Theory-driven statistical modeling for semantics & pragmatics: A case study on grammatically generated implicature readings

Michael Franke\textsuperscript{1} & Leon Bergen\textsuperscript{2}

\textsuperscript{1}University of Osnabrück
mchfranke@gmail.com

\textsuperscript{2}University of California, San Diego
lbergen@ucsd.edu
Theory-driven statistical modeling for semantics & pragmatics: A case study on grammatically generated implicature readings
Abstract

Computational probabilistic modeling is increasingly popular in linguistics, but its relationship with linguistic theory is ambivalent. We argue here for the potential benefit of theory-driven statistical modeling, based on a case study situated at the semantics-pragmatics interface. Using data from a novel experiment, we employ Bayesian model comparison to evaluate the predictive adequacy of four models which differ in the extent to and manner in which grammatically generated candidate readings are taken into account in four probabilistic pragmatic models of utterance and interpretation choice. The data provide strong evidence for the idea that the full range of potential readings made available by recently popular grammatical approaches to scalar implicature computation might be needed, and that classical Gricean reasoning may help manage the manifold ambiguity introduced by grammatical approaches to these. The case study thereby shows a way of bridging linguistic theory and empirical data with the help of probabilistic pragmatic modeling as a linking function.

Keywords: Pragmatics, computational modeling, scalar implicature, grammaticalism, Bayesian model selection
INTRODUCTION. The last two decades brought an empirical turn to several areas of theoretical linguistics, including syntax (Sprouse, 2007), semantics (Bott et al., 2011) and pragmatics (Noveck and Sperber, 2004). It is exciting when a discipline with a rich theoretical inventory becomes more experimental. But there are also challenges. An important issue concerns the precise linking of established theoretical ideas to potential experimental observations (Chemla and Singh, 2014). Based on a case study on scalar implicature, this paper argues for a benefit of casting semantic theory explicitly into a probabilistic model of pragmatic decisions by means of which empirical data can be analyzed in a theory-driven way.

The most common template for the empirical testing of a (formal or verbal) theory \( T \) is to derive specific predictions from \( T \), in the form of a hypothesis \( H \), ideally such that competitor theories (or, in the absence of such: common sense), would not make \( H \) appear very likely. An experiment is then designed to assess \( H \) in light of empirical data \( D \). The most frequent method to see whether \( H \) is corroborated by \( D \) is to use a statistical model \( M_{St} \), which is chosen from a restricted repertoire of standardized tools appropriate for the kind of experiment and data at hand (e.g., a generalized regression model). Interestingly, the original theory \( T \) does not usually inform the statistical model \( M_{St} \) with which the data are analyzed. The model \( M_{St} \) could be used in geology, medicine or economics. This is particularly troublesome for disciplines with a strong theoretical foundation which could be used to formulate more specific, theory-driven models \( M_T \) which directly link theory \( T \) to data \( D \). Indeed, researchers in several areas of science likewise argue for the integration of more theory-driven statistical modeling e.g, in ecology (McElreath, 2016), psychology (Lee and Wagenmakers, 2015), and also linguistics (Dotlačil and Brasoveanu, to appear).

To demonstrate the possible value of theory-driven modeling in an empirically oriented formal linguistics, this paper considers a case study on the familiar notion of scalar implicature (Geurts, 2010; Grice, 1975; Horn, 2004). In response to certain shortcomings of a traditional Gricean account of implicature, a recently popular and empirically successful alternative approach postulates that potential implicature-like meaning enrichments are generated in the grammar, thereby creating a massive semantic ambiguity (e.g. Chierchia, Fox, and Spector, 2012; Fox, 2007; Fox and Spector, 2018). We here address how these multiple semantic readings affect a speaker’s choice of utterance and a listener’s choice of interpretation, which is something not di-
rectly predicted by semantic theory alone, but necessary for linking its predictions to experimental data. We introduce four different models of utterance and interpretation choices, as variants of recently popular Rational Speech Act models (Frank and Goodman, 2012; Franke and Jäger, 2016; Goodman and Frank, 2016). Models differ in the way in which semantic ambiguity affects choices of utterance and interpretation (see Figure 2). The key contribution here is a novel model, the so-called Global Intentions model, in which speakers choose both an overtly observable utterance and its intended meaning (the latter stemming from the set of grammatically generated implicature meanings). This theory-driven model $M_T$ can be linked directly to experimental data from a novel small-scale experiment introduced here. We then use Bayesian model comparison (e.g. Jeffreys, 1961; Kass and Raftery, 1995), thereby demonstrating how theory-driven statistical modeling could make for an additional, helpful tool for testing linguistic theories more directly based on empirical data.

The paper is structured as follows. Section 2 introduces the grammatical approach to scalar implicature. Section 3 develops the relevant probabilistic models for utterance and interpretation choices in the light of multiple semantic ambiguity. Section 4 reports on a small-scale combined production and comprehension experiment. Section 5 uses this data to compare models.¹

2 Grammatically generated scalar implicature readings. Scalar implicatures are a special kind of Quantity implicature, so-called in reference to Grice’s Maxim of Quantity, which requires speakers to be as informative as possible, given the current goals of the conversation (Grice, 1975). For example, speakers who utter 1a may be taken to communicate 1b, because they did not utter the logically stronger, and therefore more informative alternative sentence in 1c.

(1)  
   a. I own some of Johnny Cash’s albums.  
   b. I own some but not all of Johnny Cash’s albums.  
   c. I own all of Johnny Cash’s albums.  
   d. I own some of Johnny Cash’s albums, and it’s not true that I own all.

Traditional Gricean accounts rationalize scalar implicatures as inferences a listener can draw based on reasoning about the speaker’s actions. The action to be rationalized in the example above, so the mainstream account goes, is that the speaker uttered the complete sentence in 1a.
An alternative action of the speaker is an utterance of 1c. The implicature reading in 1b therefore arises by conjoining the literal meaning of the sentence in 1a with the negation of the complete sentential alternative in 1c, resulting in 1d, which is equivalent to 1b.

This logic has difficulty explaining that an utterance of 2a intuitively conveys the meaning in 2b. This inference does not follow from 2d, which is the result of conjoining the negation of the meaning of the alternative sentence 2c with the literal meaning of 2a (but see Geurts, 2010, for an alternative solution).

(2)  a. A soldier showed some signs of malaria.
    b. A soldier showed some but not all signs of malaria.
    c. A soldier showed all signs of malaria.
    d. A soldier showed some signs of malaria, and it’s not true that a soldier showed all signs.

In response to problems like this, an alternative approach has gained prominence recently (e.g. Chierchia, Fox, and Spector, 2012; Fox, 2007; Fox and Spector, 2018). *Grammaticalism* assumes that the sentence in 2a is semantically ambiguous. It has a literal reading, where *some* just means “some and maybe all”. But it also has a reading where *some* is enriched to mean “some but not all”, in the scope of *a soldier*. Grammaticalism consequently earns its name by trying to incorporate meaning aspects that may appear to be pragmatic in nature into the grammar.

To illustrate the range of readings made available by grammaticalism, consider the set of *nested Aristotelians*, which is the focus of this paper. Nested Aristotelians are sentences in which the Aristotelian quantifiers *none*, *some* and *all* —arguably the most basic quantificational operators— appear once in outer (higher scope) position and once in inner (lower scope) position. There are nine nested Aristotelians, exemplified in 3.

(3)  { None | Some | All } of the aliens drank { none | some | all } of their water.

The following will use abbreviations for these sentences, indicating the outer and inner quantifier in order, e.g., as on the right-hand side of Table 1. For example, “AS” is short for *All of the aliens drank some of their water.*
Arguably all nested Aristotelians presuppose the existence of a plurality of relevant aliens. What matters to the truth of nested Aristotelians is whether there are (i) aliens which drank none of their water, (ii) aliens which drank some but not all, and (iii) aliens which drank all. We therefore distinguish seven kinds of world states, each of which Figure 1 provides one example of. Each situation in Figure 1 shows twelve aliens, each with its glass of water. To have a compact representation of relevant world states, we use pictures like \( \boxed{\text{full glass}} \) or \( \boxed{\text{empty glass}} \). These pictures indicate which types of aliens exist. If there are aliens which drank none of their water, the pictorial representation for this world state contains a full glass, as in \( \boxed{\text{full glass}} \); otherwise, if there are no aliens which drank none of their water (every alien drank at least some), then the corresponding picture shows no full glass, as in \( \boxed{\text{empty glass}} \). The state \( \boxed{\text{full glass}} \), for example, refers to the set of worlds in which all of the aliens drank some but not all of their water.

Grammaticalism assumes that pragmatic readings of sentences are generated by a silent exhaustionification operator \( \text{Exh} \) whose meaning contribution is similar to that of particle \textit{only} (e.g. Chierchia, Fox, and Spector, 2012). For present purposes, there are three relevant places where an \( \text{Exh} \)-operator might occur in a nested Aristotelian. As shown in 4, it can occur in matrix position thereby applying to the whole sentence and it can apply to the outer or inner quantifier.

(4) \( \text{Exh}_M[\text{Exh}_O(Q_O) \text{ of the aliens drank } \text{Exh}_I(Q_I) \text{ of their water}] \)

A \textit{parse} determines the reading of a sentence. It determines whether \( \text{Exh} \) occurs in matrix (\( M \)), in outer quantifier (\( O \)) and in inner quantifier (\( I \)) position. We use a notation where, for example, a parse of a sentence with \( \text{Exh} \) occurring only in matrix position is denoted as \( M \), a parse with \( \text{Exh} \) occurring at all three relevant positions is written as \( MOI \) and a parse without any \( \text{Exh} \) operator, thus a literal reading, is represented as \( \text{lit} \). If \( S \) is a sentence and \( p \) its parse, then \( \llbracket S \rrbracket^p \) is the reading of \( S \) under \( p \). Sentences can have multiple readings based on different parses. Table 1 lists the readings of nested Aristotelians. These readings are derived from what is perhaps the simplest instantiation of grammaticalism (see Appendix A for technical details).

In sum, grammaticalism stipulates silent grammatical operators, whose effect it is to generate a rich but conventionally predictable semantic ambiguity. It thereby predicts the availability of intuitively attractive readings for cases, like in 2, which are difficult traditional Gricean accounts. However, it remains controversial whether all of these readings are important for explaining rele-
vant empirical data (e.g. Chemla and Spector, 2011; Franke, Schlotterbeck, and Augurzky, 2017; Geurts and Pouscoulous, 2009; Geurts and van Tiel, 2013; Potts et al., 2016). Moreover, it is, to a large extent, an open issue how the massive ambiguity generated in the grammar is resolved in context, by some, perhaps pragmatic, mechanism (for some discussion see, e.g., Chierchia, Fox, and Spector, 2012; Fox and Spector, 2018). Seen in this way, it would be desirable to explore how far a combination of, on the one hand, Gricean ideas about efficient communication in the light of potential ambiguity and, on the other hand, a grammatical approach that generates potential pragmatic readings could solve each approach’s problems. This is the goal of the models introduced next.

<INSERT TABLE 1 ABOUT HERE>

3 Models. We consider four models, all of which rely on a general, Grice-inspired picture of a speaker trying to maximize the amount of information conveyed by an utterance and a listener consequently interpreting the speaker’s utterance based on such information-maximizing behavior. Each model predicts a probability with which a speaker chooses each nested Aristotelian to describe any given world state (from Figure 1) and the degree of belief a listener assigns to each world state after hearing any given nested Aristotelian. The models differ, however, in the way in which they integrate grammatically supplied ambiguity.

Figure 2 gives a condensed overview of the main conceptual differences. First, the vanilla Rational Speech Act (RSA) model of Frank and Goodman (2012) associates each sentence with its logical semantics only (Section 3.1). On the other end of the spectrum, the Global Intentions model (Section 3.4) considers all the readings listed in Table 1. In between these two extremes are two models that take increasingly large subsets of the readings listed in Table 1 into account (Sections 3.2 & 3.3). Importantly, however, models also differ in the manner in which semantic ambiguity affects speakers’ and listeners’ choices. The Lexical Uncertainty model of Potts et al. (2016) (Section 3.2) features a speaker-production part in which each speaker is assumed to have a fixed, inflexible lexicon with exactly one meaning attached to each lexical item. In addition, we also introduce two conceptually different, novel models in which all speakers are aware of the full semantic ambiguity generated by their either lexicon (Lexical Intentions model, Section 3.3) or
grammar (Global Intentions model, Section 3.4), and in which they flexibly use the potential this
ambiguity creates for the purpose of maximizing information flow in communication.

<INSERT FIGURE 2 ABOUT HERE>

3.1 The vanilla Rational Speech Act model. The RSA model defines a speaker production
rule and a listener comprehension rule, roughly as described by a classical Gricean approach (see
Franke and Jäger, 2016; Goodman and Frank, 2016, for overview). The speaker is assumed to
follow the Gricean Maxims of Quality (roughly: be truthful) and Quantity (roughly: be informa-
tive). The listener tries to infer which meaning a speaker most likely had in mind when producing
an utterance, and does so on the assumption that the speaker follows the Gricean postulates of
truthfulness and informativity.

The speaker production rule $P_S(u \mid t; \alpha)$ of the vanilla RSA model determines the probability
with which the speaker chooses an utterance $u \in U$ when the true state is $t \in T$. The model
parameter $\alpha$ determines how stringently the speaker selects more informative descriptions over
less informative ones (see below). The usual definition utilizes the notion of a literal listener ($LL$)
and looks like this:

$$P_{LL}(t \mid u) = P(t \mid \llbracket u \rrbracket) \propto P(t) \delta_{t \in \llbracket u \rrbracket} \tag{1}$$

$$P_S(u \mid t; \alpha) \propto \exp(\alpha \cdot \log P_{LL}(t \mid u)) \tag{2}$$

The literal listener, defined in Equation [1], interprets each utterance $u$ based on its semantic
meaning, by simple Bayesian updating $P(t \mid \llbracket u \rrbracket)$ of prior beliefs $P(t)$ with the proposition
$\llbracket u \rrbracket \subseteq T$ that $u$ is true.

It helps to think of the literal listener as a mere technical construct, whose purpose is to an-
chor the semantic meaning of utterances (Franke, 2009). If the prior probabilities of states are
uniform, so that $P(t) = P(t')$ for all $t, t' \in T$ (an assumption we will make throughout), Equation [2],
can be rewritten to show the two Gricean constraints of truthfulness and informativity at
play:  

\[ P_S(u \mid t \mid \alpha) \propto [P(t \mid \llbracket u \rrbracket)]^\alpha = \left[ \delta_{\text{truthfulness}} \cdot \llbracket u \rrbracket^{-1} \right]^\alpha \]

This reformulation makes clear that, as long as there is any true utterance in \( t \), any false utterance has zero probability of being selected in \( t \) (since \( \delta_{\text{truthfulness}} = 0 \)). Between two true utterances (for which the \( \delta \)-term evaluates to 1), the utterance with a stronger semantic meaning, i.e., which is true in fewer world states, will be chosen with a higher probability. The bigger \( \alpha \) the more pronounced this preference for informative utterances is.

For example, RSA’s production rule predicts that a speaker would choose utterances to communicate state \( \llbracket \text{u} \rrbracket \) with the following probabilities when we set \( \alpha = 5 \) (an arbitrary value chosen here only for illustration; Section 4’s model comparison does not hinge on specific values for \( \alpha \)):

<table>
<thead>
<tr>
<th>NN</th>
<th>NS</th>
<th>NA</th>
<th>SN</th>
<th>SS</th>
<th>SA</th>
<th>AN</th>
<th>AS</th>
<th>AA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00</td>
<td>0.00</td>
<td>0.85</td>
<td>0.13</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

In words, RSA predicts that speakers would produce only descriptions that are literally true in state \( \llbracket \text{u} \rrbracket \), namely “NA”, “SN” or “SS” (see Table 1). Among these, the speaker’s choice probabilities reflect the semantic strength of utterances (\( \llbracket \text{NA} \rrbracket = 3 \), \( \llbracket \text{SN} \rrbracket = 4 \) and \( \llbracket \text{SS} \rrbracket = 6 \)). Further example predictions for the RSA speaker rule with \( \alpha = 5 \) are shown in Figure 3.

The comprehension rule of the vanilla RSA model captures a Gricean interpreter. It is Gricean in the sense that it models pragmatic inferences derived from the assumption that the speaker adheres to the Gricean postulates of truthfulness and informativity. Concretely, for each utterance \( u \) the rule \( P_L(t \mid u \mid \alpha) \) assigns a probability to each interpretation \( t \) based on the prior probability \( P(t) \) of a state and the likelihood \( P_S(u \mid t \mid \alpha) \) that a (truthful and informative) speaker would use the observed utterance for this state, following Bayes rule:

\[ P_L(t \mid u \mid \alpha) \propto P(t) \cdot P_S(u \mid t \mid \alpha) \]

For example, consider the pragmatic listener’s belief after hearing “SS”, when fixing \( \alpha = 5 \) for illustration:
The sentence “SS” is literally false in state \( \cdot | \text{SS}; \alpha = 5 \), so this state receives zero probability in the pragmatic listener’s interpretation. But also the other states, except for \( \cdot | \text{SS} \) and \( \cdot | \text{CL} \), receive (almost) zero probability, because the pragmatic listener would expect the speaker to choose a different expression in these states with much higher likelihood (see Figure 3).

3.2 The Lexical Uncertainty model. The Lexical Uncertainty (LU) model extends the vanilla RSA model by including the listener’s potential uncertainty about the lexical meaning that the speaker assigns to certain expressions (Bergen, Levy, and Goodman, 2016). Like Potts et al. (2016) we are interested here in the case where the listener does not know which lexical entry the speaker has for the word some. The listener may nevertheless try to infer what some most likely means (literally) to the current speaker. This inference, intuitively, proceeds as follows: given that the speaker said \( u \) and \( u \) can either be interpreted in such a way that every occurrence of some means “some and maybe all” or so that every occurrence means “some but not all”, which pair \( \langle t, l \rangle \) of a state \( t \) and mental lexicon \( l \) is most likely to have caused this speaker to have produced \( u \)?

Conceptually, the idea that listeners may entertain uncertainty about the lexical meaning a speaker assigns to some is quite distinct from the idea that the grammar produces manifold pragmatic readings for sentences containing words like some by variable insertion of Exh-operators. With slight abuse of notation, however, since some is the only Aristotelian quantifier susceptible to local meaning change by an Exh-operator, we can think of a speaker who considers some to mean “some but not all” as a speaker who always, inflexibly, uses parse OI (see Table 1). Similarly, a speaker with a literal meaning for “some” can be thought of as using the parse lit everywhere.

The LU model has listeners reason about the speaker’s mental lexicon \( l \), represented here as parses \( l \in \{ \text{lit}, \text{OI} \} \). Crucially, any given speaker is assumed to have either one of the two mental
lexica, but not both (see Figure 2). The speaker selects utterances $u$ by the same mechanism as in the RSA model, but based on a semantic interpretation of utterances influenced by the speaker’s lexicon $l$:

$$P_{S_1}(m \mid t, l; \alpha) \propto [P(t \mid [m]^l)]^\alpha$$ \[4\]

The production rule of the LU model conditions the speaker’s utterance choice on that speaker’s fixed lexicon. This will be crucial when comparing the LU model to other models in the following.

The listener does not know the speaker’s mental lexicon but infers which state-lexicon pairs are likely to have caused the speaker to produce the observed utterance, using Bayes rule:

$$P_{L_1}(t, l \mid m; \alpha) \propto P(t) \cdot P(l) \cdot P_{S_1}(m \mid t, l; \alpha)$$ \[5\]

The prior $P(l)$ over lexica $l$ is here assumed to be uniform.

Equation [5] defines the listener’s beliefs for each pair of state $t$ and mental lexicon $l$. From this, we can derive the listener’s beliefs about only the state $t$, by taking the weighted sum over the mental lexica (so-called marginalization):

$$P_{L_1}(t \mid m; \alpha) = \sum_l P_{L_1}(t, l \mid m; \alpha)$$

While Potts et al. (2016) use this latter formula to explain data from a truth-value judgement task, this paper also considers data from an experimental task which is most naturally linked to a production rule. Unfortunately, the production rule $P_{S_1}$ defined above is not ideally suited for this because it only makes predictions for a given lexicon $l$. We therefore follow Lassiter and Goodman (2017) and define a speaker $S_2$ who reasons about pragmatically adequate utterance choice based on the state-interpretation of $L_1$. Finally, a pragmatic listener $L_2$ who reasons about the latter speaker’s choice of utterances is defined as before in terms of Bayes rule. Consequently, the
The final definition of the LU models’ production and comprehension rules are:

\[
P_{S_2}(m \mid t; \alpha) \propto [P_{L_1}(t \mid m; \alpha)]^\alpha
\]

\[
P_{L_2}(t \mid m; \alpha) \propto P(t) \cdot P_{S_2}(m \mid t; \alpha)
\]

Figure 3 shows the predictions of the LU model’s production rule (for \(\alpha = 5\)). These clearly differ from those of the vanilla RSA model in several places. This also affects interpretation of utterances. For example, the sentence “SS” is interpreted as follows:

<p>| | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
P_{L_2}(\cdot \mid “SS”; \alpha = 5) = 0.00, 0.04, 0.00, 0.90, 0.00, 0.06, 0.00
\]

The most likely interpretation of “SS” is [I], which is a maximally informative utterance ins [O] for a speaker with lexicon 0I.

### 3.3 The Lexical Intentions Model

The LU model introduced in the previous section treats potential in situ enrichments of *some* as a consequence of a speaker’s lexicalization of a “some but not all” meaning. This entails that any given speaker assigns to all occurrences of *some* the same lexical meaning (see Figure 2). Consequently, the LU model considers two possible meanings for the sentence “SS”: the standard literal reading \([SS]^{\text{lit}}\) and the reading \([SS]^{\text{OI}}\) in 5.

\[
(5) \text{ Some but not all of the aliens drank some but not all of their water.} \quad ([SS]^{\text{OI}})
\]

But it is also conceivable that some speakers, when uttering a sentence with multiple occurrences of *some*, like “SS”, might mean to convey a reading that results from different lexical meanings for different occurrences, such as in 6a or 6b, which correspond to readings \([SS]^{\text{I}}\) and \([SS]^{\text{O}}\) respectively.

\[
(6) \quad \text{a. Some (and maybe all) of the aliens drank some but not all of their water.} \quad ([SS]^{\text{I}})
\]

\[
(6) \quad \text{b. Some but not all of the aliens drank some (and maybe all) of their water.} \quad ([SS]^{\text{O}})
\]

The LU model does not contain speakers of this kind. To model speakers that associate “SS” with any of the readings in 6, it is not enough to just include parses I and O as additional values for
the variable $l$ in the LU model. The result would be conceptually highly implausible. By Equation [4], the speaker has a fixed $l$ and invariably applies it to interpret whatever sentence comes along. That does make sense when $l$ is instantiated with parses $1i\ t$ and $0I$, because we can interpret it as speakers who have lexicalized a particular meaning of some and apply it invariably. But if $l$ is a parse like 0, Equation [4] would model a speaker who inflexibly assigns a logical meaning to some in inner position, and a strengthened lexical meaning in outer position, no matter what the sentence and the state to be communicated. This is not a conceptually plausible model of a speaker’s general behavior.

Alternatively, we can treat parses as a choice of the speaker. Suppose that all speakers have an ambiguous lexical entry for some: it might mean “some but not all” or “some and maybe all”. Speakers can then choose to mean some in this way or that, depending on whether they deem this beneficial from a communicative point of view. In other words, when carrying an ambiguous lexicon, speakers may utter sentences with different lexical intentions, where the intended meaning for each occurrence of a word might be different. Pragmatic listeners can then try to recover how a speaker may have chosen to mean any single occurrence of some based, as usual, on a model of the speaker’s strategy of choosing utterances and lexical intentions (see Figure 2).

The Lexical Intentions (LI) model formalizes these intuitions. It resembles the LU model superficially, but is conceptually different and also simpler. We assume that $l \in \{1i\ t, I, 0, 0I\}$ and define:

$$P_S(m, l \mid t; \alpha) \propto [P(t \mid [m])^l]$$  
$$P_L(t, l \mid m; \alpha) \propto P(t) \cdot P_S(m, l \mid t; \alpha)$$

As before, we then take weighted sums to retrieve predictions of utterance and interpretation choice probabilities:

$$P_S(m \mid t; \alpha) = \sum_l P_S(m, l \mid t; \alpha)$$
$$P_L(t \mid m; \alpha) = \sum_l P_L(t, l \mid m; \alpha)$$

The main difference between Equations [4] and [6] is the position of the lexicon parameter $l$. 
In the LI model the lexical meaning \( l \) is treated not as an argument to be passed into the speaker function, like in the LU model, but as an output of it. Conceptually, this means that the LI model does not model speakers who invariably assign a particular lexical meaning to each occurrence of a word, but speakers who choose utterances and their meanings in tandem. They do so, as usual, in such a way as to maximize the informativity of their utterances. Consequently, speakers choose a pair \( \langle u, l \rangle \) as a description of \( t \) only if \( t \in \llbracket u \rrbracket^{l} \) and they make this choice with a probability proportional to the relative informativity of \( \llbracket u \rrbracket^{l} \), i.e., they prefer pairs \( \langle u, l \rangle \) for which \( \llbracket u \rrbracket^{l} \) is small.\(^6\)

The predictions of LI are subtly different from the two previous models (see Figure 3). For example, the LI model predicts that the listener’s interpretation for the sentence “SS” is:

\[
P_L(·|“SS”; \alpha = 5) = \begin{bmatrix} 0.00 & 0.31 & 0.15 & 0.38 & 0.00 & 0.16 & 0.00 \end{bmatrix}
\]

According to the LI model, “SS” should be associated with interpretation much more readily than predicted by vanilla RSA or the LU model.

3.4 The Global Intentions model. The LI model assumes that speakers have ambiguous lexical entries for \textit{some}. When pondering the choice of a sentence to communicate the given state, speakers actively choose readings of sentences which makes them true and informative. From here it is only a very small step towards a model that includes all readings supplied by grammaticalism. The Global Intentions (GI) model is exactly like the LI model, but integrates the all parses \( p \) considered in Table 1 (see Figure 2):

\[
P_S(m, p | t ; \alpha) \propto [P(t | \llbracket m \rrbracket^{p})]^{\alpha} \\
P_L(t, p | m ; \alpha) \propto P(t) \cdot P_S(m, p | t ; \alpha)
\]

The GI model goes far beyond the LU model of Potts et al. (2016). While the LU model captures certain phenomena, like locally embedded scalar implicatures, in an RSA-style reasoning framework, it does not directly engage with or adopt grammaticalism. The GI model, instead, does. Including the full set of grammaticalist readings into the LU model makes little conceptual sense for the same reasons laid out in the context of the LI model: it would be quite unnatural.
speakers who cannot rise above their language’s ambiguity, invariably applying a particular pattern of insertions of $Exh$-operators, irrespective of the sentence under consideration and the state to be described. The GI model, instead, models speakers who are masters of the rich ambiguities provided by their language’s grammar, using these ambiguities to flexibly intend to express this reading or another, depending on what serves communication.

The GI model also makes distinct empirical predictions (Figure 3). For example, the speaker is predicted to use the sentence “SS” with a very high probability in state $\text{[\text{[}SS\text{][}]}}$. This is because the reading $[[SS]]^M$ is available for the speaker, which yields a very strong reading that uniquely singles out this world state. Consequently, the listener’s interpretation of “SS” also puts substantial probability on the interpretation $\text{[\text{[}SS\text{]}]}$:

| $P_L(\cdot \mid “SS”; \alpha = 5)$ | 0.00 | 0.70 | 0.05 | 0.21 | 0.00 | 0.03 | 0.00 |

4 Experiment.

Design. The models just introduced differ mostly in their quantitative predictions about likelihood of expression and interpretation choices. Introspection is not reliable enough to assess such fine-grained probabilistic predictions, but experimental data may be. As all models make predictions about expression-choice and state-interpretation probabilities, we collected data from tasks that probe these model predictions directly. We will refer to these tasks as a production and an interpretation task.

Participants. 100 participants with US IP-addresses were recruited via Amazon’s Mechanical Turk, using psiTurk (Gureckis et al., 2016). We excluded data from two participants who did the experiment twice and another three who did not self-identify as native speakers of English.

Materials. The experiment was couched in a cover story about friendly aliens visiting earth. The relevant test sentences are the nested Aristotelians in 3. We used quantification over a mass term (water that the aliens drank) for the inner quantifier so as to avoid typicality effects (e.g. Degen and Tanenhaus, 2015; van Tiel, 2014). The seven relevant states were displayed using pic-
tures like in Figure 1. All pictures contained twelve aliens with full, half-full or empty glasses. For situations corresponding to a “some but not all” reading of the outer quantifier we used four or six aliens (as appropriate) out of the total twelve, both of which are fairly natural or typical numbers to be denoted by \textit{some} (Degen and Tanenhaus, 2015; van Tiel, 2014).

Procedure. Participants were first told that the experiment consisted of two parts. Each part started with a background story, which served to introduce the alien scenario, and what was expected of participants, as well as our sentence and picture material. Participants completed seven production trials, one for each world state, in random order. Each trial displayed a picture of the state and participants selected the outer and inner quantifier from a dropdown menu (see Figure 4a). Participants then completed nine interpretation trials, one for each sentence, in random order. Each trial displayed all of the seven states with a short repetition of the task question and a slider bar next to the picture (see Figure 4b). Participants rated how likely they thought the displayed situation is what the speaker had observed.

Results. Figure 5 shows the frequencies with which sentences were selected for each state. Interestingly, sentences with \textit{none} as outer quantifier were used very infrequently. Another interesting point is that sentence “AS” was the most frequently used utterance in state [ ]; which seems to conflict with the predictions of the RSA model plotted in Table 1. Finally, we see that the most frequently used utterance in state [ ] is “SS,” which goes against the predictions of the RSA and the LU model plotted in Table 1, but is consistent with the GI model.

Each trial in the comprehension task returns nine slider ratings, one for each state. We normalized each of these nine-placed vectors. Figure 6 shows averages over these normalized vectors. The interpretation of “AS” sentences reflects the production data in the sense that the (on average) most likely interpretation was state [ ], followed by state [ ] and finally [ ], thereby replicating previous results on interpretation preferences for these sentences (Chemla and Spector, 2011; Franke, Schlotterbeck, and Augurzky, 2017). Finally, the interpretation of the sentence
“SS” seems inconsistent with the interpretation predicted by the RSA model (for $\alpha = 5$), but it is hard to assess by visual inspection whether this case might provide strong evidence for or against any of the other models. This is why we turn to formal model comparison next.

Model comparison. In order to quantify the relative evidence provided by the experimental data for or against the models introduced in Section 3 we look at Bayes factors (Jeffreys, 1961; Kass and Raftery, 1995). From a Bayesian point of view, a model $M$ consists of a prior $P(\theta | M)$ over vectors $\theta$ of values for its parameters and a likelihood function $P(D | \theta, M)$, which assigns a likelihood to the observed data $D$ for each vector $\theta$. The marginalized likelihood for model $M$ given data $D$ quantifies how likely $D$ is a priori for any parameter value:

$$P(D | M) = \int P(\theta | M) \cdot P(D | \theta, M) \, d\theta.$$ 

The Bayes factor in favor of model $M_1$ over $M_2$ is the ratio of marginalized likelihoods: $P(D | M_1) / P(D | M_2)$. It quantifies the factor by which our beliefs should shift in favor of $M_1$, relative to $M_2$, given that we have observed $D$, since by Bayes rule:

$$\frac{P(M_1 | D)}{P(M_2 | D)} = \frac{P(D | M_1) \cdot P(M_1)}{P(D | M_2) \cdot P(M_2)}.$$ 

Bayes factors are therefore independent of the prior odds of models $P(M_1)/P(M_2)$, but depend on the priors over parameter values $P(\theta | M_i)$ for each model $M_i$.

We have production $D_p$ and comprehension data $D_c$ and compare models based on how well they explain their conjunction: $P(D | \theta, M) = P(D_p | \theta, M) \cdot P(D_c | \theta, M)$. The following explains how the models defined in Section 3 give rise to a likelihood function for $D_p$ and $D_c$ respectively.

The production rule of model $M$ defines the likelihood $P_S(u | t, \theta, M)$ of a single choice of expression for a state. All of the models from Section 3 use a single parameter $\alpha$. On top of these, we include two more parameters in each model in order to accommodate for two general observations about $D_p$. First, our models predict probability zero for utterances that are false for a given state. However, we do observe false utterance choices (see Figure 5). We therefore add a constant
error term $\epsilon$ to all predicted utterance choice probabilities. Second, since no model accounts by itself for the low choice rates of sentences starting with none, we include a cost term $c(u)$ which is fixed to zero for utterances starting with some or all, but may be positive for utterances starting with none. These costs capture a general dispreference for particular expressions, and are not to be confused with “processing costs” from the psycholinguistic literature. Costs are subtracted, following standard practice, so that, e.g., for vanilla RSA we obtain:

$$P_S(u \mid t; \alpha, \epsilon, c) \propto [P_{LL}(t \mid u) - c(u)]^\alpha + \epsilon$$

Finally, if $D_p$ consists of counts $n_{ij}$ of the number of times utterance $u_j$ was chosen in state $t_i$, the likelihood $P(D_p \mid \theta, M)$ is:

$$P(D_p \mid \theta, M) \propto \prod_{i=1}^7 \text{Multinomial} (\langle n_{i1}, \ldots, n_{i9} \rangle, \langle P_S(m_1 \mid t_i, \theta, M), \ldots, P_S(m_9 \mid t_i, \theta, M) \rangle) .$$

Comprehension rules give a probability distribution over states for each given utterance: $P_L(t \mid u, \theta, M)$. Since these are defined in terms of the speaker choice probabilities, the parameterization of the listener rules are the same. The comprehension data $D_c$ consists of probability vectors $c_{ij}$, where $c_{ij}$ is the average of normalized ratings assigned to state $j$ for utterance $i$ (as plotted in Figure 6). We think of the probability vector $\langle c_{i1}, \ldots, c_{i7} \rangle$ as a sample from a Dirichlet distribution whose modal value is the model’s prediction $\langle P_L(t_1 \mid u_i, \theta, M), \ldots, P_L(t_7 \mid u_i, \theta, M) \rangle$. To allow for more or less deviation in the realization of observed ratings, we introduce a parameter $w$, where the higher $w$ the more we expect observations that are very close to the model’s predictions:

$$P(D_c \mid \theta, w, M) \propto \prod_{i=1}^9 \text{Dirichlet}(\langle c_{i1}, \ldots, c_{i7} \rangle, w \cdot \langle P_L(t_1 \mid u_i, \theta, M), \ldots, P_L(t_7 \mid u_i, \theta, M) \rangle) .$$

As for priors over model parameters, we assume that all parameters are independently sam-
pled from flat priors with a sufficiently large support. Priors are the same for all models:

\[
\begin{align*}
\alpha & \sim \text{Uniform}(0, 6) \\
\epsilon & = 0.045 \\
\end{align*}
\]

Notice that all models fix the error parameter \(\epsilon\) to single value, which is chosen to be close to the average of each model’s maximum likelihood estimate for \(\epsilon\) over all models.

Estimates of marginal likelihoods were obtained by grid approximation on a grid size of 20 for each parameter (e.g., Kruschke, 2015, Chapter 10). Figure 7a shows the resulting Bayes factor approximations in favor of each model when compared to the RSA model. We see that all models are better than the baseline vanilla RSA model. The best model is GI. The Bayes factor in favor of GI when compared against the second best model, LI, is approximately 29; and when compared against the third best model, LU, it is ca. 95. From these results we can also give approximations of modeller’s posterior beliefs after conditioning unbiased priors over models with the observed data:

<table>
<thead>
<tr>
<th>model (M_i)</th>
<th>RSA</th>
<th>LU</th>
<th>LI</th>
<th>GI</th>
</tr>
</thead>
<tbody>
<tr>
<td>(P(M_i</td>
<td>D))</td>
<td>0</td>
<td>0.01</td>
<td>0.033</td>
</tr>
</tbody>
</table>

In words, if each model is a priori equally likely, the posterior probability, after seeing the data, of the GI model is around 0.956. The combined data from production and comprehension provide strong evidence in favor of the GI model, suggesting that, within the extensions of RSA-style pragmatic reasoning models, it is best to include the full range of grammatically generated implicature readings.

Since Bayes factors depend on priors over parameter values, Figure 7b additionally shows the results of model comparison using the Bayesian Information Criterion (BIC) (Schwarz, 1978), which relies on the maximum likelihood estimates of the parameters and is therefore independent of priors over parameter values. A model is better the lower its BIC score. Consequently, we retrieve the same ordinal result, as for Bayes factor comparison.
To understand the results better, we can look at production and comprehension data separately. For only production data, the Bayes factor for the GI model is about 30600; the LI model is absolutely no competition. If we look at comprehension data only, the Bayes factor in favor of the GI model is ca. 0.011, i.e., the LI model is roughly 100 times more likely \textit{a posteriori}, if we start from unbiased prior beliefs. Consequently, the main advantage of the GI model comes from its superior predictions for the production data, which outweigh its weaker predictions for comprehension.

Zooming in to look at each individual condition from production (i.e., the utterance choices at a single state) and comprehension (i.e., the average slider ratings for a single sentence), Figure 7c shows that the main evidence in favor of the GI model comes from its superior predictions for the production data in state \[\text{a} \]. Figure 8 plots the \textit{prior predictive distributions} for production data under both models. The prior predictive distribution gives the marginalized likelihood for each possible data observation. We see that in most conditions the prior predictives of models coincide, and that the main difference in prior predictions is indeed that the LI model underpredicts the choice of sentence “SS” and overpredicts choice of “SN” for condition \[\text{b} \]. The LI model makes better predictions for states \[\text{c} \] and \[\text{d} \] but, as becomes apparent from Figure 8, its predictive advantage over the GI model is less pronounced than GI’s advantage in condition \[\text{e} \]. The reason why the GI model predicts a high frequency of “SS” choices in state \[\text{f} \] is because it makes it possible that a rational speaker utters “SS” while intending a global reading. The global reading with \[\text{M} \] is not available to the LI model, but it serves to perfectly single out state \[\text{a} \] from all other states (see Table 1 in main paper) and so it is rational for a speaker to say “SS” and mean \[\text{a} \].

\[<\text{INSERT FIGURE 8 ABOUT HERE}>\]

6 Conclusions. We addressed the general question of how multiple ambiguity affects a speaker’s pragmatic choice of utterance and a listener’s interpretation. Focusing on implicature readings of complex sentences, we showcased theory-driven statistical modeling as a method of directly linking linguistic theory to empirical data. We formulated the \textit{Global Intentions} model, which assigns a conceptually plausible role to the full richness of meanings generated by grammatical approaches to scalar implicature inside of a Gricean model of probabilistic pragmatic reasoning (see
Figure 2). The GI model is not an innocuous synthesis of traditional Gricean and grammatical approaches to scalar implicatures. It is a new breed of its own. It puts general pragmatic reasoning first and embeds grammaticalism as one potential mechanism of generating conventionally associated readings for sentences. It thereby also provides a possible link hypothesis, i.e., a way of deriving precise quantitative predictions for empirical data for a wide range of cases. Statistical model comparison in terms of Bayes factors indeed suggests that the GI model might be empirically superior to the investigated alternatives.

Theory-driven statistical modeling of the kind executed here can inform linguistic theory in multiple ways. We might, for example, fix the GI model but compare different instantiations of grammaticalism embedded inside of it. Consider Gotzner and Romoli (2018)’s proposal of a different construction of the set of sentential alternatives Alt(S, p) than used here. Without going into the details of their account, the Bayes factor in favor of an GI model based on the simpler construction used here (see Appendix A), when compared to Gotzner and Romoli’s more elaborate alternative, is about 1530, suggesting that the simpler construction is empirically superior. This is not meant to be a decisive argument against Gotzner and Romoli’s approach, but as an example for how theory-driven modeling opens up new ways to quantitatively compare elaborate and fine-grained theoretical proposals in the light of experimental data.

The main reason for the GI model’s predictive success was identified as the availability of a reading \([SS]^{\text{M}} = \{\}\) of the sentence “SS” obtained from inserting an Exh-operator at matrix position. This provides suggestive evidence for the idea that the full set of grammatically-induced readings is needed. Indeed, the vanilla RSA model does not associate “SS” with \([\]\), because “SN” is always a better choice in state \([\]\) than “SS” for the speaker. This highlights an important implication of the GI model: it uses two different types of sentential alternatives. For one, there are what we could call grammatical sentential alternatives, which are used to derive the grammatically-derived readings of candidate utterances, as detailed in Appendix A. For example, by common construction “SN” is not a grammatical alternative to “SS”, so that the reading \([SS]^{\text{M}} = \{\}\) can be derived. On the other hand, there is the set of utterance alternatives, which we here equated, naturally, with the set of utterances offered in the experimental design (exactly like Potts et al., 2016). The set of utterances alternatives to “SS” does include “SN”. It is not necessary to equate these two sets of alternatives, because it is conceivable, for example, that
the grammatical module that generates readings for sentences is (fairly) encapsulated from context. Utterance alternatives might be much more flexibly adapted in the context, as they provide the set of actual utterances the speaker is currently aware of in her choice. Future work should explore the potential generated by this distinction of grammatical vs. utterance alternatives further.

Future work should also scrutinize different models of pragmatic reasoning. For example, all of the models considered here assume that speakers do not have a veridical view of the listener’s interpretation; they base their choices on a simple Gricean heuristic of preferring semantically stronger utterances; they do not engage actively in audience design. Clearly, exploring other models of pragmatic reasoning in connection with grammatically generated implicature readings, and testing them in experimental designs which take the possibility of active audience design into account, is clearly a worthwhile enterprise for future research.

Although our focus here was conceptual and methodological, more work is clearly necessary on the empirical side as well. The scope of our experimental manipulations is admittedly rather small. Recent experimental papers on scalar implicatures (e.g. Chemla and Spector, 2011; Franke, Schlotterbeck, and Augurzky, 2017) in complex sentences have focused attention on scalar implicature triggers in non-monotonic environments like in (7).

(7) Exactly one of the aliens drank some of its water.

Our design was more restricted in order to keep the set of alternative utterances for global Gricean reasoning confined to what is hopefully a maximally uncontroversial selection. For instance, the modeling of Potts et al. implicitly assumes that listeners also include 7 as a speaker’s alternative utterance during the interpretation of any nested Aristotelian. To sidestep such controversial assumptions, at least for the time being, this paper’s experimental set-up is deliberately minimal. Nonetheless, our data offer enough grip for a rational experimenter to quite substantially update their beliefs about which model could likely be true.

A Grammatically-generated implicature readings. When \( Exh \) applies directly to a quantifier \( Q \), the resulting meaning \( Exh(Q) \) is that of \( Q \) conjoined with the negation of all lexical alternatives, which are strictly stronger, i.e., which entail \( Q \) but are not entailed by it. As usual,
we assume that *some* and *all* are lexical alternatives of each other. We also assume, following standard practice, that *none* and *not all* are, since we analyze *none* as *not some* (e.g. Levinson, 2000, p. 80). Consequently, the effect of applying *Exh* to *Q* will be vacuous for *none* and *all*, but *Exh(some)* gives an *in situ* enrichment to “some but not all.” For example, under a parse I the sentence “SS” gets the reading $\llbracket SS \rrbracket^I$ in 8 that will be true in any state whose pictorial representation includes a half-full glass (see Table 1).

(8) Some of the aliens drank some but not all of their water.

When *Exh* applies in matrix position to a sentence *S* it takes the meaning of *S* obtained from insertions of *Exh* below matrix position, and conjoins that meaning with the negation of suitable sentential alternatives. Let *S* be a sentence and *p* its parse. Let $\llbracket S \rrbracket^{p-\neg}$ be the meaning of *S* under the parse $p_{\neg-\neg}$, which is like *p* with respect to outer and inner quantifier, but has no *Exh* in matrix position. When *Exh* applies in matrix position to *S* whose parse is *p*, the meaning of *Exh(S, p)* is obtained by conjoining $\llbracket S \rrbracket^{p-\neg}$ of *S* with the negation of all relevant sentential alternatives $Alt(S, p)$ of *S*, if this operation is non-contradictory (but see, e.g., Gotzner and Romoli, 2018, for discussion of more elaborate definitions):

$$\llbracket Exh(S, p) \rrbracket = \begin{cases} \llbracket S \rrbracket^{p-\neg} \cap_{S' \in Alt(S, p)} \llbracket \text{not } S' \rrbracket^{1\neg t} & \text{if non-contradictory} \\ \llbracket S \rrbracket^{p-\neg} & \text{otherwise.} \end{cases}$$

We will here assume that the set $Alt(S, p)$ is obtained in two steps, a generation and a filtering step. First, we generate a set $Alt^*(S)$ of potential alternatives by blindly replacing any occurrence of a scalar item in *S* with any of its lexical alternatives, as defined above. A second step then filters out sentences from $Alt^*(S)$ which do not entail *S*.

$$Alt^*(S) = \{ S' \mid S' \text{ is derived from } S \text{ by replacements of lexical alternatives} \}$$

$$Alt(S, p) = \{ S' \in Alt^*(S, p) \mid \llbracket S' \rrbracket^{1\neg t} \subseteq \llbracket S \rrbracket^{p-\neg} \}$$

For example, consider the sentence “SS” under a parse $M$. By replacing occurrences of *some* with its lexical alternatives, we obtain $Alt^*(SS) = \{ SS, AS, SA, AA \}$. Of these “SS” is filtered out. Applying matrix-*Exh* consequently yields the conjunction of the negation of all sentences “AS”,...
“SA” and “AA” with the literal meaning of “SS” as paragraphed is 9. This is a very strong reading, as the only world state that makes this reading true is $\mathbf{w}$. 

(9) Some of the aliens drank some of their water, and it is not true that some aliens drank all or that all drank some.

As a final example, consider again the sentence “SS” but now under a parse MI. We first consider $\llbracket SS \rrbracket^1$, which gives the set of situations in which there is at least one half-full glass. The set $\text{Alt}^*(SS)$ is as before, but $\text{Alt}(S, p)$ is now empty because none of the sentences in $\text{Alt}^*(SS)$ are, under parse 1i t, strictly stronger than “SS” under parse 1. Consequently, additional insertion of $Exh$ in matrix position is vacuous in this case.
References


Potts, Christopher et al. (2016). Embedded implicatures as pragmatic inferences under compositional lexical uncertainty. *Journal of semantics* 33.4, pp. 755–802.


1Materials, data and modeling code can be retrieved from https://osf.io/fuary/.

2The vanilla RSA model and all other models considered here assume that the speaker knows the true world state; an assumption which is arguably warranted by the design of the experiment described in Section 4. See Goodman and Stuhlmüller (2013) and Herbstritt and Franke (2019) for extensions to cases with (higher-order) uncertain speakers.

3Here, $\delta_{\text{boolean}}$ is the delta function which returns 1 if the supplied boolean expression is true and 0 otherwise. The symbol “$\propto$” (for “proportional to”) allows us to define probability distributions more compactly by leaving the normalizing constant implicit: writing $P(x) \propto F(x)$ for some function $F(x)$ with domain $X$ is shorthand for $P(x) = \frac{F(x)}{\sum_{x' \in X} F(x')}$. 

4The speaker’s choice of expression, defined in Equation [2], is derived from two assumptions (see, e.g., Franke and Jäger, 2016, for details). Firstly, the speaker seeks to minimize the distance (in terms of Kullback-Leibler divergence) between her probabilistic belief and that of the literal listener. Secondly, the speaker tends towards optimal choices probabilistically (modelled with a so-called soft-max function with rationality parameter $\alpha$).

5Equation [2] can be rewritten like so: $P_S(u \mid t; \alpha) \propto \exp(\alpha \cdot \log P_{LL}(t \mid u)) = [P_{LL}(t \mid u)]^\alpha$, which by definition of the literal listener expands to: $\left[ \frac{P_{LL}(t \mid u)}{\sum_{t'} P_{LL}(t' \mid u)} \right]^\alpha$. Since $P(t) = P(t')$ for all $t, t'$ the prior term cancels out and we retrieve: $\left[ \frac{\delta_{u=t}}{\sum_{t'} \delta_{u=t'}} \right]^\alpha$.

6Notice that the LI model, and the following GI model, assumes that the literal listener “miraculously” recognizes the speaker’s intended meaning. Since the literal listener is a technical construct, the implication is that speakers are not assumed to take into account a realistic model of their interlocutor. Speakers are modelled as using a (myopic, conventional) heuristic manner of choosing utterances and intended meanings, which, however, can then be recognized by pragmatic listeners to retrieve the speaker’s intended meaning.

7The only effect of considering not all a lexical alternative to none is that the sentence “NN” gets an additional reading when Exh applies in matrix position. But this has no noteworthy effect on anything of relevance to this paper’s main concerns.
LIST OF FIGURES

Examples of world states relevant for the interpretation of nested Aristotelians. Similar pictures are used in the experiment reported in Section 4. 30

2 Schematic representation of main conceptual differences in the speaker production part of the four models compared here. 31

3 Predictions of the production rules of different models for parameter value $\alpha = 5$. Rows represent states, columns utterances. An entry in a cell gives the probability assigned to the speaker’s choice of the column-utterance when trying to communicate the row-state. 32

4 Illustrations of experimental trials. (NB: in interpretation trials the pictures where actually shown on the left of the text and slider bar, but are here shown vertically stacked for ease of presentation.) 33

5 Results from production trials. The plot shows the frequencies with which each sentence (columns) was selected for each state (rows). The black outlines indicate cases where a sentence is true in a given state under at least one of the pragmatic readings from Table 1. 34

6 Results from interpretation trials. The plot shows that the average normalized slider ratings (see main text) that participants assigned to each state (columns) as an interpretation for a given sentence (row). The black outlines show truth under at least one reading. 35

7 Results from model comparison 36

8 Prior predictive distribution for the production data. Black vertical bars indicate the observed counts for each condition. 37
Figure 1: Examples of world states relevant for the interpretation of nested Aristotelians. Similar pictures are used in the experiment reported in Section 4.
Vanilla RSA
literal semantics for all speakers

Lexical Uncertainty
different speakers may have different lexical meanings
each speaker has a single fixed lexical meaning

## Internal meaning representations

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>${ [\text{AS}]^{\text{lit}} }$</td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>${ [\text{SS}]^{\text{lit}} }$</td>
<td></td>
</tr>
<tr>
<td>$\ldots$</td>
<td>${ [\text{...}]^{\text{lit}} }$</td>
<td></td>
</tr>
</tbody>
</table>

Speaker’s choice

## Lexical Intentions

all speakers aware of all grammar-induced lexical enrichments
utterances come with a speaker-intended lexical meaning

## Global Intentions

all speakers aware of all grammar-induced sentence readings
utterances come with a speaker-intended pragmatic reading

## Internal meaning representations

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AS</td>
<td>${ [\text{AS}]^{\text{lit}}, [\text{AS}]^{\text{MOI}}, [\text{AS}]^{\text{I}}, \ldots }$</td>
<td></td>
</tr>
<tr>
<td>SS</td>
<td>${ [\text{SS}]^{\text{lit}}, [\text{SS}]^{\text{MOI}}, [\text{SS}]^{\text{I}}, \ldots }$</td>
<td></td>
</tr>
<tr>
<td>$\ldots$</td>
<td>${ [\text{...}]^{\text{lit}}, [\text{...}]^{\text{MOI}}, [\text{...}]^{\text{I}}, \ldots }$</td>
<td></td>
</tr>
</tbody>
</table>

Speaker’s choice

**Figure 2:** Schematic representation of main conceptual differences in the speaker production part of the four models compared here.
Figure 3: Predictions of the production rules of different models for parameter value $\alpha = 5$. Rows represent states, columns utterances. An entry in a cell gives the probability assigned to the speaker’s choice of the column-utterance when trying to communicate the row-state.
Figure 4: Illustrations of experimental trials. (NB: in interpretation trials the pictures were actually shown on the left of the text and slider bar, but are here shown vertically stacked for ease of presentation.)
Figure 5: Results from production trials. The plot shows the frequencies with which each sentence (columns) was selected for each state (rows). The black outlines indicate cases where a sentence is true in a given state under at least one of the pragmatic readings from Table 1.
Figure 6: Results from interpretation trials. The plot shows that the average normalized slider ratings (see main text) that participants assigned to each state (columns) as an interpretation for a given sentence (row). The black outlines show truth under at least one reading.
(a) Bayes factors in favor of a model when compared to RSA, plotted on a log scale.

(b) Model comparison based on BICs (with maximum likelihood estimates for parameters)

<table>
<thead>
<tr>
<th>parameter</th>
<th>model</th>
<th>$\alpha$</th>
<th>$c$</th>
<th>$w$</th>
<th>BIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSA</td>
<td>1.43</td>
<td>1.56</td>
<td>43.85</td>
<td>259.73</td>
<td></td>
</tr>
<tr>
<td>LU</td>
<td>0.94</td>
<td>2.2</td>
<td>39.75</td>
<td>202.06</td>
<td></td>
</tr>
<tr>
<td>LI</td>
<td>1.27</td>
<td>1.52</td>
<td>41.22</td>
<td>200.16</td>
<td></td>
</tr>
<tr>
<td>GI</td>
<td>1.3</td>
<td>1.71</td>
<td>31.87</td>
<td>195.69</td>
<td></td>
</tr>
</tbody>
</table>

(c) Bayes factors in favor of GI over LI per production (left) and comprehension condition (left)

**Figure 7:** Results from model comparison
Figure 8: Prior predictive distribution for the production data. Black vertical bars indicate the observed counts for each condition.
LIST OF TABLE Readings produced by the instantiation of grammaticalism described in Appendix A. For each nested Aristotelian, the table shows whether this sentences is true (1) or false (0) in each world state under a given parse. E.g., the first line of the table gives the truth conditions of sentence NN for parses without Exh in matrix position, all of which happen to be equivalent.
Table 1: Readings produced by the instantiation of grammaticalism described in Appendix A. For each nested Aristotelian, the table shows whether this sentences is true (1) or false (0) in each world state under a given parse. E.g., the first line of the table gives the truth conditions of sentence NN for parses without *Exh* in matrix position, all of which happen to be equivalent.