Proceedings of SALT 31: 570–605, 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. 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. We investigate the inferences that constitute each pattern as well as the distributional properties these patterns correlate with.

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).

* 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

(1) A predicate v triggers...

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

b. ...*neg(ation)-raising inferences* in a sentence np *not* v ((to) np) s iff the use of np *not* v ((to) np) s can trigger the inference that np v ((to) np) *not* s in some contexts.

c. ...{*doxastic*, *antidoxastic*} *inferences* about the role associated with position *i* in a sentence np_1 (*not*) v ((to) np_2) s iff the use of np_1 (*not*) v ((to) np_2) s consistently triggers the inference that {np_i *believes* s, np_i *believes not* s} across contexts.

d. ...{*bouletic*, *antibouletic*} *inferences* about the role associated with position *i* in a sentence np_1 (*not*) v ((to) np_2) s iff the use of np_1 (*not*) v ((to) np_2) s consistently triggers the inference that {np_i *wants* s, np_i *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 1982; 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 2021) 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 logically possible inference patterns vast, even the cleanest measurements of at least veridicality and neg-raising using inference judgment tasks display substantial gradience (White & Rawlins 2018b; An & White 2020). This gradience makes it difficult to ascertain which inference patterns are attested because it makes it difficult to determine (i) whether a particular sentence should be considered to trigger a particular inference; and (ii) whether there exist patterns consisting partly or wholly of non-necessary inferences. Such difficulties may be unavoidable—e.g. because gradience indicates that no formally represented lexical property controls whether a particular inference is triggered (see Tonahuser & Degen under review). But there are at least two other (non-exclusive) possibilities: (i) apparent gradience is partly or wholly a product of the methods often used to collect inference judgments and that there are discrete, formally represented lexical properties that are active in triggering the necessary inferences;[1] and/or (ii) that there are discrete, formally represented lexical properties active in triggering non-necessary inferences. We argue that assessing these possibilities requires a lexicon-scale approach through which an inference pattern taxonomy can be derived.

Our aim in this paper is to derive such a lexicon-scale 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 2018b) 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 and resulting in the MegaIntensionality dataset. 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

1 This possibility is strong, given what is known about gradience in acceptability judgments (Schütze & Sprouse 2014).

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 describe a method developed by Gantt, Kane & White (in prep) for capturing role-anchored inferences while avoiding typicality effects and how we deploy it to collect a lexicon-scale dataset covering a large swath of the English predicates that take finite clauses.

After describing our lexicon-scale data collection, we turn to the model we use to discover inference patterns and the verbs that trigger them in §4. To improve the chances that these clusters correspond to some formally represented lexical properties, 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 discuss the inference patterns that we discover, how they relate to syntactic distribution, and what generalizations can be made about their structure before concluding in §5.

2 Existing lexicon-scale datasets

Our work builds on two previous lexicon-scale datasets capturing veridicality (the MegaVeridicality dataset; White & Rawlins 2018b) and neg-raising inferences (the MegaNegRaising dataset; An & White 2020) across a variety of clause-embedding verbs in a variety of syntactic contexts. Both datasets include sentences selected on the basis of their acceptability as measured in the MegaAcceptability dataset (White & Rawlins 2016, 2020), which contains acceptability judgments for 1,000 English clause-embedding verbs in 50 syntactic contexts.

A major challenge to collecting such inference judgments at scale lies in disentangling the lexical effects of a sentence's matrix predicate on the inference from the potentially confounding effects of world knowledge. For instance, if a respondent were to indicate that (6b) follows from (6a), it would be difficult as an 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.

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 noun phrases in a syntactic context with indefinite pronouns and by replacing all verbs except for the target verb with *do*, *have*, or *happen*, as appropriate. White & Rawlins capture factivity in this paradigm by manipulating the negation on the verb and seeing whether participants judge that the inferences goes through in both conditions. The resulting MegaVeridicality dataset contains veridicality judgments for 517 finite clause-embedding English verbs in both transitive (passivized) and intransitive contexts.

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 (*not likely at all*) to 1 (*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 2018b, the authors select predicates based on their acceptability in different frames and with different tenses based on data from MegaAcceptability. The resulting MegaNegRaising dataset contains judgments for 925 clause-embedding verbs in six syntactic contexts (including multiple involving infinitival complements) in both past and present tenses, and with both first and third person subjects—e.g. the past analogue of (8) with a third person subject is (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*?

We focus only on the items in these datasets that contain finite embedded clauses.

3 Lexicon-scale data collection

In addition to the veridicality and neg-raising inferences already captured by MegaVeridicality and MegaNegRaising, we aim to capture patterns of doxastic and bouletic inferences at scale. As with veridicality and neg-raising inferences, a major obstacle to collecting inference 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 is defeasible in cases where the boaster is 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 bouletic 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. To mitigate this issue, we deploy a *templatic* variant of the bleaching method.

3.1 Method

In the *templatic bleaching method*, participants are presented with *templatic items* consisting of a *templatic antecedent*, as in (12), and a *templatic consequent*, as in (13), and are asked to judge the likelihood that the consequent is true given the antecedent using a slider with *extremely unlikely* on the left and *extremely likely* on the right (Gantt et al. in prep).[2]

(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.

3.2 Materials

We select 725 unique predicates for use in this experiment based on their normalized acceptability score in the MegaAcceptability dataset.[3] We focus on the 12 frames found in Table 1, 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 (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 (14a), future (14b), and tenseless (14c)—to ensure good coverage of bouletic predicates, like *hope*, and deontic predicates, like *demand*; and we manipulate DO/PP-taking to ensure good coverage of communicative predicates, like *say*.

(14) 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

2 In Appendix A, we reanalyze the data collected by Gantt et al. (in prep), demonstrating that this method can be used to accurately capture the doxastic and bouletic inferences triggered by a specific predicate, enabling us to accurately scale data collection for inference judgments to the entire lexicon.

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

1.5. These thresholds were 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.[4] The number of predicates that lie above this threshold for each frame can be found in Table 1. We additionally manipulate the polarity of the matrix clause in templatic antecedents, as in (15).

(15) A {wanted, didn't want} C to have happened.

And we construct templatic consequents for each antecedent that are conditioned on the tense/modality of the embedded clause—(16) for antecedents with embedded past tense and (17) for antecedents with tenseless or future embedded clauses.

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

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

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

(18) a. A/B believed that C happened.
b. A/B wanted C to have happened.

We sort the resulting items into lists of 32, aiming to constrain the construction of these lists such that the distribution over the expected responses for the items list has a mean of approximately 0.5 and has 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. For instance, if the underlying distribution of inferences for items in a list were to result in a heavy bias toward extremely high likelihood responses, any responses that might otherwise be moderately low likelihood or middling likelihood responses might be assigned extremely low likelihood in comparison to the majority of the other inferences in the list.

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

4 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.

templatic antecedent given by SUBTLEX (Brysbaert & New 2009).[5]

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 ($\geq$3).

3.3 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.

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

Figure 1 Distribution of verbs with respect to normalized doxastic and bouletic inference judgments.

3.4 Results

Figure 1 plots the normalized judgments for the doxastic and bouletic inferences (for both subject target and object target, when applicable), with select predicates labeled. To obtain these normalized judgments for each item, we use a mixed effects beta model-based normalization procedure.[6]

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 correspond well with intuitions about a variety of commonly discussed predicates. For instance, cognitive predicates (such as *think* and *know*) and communicative predicates that trigger doxastic inferences about the

6 See Appendix B for details.

recipient (such as *convince* and *persuade*) show up in the top left quadrant, indicating 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 suggests that no predicate has a positive bouletic inference when under both positive and negative matrix polarity. This pattern is consistent with the hypothesis that bouletic inferences (if present) are always at issue and targeted by negation (Anand & Hacquard 2014).

The bottom two subplots show judgments for the same items with embedded future (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 embedded clause but only a weak doxastic inference with a future embedded clause.

4 Discovering inference patterns

With the requisite data in hand, we now turn to the task of discovering inference patterns. To do this, we develop a model that 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 doxastic and bouletic inference judgments in MegaIntensionality. Each cluster is characterized jointly by the predicate-frame pairs that fall into it and the prototypical pattern of veridicality,

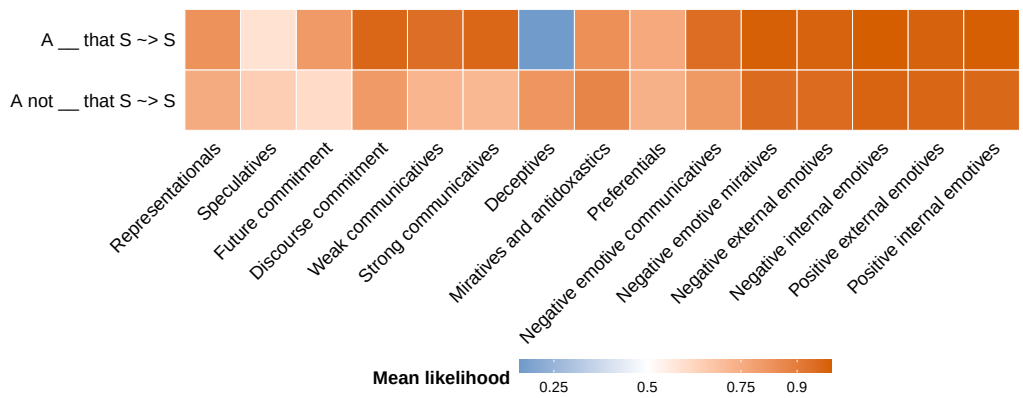

Figure 2 Prototypical veridicality inference patterns for each cluster.

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 predicates' morphosyntactic distribution as measured in the MegaAcceptability dataset.

4.1 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* (the three datasets) 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 Gaussian 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.[7]

7 See Appendix C for the full model specification.

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 from which the datapoint is drawn. We fit this model using variational inference as implemented in `pyro` (Bingham, Chen, Jankowiak, Obermeyer, Pradhan, Karaletsos, Singh, Szerlip, Horsfall & Goodman 2018).

4.2 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 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 has not 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.

4.3 Results

Using the approach described in §4.2, 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. In the remainder of this section, We investigate the inference patterns (§4.3.1) and distributional properties associated with each cluster (4.3.2). Further analysis of the inference patterns, including initial results on discovering generalizations governing them, can be found in Appendix D.

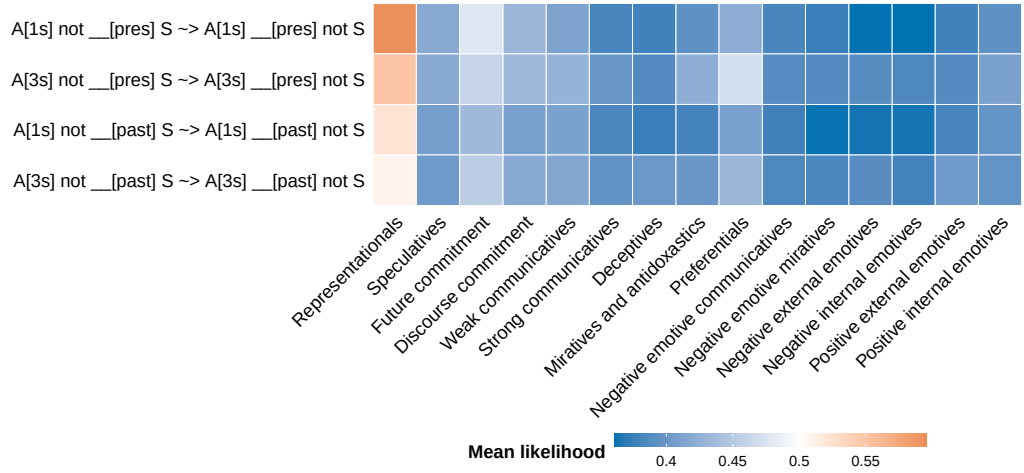

Figure 3 Prototypical neg-raising inference patterns for each cluster.

4.3.1 Inference Patterns

Figures 2–4 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 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: not all predicates in the cluster will share those prototypical characteristics. (19)–(33) give examples for each cluster.

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

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

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

(22) *Discourse commitment:* communicative acts involving the source's commitment

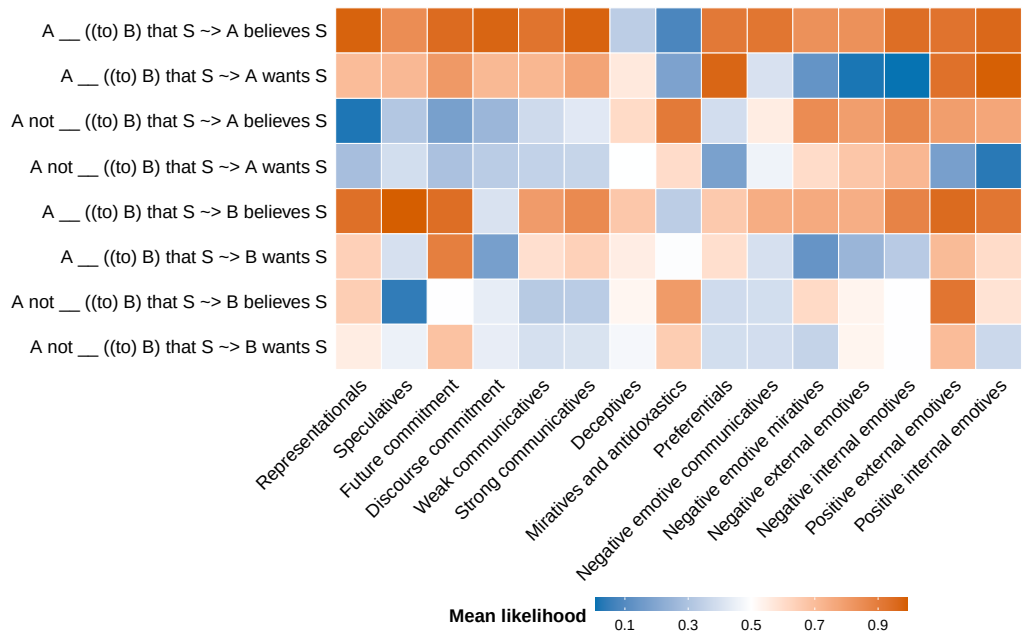

Figure 4 Prototypical doxastic and bouletic inference patterns for each cluster.

to the truth of some propositional content.
A {maintained, remarked, swore} that C would happen.

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

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

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

(26) *Miratives and antidoxastics:* states of surprise or disbelief.
A was {surprised, stunned, startled} that C would happen.

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

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

(29) *Negative emotive communicatives:* communicative acts with negative valence.
A {screamed, ranted, growled} to B that C would happen.

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

(31) *Positive external emotives:* externalized expressions of positive emotion.
A was {congratulated, praised, fascinated} that C happened.

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

(33) *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 2—is largely confined to emotive classes, with cognitive and non-emotive communicatives generally being veridical—the top cell only being dark orange—nonveridical—both cells being light orange or white—or *antiimplicative*—the top cell being blue and the bottom cell being orange. The one nonemotive class that comes close to the emotives in terms of factivity is the discourse commitment class (22). This class largely involves communicative predicates, but it also contains canonical cognitive (semi)factives, like *know*. This pattern is potentially consistent with Anand & Hacquard's (2014) suggestion that communicatives are never factive.

Neg-raising is even more constrained than factivity, only showing up among the representationals—that class being the only one with any orange in Figure 3. 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 4) are common to the majority of clusters, with only the deceptives and miratives/antidoxastics showing 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).

4.3.2 Relationship between clusters and syntactic structures

A plot of regression coefficients from the cluster selection procedure is shown in Figure 5. Each row of this plot displays the mean coefficients (across folds of the cross-validation described in §4.2) obtained from regressing the acceptability judgments for each predicate in each syntactic frame on the cluster membership probabilities $\boldsymbol{\phi}_v$ for each predicate. A darker orange box for a particular cluster-frame combination indicates that verbs within that cluster tend to be more acceptable within that frame (adjusting for all other clusters those predicates might fall into), whereas a darker blue box indicates that verbs within the cluster tended to be unacceptable within that frame (again, adjusting for all other clusters those predicates might fall into). Coefficients are set to 0 (white) when the bootstrapped Bonferroni-corrected confidence interval for that coefficient include 0.

The coefficients we obtain are broadly consistent with distributional patterns expected from prior literature (see White 2015; White, Hacquard & Lidz 2018 for references and a review), though a number of the finer-grained patterns we observe are novel to our knowledge. The clearest high-level distinctions we observe are in NP argument-taking behavior. One such distinction is that between the emotive and non-emotive classes: emotives tend to prefer a direct object when passivized—e.g. NP was Ved that S—but not in the active—e.g. NP Ved NP that S—though there is some variability in terms of how strong this preference is—e.g. the negative internal emotives tend to be better in active constructions than the positive internal emotives. This pattern is likely indicative of the fact that many emotives—e.g. *surprise*—are better in expletive subject + experiencer object constructions (34b) than with transitives (34c), which the passive constructions (34a) do not distinguish.[8]

(34) a. Someone was surprised that something happened.
b. It surprised someone that something happened.
c. Someone surprised someone that something happened.

Another such distinction in NP-taking behavior is observed between the communicative classes and the non-communicative classes: communicatives—such as the weak and strong communicatives, discourse and future commitments, and deceptives (though interestingly, not the speculatives)—tend to be more acceptable in active

8 The use of passive constructions in MegaAcceptability is designed to capture whether predicates prefer expletive subjects: White & Rawlins (2020) find that obtaining acceptability judgments for expletive subject constructions such as (34b) yields higher disagreement than using active and passive constructions, such as (34a) and (34c), to diagnose a preference for expletive subjects, probably because *it* is ambiguous between a referential and expletive pronoun.

Figure 5 Mean regression coefficients over 10 folds of cross-validation from a multivariate ridge regression on acceptability scores. Orange implies a positive coefficient and blue implies a negative coefficient. Coefficients are set to 0 (white) when the bootstrapped Bonferroni-corrected confidence interval for that coefficient include 0.

frames with a direct or indirect object than non-communicatives. This pattern is likely indicative of the fact that communicatives tend to entail (intended) transfer of information from a source (usually characterized by the subject) to a goal (usually characterized by the direct or indirect object).

Among the communicative classes there is a further distinction between classes that strongly prefer to realize the goal as an indirect object—e.g. the weak and strong communicatives and (to some extent) the discourse commitments—while others do not show as strong a preference for direct or indirect objects—e.g. the deceptives.[9] Interestingly, we do not observe a communicative class that prefers direct object objects to the exclusion of indirect objects—though the mirative and antidoxastic class does show something approximating such a pattern. This pattern appears to arise because, while this class is largely constituted by predicates of internal states like *surprise*, at least some of these internal state-characterizing predicates can at least marginally occur in active transitive constructions, such as (35).[10]

(35) Someone {doubted, disbelieved, distrusted} someone that something happened.

In addition to NP-taking behavior, we also see clear trends in the relationship to features of clausal complements. For instance, the representationals tend to prefer finite complements, and the preferentials tend to prefer tenseless complements—and particularly, infinitivals with an overt subject, such as NP Ved for NP to VP and NP Ved NP to VP (see Bolinger 1968 et seq).[11] This trend is expected from prior literature, though interestingly it is nowhere near categorical: representationals can occur with some kinds of infinitival complements and preferentials can occur with some kinds of finite complements. The compatibility of representationals with infinitival complements, however, tends to be biased toward stative infinitival complements, while preferentials tend not to be. This pattern may indicate a structural distinction between the sorts of infinitivals that representationals take and those that preferentials take (see Wurmbrand 2014).

9 That our model distinguishes weak and strong communicatives is somewhat surprising, since they show very close overlap both in terms of their syntactic profile in Figure 5 as well as their inference patterns. Furthermore, these clusters show high overlap in terms of the predicates themselves—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 profiles of these two clusters do differ slightly, as evidenced for example by the NP was Ved that S frame in Figure 5, indicating that some distinction does seem to exist between these two clusters.

10 This class also includes a small number of communicative predicates like *deny*, though these predicates do not appear to drive the preference for direct objects, since these communicative predicates actually tend to be ones that are best with indirect objects. But due to their small number relative to the size of the class, any preference for indirect objects in addition to direct objects is weak.

11 What we have called representationals here are also compatible with S-lifting (Ross 1973) and *so* anaphora, which indicate that this class is further constrained to "assertive" predicates (Hooper 1975).

One notable difference we observe in comparing our results to previous literature is a lack of clear distinction in the distributions of declaratives and interrogatives. This lack of clear distinction might initially seem surprising, in light of the fact that there are a wide variety of apparently successful proposals relating the inference types we consider here with the distribution of interrogatives and declaratives (Hintikka 1975; Zuber 1982; Egré 2008; Theiler et al. 2017, 2019; Uegaki & Sudo 2019; Roberts 2019). For instance, even though the representational cluster contains many predicates that are predicted by these proposals not to take interrogatives—e.g. *think* and *believe*—that class still has a reasonably strong preference for interrogatives—e.g. as in NP Ved whether to VP and NP Ved whether S—and certainly not a dispreference, as many of those proposals predict. But this pattern is consistent with recent work that tests generalizations relating these inference types to declarative- and interrogative-taking, finding that they do not in fact hold (White 2021).

Together with the fact that the major distinctions lie in the domain of NP-taking and that NP-taking elsewhere in the lexicon track aspectual/event structural behavior (see Levin & Rappaport Hovav 2005 and references therein), we follow White & Rawlins 2018a,b and White 2021 in suggesting that rather than searching for generalizations that relate the inference types investigated here directly to syntactic distribution, it is likely more fruitful to search for aspectual/event structural properties—e.g. transfer—that mediate the relationship.

5 Conclusion

We have investigated which patterns of lexically triggered doxastic, bouletic, negraising, and veridicality inferences are attested across finite clause-embedding verbs in English. We 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 only background belief inferences and for communicatives to only be at most veridical—while others point to novel ones—e.g. a variety of patterns involving negative inferences—that are not commonly assumed to have distributional correlates in English.

These findings lay the groundwork for further investigations of the relationship between inference patterns and distributional properties. With the aim of understanding why some (a relatively small number of) inference patterns are attested while (a substantial number of) others are not, we pursue two directions in ongoing work: (i) attempting to discover the lexical components that underlie both the inference patterns themselves and their relationship to distributional properties using models that simultaneously learn abstractions on both; and (ii) expanding MegaIntensionality to predicates that embed nonfinite subordinate clauses with the aim of assessing how much our inventory of patterns must expand to cover this area of the lexicon.

References

An, Hannah & Aaron White. 2020. The lexical and grammatical sources of neg-raising inferences. *Society for Computation in Linguistics (SCiL)* 3(1). 220–233. doi:10.7275/yts0-q989.

Anand, Pranav & Valentine Hacquard. 2013. Epistemics and attitudes. *Semantics and Pragmatics* 6(8). 1–59. doi:10.3765/sp.6.8.

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.

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

Bingham, Eli, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall & Noah D. Goodman. 2018. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research* .

Blei, David M., Andrew Y. Ng & Michael I. Jordan. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research* 3. 993–1022. doi:10.5555/944919.944937.

Bolinger, Dwight. 1968. Postposed main phrases: An English rule for the Romance subjunctive. *Canadian Journal of Linguistics* 14(1). 3–30. doi:10.1017/S0008413100019629.

Brysbaert, Marc & Boris New. 2009. Moving beyond Kučera and Francis. *Behavior research methods* 41(4). 977–990. doi:10.3758/BRM.41.4.977.

Egré, Paul. 2008. Question-embedding and factivity. *Grazer Philosophische Studien* 77(1). 85–125. doi:10.1163/18756735-90000845.

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. doi:10.1093/jos/ffx008.

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

Gantt, William, Benjamin Kane & Aaron Steven White. in prep. The interaction of lexical knowledge and world knowledge in role-anchored inferences.

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. doi:10.1007/BF00983299.
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. doi:10.1093/jos/9.3.183.
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. doi:10.1007/BF00351935.
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.
Levin, Beth & Malka Rappaport Hovav. 2005. *Argument Realization*. Cambridge: Cambridge University Press.
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.
Ross, John Robert. 1973. Slifting. In Maurice Gross, Morris Halle & Marcel-Paul Schützenberger (eds.), *The Formal Analysis of Natural Languages*, 133–169. The Hague: Mouton de Gruyter.
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. doi:10.3765/salt.v27i0.4125.
Theiler, Nadine, Floris Roelofsen & Maria Aloni. 2019. Picky predicates: why

believe doesn't like interrogative complements, and other puzzles. *Natural Language Semantics* 95–134. doi:10.1007/s11050-019-09152-9.

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.

Villalta, Elisabeth. 2000. Spanish subjunctive clauses require ordered alternatives. *Semantics and Linguistic Theory* 10. 239–256. doi:10.3765/salt.v10i0.3123.

Villalta, Elisabeth. 2008. Mood and gradability: an investigation of the subjunctive mood in Spanish. *Linguistics and Philosophy* 31(4). 467–522. doi:10.1007/s10988-008-9046-x.

White, Aaron Steven. 2015. *Information and incrementality in syntactic bootstrapping*. College Park, MD: University of Maryland PhD Thesis.

White, Aaron Steven. 2021. On believing and hoping whether. *Semantics and Pragmatics* 14. 6. doi:10.3765/sp.14.6.

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

White, Aaron Steven & Kyle Rawlins. 2018a. Question agnosticism and change of state. In Robert Truswell, Chris Cummins, Caroline Heycock, Brian Rabern & Hannah Rohde (eds.), *Sinn und Bedeutung*, 1325–1342. University of Edinburgh.

White, Aaron Steven & Kyle Rawlins. 2018b. The role of veridicality and factivity in clause selection. In Sherry Hucklebridge & Max Nelson (eds.), *Annual Meeting of the North East Linguistic Society (NELS)*, 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.

Wurmbrand, Susi. 2014. Tense and aspect in English infinitives. *Linguistic Inquiry* 45(3). 403–447.

Zuber, Richard. 1982. Semantic restrictions on certain complementizers. In S. Hattori & K. Inoue (eds.), *International Congress of Linguists (ICL)*, 434–436. Tokyo.

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

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

A Materials and Methods for Norming and Validation Studies

Here, we analyze data collected in Gantt et al. (in prep) in order to validate that our use of templatic items in our lexicon-scale study captures the same information about lexically triggered inferences as we would obtain using contentful items, while controlling for interactions between inference judgments and world knowledge. This data comprises judgments from three distinct tasks:

i. **Templatic inference task**. Given a *templatic* antecedent (e.g. (36a)) and a *templatic* consequent (e.g. (36b)), participants are asked to judge the likelihood that the consequent is true given that the antecedent is true.

ii. **Contentful inference task**. Given a *contentful* antecedent (e.g. (37a)) and a *contentful* consequent (e.g. (37b)), participants are asked to judge the likelihood that the consequent is true given that the antecedent is true.

iii. **Norming task**. For each *contentful* consequent in (ii), participants are asked to judge the *a priori* likelihood that it is true — i.e., without considering any associated antecedent. Specifically, these items ask about the likelihood that the subject or object of the antecedent would believe or desire the content of the embedded clause, as in (38).

(36) a. A doubted that C happened.
b. A believed that C happened.

(37) a. The publisher doubted that the author had finished her manuscript on time.
b. The publisher believed that the author had finished her manuscript on time.

(38) What proportion of publishers generally believe authors finish their manuscripts on time?

At a high level, we show that the judgments from (i) correlate strongly with the judgments from (ii), after controlling for potential effects of world knowledge using the judgments from (iii), thus validating the templatic method for our lexicon-scale study. For clarity, we describe the way items were constructed for all three studies below. Next, we present an analysis of the data from (iii)—our *norming* study—followed by an analysis of the data from (i) and (ii)—our *validation* study. Descriptions of the materials and methods used to produce this data are included simply to aid understanding and draw heavily on Gantt et al. (in prep). However, we emphasize that readers interested in this data, or in the methods used to collect it, should consult that paper.

A.1 Verb Selection and Item Construction

We select 24 predicates for the validation study (see Figure 7 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 doxastic and bouletic 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 doxastic and bouletic 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 Table 2.

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 doxastic and bouletic 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, (37a) gives the past variant of one scenario for *doubt*, (39a) gives the negated variant, and (39b) gives the future variant.

(39) 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 (37b)—yielding

768 total contentful items = 384 contentful antecedents × 2 contentful consequents. Table 2 gives a complete list of the verbs used in the validation study.

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 doxastic and bouletic 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. − = *strong negative inference*, ⇜ = *weak negative inference*, × = *no inference*, ⇝ = *weak positive inference*, + = *strong positive inference*. Corresponding contentful items can be found in Appendix A.

A.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.

A.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 the appendices of Gantt et al. (in prep) or online at megaattitude.io.

A.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. (38) gives the generic variants corresponding to (37b). 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.

A.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 et al. 2018 to ensure they were a native speaker of English.

A.2.4 Results

Figure 6 shows the mean judgments across annotators for each item, with color corresponding to the intended valence for the bouletic inferences. As intended from the way we constructed the scenarios, the means for the *want* items are bimodally distributed, thus satisfying our aim of high variability in the judgments across

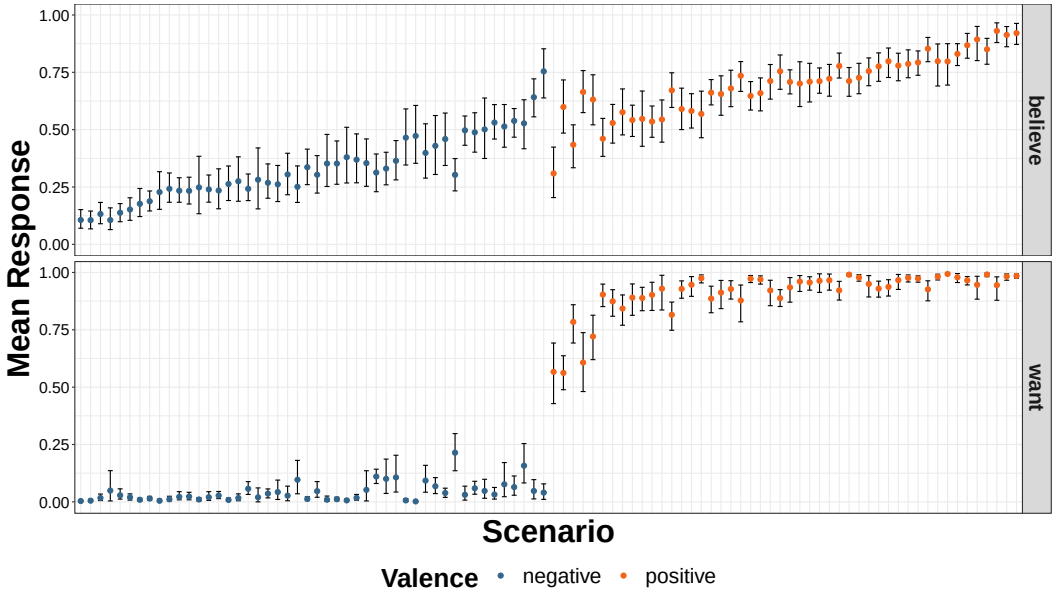

Figure 6 Mean judgments from the norming study.

those items. The means for the doxastic inferences are unimodally distributed—likely because we did not explicitly manipulate the prior probability of the doxastic inferences—but similarly show the desired high variability.

A.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.

A.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 (predicates) in the antecedents as an approach that uses contentful items and adjusts

for world knowledge.

A.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 Gantt et al. (in prep) or online at megaattitude.io.

A.3.2 Materials

Contentful items were constructed in the way described at the beginning of this appendix. 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.

A.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.

A.3.4 Results

Figure 7 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*,

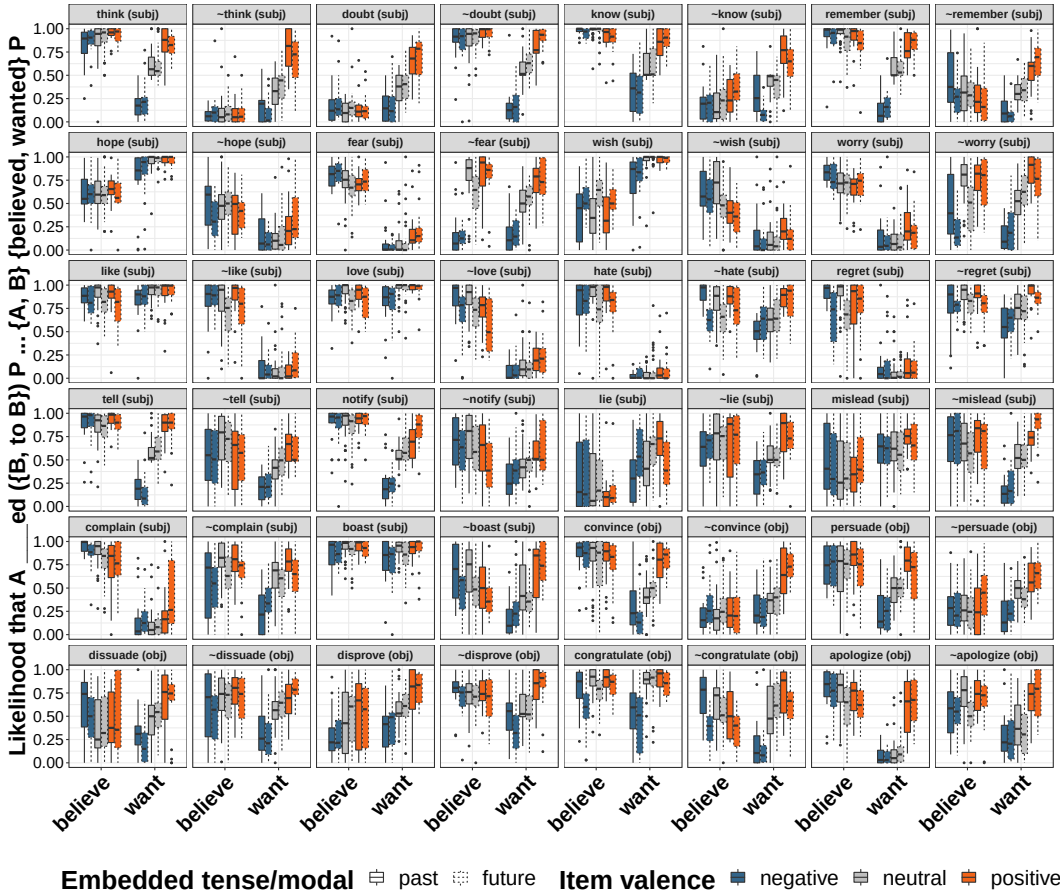

Figure 7 Distributions of inference judgments for all 24 predicates in the validation study, with and without negation on the predicate and for both past and future tense. Distributions for templatic judgments are shown in gray, and distributions for contentful items with predicted positive and negative inferences are shown in orange and blue, respectively.

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 normal-

ized 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, 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.

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 $\beta^{(\nu)}$, a fixed shifting term for each item $\beta_i^{(\mu)}$, a random scaling term $\rho_p^{(\nu)}$ for each participant p, and a random shifting term $\rho_p^{(\mu)}$ for each participant.

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

The normalizer optimizes these parameters against the log-likelihood of the data. The normalized scores for each item i in the MegaIntensionality dataset (plotted in Figure 1) correspond to $\text{logit}^{-1}\left(\beta^{(\mu)}_i\right)$.

C Model Details

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.

$$\begin{aligned}
\alpha &\sim \log\mathcal{N}(0, 100) \\
\boldsymbol{\gamma} &\sim \text{Dirichlet}(\mathbf{1}_K) \\
\boldsymbol{\theta}_{\langle v,f\rangle} &\sim \text{Dirichlet}(\alpha\boldsymbol{\gamma}) \\
c_i &\sim \text{Categorical}\left(\boldsymbol{\theta}_{\langle \text{verb}(i),\text{frame}(i)\rangle}\right) \\
\boldsymbol{\beta}_{\langle \text{view}(i),c_i\rangle} &\sim \mathcal{N}(\mathbf{0}, 100\mathbf{I}) \\
y_i &\sim f_{\text{view}(i)}\left(\boldsymbol{\beta}_{\langle \text{view}(i),c_i\rangle} \cdot \mathbf{x}_i, \boldsymbol{\rho}_{\text{participant}(i)}\right)
\end{aligned}$$

The beta mixed effects model for unit responses is shown below. $\beta^{(\mu)}$ and $\beta^{(\nu)}$ represent fixed effects: shifting terms for each cluster c_i and a scaling term, respectively. $\rho^{(\mu)}_p$ and $\rho^{(\nu)}_p$ represent shifting and scaling random effects for each participant p. We place a zero-mean log-normal prior on each scaling term, and a zero-mean normal prior on each shifting term.

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

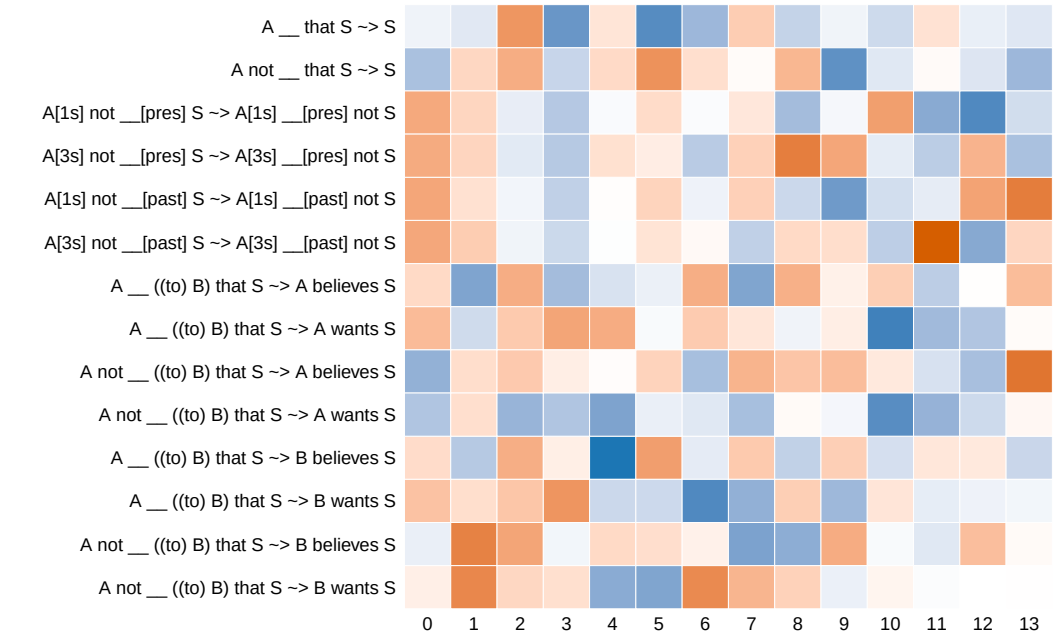

Figure 8 PCA loadings for each inference type.

The ordinal mixed effects model assumes that each item with a particular cluster c_i maps to some real-valued score β_{c_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 $\left(-\infty, \rho_{p1}\right]$, the best rating to bin $\left(\rho_{p(n-1)}, \infty\right)$, and all other ratings r to bins $\left(\rho_{p(r-1)}, \rho_{pr}\right)$.

The probability of a particular item i (with participant p_i and assigned cluster c_i) getting ordinal rating r is defined based on these cutpoints:

$$\begin{aligned}
\mathbb{P}(y_i \leq r) &= \text{logit}^{-1}(c_{p_i,r} - \beta_{c_i,k_i}) \\
\mathbb{P}(y_i = r) &= \theta_{ir} = \mathbb{P}(y_i \leq r) - \mathbb{P}(y_i \leq (r-1)) \\
y_i &\sim \text{Categorical}(\boldsymbol{\theta}_i)
\end{aligned}$$

D Generalizations about the attested inference patterns

With the aim of laying the groundwork for future investigations, we attempt to derive a set of generalizations about the clusters described above by applying principal component analysis (PCA) to the parameters of the probability distributions associated with each cluster (as visualized in Figures 2–4). The idea behind this method is

to decompose the inference patterns associated with each cluster into *components* capturing which inferences are correlated across clusters. Each component is thus associated with an inference subpattern. One way to view these subpatterns is as characterizations of some set of *foundational* patterns from which the clusters' actual patterns might be constructed; another is to view them 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 subpatterns are visualized in Figure 8. Orange tiles and blue tiles correspond to positive and negative loadings, respectively. There are two important things to note about interpreting these patterns. First, they are holistic in the sense that "pieces" of one subpattern can "cancel" or "augment" pieces of another. Second, it is the relative polarity, rather than the absolute polarity, that is important for interpreting these subpatterns because a particular cluster can be either positively or negatively associated with a component. This means that there are always two "sides" to any generalization based on the component: the positive and the negative. 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* antifactivity, since it suggests that the inference is robust under negation, but its polarity could be inverted. (More generally, we treat robustness under negation for any inference type as evidence of presupposition for that type.) In 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 these interpretations for the first six components below. Together, these six explain 95% of the information in the inference patterns (as measure by variance), and so there are likely to be diminishing returns in terms of explaining the remainder.

(40) The polarity of veridicality and doxastic inferences under negation is anti-correlated with neg-raising.

(41) The polarity of a doxastic presupposition about a recipient is correlated with the polarity of a bouletic presupposition.

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

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

(44) Bouletic inferences about the source in a communication are anticorrelated with doxastic inferences about the target.

(45) Veridicality is correlated with doxastic inferences in the target of a communication but anticorrelated with bouletic inferences.

Some of these generalizations are unsurprising from the perspective of prior work. For instance, the first of these generalizations (40) characterizes the behavior of emotive inference patterns: these classes tend to be associated with veridicality and doxastic presuppositions and have long been known not to be neg-raising.

Others of these generalizations are more surprising—at least in part because they concern the relationship between doxastic and bouletic inferences about sources and targets, which are relatively understudied. For instance, the second generalization (41) characterizes both the source and the recipient in a communicative event, pitting the beliefs of the source against the beliefs and desires of the recipient. Indeed, many of the above generalizations target both the source and the recipient of a communication, potentially suggesting intricate structure in the lexical representation of communicatives in need of further exploration.