.

Conversely, *doubt* tends not to trigger desire inferences; but if given a sentence like (11) and asked how likely it is that Trump wants to have won the 2020 election, one would likely answer that it is highly likely—mainly on the basis of prior knowledge.

There are two ways to mitigate this issue: (i) ensure that the attitude holder is not opinionated with respect to the content of the embedded clause; or (ii) explicitly adjust for the possibility that judgments are due to prior knowledge. Either option would require the collection of norming data regarding the prior probability of the belief or desire relevant to an item; and while this approach appears sound, it is clearly infeasible for lexicon-scale tasks when considering the number of items that would need to be hand-written as well as the additional cost of norming those items.

In this section, we validate an alternative, efficiently scalable technique based on the bleaching method described in §2 that was proposed by White & Rawlins (2016, 2020) for acceptability judgment tasks and subsequently modified by White & Rawlins (2018), White, Rudinger, Rawlins & Van Durme (2018b), and An & White (2020) for use in inference judgment tasks. In our variant of this method, we provide participants with *templatic items* consisting of a *templatic antecedent*, as in (12), and a *templatic consequent*, as in (13), and ask them to judge the likelihood that the consequent is true given the antecedent.

(12) a. A boasted to B that C happened.
b. A doubted that C happened.

(13) a. A believed that C happened.
b. A wanted C to have happened.

To demonstrate that this technique provides the same information about the inferential properties of the lexical items in the antecedents as a more standard approach with adjustment for world knowledge, we compare participants' judgments to these templatic items against their judgments to *contentful items*, consisting of a *contentful antecedent* (14) and *contentful consequent* (15).

(14) The publisher doubted that the author had finished her manuscript on time.

(15) a. The publisher believed that the author had finished her manuscript on time.
b. The publisher wanted the author to finish her manuscript on time.

To make the comparison fair, we explicitly adjust for world knowledge about the preferences and beliefs using norming data collected about the complements of the contentful consequents. Specifically, the norming study asks participants about the *a priori* likelihood that the contentful subject or object would believe or desire the content of the embedded clause, as in (16).

(16) a. What proportion of publishers generally believe authors finish their manuscripts on time?
b. What proportion of publishers generally want authors to finish their manuscripts on time?

3.1 Predicate Selection and Item Construction

We select 24 predicates for the validation study (see Figure 2 for a complete list). We choose the predicates such that half (12) are prototypically cognitive—e.g. *think*, *know*, *doubt*— and intransitive, and half (12) are prototypically communicative—e.g. *tell*, *notify*, *lie*—and either

transitive or PP-taking. For half (six) of the communicative predicates, we target belief and desire inferences about the referent of the predicate's subject—e.g. based on our own judgments, *lie* triggers the inference that the liar does not believe the content of the lie and *complain* triggers the inference that the complainer does not want the content of the complaint to be true. For the other half (six) of the communicative predicates, we target belief and desire inferences about the referent of the predicate's direct/prepositional object—e.g. based on our own judgments, *convince* triggers the inference that the convincee believes the content of the convincing (after the event) and *congratulate* triggers the inference that the congratulatee wants the content of the congratulations to be true. Finally, in an attempt to avoid inducing response biases in participants, we aim to ensure that the average judgment across pairings of a predicate, a target (subject or object), and a consequent verb (*believe* or *want*) is roughly neutral, balanced between extreme positive inferences, extreme negative inferences, and neutral inferences. Our own judgments, on which the selection criteria above are based, can be found in Appendix A.4.

For each predicate, we construct four *scenarios* that pair an entity (under a description) with a propositional content. We construct two of these scenarios to involve what we judged to have strong *positive valence*—e.g. publishers generally want authors to finish their manuscripts on time—and half involve what we judged to have strong *negative valence*—e.g. CFOs generally don't want their companies to lose money. Our aim here is to introduce high variability in the prior probability of the belief and desire inferences so that any effect of that prior probability will be revealed in the judgments, thereby allowing us to estimate and adjust for its effect.

For each such scenario we instantiate four contentful antecedents, differing only in the polarity of the matrix clause and the tense/modality of the embedded clause: one with the past or past perfect and another with a past future modal. For instance, (14) gives the past variant of one scenario for *doubt*, (17a) gives the negated variant, and (17b) gives the future variant.

(17) a. The publisher didn't doubt that the author had finished her manuscript on time.
b. The publisher doubted that the author would finish her manuscript on time.

This method yields 384 total contentful antecedents = 24 predicates × 4 scenarios per predicate × 2 matrix polarities × 2 embedded tenses. Each contentful antecedent is associated with two possible contentful consequents—e.g. those in (15)—yielding 768 total contentful items = 384 contentful antecedents × 2 contentful consequents.

3.2 Norming Study

The goal of our norming study is to obtain a measurement of the prior probability that each contentful consequent is true, irrespective of its corresponding contentful antecedent.

3.2.1 Methodology

Participants are asked to judge what proportion of entities belonging to some group generally believe or desire something by using a bounded slider, where the leftmost end is marked by *none* (0%) and the rightmost end by *all* (100%). The task instructions can be found in Appendix A.1.

3.2.2 Materials

For each predicate and scenario, we construct generic variants of the corresponding contentful consequents, wherein the matrix subject is a bare plural and the embedded clause expresses a generic proposition—e.g. (16) gives the generic variants corresponding to (15). We do not manipulate polarity or tense for these generic statements, and thus we obtain 192 norming items = 24 predicates $\times$ 4 scenarios per predicate $\times$ 2 contentful consequents. We divide these norming items into four lists of 48, constructed such that (i) half of the questions are about belief and half about desire; (ii) the content of half of the sentences has positive valence and half has negative valence; (iii) the same sentence does not appear twice in a list; and (iv) for each sentence, a list contains *either* the belief question *or* the desire question, but not both.

3.2.3 Participants

We recruited 100 participants on Amazon Mechanical Turk, with 25 distinct participants responding to each list. Each participant was required to pass the qualification test described in White, Hacquard & Lidz 2018a to ensure they were a native speaker of English.

3.2.4 Results

Figure 1 shows the mean judgments from the norming study for each item, with color corresponding to the intended valence for the desire inferences. As intended from the way we constructed the scenarios, the means for the *want* items are bimodally distributed, and thus the items satisfy our aim of high variability in the judgments across those items. The means for the belief inferences are unimodally distributed—likely because we did not explicitly manipulate the prior probability of the belief inferences—but similarly show the desired high variability.

3.2.5 Normalization

To obtain a single score for each item that adjusts for annotator differences, we fit a mixed effects beta regression to the responses, with fixed effects for VALENCE (*positive*, *negative*), CONSEQUENT VERB (*believe*, *want*), and their interaction, by-scenario random intercepts and random slopes for CONSEQUENT VERB, and by-participant random intercepts and random slopes for VALENCE, CONSEQUENT VERB, and their interaction. We then use this model to predict a judgment for each item, setting the participant random effects to 0 and the predicate random effects to their best linear unbiased predictors. These predictions can be thought of as those that the "average" participant would give.

3.3 Validation Study

The aim of the validation study is to ensure that gathering data using templatic items yields the same information about the inferential properties of the lexical items in the antecedents as an approach that uses contentful items and adjusts for world knowledge.

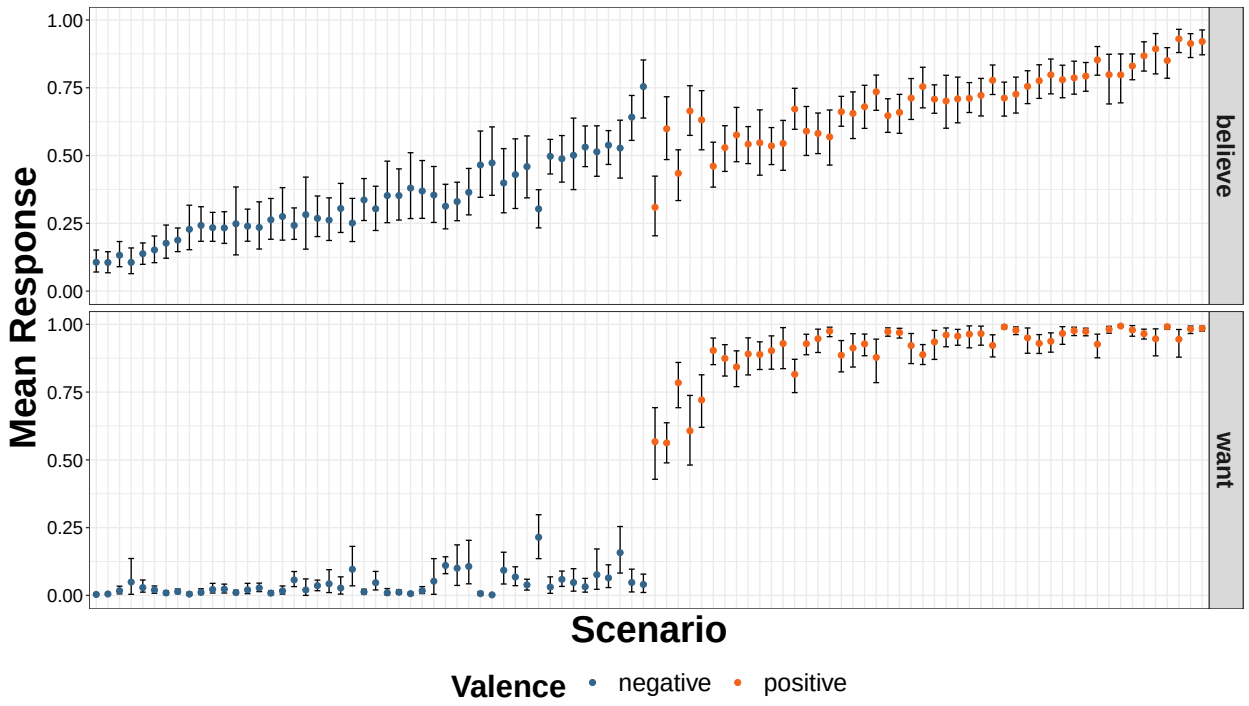

Figure 1 Mean judgments from the norming study.

3.3.1 Methodology

Participants are presented with either two sentences or two sentence templates and asked about the likelihood that the second sentence (the *consequent*) is true, assuming that the first sentence (the *antecedent*) is true. They respond using a bounded slider marked *extremely unlikely* on the left and *extremely likely* on the right. The instructions can be found in Appendix A.

3.3.2 Materials

Contentful items were constructed in the way described in §3.1. These items as well as the templatic items were organized into lists of 48, each containing only contentful or only templatic items. For both templatic and contentful items, we ensure that each list contains (i) half positive and half negative polarity items; (ii) half past and half future items; and (iii) half *believe* and half *want* items. For the contentful items, we additionally ensure that the lists contain half positive valence and half negative valence items.

3.3.3 Participants

We recruited 320 participants on Amazon Mechanical Turk, with 10 distinct participants responding to each list. The same qualification test as in the norming study was used.

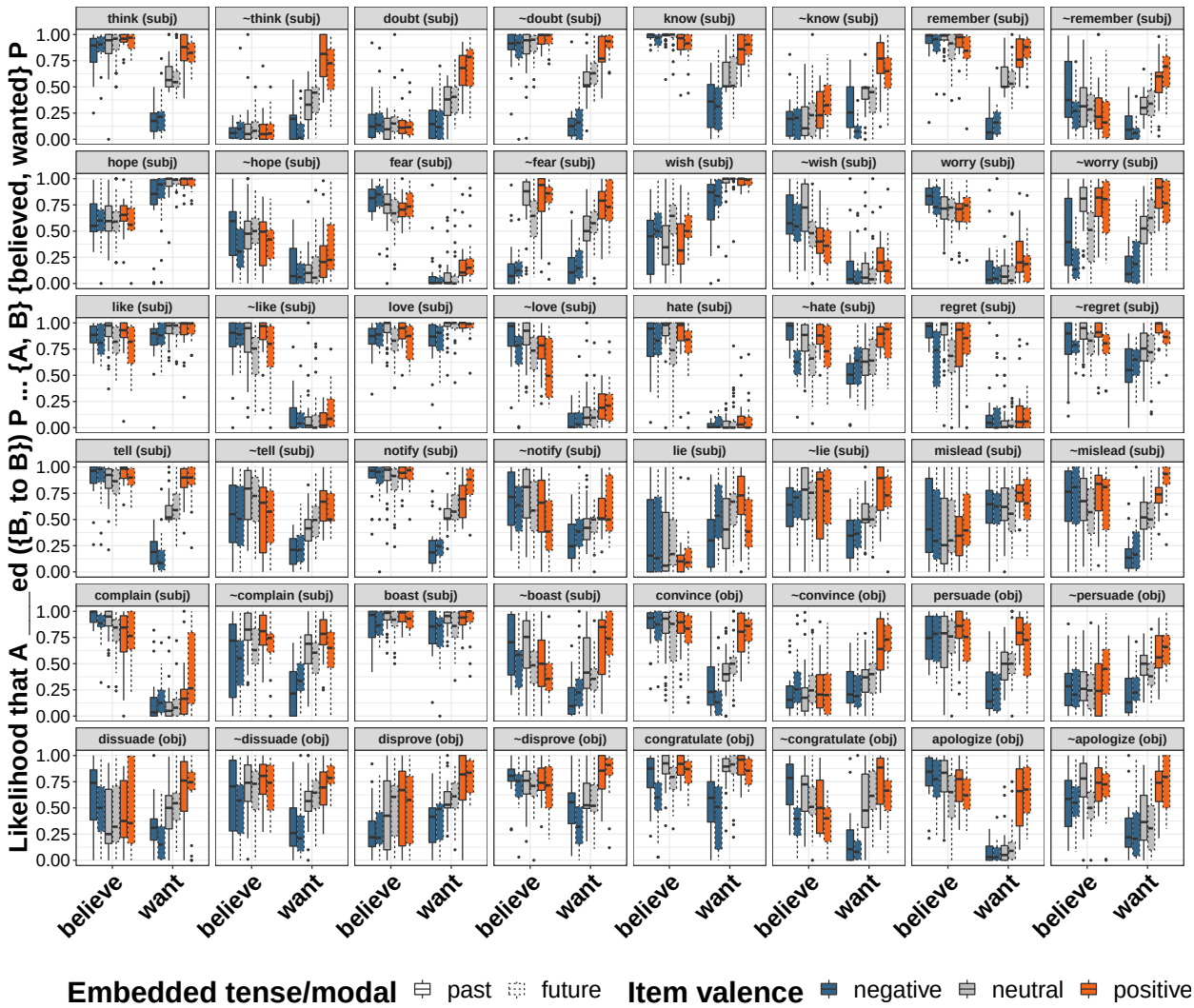

Figure 2 Distributions of inference judgments for all 24 predicates in the validation study, with and without negation on the predicate and for both past and future tense. Distributions for templatic judgments are shown in gray, and distributions for contentful items with predicted positive and negative inferences are shown in orange and blue, respectively.

3.3.4 Results

Figure 2 shows the distribution of judgments for each predicate. The positive and negative valence items (in orange and blue, respectively) are contentful and the neutral valence items (in gray) are templatic. To quantitatively assess how well responses to templatic items capture the same information about the inferential properties of the lexical items as an approach that uses contentful items and adjusts for world knowledge, we assess how well the judgments for contentful items can be predicted given just the normalized scores from the norming experiment and the judgments for the templatic items.

We first construct a normalized score for each templatic item in the same way as for the norming items. We fit a mixed effects beta regression to the templatic responses, with fixed effects for POLARITY (*positive*, *negative*), TENSE (*past*, *future*), CONSEQUENT VERB (*believe*, *want*), TARGET (*intransitive subject*, *transitive subject*, *transitive object*), and all possible interactions; by-predicate random intercepts and random slopes for POLARITY, TENSE, CONSEQUENT VERB, and all possible interactions; and by-participant random intercepts and random slopes for POLARITY, TENSE, CONSEQUENT VERB, TARGET, and all possible interactions. We then use this model to predict a judgment for each item, setting the participant random effects to 0 and the predicate random effects to their best linear unbiased predictors. As in the norming study, these predictions can be thought of as those that the "average" participant would give.

Next, we combine both the normalized norming scores (NORM) and the normalized templatic scores (TEMPLATIC) to ask the following: if we knew only the prior probability of a particular scenario and the normalized response to some templatic items without knowing the identity of the predicate itself, how well could we predict the contentful responses. We answer this queston using cross-validation: for each predicate, we remove the responses corresponding to that predicate, fit a model predicting the contentful responses to the remaining predicates given NORM and TEMPLATIC, then ask how well that model predicts the predicate it wasn't trained on. We fit a mixed effects beta regression, with fixed effects for NORM, TEMPLATIC, and their interaction; by-predicate random intercepts and random slopes for NORM, TEMPLATIC, and their interaction; and by-participant random intercepts and random slopes for NORM, TEMPLATIC, and their interaction. To measure the quality of the prediction of the held-out data, we compute the correlation between the predicted values and the true values.

We obtain a mean correlation across predicates of 0.70 (95% bootstrapped CI=[0.65, 0.75]). This is significantly better than the mean correlation between judgments given by participants who did the same contentful list (ρ=0.58, 95% bootstrapped CI = [0.57, 0.59]), suggesting that the templatic items correctly capture the average inferences triggered by a particular predicate.

4 Lexicon-scale data collection

We extend the methods in the validation experiment to a lexicon-scale experiment, gathering judgments for 725 predicates from the MegaAcceptability dataset (White & Rawlins 2016, 2020).

4.1 Methodology

Participants are presented with two templatic sentences and asked about the likelihood that the consequent is true if the antecedent is true. As in the validation study, participants respond using a bounded slider with *extremely unlikely* on the left and *extremely likely* on the right. The instructions that participants received can be found in Appendix B.1.

4.2 Materials

We select 725 unique predicates for use in this experiment based on their normalized acceptability score in the MegaAcceptability dataset.[1] We focus on the 12 frames found in Table 1, thereby manipulating (i) embedded tense/modality; (ii) the presence of a direct object (DO) or *to*-PP; and (iii) whether the matrix clause is passivized or not (in order to naturally capture predicates that take expletive subjects and direct objects, such as *amaze*, *surprise*, etc.). We manipulate tense/modality of the embedded clause—past (18a), future (18b), and tenseless (18c)—to ensure good coverage of future-oriented predicates, like *hope*, and preferential predicates, like *demand*; and we manipulate DO/PP-taking to ensure good coverage of communicative predicates, like *say*.

(18) a. A knew that C happened.
b. A hoped that C would happen.
c. A demanded that C happen.
d. A said to B that C happened.

In all of these frames the matrix predicate is in the simple past (for the active frames) or past participial form (for the passive frames). For each frame except the embedded tenseless ones, we select predicates with a normalized acceptability score in that frame of ≥0.2. For embedded tenseless frames, we set the threshold at 1.5. This threshold was determined by manual inspection of the least acceptable items that would be included. The 0.2 threshold roughly corresponds to an average rating of approximately 4.5 on the original ordinal scale (1-7), and the 1.5 threshold corresponds to an average rating of approximately 6.[2] The number of predicates that lie above this threshold for each frame can be found in Table 1.

As in the validation study, we manipulate the polarity of the matrix clause in templatic antecedents and construct templatic consequents for each antecedent conditioned on the tense/modality of the embedded clause—(19) for antecedents with embedded past tense and (20) for antecedents with tenseless or future embedded clauses.

(19) a. A believed that C happened.
b. A wanted C to have happened.

(20) a. A believed that C would happen.
b. A wanted C to happen.

We additionally construct two sets of templatic consequents for each antecedent containing a DO/PP.

(21) a. {A, B} believed that C happened.
b. {A, B} wanted C to have happened.

We sort the resulting items into lists of 32. We aim to constrain the construction of these lists similarly to how we constrain the lists in the validation study—attempting to obtain a distribution over expected slider responses for each list with a mean of approximately 0.5 and with high variance. The idea here is to avoid introducing "warping" into the participants' use of the response scale due to the underlying distribution of inferences associated with items in a particular list.

1 These normalized scores are described in White & Rawlins 2020 and are available at megaattitude.io.

2 We use a distinct threshold for the embedded tenseless frames because we found that the acceptability scores are significantly noisier for them than in other frames, resulting in many unnatural tenseless items being included when the threshold is set to 0.2. We believe this noise may be due to some MegaAcceptability participants missing the subtle difference between the embedded simple past and tenseless items—namely, the presence or lack of an *-ed* suffix.

Embedded Past		Embedded Future		Embedded Tenseless	
A ___ that C happened	534	A ___ that C would happen	498	A ___ that C happen	50
A ___ to B that C happened	156	A ___ to B that C would happen	160	A ___ to B that C happen	7
A was ___ that C happened	238	A was ___ that C would happen	213	A was ___ that C happen	24
A ___ B that C happened	33	A ___ B that C would happen	31	A ___ B that C happen	1

Table 1 Counts of unique predicates for each frame in the lexical scale experiment.

An obvious difficulty in achieving this goal is that we do not have access to the underlying distribution of inferences. Rather, this is exactly what we aim to measure, and so we must use proxy measures that are plausibly correlated. We use four such measures: the normalized veridicality and neg-raising scores from MegaVeridicality and MegaNegRaising, respectively; the normalized acceptability judgments from MegaAcceptability; and the frequency counts for the predicate in a particular item's templatic antecedent given by SUBTLEX (Brysbaert & New 2009).[3]

We perform PCA on these scores and then bin items based on their score on each component. This binning was done sequentially: we first derive two bins based on a median split of scores on the first component; then for each of those bins, we derive two bins based on a median split of scores on the second component; continuing similarly for the remaining two components. This procedure results in 16 equally-sized bins. We construct lists such that every combination of PCA bin and consequent verb appear exactly once in each list. To fill a list with items, we choose frames and antecedent verbs proportionally to their frequency in that respective PCA bin, also enforcing a hard constraint that each antecedent verb appears no more than once in a list. For transitive frames, the subject of the consequent is toggled each time an item with a DO or PP in the templatic antecedent is added to a list, ensuring that we also obtain a balance of subject and object targets among the DO/PP frames in each list.

Finally, we add four sanity check questions to each list to verify participant reliability. These items are constructed in pairs, with one item in the pair having a clear-cut 0 response and the other having a clear-cut 1 response. Each item in the pair uses the same verb in both the antecedent and the consequent, with one item having a negated antecedent (creating a contradiction) and the other having a positive antecedent (creating a tautology). All such items use the A ___ that C happened frame, and we only use predicates with a very high acceptability in this frame (≥3).

4.3 Participants

We recruited 272 native American English speakers on Amazon Mechanical Turk. Participants were allowed to respond to at most 20 lists, and each list was rated by 10 unique participants.

Figure 3 Distribution of verbs with respect to normalized belief and desire judgments.

4.4 Results

Figure 3 plots the normalized belief and desire judgments (for both subject target and object target, when applicable), with select predicates labelled. To obtain a single aggregate score for each item, we use a mixed effects beta model-based normalization procedure (see Appendix B.2 for details).

We examine the top two subplots—those showing judgments for items with embedded past tense—first. The top left subplot shows judgments for doxastic inferences. We observe that our results correctly capture a variety of commonly discussed predicates. For instance, cognitive predicates (such as *think* and *know*) and communicative predicates that trigger doxastic inferences about the recipient (such as *convince* and *persuade*) show up in the top left quadrant, indicating

3 We use the normalized veridicality and neg-raising scores described in White & Rawlins 2018 and An & White 2020, respectively. In cases where no veridicality or neg-raising score exists for a particular item, we randomly impute it.

doxastic components that are "foregrounded" and targeted by negation.

In contrast, emotive predicates such as *love*, *like*, and *hate* appear in the top right, indicating that the doxastic inferences are "backgrounded" and persist under negation. The center of the subplot shows predicates that don't trigger doxastic inferences, e.g., *wish* and *hope*. We also observe predicates that yield "backgrounded" negative doxastic inferences (deceitful communicatives, such as *lie* and *pretend*), and predicates that yield "foregrounded" negative doxastic inferences that are targeted by negation, such as *miss* and *doubt*.

The top right subplot shows judgments for bouletic inferences. As expected, we observe various positive emotive and preferential predicates such as *love*, *like*, *wish*, and *hope* in the top left quadrant, and negative emotive predicates such as *hate*, *regret*, *worry*, and *fear* in the bottom right quadrant. These indicate predicates that yield positive and negative bouletic inferences, respectively, that are "foregrounded" and targeted by negation. Many predicates are clustered around the center, indicating weak positive or negative bouletic inferences (such as *pretend* and *doubt*, respectively), or a lack of a bouletic component, as in the case of most cognitive predicates (e.g., *know* and *think*) and communicatives (e.g., *tell* and *say*). Notably, the overall pattern observed in this subplot shows that no predicate has a positive bouletic inference when under both positive and negative matrix polarity. This attests to the hypothesis that the bouletic component is always at issue and targeted by negation, if present at all (Anand & Hacquard 2014).

The bottom two subplots show judgments for the same items with embedded future tense (when applicable). For the bouletic inferences for these items (bottom right subplot), the judgments we obtain are approximately the same as the corresponding items with embedded past tense. However, the doxastic inferences from these items (bottom left subplot) weaken significantly for any predicates where the doxastic inferences are backgrounded, relative to the embedded past tense items — e.g., *love* yields a strong doxastic inference with a past tense embedding, but only a middling doxastic inference with a future-oriented embedding.

5 Discovering inference patterns

Our multiview mixture model attempts to find a soft clustering of templatic antecedents—identified by the predicate-frame pair instantiated by that antecedent—based simultaneously on the veridicality judgments in MegaVeridicality, the neg-raising judgments in MegaNegRaising, and the belief and desire judgments in MegaIntensionality. Each cluster is characterized jointly by the predicate-frame pairs that fall into it and the prototypical pattern of veridicality, neg-raising, doxastic, and bouletic inferences associated with it. One way to think of what this model aims to do is to find dense areas of inference pattern space by looking for commonly occurring patterns. The most common inference patterns in each of these dense areas can be thought of as the prototypical pattern associated with the cluster. Predicate-frame pairs whose inference pattern does not exactly match a prototypical pattern are then associated with the inference pattern they are most similar to.

Because we don't know *a priori* how many inference patterns there are, it is important to determine a method for selecting the number we direct the model to find. Our main goal in designing this procedure is to ensure that the inference patterns we find are plausibly formally represented. Thus, to select this number, we choose the clustering that allows us to best predict

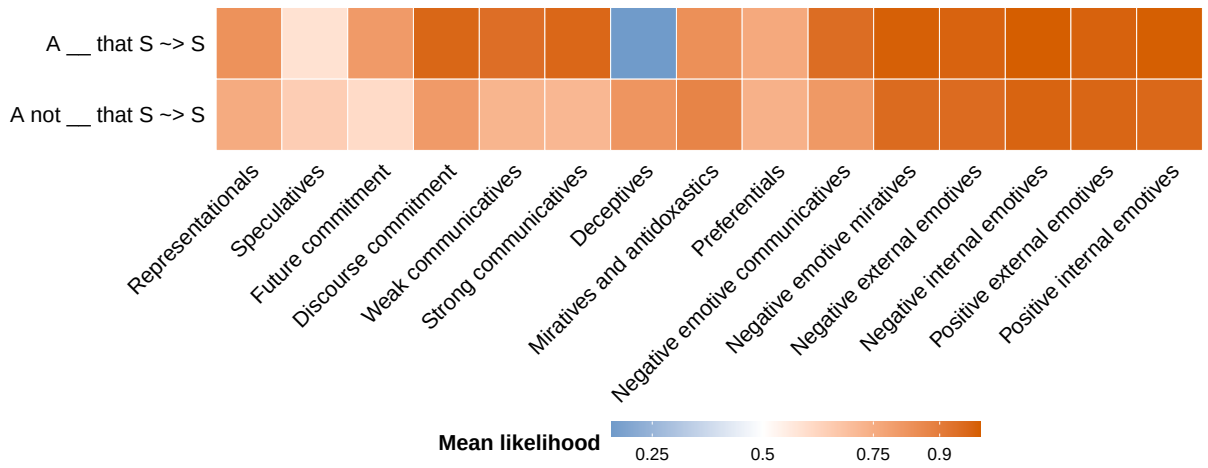

Figure 4 Prototypical veridicality inference patterns for each cluster.

predicates' morphosyntactic distribution as measured in the MegaAcceptability dataset.

5.1 Model description

We use a *multiview mixed effects mixture model* to perform soft clustering. This model is a *multivew* mixture model in that it combines multiple different *views* of the predicate frame pairs we are clustering within the same model. It is a *mixed effects* mixture model in the sense that, rather than simply associating each cluster of the mixture model with the parameters of a simple distribution—e.g. as in a Guassian mixture model—it associates each with the fixed effect parameters of a distinct mixed effects model for each task. For MegaVeridicality, which contains ordinal inference judgments, we use an ordinal mixed effects model; and for MegaNegRaising and MegaIntensionality, we use a beta mixed model.[4].

This mixture model defines a soft clustering similarly to how Latent Dirichlet Allocation associates *topic distributions* with documents (Blei, Ng & Jordan 2003): a categorical distribution over cluster assignments with parameter $\boldsymbol{\theta}_{\langle v,f\rangle}$ is associated with each predicate-frame pair $\langle v,f\rangle$. We then assume a cluster assignment $c_i \sim \text{Categorical}\left(\boldsymbol{\theta}_{\langle \text{verb}(i),\text{frame}(i)\rangle}\right)$ is sampled for each datapoint i such that the templatic antecedent associated with i instantiates $\langle v,f\rangle$. The response y_i is then sampled from some mixed effects model with fixed effects $\boldsymbol{\beta}_{\langle \text{view}(i),c_i\rangle}$, where view$(i)$ is the particular dataset (MegaVeridicality, MegNegRaising, or MegaIntensionality).

5.2 Selecting the optimal number inference patterns

To implement the selection procedure, we fit the model with different numbers of clusters $|C| \in \{2,...,20\}$. For each such clustering, we extract $\boldsymbol{\theta}_{\langle v,f\rangle}$ for each predicate-frame pair and then, for each predicate, we construct a vector $\boldsymbol{\phi}_v$ such that $\phi_{vk} = 1 - \prod_f 1 - \theta_{\langle v,f\rangle k}$. These vectors track the

4 See Appendix C.1 for the full model specification.

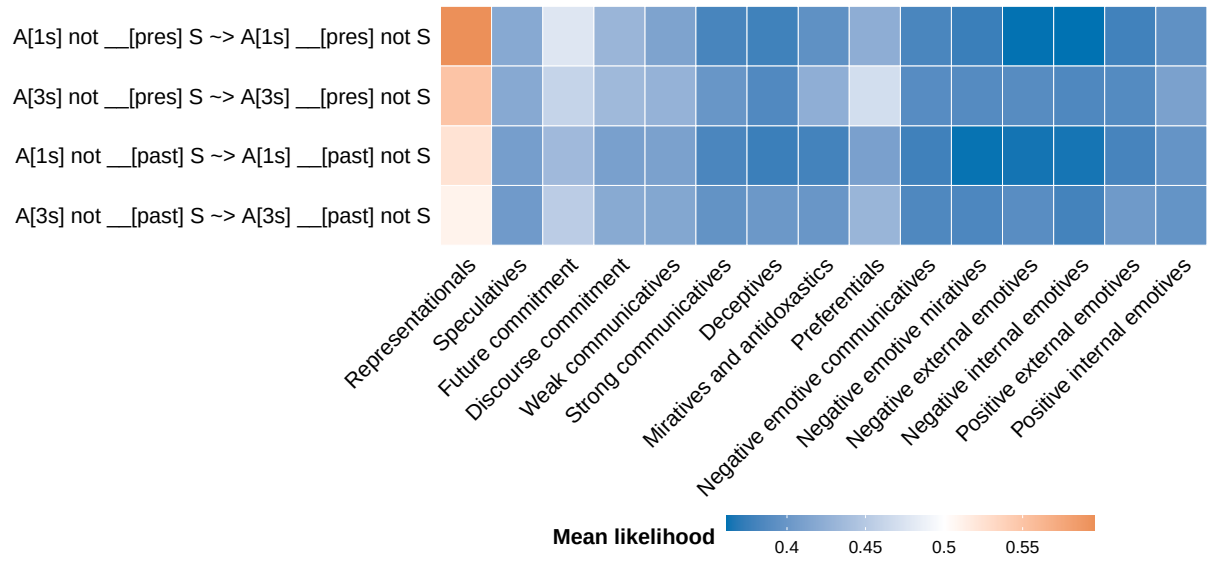

Figure 5 Prototypical neg-raising inference patterns for each cluster.

probability that a particular predicate v falls into a particular cluster k in any frame it occurs in.

Next, we regress the normalized acceptability for each predicate v in MegaAcceptability on these $\boldsymbol{\phi}_v$s using multivariate ridge regression, with the regularization parameter selected by 5-fold cross-validation. To evaluate the performance of this regression on data the model hasn't seen before, we use a separate ("outer") 10-fold cross-validation, breaking the data into 10 equally sized bins, then for each bin, training on all the other bins and testing on that bin. On the held-out data, we compute per-item squared error for each model and compute Bonferroni-corrected 95% confidence intervals for the mean difference in squared error between each pairing of models via non-parametric bootstrap over items. We then choose $|C_{\text{best}}|$ as the minimal number of clusters such that no model with a greater number of clusters performs reliably better.[5]

5.3 Results

Using the approach described in §5.2, we find $|C| = 15$ to be optimal. This model obtains $R^2 = 0.34$ (95% CI = [0.32, 0.35]) on the acceptability judgments, which is (perhaps surprisingly) superior to the corpus subcategorization frame frequency-based models described in White & Rawlins 2020. Figures 4–6 show the prototypical inference patterns for veridicality inferences, negation-raising inferences, and doxastic and bouletic inferences, respectively. These prototypical inference patterns are computed from the fixed effects for each cluster and view: for the views with unit data (MegaNegRaising and MegaIntensionality), we compute the expected unit response, setting by-participant random effects to 0; and for the ordinal data, we compute the expect distribution over ordinal values using the average by-participant cutpoints then computing a weighted mean based on

5 For reasons of space, we do not discuss the relationship between clusters and the syntax here. For the interested reader, the coefficients for the best model are plotted in Appendix C.2.

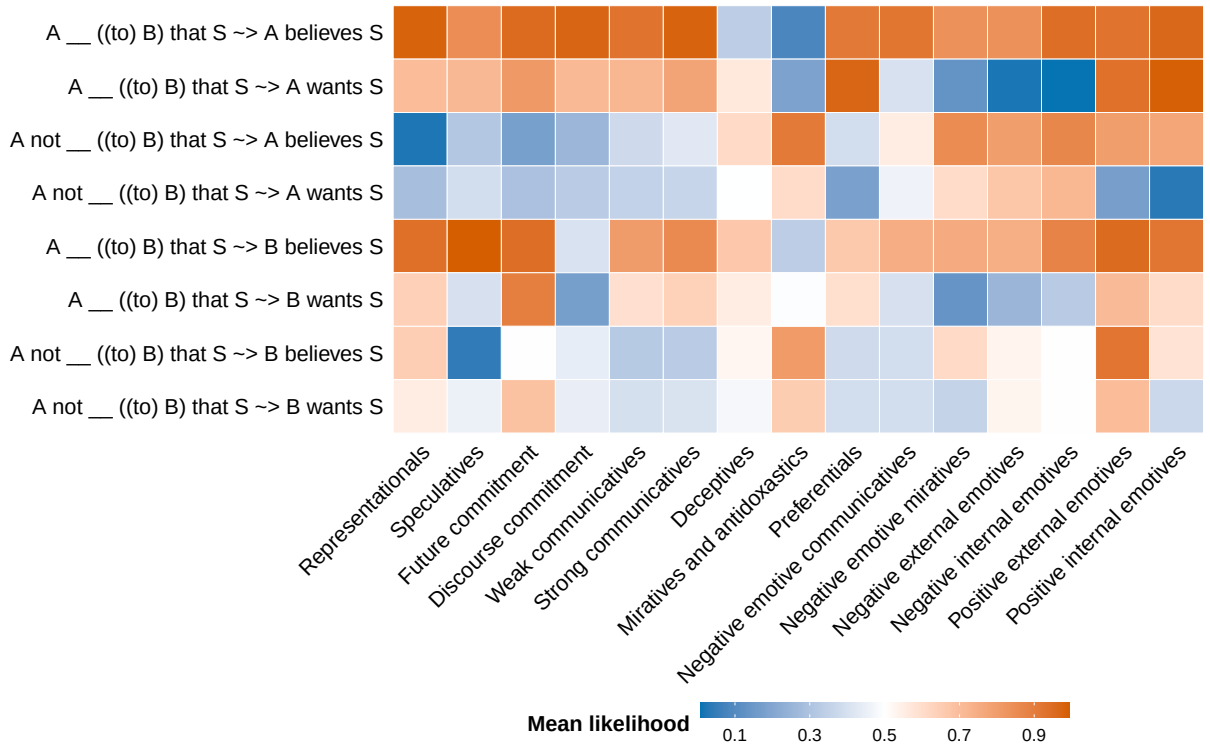

Figure 6 Prototypical doxastic and bouletic inference patterns for each cluster.

those distributions, assigning *yes* to 1, *maybe or maybe not* to 0.5, and *no* to 0.

To facilitate analysis of the clusters, we assign labels to them based on (i) our interpretations of the prototypical inference pattern for that cluster and (ii) the predicate-frame pairs with the highest weight for that cluster. Importantly, these labels are intended to reflect the prototypical characteristics of the predicates associated with a particular cluster: as noted above, not all predicates in the cluster will share those prototypical characteristics. (22)–(35) give examples for each cluster.

(22) *Representiationals:* doxastic mental states and mental processes.
A {thought, believed, suspected} that C happened.

(23) *Preferentials:* expressions of preference for particular (future) states of affairs.
A {hoped, wished, demanded, recommended} that C (would) happen.

(24) *Speculatives:* communication of uncertain beliefs.
A {ventured, guessed, gossiped} that C happened.

(25) *Future commitment:* expressions of commitment to some future action or result.
A {promised, ensured, attested} that C would happen.

(26) *Negative internal emotives:* negative emotional states.
A was {frightened, disgusted, infuriated} that C happened.

(27) *Negative emotive communicatives:* communicative acts with broadly negative valence.

A {screamed, ranted, growled} to B that C would happen.

(28) *Strong communicatives:* communicative acts with strong doxastic inferences about the source.
A {confessed, admitted, acknowledged} that C happened.

(29) *Weak communicatives:* communicative acts with weak doxastic inferences about the source.[6]
A {reported, remarked, yelped} to B that C happened.

(30) *Deceptives:* actions involving dishonesty, deceit, or pretense.
A {lied, misled, faked, fabricated} ((to) B) that C happened.

(31) *Discourse commitment:* communicative acts committing the source to the content's truth.
A {maintained, remarked, swore} that C would happen.

(32) *Negative external emotives:* expressions of negative emotion with behavioral correlates.
A {whined, whimpered, pouted} to B that C would happen.

(33) *Positive external emotives:* expressions of positive emotion with (external) behavioral correlates.
A was {congratulated, praised, fascinated} that C happened.

(34) *Positive internal emotives:* positive emotional states.
A was {pleased, thrilled, enthused} that C happened.

(35) *Negative emotive miratives:* expressions of surprise with negative valence.
A was {dazed, flustered, alarmed} that C would happen.

We see that factivity—both cells being dark orange for a particular class in Figure 4—is largely confined to emotive classes, with cognitive and non-emotive communicatives generally being veridical—the top cell only being dark orange in Figure 4—nonveridical—both cells being light orange or white—or *antiimplicative*—the top cell being blue and the bottom cell being orange. This finding is consistent with Anand & Hacquard's (2014) proposal that veridical communicatives are never factive.

Neg-raising is even more constrained than factivity, only showing up among the representationals, as indicated by that class being the only one with any orange in Figure 5. One might have expected at least preferentials to also allow neg-raising, since one of the most commonly discussed neg-raising predicates, *want*, would presumably fall into this class. Note, however, that because we focus only on predicates them embed *that*-clause complements, we likely have a biased sample of preferentials—indeed, one not including predicates like *want*. A wider sample, including infinitival-taking predicates might yield a preferential subclass that does allow neg-raising.

The inference patterns are substantially more varied among doxastic and bouletic inferences. Positive doxastic inferences that target the source of a communication (top row of Figure 6) are common to the majority of clusters, with only the deceptives and miratives/antidoxastics showing

6 The inference patterns for the strong and weak communicatives are extremely similar, with the "weakness" of the weak communicatives not being particularly weak. These classes also show high overlap in predicates—e.g. one class containing a predicate with an embedded clause in the past tense and the other containing the same predicate with an embedded clause containing a future modal. Nonetheless, the syntactic profile of these classes does differ slightly as can be seen in the plot in Appendix C.2, and so some distinction does seem to exist between these clusters.

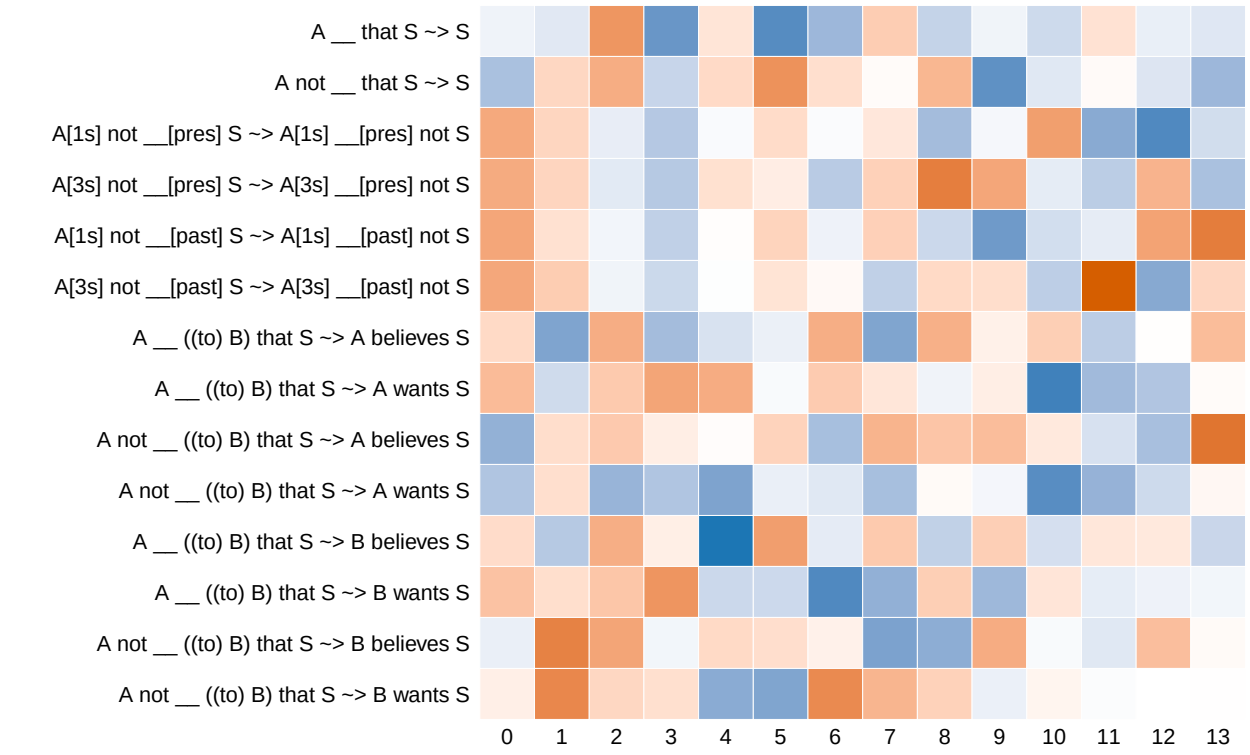

Figure 7 PCA loadings for each inference type.

mean likelihoods below 0.5 for that inference type. Positive doxastic inferences that target the *recipient* of a communication (fifth row from the top) are similarly widely distributed, but are most highly concentrated among the speculatives.

Unsurprisingly, positive bouletic inferences are likeliest among the preferentials and the positive internal and external emotives, and negative bouletic inferences among the negative internal and external emotives. Moreover, all four of these clusters capture (to varying degrees) doxastic presuppositions, but none of them—and indeed, no other clusters besides—show evidence of bouletic presuppositions. This is consistent with the observation that negation preferentially targets the emotive component of a predicate when it is present (Anand & Hacquard 2014).

5.4 Extracting generalizations about the attested inference patterns

To determine a set of generalizations about these clusters, we apply principal component analysis (PCA) to the parameters of the property-specific probability distributions for each cluster. The idea behind this method is to decompose the inference patterns associated with each cluster into *components* capturing which inferences are correlated across clusters. These components can be thought of as capturing violable biconditional statements that are strictly ordered in terms of importance, with the most important components capturing the most variability in the patterns.

The PCA loadings across properties for each cluster are shown in Figure 7, with orange tiles and

blue tiles corresponding to positive and negative loadings, respectively. The fact that the loadings can take on both positive and negative values suggests a particular basis for stating generalizations about inference patterns—namely, as (anti-)correlations between inference types that lie along continua defined by the range of their loadings, with strongly positive inferences at one end and strongly negative inferences at the other. Such a framing thus enables us to define the different inference types in terms of these loadings. For instance, if a component has high positive or high negative loadings for both the A ___ that S ⇝ S and A not ___ that S ⇝ S inferences, we interpret that to be indicative of factivity or anti-factivity, since it suggests that the inference is robust under negation.[7] By contrast, if a component has high positive or high negative loading for only one of those inferences and a near-zero loading for the other, we interpret that as evidence of veridicality or anti-veridicality, as the inference is *not* robust under negation. Finally, if the loading is strongly positive for one of these inferences and strongly negative for the other, we interpret this as (anti-)implicativity. Our interpretations of neg-raising, doxastic, and bouletic inference patterns follow the same logic.

We give our interpretation for the first six components below.

(36) The polarity of a veridicality presupposition is anti-correlated with neg-raising.

(37) The polarity of a belief presupposition about a recipient is correlated with the polarity of a desire presupposition.

(38) The valence of an emotive communicative is anticorrelated with veridicality.

(39) Bouletic inferences about the source and the target of a communication are anticorrelated with veridicality.

(40) Desire inferences about the source in a communication are anticorrelated with belief inferences about the target.

(41) Veridicality is correlated with belief inferences in the target of a communication but anticorrelated with desire inferences.

Naturally, these generalizations demand theoretically-motivated explanations. However, our aim here is merely to identify the patterns that emerge when considering the interaction between different inference types at the scale of the lexicon, and we reserve the project of presenting an account of those patterns for future work.

6 Conclusion

We have investigated which patterns of lexically triggered inferences are attested or unattested across clause-embedding verbs in English. Our investigation has taken a more holistic view of the problem than has previously been attempted by simultaneously considering doxastic, bouletic, neg(ation)-raising, and veridicality inferences. In particular, we have identified 15 inference patterns that correlate with predicates' morphosyntactic distribution. Many of these patterns corroborate past observations from the literature—e.g. a tendency for emotive predicates to be factive—while

7 More generally, we treat robustness under negation for any inference type as evidence of presupposition for that type.

others point to novel ones—e.g. a correlation between mirativity and antiveridicality—that are less commonly discussed. These findings lay the groundwork for further work investigating fine-grained aspects of the relationship between these inference patterns and the syntax as well as expansion of our inference pattern inventory beyond predicates that embed finite clauses.

References

An, Hannah & Aaron White. 2020. The lexical and grammatical sources of neg-raising inferences. *Proceedings of the Society for Computation in Linguistics* 3(1). 220–233. doi:https://doi.org/10.7275/yts0-q989. https://scholarworks.umass.edu/scil/vol3/iss1/23.

Anand, Pranav & Valentine Hacquard. 2013. Epistemics and attitudes. *Semantics and Pragmatics* 6(8). 1–59.

Anand, Pranav & Valentine Hacquard. 2014. Factivity, belief and discourse. In Luka Crnič & Uli Sauerland (eds.), *The Art and Craft of Semantics: A Festschrift for Irene Heim*, vol. 1, 69–90. Cambridge, MA: MIT Working Papers in Linguistics.

Barwise, Jon & Robin Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy* 4(2). 159–219. doi:10.1007/BF00350139. https://doi.org/10.1007/BF00350139.

Berman, S.R. 1991. *On the semantics and logical form of wh-clauses*. Amherst, MA: University of Massachusetts at Amherst PhD dissertation.

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3. 993–1022.

Bolinger, Dwight. 1968. Postposed main phrases: An English rule for the Romance subjunctive. *Canadian Journal of Linguistics* 14(1). 3–30.

Brysbaert, Marc & Boris New. 2009. Moving beyond Kučera and Francis. *Behavior research methods* 41(4). 977–990.

Egré, Paul. 2008. Question-embedding and factivity. *Grazer Philosophische Studien* 77(1). 85–125.

Elliott, Patrick D, Nathan Klinedinst, Yasutada Sudo & Wataru Uegaki. 2017. Predicates of relevance and theories of question embedding. *Journal of Semantics* 34(3). 547–554.

Farkas, Donka. 1985. *Intensional Descriptions and the Romance Subjunctive Mood*. New York: Garland Publishing.

George, Benjamin Ross. 2011. *Question Embedding and the Semantics of Answers*: University of California Los Angeles PhD dissertation.

Giannakidou, Anastasia. 1997. *The Landscape of Polarity Items*: University of Groningen PhD dissertation.

Giannakidou, Anastasia & Alda Mari. 2021. *Truth and Veridicality in Grammar and Thought*. Chicago: University of Chicago Press.

Ginzburg, Jonathan. 1995. Resolving questions, II. *Linguistics and Philosophy* 18(6). 567–609.

Giorgi, Alessandra & Fabio Pianesi. 1997. *Tense and Aspect: Form Semantics to Morphosyntax*. Oxford: Oxford University Press.

Heim, Irene. 1992. Presupposition projection and the semantics of attitude verbs. *Journal of Semantics* 9(3). 183–221.

Hintikka, Jaakko. 1975. Different Constructions in Terms of the Basic Epistemological Verbs: A

Survey of Some Problems and Proposals. In *The Intentions of Intentionality and Other New Models for Modalities*, 1–25. Dordrecht: D. Reidel.

Hooper, Joan B. 1975. On assertive predicates. In John P. Kimball (ed.), *Syntax and Semantics*, vol. 4, 91–124. New York: Academy Press.

Horn, Laurence. 1972. *On the Semantic Properties of Logical Operators in English*: UCLA PhD dissertation.

Karttunen, Lauri. 1977. Syntax and semantics of questions. *Linguistics and Philosophy* 1(1). 3–44.

Lahiri, Utpal. 2002. *Questions and Answers in Embedded Contexts*. Oxford University Press.

Levin, Beth & Malka Rappaport Hovav. 1991. Wiping the slate clean: A lexical semantic exploration. *Cognition* 41(1-3). 123–151. doi:10.1016/0010-0277(91)90034-2. http://www.sciencedirect.com/science/article/pii/0010027791900342. Publisher: Elsevier.

Portner, Paul. 1992. *Situation Theory and the Semantics of Propositional Expressions*: University of Massachusetts, Amherst PhD dissertation.

Quer, Josep. 1998. *Mood at the Interface*: Utrecht Institute of Linguistics, OTS PhD dissertation.

Roberts, Tom. 2019. I can't believe it's not lexical: Deriving distributed veridicality. *Semantics and Linguistic Theory* 29(0). 665–685. doi:10.3765/salt.v29i0.4634. https://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/29.665. Number: 0.

Schütze, Carson T. & Jon Sprouse. 2014. Judgment data. In Robert J. Podesva & Devyani Sharma (eds.), *Research Methods in Linguistics*, 27–50. Cambridge University Press.

Theiler, Nadine, Floris Roelofsen & Maria Aloni. 2017. What's wrong with believing whether. *Semantics and Linguistic Theory* 27. 248–265.

Theiler, Nadine, Floris Roelofsen & Maria Aloni. 2019. Picky predicates: why believe doesn't like interrogative complements, and other puzzles. *Natural Language Semantics* .

Tonahuser, Judith & Judith Degen. under review. Are there factive predicates? an empirical investigation. Https://ling.auf.net/lingbuzz/005360.

Uegaki, Wataru. 2015. *Interpreting Questions under Attitudes*: Massachusetts Institute of Technology PhD dissertation.

Uegaki, Wataru & Yasutada Sudo. 2019. The *hope-wh puzzle. *Natural Language Semantics* 27(4). 323–356. doi:10.1007/s11050-019-09156-5. https://doi.org/10.1007/s11050-019-09156-5.

Villalta, Elisabeth. 2000. Spanish subjunctive clauses require ordered alternatives. *Semantics and Linguistic Theory* 10. 239–256.

Villalta, Elisabeth. 2008. Mood and gradability: an investigation of the subjunctive mood in Spanish. *Linguistics and Philosophy* 31(4). 467–522.

White, Aaron Steven, Valentine Hacquard & Jeffrey Lidz. 2018a. Semantic Information and the Syntax of Propositional Attitude Verbs. *Cognitive Science* 42(2). 416–456. doi:https://doi.org/10.1111/cogs.12512.

White, Aaron Steven & Kyle Rawlins. 2016. A computational model of S-selection. *Semantics and Linguistic Theory* 26. 641–663. doi:10.3765/salt.v26i0.3819. https://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/26.641. Number: 0.

White, Aaron Steven & Kyle Rawlins. 2018. The role of veridicality and factivity in clause selection. In Sherry Hucklebridge & Max Nelson (eds.), *Proceedings of the 48th Annual Meeting of the North East Linguistic Society*, 221–234. Amherst, MA: GLSA Publications.

White, Aaron Steven & Kyle Rawlins. 2020. Frequency, acceptability, and selection: A case study of clause-embedding. *Glossa: a journal of general linguistics* 5(1). 105. doi:10.5334/gjgl.1001. http://www.glossa-journal.org//article/10.5334/gjgl.1001/. Number: 1 Publisher: Ubiquity Press.

White, Aaron Steven, Rachel Rudinger, Kyle Rawlins & Benjamin Van Durme. 2018b. Lexicosyntactic Inference in Neural Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4717–4724. Brussels, Belgium: Association for Computational Linguistics. doi:10.18653/v1/D18-1501. https://www.aclweb.org/anthology/D18-1501.

Zuber, Richard. 1983. Semantic restrictions on certain complementizers. In S. Hattori & K. Inoue (eds.), *Proceedings of the 12th International Congress of Linguists, Tokyo*, 434–436.

Benjamin Kane
Department of Computer Science
University of Rochester
500 Joseph C. Wilson Blvd.
Rochester, NY, USA 14627
bkane2@cs.rochester.edu

Aaron Steven White
Department of Linguistics
University of Rochester
500 Joseph C. Wilson Blvd.
Rochester, NY, USA 14627
aaron.white@rochester.edu

William Gantt
Department of Computer Science
University of Rochester
500 Joseph C. Wilson Blvd.
Rochester, NY, USA 14627
wgantt@cs.rochester.edu

A Materials for Norming and Validation Studies

A.1 Norming instructions

In this experiment, you will be asked about the stereotypical beliefs and desires of particular groups of people. Specifically, we're interested in getting your best guess about what proportion of a certain group generally **believes** or **wants** particular things. Your task will be to respond using a slider that ranges from 0 (0 percent) on the left to 1 (100 percent) on the right.

Please carefully look over the examples below before beginning.

Believe Questions

Example 1 For the **believe** questions, you might get one like the following: *What proportion of dentists generally believe that patients should brush their teeth every day?* Dentists almost universally recommend daily (indeed, twice daily) teeth brushing, so your answer in this case should probably be close to the right end of the scale, around 1.

Example 2 If the question were instead *What proportion of dietitians believe that their clients should eat candy for every meal?*, your answer should be close to, if not exactly equal to, 0, since candy is generally assumed to be bad for one's health in large quantities.

Example 3 Many questions will not be as clear-cut as the two examples above. You may, for example, get a question like *What proportion of voters generally believe that their state government will raise taxes?* Any given voter's beliefs will depend on a variety of factors, including their politics, their state, and who the state officials are. It may thus be hard to know what's true of voters' beliefs in general. In this case, you should just answer as best you can.

Want Questions

Example 1 For **want** questions, you might get one like the following: *What proportion of pilots generally want their flights to be free of turbulence?*. Since very few, if any, pilots enjoy flying through turbulence your response should be very close to, if not equal to, 1.

Example 2 However, if the question were instead *What proportion of gardeners generally want weeds to overrun their gardens?*, your answer should be close to 0, since weeds overruning one's garden is generally undesirable.

Example 3 Similar to the **believe** questions, some of the **want** questions may be more difficult than the preceding examples. You may see a question like *What proportion of family members generally want to go to family reunions?* Naturally, this will depend on the family member and their relationship with their family. Once again, you should answer as best you can.

More Information

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

This research is being carried out by Dr. Aaron White at the University of Rochester. You can read more about the study here.

A.2 Contentful validation instructions

In this experiment, you will be given a statement and asked about the likelihood that a second statement is true, assuming that the first statement is true. Your task will be to respond using a slider with *extremely **un**likely* to the left and *extremely likely* to the right.

High Likelihood

For instance, you might get the statement **the teenager remembered to buy milk** and the question **How likely is it that the teenager bought milk?**. In this case, the second statement is very likely to be true assuming the first statement is true: if someone remembered to do something, that person has to have done that thing. So you would slide the slider fairly far to the right (toward *extremely likely*).

Low Likelihood

If the statement were **the teenager forgot to buy milk** and the question were the same, you would slide the slider fairly far to the left (toward *extremely **un**likely*).

Middling Likelihood

And if the statement were **the teenager wanted to buy milk** and the question were the same as before, you might leave the slider in the middle or slide it slightly toward the right (*extremely likely*) to signal that wanting to do something makes it more likely that someone will do it (all other things being equal).

Similarly, if the statement were **the teenager didn't want to buy milk** and the question were the same, you might leave the slider in the middle or slide it slightly toward the left (*extremely **un**likely*) to signal that not wanting to do something makes it less likely that someone will do it (all other things being equal).

Changes in Likelihood

If the first sentence describes a situation where something changes, the question should be answered assuming the change has happened. For instance, if the statement were **the teenager decided to buy milk** and the question were **How likely is it that the teenager intended to buy milk?**, the

second statement is very likely to be false before the decision, but true afterward, so you would slide the slider fairly far to the right (toward *extremely likely*).

More Information

Try to answer the questions as quickly and accurately as possible. Some of the statements may sound odd. In these cases, try your best to think about a plausible situation it might describe.

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

This research is being carried out by Dr. Aaron White at the University of Rochester. You can read more about the study here.

A.3 Templatized validation instructions

In this experiment, you will be given a statement and asked about the likelihood that a second statement is true, assuming that the first statement is true. Your task will be to respond using a slider with *extremely **un**likely* to the left and *extremely likely* to the right.

High Likelihood

For instance, you might get the statement **A remembered to do B** and the question **How likely is it that A did B?**, where **A** just stands for some person and **B** stands for some action. In this case, the second statement is very likely to be true assuming the first statement is true: if someone remembered to do something, that person has to have done that thing. So you would slide the slider fairly far to the right (toward *extremely likely*).

Low Likelihood

If the statement were **A forgot to do B** and the question were the same, you would slide the slider fairly far to the left (toward *extremely **un**likely*).

Middling Likelihood

And if the statement were **A wanted to do B** and the question were the same as before, you might leave the slider in the middle or slide it slightly toward the right (*extremely likely*) to signal that wanting to do something makes it more likely that someone will do it (all other things being equal).

Similarly, if the statement were **A *didn't* want to do B** and the question were the same, you might leave the slider in the middle or slide it slightly toward the left (*extremely **un**likely*) to signal that not wanting to do something makes it less likely that someone will do it (all other things being equal).

Changes in Likelihood

If the first sentence describes a situation where something changes, the question should be answered assuming the change has happened. For instance, if the statement were **A decided to do B** and the question were **How likely is it that A intended to do B?**, the second statement is very likely to be false before the decision (since the whole point of a decision is to form an intention), but true afterward, so you would slide the slider fairly far to the right (toward *extremely likely*).

More Information

Try to answer the questions as quickly and accurately as possible. Some of the statements may sound odd, even setting aside that we replace words for specific people and actions with letters (A, B, etc.). In these cases, try your best to think about the sort of situation the statements might describe.

A.4 Validation study verbs

A list of verbs used for the validation experiment is shown in Table 2.

B Materials for Lexicon-Scale Data Collection

B.1 Instructions

In this experiment, you will be given a statement and asked about the likelihood that a second statement is true, assuming that the first statement is true. Your task will be to respond using a slider with *extremely **un**likely* to the left and *extremely likely* to the right.

High Likelihood

For instance, you might get the statement **A remembered to do B** and the question **How likely is it that A did B?**, where **A** just stands for some person and **B** stands for some action. In this case, the second statement is very likely to be true assuming the first statement is true: if someone remembered to do something, that person has to have done that thing. So you would slide the slider fairly far to the right (toward *extremely likely*).

Low Likelihood

If the statement were **A forgot to do B** and the question were the same, you would slide the slider fairly far to the left (toward *extremely **un**likely*).

Middling Likelihood

And if the statement were **A wanted to do B** and the question were the same as before, you might leave the slider in the middle or slide it slightly toward the right (*extremely likely*) to signal that

wanting to do something makes it slightly more likely that someone will do it (all other things being equal).

Similarly, if the statement were **A *didn't* want to do B** and the question were the same, you might leave the slider in the middle or slide it slightly toward the left (*extremely **un**likely*) to signal that not wanting to do something makes it slightly less likely that someone will do it (all other things being equal).

Changes in Likelihood

If the first sentence describes a situation where something changes, the question should be answered assuming the change has happened. For instance, if the statement were **A decided to do B** and the question were **How likely is it that A intended to do B?**, the second statement is very likely to be false before the decision (since the whole point of a decision is to form an intention), but true afterward, so you would slide the slider fairly far to the right (toward *extremely likely*).

More Information

Try to answer the questions as quickly and accurately as possible. Some of the statements may sound odd, even setting aside that we replace words for specific people and actions with letters (A, B, etc.). In these cases, try your best to think about the sort of situation the statements might describe.

B.2 Normalization

Prior to normalization, the lexical-scale data had an inter-annotator agreement (Krippendorff's alpha) of 0.58. To aggregate annotator responses into a single score for each item, we fit a mixed effects Beta regression. This model incorporates a fixed scaling term β^{ν}, a fixed shifting term for each item β_i^{μ}, a random scaling term ρ_p^{ν} for each participant p, and a random shifting term ρ_p^{μ} for each participant.

$$\begin{aligned} \nu_i &= \left(\beta^{\nu} + \rho^{\nu}_{p_i}\right)^2 \\ \mu_i &= \text{logit}^{-1}\left(\beta_i^{\mu} + \rho^{\mu}_{p_i}\right) \\ a_i;\ b_i &= \nu_i\mu_i;\ \nu_i(1-\mu_i) \\ \hat{r}_i &\sim \text{Beta}(a_i, b_i) \end{aligned}$$

The normalizer optimizes these parameters against the loss function $\mathcal{L} = \sum_i D_{KL}(r_i \parallel \hat{r}_i)$. The normalized scores for each item i in the MegaIntensionality dataset (plotted in Figure 3) correspond to $\text{logit}^{-1}(\beta_i^{\mu})$.

C Model Details

C.1 Model Definition

The full model definition is shown below. We place a Dirichlet prior on the predicate-frame weight parameters, parameterized by a scaled mean $\alpha\boldsymbol{\gamma}$, where the mean is drawn from a uniform Dirichlet hyperprior. The scaling term controls the dispersion of the cluster weights, and is drawn from a zero-mean log-normal prior.

$$\begin{aligned}
\alpha &\sim \log\mathcal{N}\left(0, \sigma^2\right) \\
\boldsymbol{\gamma} &\sim \text{Dirichlet}\left(\mathbf{1}_K\right) \\
\boldsymbol{\theta}_{\langle v,f\rangle} &\sim \text{Dirichlet}\left(\alpha\boldsymbol{\gamma}\right) \\
c_i &\sim \text{Categorical}\left(\boldsymbol{\theta}_{\langle \text{verb}(i),\text{frame}(i)\rangle}\right) \\
y_i &\sim f_{\text{task}(i)}\left(\boldsymbol{\beta}_{\langle \text{task}(i),c_i\rangle} \cdot \mathbf{x}_i, \boldsymbol{\rho}_{\text{participant}(i)}\right)
\end{aligned}$$

The Beta mixed effects model for unit responses is shown below. β^μ and β^ν represent fixed effects: shifting terms for each cluster c_i and sentential context k_i and a scaling term, respectively. ρ^μ and ρ^ν represent shifting and scaling random effects for each participant. We place a zero-mean log-normal prior on each scaling term, and a zero-mean normal prior on each shifting term.

$$\begin{aligned}
\nu_i &= \beta^\nu + \rho^\nu_{p_i} \\
\mu_i &= \text{logit}^{-1}\left(\beta^\mu_{c_i,k_i} + \rho^\mu_{p_i}\right) \\
a_i;\ b_i &= \nu_i\mu_i;\ \nu_i(1-\mu_i) \\
y_i &\sim \text{Beta}(a_i, b_i)
\end{aligned}$$

The Ordinal mixed effects model assumes that each item with a particular cluster c_i and sentential context k_i maps to some real-valued score β_{c_i,k_i}, and that each participant p has a different way of binning these scores into ordinal rankings. These bins are defined by random effects $\boldsymbol{\rho}_p$ that represent cutpoints, such that the worst rating corresponds to bin $\left(-\infty, \rho_{p,1}\right]$, the best rating to bin $\left(\rho_{p,n}, \infty\right)$, and all other ratings to bins $\left(\rho_{p,n-1}, \rho_{p,n}\right)$.

The probability of a particular item i (with participant p_i, sentential context k_i, and assigned cluster c_i) getting ordinal score ℓ is defined based on these cutpoints:

$$\begin{aligned}
\mathbb{P}(y_i \le \ell) &= \text{logit}^{-1}(c_{p_i,\ell} - \beta_{c_i,k_i}) \\
\mathbb{P}(y_i = \ell) &= \mathbb{P}(y_i \le \ell) - \mathbb{P}(y_i \le (\ell - 1)) \\
y_i &\sim \text{Categorical}(\mathbb{P}(y_i))
\end{aligned}$$

C.2 Cluster Regression Coefficients

A plot of regression coefficients from the cluster selection procedure is shown in Figure 8.

Verb	Matrix	Template	Target	Believe	Want
think	+	A thought that C happened.	A	+	×
	−	A didn't think that C happened.	A	−	×
doubt	+	A doubted that C happened.	A	−	×
	−	A didn't doubt that C happened.	A	+	×
know	+	A knew that C happened.	A	+	×
	−	A didn't know that C happened.	A	−	×
remember	+	A remembered that C happened.	A	+	×
	−	A didn't remember that C happened.	A	−	×
hope	+	A hoped that C happened.	A	×	+
	−	A didn't hope that C happened.	A	×	⇜
fear	+	A feared that C happened.	A	×	−
	−	A didn't fear that C happened.	A	×	×
wish	+	A wished that C happened.	A	−	+
	−	A didn't wish that C happened.	A	×	⇜
worry	+	A worried that C happened.	A	×	−
	−	A didn't worry that C happened.	A	⇝	×
like	+	A liked that C happened.	A	+	+
	−	A didn't like that C happened.	A	+	−
love	+	A loved that C happened.	A	+	+
	−	A didn't love that C happened.	A	+	⇜
hate	+	A hated that C happened.	A	+	−
	−	A didn't hate that C happened.	A	+	⇝
regret	+	A regretted that C happened.	A	+	−
	−	A didn't regret that C happened.	A	+	⇝
tell	+	A told B that C happened.	A	×	×
	−	A didn't tell B that C happened.	A	⇝	×
notify	+	A notified B that C happened.	A	+	×
	−	A didn't notify B that C happened.	A	⇝	×
lie	+	A lied to B that C happened.	A	−	×
	−	A didn't lie to B that C happened.	A	⇝	×
mislead	+	A misled B that C happened.	A	−	×
	−	A didn't mislead B that C happened.	A	⇝	×
complain	+	A complained to B that C happened .	A	+	−
	−	A didn't complain to B that C happened .	A	+	⇜
boast	+	A boasted to B that C happened .	A	+	+
	−	A didn't boast to B that C happened .	A	⇝	⇝
convince	+	A convinced B that C happened.	B	+	×
	−	A didn't convince B that C happened.	B	⇜	×
persuade	+	A persuaded B that C happened.	B	+	×
	−	A didn't persuade B that C happened.	B	⇜	×
dissuade	+	A dissuaded B that C happened.	B	−	×
	−	A didn't dissuade B that C happened.	B	+	×
disprove	+	A disproved to B that C happened.	B	−	×
	−	A didn't disprove to B that C happened.	B	×	×
congratulate	+	A congratulated B that C happened.	B	+	+
	−	A didn't congratulate B that C happened.	B	+	+
apologize	+	A apologized to B that C happened.	B	+	−
	−	A didn't apologize to B that C happened.	B	+	−

Table 2 Verbs used in the validation experiment, along with predicted belief and desire inferences (based on authors' judgments) for referent of target argument (A = subject, B = non-subject), either with or without matrix negation. For conciseness, only past-tense items are shown. $\underline{-}$= *strong negative inference*, ⇜ = *weak negative inference*, × = *no inference*, ⇝ = *weak positive inference*, + = *strong positive inference*. Corresponding contentful items can be found in Appendix A.

Figure 8 Regression coefficients of best model from multivariate ridge regression on acceptability scores.