Pragmatic Alternatives and Scalar Implicature

Brandon Waldon

LINGUIST 230B Final Project - June 13, 2018

1 Introduction

This literature review is an attempt to identify, recapitulate, and critically assess previous work in the pragmatics literature on questions related to *pragmatic alternatives* in scalar implicature. By what mechanism(s) is/are alternatives to a given utterance computed? How are these alternatives activated in scalar inference? In what ways have researchers in pragmatics attempted to circumvent the so-called *symmetry problem*? After a brief discussion of the import of this topic for both Gricean and grammaticist theories of scalar implicature, I will review past theories of pragmatic alternatives, including theories centered around scalar conventions (Horn 1972; Horn and Abbott 2012; Gazdar 1979) as well as theories which rely chiefly on structurally-defined principles (Katzir 2007; Fox and Katzir 2011). A discussion of the status of alternatives in Rational Speech Act approaches to scalar implicature (Goodman and Stuhlmüller 2013) follows. In the final section, I review recent psycholinguistic research on the computation and activation of pragmatic alternatives in scalar inference.

2 The role of alternatives in scalar implicature

In what follows, I discuss how two existing strands of theory of scalar implicature - Gricean and grammaticist theories - differ as to what exactly constitutes scalar implicature as well as to how scalar inferences come about. A definition of scalar implicature which is neutral with respect to the Gricean/grammaticist divide is offered by Potts (2015), who makes clear the centrality of 'alternatives' regardless of the choice we make between the two strands of theory:

(1) An utterance U conveys a scalar conversational implicature iff there are alternative utterances U' that are at least as relevant as U in the discourse and that are communicatively stronger[1] than U. (The content of this implicature will depend on the context, the nature of the utterance competition, and other pragmatic factors.) [Potts 2015: 179]

[1]Note that this definition leaves underspecified the exact type of communicative 'strength' that we care about in the case of scalar implicature. As discussed below, many theorists have operationalized strength as logical entailment, but there are competing proposals (cf. Hirschberg 1985).

On theories which follow the framework of Grice (1975) in representing conversational implicature in general as a confluence of linguistic competence and social cognition, a scalar implicature is conveyed when a hearer engages in counterfactual reasoning about possible stronger alternative utterances a speaker could have said but did not. This counterfactual reasoning depends on the hearer's belief that the speaker is behaving cooperatively - as well as on the hearer's belief that the speaker is aware that the hearer believes the speaker to be cooperative. The following Gricean-inspired analysis of scalar implicature is adapted from Potts (2015) and Sebastian Schuster's presentation on implicature in Stanford's LINGUIST 230A course (Februrary 13, 2018):

(2)
- Let q = the proposition denoted by *John ate some of the cookies*
- Assume a speaker Ellen asserts *John ate some of the cookies* to a listener Kyle

i. *Contextual premise:* Ellen and Kyle know that John is able to eat any portion of some contextually-salient batch of cookies.

ii. *Contextual premise:* Ellen is maximally-informed with respect to how many cookies John ate.

iii. Assume that Ellen is cooperative and obeys the Gricean maxims of quality and quantity.[2]

iv. Then she will assert what is maximally relevant, informative and true.

v. An alternative proposition p (*John ate all of the cookies*) is more informative and (at least) as relevant in this context than is q.

vi. Therefore, Ellen must be experiencing a clash between the maxims: she cannot assert p because she lacks sufficient evidence for p.

vii. By (ii), she must lack evidence because it is false.

Implicature: it is not the case that John ate all of the cookies.

The aforementioned analysis of scalar implicature contrasts with more recent grammaticist analyses, whereby scalar implicature is associated with an inference which arises from a semantic operator whose input is a proposition p and whose output is a proposition computed from p and a set of alternatives to p. The following analysis of scalar implicature is adapted from one proponent of the grammaticist view, Chierchia (2006). The analysis must first define a function $||\cdot||^{ALT}$ which can map propositions to sets of propositions[3]; for example,

(3) $||$John ate some of the cookies$||^{ALT}$ = {*John ate some of the cookies*, *John ate all of the cookies*}

[2]Grice's maxim of quantity: "Make your contribution as informative as is required. Do not say more than is required." The maxim of quality: "Contribute only what you know to be true. Do not say false things. Do not say things for which you lack evidence."

[3]One account offered by Chierchia (2006), $||\cdot||^{ALT}$ is the function that "associates any item [not necessarily a proposition] with its scalar alternatives" (Chierchia 2006: 546). The ability to specify alternatives for subsentential items - and thus to recursively define propositional alternatives - is necessary to capture embedded scalar inferences within Chierchia (2006)'s framework. But crucially, Chierchia leaves underspecified how exactly the alternatives of any item are computed.

The aforementioned semantic operator takes as its input a proposition and returns a proposition which is a conjunction of the input proposition and the statement that the input proposition is the strongest true alternative to itself:

(4) $\mathrm{O}_C(\mathrm{p}) = p \wedge \forall q \in C[q \rightarrow p \subseteq_c q]$, where C = the output of $||\mathrm{p}||^{ALT}$

Given how we specify the semantic operator O_C in (4) and how we specify the alternatives to *John ate some of the cookies* in (3), the analysis predicts that *John ate some of the cookies* under the scope of O_C should give rise to the inference that it is not the case that John ate all of the cookies:

(5) O_C(*John ate some of the cookies*)
$= \Big[$ *John ate some of the cookies* $\wedge$
$\forall p : p \in ||$John ate some of the cookies$||^{ALT} \wedge \mathrm{p} \rightarrow$
John ate some of the cookies $\subseteq \mathrm{p} \Big]$
Pragmatically-enriched meaning: *John ate some of the cookies* $\wedge \neg$(*John ate all of the cookies*)

One commonality between both approaches is that neither, on its own, is a fully-specified theory of how to compute the alternatives of a given utterance. While both theories claim that a scalar implicature will be conveyed by an utterance in case there are pragmatically-active stronger alternatives to that utterance, this alone is not sufficient to avoid what Fox and Katzir (2011), following von Fintel (2002), call the *symmetry problem*. The symmetry problem is illustrated as follows (discussion adapted from Fox and Katzir 2011):

(6)
- Let q = the proposition denoted by *John ate some of the cookies*. Let p_1 = the proposition denoted by *John ate all of the cookies*, which is stronger with respect to logical entailment than q.
- Let p_2 = the proposition denoted by *John ate some but not all of the cookies*, which is also stronger with respect to logical entailment than q.
- $\mathrm{q} \rightarrow p_1 \vee p_2$.
- p_1 and p_2 contradict one another.
- q seems to implicate $\neg p_1$, but this is only possible if p_2 is not activated as an alternative to q.

By virtue of the fact that p_1 and p_2 contradict one another, we call p_1 and p_2 *symmetric alternatives* of q. The symmetry problem is twofold. First, we want a way to favor p_1 as an alternative to q over p_2: $q \wedge \neg p_1$ corresponds to the desired 'some but not all' strengthening of q in the context of utterance provided in, e.g., (2), while $q \wedge \neg p_2$ expresses a proposition that is truth-conditionally equivalent to *John ate all of the cookies* (clearly not the desired strengthening of q in the context of 2)[4]. Second, the utterance of q cannot implicate both

[4]Note furthermore that there is nothing inherent to scalar implicature as we have defined it in (1) that seems to favor p_1 as an alternative over p_2; for example, p_1 being more relevant than p_2 in the context of utterance. von Fintel (2002) [`http://web.mit.edu/24.954/www/files/24.954.lecturenotes.pdf`] remarks that in a context where *How many of the boys are at the party?* is under discussion, *all of the boys are at the party* seems to be just as relevant as *some but not all of the boys are at the party.*

$\neg p_1 \wedge \neg p_2$ because this conjunction is a contradiction. The avoidance of these two issues - conveying illicit scalar implicatures and conveying contradictory meanings - is a main desideratum of any theory of alternatives. That is, a successful theory of alternatives **breaks** symmetry in a manner which delivers the right predictions regarding the nature of the scalar implicature(s) conveyed by a given utterance.

The next two sections summarize ways in which pragmaticists have attempted to break symmetry by positing that any utterance U is associated with a specified, finite set of alternatives U' which may be activated in pragmatic competition with U. For the case of *some*, *some but not all* is simply blocked as a possible alternative to *some*. I then review a more recent formalism of pragmatic competence - the Rational Speech Act model - which affords different options for breaking symmetry.

3 Computing alternatives from scales

Horn (1972) proposes an account of scalar implicature whereby the calculation of alternatives is driven by conventionalized scalar relationships between lexical items. According to Horn, these scales are always ordered according to the relative logical strength of the lexical items on the scale. Utterance alternatives are then computed by taking the original utterance and replacing an lexical item from that utterance with one of that lexical item's stronger scale-mates. This provides a straightforward way of understanding why *John ate all of the cookies* may be an alternative to *John ate some of the cookies*: an item in the original utterance, *some*, is replaced with another item, *all*, which stands in an asymmetric logical entailment relationship to *some*. That is, Horn proposes that <some, all> is a conventionalized lexical scale from which alternatives may be computed.

Horn proposes several other lexical scales which are similar to the <some, all> scale in that they represent an ordering of logical entailment. These scales include the scale of cardinal numerals, <or, and>, <might, must>, <pretty, beautiful>, and <warm, hot>. Importantly, <some, some but not all> is not taken by Horn to be a scale, even though the two items stand in an asymmetric logical entailment relationship to one another. The same can be said of potentially problematic scalar orderings such as <or, either-or> or <N, exactly N> (for any cardinal number N), all of which would lead to a similar symmetry problem as outlined above. Symmetry is thus broken on the grounds that only select alternatives are activated in pragmatic inference - and the alternatives which we happen to activate lead to empirically satisfying results (*some* implicates not *all*, *or* implicates not *and* and does not implicate not *either-or*, etc).

Though this may at first seem to have the air of stipulation, Horn provides some support for the idea that while, for example, <some, all> has conventionalized as a lexical scale, <some, some but not all> has not. Evidence comes from focus or contrastive discourse environments of the form in (7), where a scalar inference of the form $\neg Y$ is canceled, suspended, or reinforced in the context of asserting X (discourse frames adapted from Horn and Abbott 2012 as cited in Collins 2017):

(7) i. not only X but Y

ii. X and for all I know Y

iii. X if not Y

iv. X or even Y

v. X, indeed/in fact Y

vi. not even X, let alone Y

vii. Y, or at least X

For any two items X and Y, their acceptability in environments such as those in (7) is evidence that those two items are, to borrow the language of Horn and Abbott (2012), "natural paradigmatic alternatives" (334). The acceptability of (8) and (9) supports this analysis for *some* with respect to *many* and *all*, while (10) points against a similar analysis for *some* and *some but not all*:

(8) i. not only some but all

ii. some and for all I know all

iii. some if not all

iv. some or even all

v. some, indeed/in fact all

vi. not even some, let alone all

vii. all, or at least some

(9) i. not only some but many

ii. some and for all I know many

iii. some if not many

iv. some or even many

v. some, indeed/in fact many

vi. not even some, let alone many

vii. many, or at least some

(10)# i. not only some but some but not all

ii. some and for all I know some but not all

iii. some if not some but not all

iv. some or even some but not all

v. some, indeed/in fact some but not all

vi. not even some, let alone some but not all

vii. some but not all, or at least some

Application of this test with *or/either-or* as well as *N/exactly-N* similarly results in infelicity, as in (10).

One major revision to this account comes from Gazdar (1979), who argues that the scales from which alternatives are computed should rather be understood as conventionalized scales

of *semantic representations* rather than lexical scales. One of Gazdar's key observations was that by positing scales at the level of the lexicon rather than at a deeper level of representation, the analysis put forward by Horn (1972) fails to account for the *nondetachability* of scalar inferences: independent of the individual lexical choices one makes in forming an utterance, we expect regularity in the implicatures that semantically-equivalent utterances will give rise to. For example, in (11), both response (a) and response (b) may give rise to the implicature that for all A knows, *from both courses* is not certain (example from Hirschberg 1985: 70, her example 73):

(11) A: I think you would have to get it from the instructor for the course...
B: For which course?
a. A: Possibly from both courses.
b. A: Maybe from both courses.

If scales are lexical, as Horn (1972) takes them to be, then we miss the generalization that non upper-bounded expressions of possibility such as *possibly* and *maybe* are associated with inferences of non-maximal certainty. Rather, we are put in a position of having to stipulate that *possibly* and *maybe* each reside on a lexical scale with *certain*. Gazdar's solution remedies this problem, but opens a new one: at *which level* of semantic representation are the scales derived? Gazdar is hesitant to claim that scales are derived at the level of semantic interpretation, on the grounds that conversational implicatures in general may vary, even between truth-conditionally equivalent sentences. For example, a sentence involving the verb phrase connective *and* may give rise to different temporal conversational implicatures depending on the order of the conjuncts. Thus, for Gazdar, an intermediate notion of semantic representation - somewhere between semantic interpretation and lexical representation (where, for example, information such as order of verb phrase conjuncts is preserved) - is the level at which scales should be derived.

However, Hirschberg (1985) observes that in a number of cases (including upper-bounding inferences with *some*, *or*, and cardinal numbers), truth-conditional semantic interpretation is an acceptable level at which to compute utterance alternatives. Moreover, temporal conversational implicatures involving *and* are qualitatively different from scalar implicatures in that the latter, but not the former, presumably relies on a clash in the Gricean maxims of quantity and quality in order for the inference to come about: the temporal *and*-implicature is presumably a manner implicature, and not a quantity one. While Hirschberg concedes that "a general representation of conversational implicature must accommodate conversational implicatures that rely upon the Maxim of Manner for their interpretation" (70), she remarks that "Gazdar's specific claim about scalar quantity implicature" - in particular the need to compute scalar and non-scalar alternatives at the same level of semantic representation - "is unclear" (ibid).

Hirschberg's second observation with respect to the analysis put forth by Gazdar (1979) is that he, like Horn, must continue to operate on the assumption that scales are somehow "simply given" (70). That is, there remains a lack of a fully descriptively-adequate analysis of where scales come from and of how to characterize their constraints. Though logical entailment appears to be an operable notion in scalar inference, Hirschberg observes that this not a sufficient - or even necessary - condition for deriving scales. Hirschberg offers a

number of semantic orderings which support scalar inferences and notes that "[w]hile many might be defined by some notion of entailment, a number cannot" (84).

For example, orderings such as <dating, married, engaged>, <sick, dying>, <condone, promote> (e.g. violence), and <misdemeanor, felony> all appear to be pragmatically-active scales, which is supported by the fact that they licit in the discourse environments from (7).[5] However, there are no logical entailment relations **between** any of the items in these orderings. Rather, Hirschberg argues that a more general and more context-sensitive notion of communicative strength is needed.

Horn (1989), cited in Matsumoto (1995), presents an analysis specifically designed to constrain scales such that orderings such as <some, some but not all>, <or, either-or> and <N, exactly-N> are ruled out for the purposes of pragmatic inference. The constraint proposed by Horn builds off an observation that linguistic expressions may have one of three monotonicity properties: they may be upward monotone, as in the case of the generalized quantifier *some men*; downward monotone, as in the case of *no men*; or nonmonotone, as in the case of *exactly three men*. Horn (1989) introduces a constraint on alternatives whereby alternatives may not be calculated from scales where the expressions are nonuniform in their monotonicity behavior; moreover, nonmonotone expressions may never participate in scales.

Matsumoto caches out Horn (1989)'s observation as a more generalized conversational condition on pragmatic competition, as articulated below (from Matsumoto 1995: 25):

(12) Conversational Condition: The choice of W instead of S must not be attributed to the observance of any information-selecting Maxim of Conversation other than the Quality Maxims and the Quantity-1 Maxim (i.e., the Maxims of Quantity-2, Relation, and Obscurity Avoidance, etc.).[6]

As illustration of this principle in action, Matsumoto provides examples such as the following (taken from Matsumoto 1995: 45-46):

(13) a. "Three men came."
b. '(The speaker believes) it is not the case that exactly three men came.'
c. '(The speaker believes) exactly three men came.'
d. "Exactly three men came."

(14) a. "Bill met John on the way home."
(where the speaker knows that meeting Mary is of concern to the hearer)

[5]Hirschberg, for her part, takes issue with Horn's diagnostics for scales which support scalar inference. As she notes, Horn's discourse environments "will identify scales such as *only vote/ did vote* from [(1-a)] and *bald/exists* from [(1-b)]", which are clearly not pragmatically-active orderings:

(1) a. Only Muriel voted for Humphrey if even she did.
b. The king of France is bald if there is a king of France.

Of course, if we continue to assume that scales must be orderings of logical entailment, then the data in (1) do not pose a problem for Horn.

[6]Matsumoto makes a distinction between two quantity maxims: a Quantity-1 maxim, whereby speakers are directed to "make your contribution as informative as is required", and a Quantity-2 maxim: "do not make your contribution more than is required in the context of the exchange". That is, the Quantity-1 and Quantity-2 maxim taken together simply amount to Grice's original quantity maxim.

b. '(The speaker believes) is not the case that Bill met John but not Mary on the way home.'

c. '(The speaker believes) Bill met only John on the way home.'

d. "Bill met John but not Mary (or any other person) on the way home."

Matusmoto remarks that while the (d) sentences asymmetrically entail the (a) sentences above and are thus stronger than those sentences, the (b) sentences are not available implicatures. Rather, it seems that only the (c) sentences are available inferences. Rather than stipulate a monotonicity constraint on scales in the vein of Horn (1989), Matsumoto offers the following explanation: in asserting (13-a), the only utterances which may act as alternatives are those which themselves would license quantity implicatures had they been asserted (e.g. *four men came*, *five men came*, etc.). (13-d) is not such a candidate, as *exactly N* does not license scalar implicatures. A similar pattern of reasoning holds for (14): *Bill met Mary* may license a quantity implicature, unlike *Bill met John but not Mary*, hence why the former, but not the latter, may be activated as an alternative to the original utterance.

4 Computing alternatives via structurally-defined principles

One commonality shared among *some* with respect to *some but not all*, *or* with respect to *either-or*, and *N* with respect to *exactly N* is that the latter is morphosyntactically more complex than the former. It may thus be tempting to rule out these pairings as pragmatically-active contrasts simply on the basis that alternatives may not be structurally more complex than the original utterance. However, Matsumoto (1995) presents data which problematizes this view, as in the following example (taken from Matsumoto 1995: 44):

(15) a. It was warm yesterday, and it is a little bit more than warm today.

b. It was a little bit more than warm yesterday, and it is a little bit more than warm today.

In the below example, the (b) sentence is intuitively more complex than the (a) sentence; but nonetheless, (b) appears to be a salient alternative to (a), in that (a) appears to give rise to the inference that it is not the case that it was a little bit more than warm yesterday. Matsumoto takes this as evidence that alternatives cannot be ruled out due to relative structural complextity.

Katzir (2007) responds in turn that a theory of alternatives computed from scales - even when those scales are constrained following Matsumoto (1995) or Horn (1989) - appears to yield inadequate empirical coverage. In support of this claim, Katzir first offers data such as the following as evidence that a scale-based account undergenerates (examples from Katzir 2007: 677):

(16) a. I doubt that exactly three semanticists will sit in the audience.

b. I doubt that three semanticists will sit in the audience.

(17) a. If we meet John but not Mary it will be strange.

b. If we meet John it will be strange.

(18) a. Everyone who loves John but not Mary is an idiot.

b. Everyone who loves John is an idiot.

According to Katzir (2007), in (16)-(18), the assertion of the (a) sentence seems to give rise to the inference that the (b) sentence is unassertable (if not necessarily untrue). However, the scales such as <three, exactly three> and <John, John but not Mary> are ruled out on either Matsumoto (1995)'s or Horn (1989)'s analyses. Moreover, Katzir offers data such as the following to suggest that a scale-based account sometimes overpredicts the prevalance of scalar inferences; for example, the (a) sentences below do not implicate the negation of the respective (b) sentence (examples from Katzir 2007: 684):

(19) a. A man came to every party.

b. A tall man came to every party.

(20) a. Each reporter talked to a candidate.

b. Each reporter talked to a candidate who sang.

(21) a. John is sure that many dogs will be sold.

b. John is sure that many dogs with long tails will be sold.

In (19) through (21), we have examples of potential orderings - including <a man, a tall man>, <a candidate, a candidate who sang>, <many dogs, many dogs with long tails> - which are neither blocked by the uniformity of monotonicity principle proposed by Horn (1989) nor by the Conversational Condition proposed by Matsumoto (1995). However, as Chris Potts (p.c.) points out, it does not necessarily follow that the analyses put forward by these authors predict illicit scalar inferences in these cases, provided that we assume listeners reason about only those alternatives that are at least as relevant in context as the original utterance (and it is easy to imagine a context where, for example, interlocutors are interested in communicating whether and how many men came to a party, with no regard for the tallness of those men). In fact, on Katzir's own account, alternatives to a given utterance should only be those that are *weakly assertable* in context, where weak assertability is defined partly in terms of discourse relevance. So it is not clear whether (19) through (21) motivate a new analysis of alternatives.

Katzir (2007), for his part, takes the data in (16) through (21) to be a challenge for existing scale-based accounts of alternative computation. Instead, he proposes the following principles from which alternatives are calculated:

(22)
- SUBSTITUTION SOURCE: Let ϕ be a parse tree. The substitution source for ϕ, written as $\mathrm{L}(\phi)$, is the union of the lexicon of the language with the set of all subtrees of ϕ.
- STRUCTURAL COMPLEXITY: Let ϕ, ψ be parse trees. If we can transform ϕ into ψ by a finite series of deletions, contractions, and replacements of constituents in ϕ with constituents of the same category taken from $\mathrm{L}(\phi)$, we will write $\psi \lesssim \phi$ If $\psi \lesssim \phi$ and $\phi \lesssim \psi$ we will write $\psi \sim \phi$ If $\psi \lesssim \phi$ but not $\phi \lesssim \psi$ will write $\psi < \phi$.

- STRUCTURAL ALTERNATIVES: Let ϕ be a parse tree. The set of structural alternatives for ϕ, written as $A_{str}(\phi)$,, is defined as $A_{str}(\phi) := \{\phi' : \phi' \lesssim \phi\}$

In short, Katzir's analysis allows for the alternatives of an utterance to be those structures of equal or lesser structural complexity than the original utterance. Katzir's definition of a *SUBSTITUTION SOURCE* provides that alternatives derived from substituting material in the original utterance with other material from that same utterance is a neutral operation with regards to structural complexity. This circumvents the original problem posed by Matsumoto (1995) in example (15): because *a little bit more than warm* is a subtree of (15-a), it is in the substitution source with respect to that sentence. This allows for the construction of (15-b) as an alternative, where *a little bit more than warm* simply replaces *warm* in the first conjunct of the sentence.[7]

5 Breaking symmetry with probabilistic pragmatics

In the accounts discussed above, symmetry is broken by restricting the set of alternatives against which an utterance is compared. In this section, I review a probabilistic formalism of pragmatic competence - the Rational Speech Act model - and discuss how it affords a wider range of options for symmetry breaking.

As it is presented by Goodman and Stuhlmüller (2013) in those authors' analysis of scalar implicature, the Rational Speech Act model is an iterative, probabilistic model of pragmatic competence in which utterance interpretation is modeled as a distribution of possible meanings given an observation of that utterance. Just as with more traditional Gricean accounts, scalar implicature in the RSA framework is, at its core, the product of counterfactual reasoning about alternative utterances that a speaker might produce (but does not, on the assumption that the speaker is a cooperative interlocutor). However, the RSA framework explicitly models cooperative interlocutors as agents whose language production and comprehension is a function of Bayesian probabilistic inference regarding other interlocutors' expected behavior in a discourse context.

Specifically, in the RSA framework the probability with which L_1 concludes meaning s from an utterance u depends upon a prior probability distribution of potential states of the world P_w, and upon reasoning about the communicative behavior of a speaker S_1. S_1 in turn is modeled as a continuous probabilistic distribution over possible utterances given an intended meaning the speaker intends to communicate. This distribution is sensitive to a rationality parameter α, the production cost C of potential utterances, and the informativeness of the utterance, quantified via a representation of a literal listener L_0 whose interpretation of an utterance is in turn a function of that utterance's truth conditional semantics $[[u]](s)$ and her prior expectations about the state of the world $P_w(s)$.

(23) $P_{L_1}(s|u) \propto P_{S_1}(u|s) * P_w(s)$
$P_{S_1}(u|s) \propto exp(\alpha(log(L_0(s|u)) - C(u)))$
$P_{L_0}(s|u) \propto [[u]](s) * P_w(s)$

[7]See Fox and Katzir (2011) for an attempt to extend this theory to one of alternative computation in focus constructions.

One consequence is that in Goodman and Stuhlmüller (2013)'s RSA analysis of scalar implicature, individuals never categorically draw (or fail to draw) scalar inferences. For example, upper-bounded readings of *some* are represented in RSA as relatively lower posterior conditional probability of an *all* meaning on the P_L distribution given an utterance of *some*, compared to the prior probability of that meaning.

5.1 An example: exclusive *or*

The RSA model of pragmatic competence models exclusivity inferences with *or* as a negative change from the prior to posterior probability of an inclusive *and* meaning given an observation of an utterance of *or*. The model first declares a possible set of possible world states. For simplicity, I assume that the set of possible worlds contains only a world in which some statement X holds, a world in which some statement Y holds, and a third world in which both statements are jointly true. We assume that listeners have a uniform prior expectation of being in any of these three world states:[8]

```
var states = [['X'],['Y'],['X','Y']]

var statePrior = function() {
  return uniformDraw(states);
};
```

Next, we add a space of possible utterances that a speaker might produce. In this example, speakers are only capable of producing utterances of the form X *and* Y or X *or* Y; that is, we have implicitly constrained the model such that the only possible alternative to *or* is *and*. Furthermore, these two utterances are assumed to have a uniform cost, such that the prior expectation of hearing either of these utterances is uniform:

```
var utterances = ["or", "and"];

var cost = {
  "or": 1,
  "and": 1,
};

var utterancePrior = function() {
 var uttProbs = map(function(u) {return Math.exp(-cost[u]) }, utterances);
 return categorical(uttProbs, utterances);
};
```

A truth conditional semantics of the two utterances is then specified. *Or* assumes the meaning of logical disjunction (i.e. it is logically consistent with conjunction), while *and* assumes the meaning of conjunction:

[8]The code presented here is excerpted from a model I've coded in WebPPL, which is accessible here: `https://bwaldon.github.io/files/bareor.js`. The model is executable in the online WebPPL editor: `http://webppl.org/`.

```
var literalMeanings = {
    and: function(state) { return state.includes("X") &&
    state.includes("Y"); },
    or: function(state) { return state.includes("X") ||
    state.includes("Y"); },
};
```

Lastly, models of a pragmatically-competent listener and speaker, as well as of a literal listener are specified as according to the equations in (23). For the pragmatic speaker, we set the rationality term α to 1 and keep it constant for the following discussion. On these assumptions, a pragmatically competent listener's expectation of being in the inclusive ['X','Y'] world state shifts from 33.33% to roughly 11.11% upon observation of an utterance of *or*. This negative change from prior to posterior expectation of the inclusive world state is how RSA captures the pragmatic exclusivity inference of *or*.

Note that in the above example, we have stipulated that there be only one alternative to *or* (namely *and*), but this need not be the case in order to break symmetry. Consider a second model, in which listeners also reason about a speaker's potential use of the exclusive disjunctive construction *either-or*.[9]The model differs minimally from the first model, where no *either-or* is present. The major changes are listeners have equal expectation of seeing any of three utterances (*or*, *and*, and *either-or*), and we must also specify the semantics of *either-or* (the rest of the model remains the same):

```
var utterances = ["or", "eitheror", "and"];

var cost = {
  "or": 1,
  "eitheror": 1,
  "and": 1,
};

var literalMeanings = {
    and: function(state) { return state.includes("X") &&
    state.includes("Y"); },
    or: function(state) { return state.includes("X") ||
    state.includes("Y"); },
    eitheror: function(state) { return (state.includes("X") ||
    state.includes("Y")) &&
    !(state.includes("X") && state.includes("Y"));
};
},
```

On these assumptions, a pragmatically competent listener's expectation of being in the inclusive ['X','Y'] world state shifts from 33.33% to roughly 23.8% upon observation of an

[9]This second WebPPL model is accessible here: `https://bwaldon.github.io/files/eitheror.js`

utterance of *or*. This negative change from prior to posterior expectation of the inclusive world state is smaller than the change observed in the previous model; however, on these assumptions, symmetry is broken even if *either-or* is allowed to enter into pragmatic competition with *or*. The change in prior to posterior expectation can be made greater by changing the cost term on utterances, such that the morphosyntactically more complex *either-or* is costlier than *or*.[10]

The flexibility of options afforded to the Rational Speech Act model in breaking symmetry - by restricting the set of alternatives and/or by assigning variable weights to the costs of alternatives - illustrates that the RSA model does not, in itself, commit us to any specific theory of alternatives. However, the quantitative predictions made by the model - when coupled with informed linking hypotheses from RSA's representation of pragmatic competence to observed linguistic behavior - provide a fruitful avenue of future research. For example, in future work, perhaps it is revealed that the best RSA model (on some clearly defined metric of model quality) is one where *either-or* is not allowed to enter into pragmatic competition with *or*. There is much more to be said here which must be left to future theoretical and empirical investigations.

For the time being, however, I note that in the RSA model of scalar implicature as it is conceived by Goodman and Stuhlmüller (2013) allows that the alternatives "**could be all possible utterances** or could be a limited set generated by replacing key words in the actual utterance with related words" (176, my emphasis), a possibility unavailable to the Gricean or grammaticist analyses of scalar implicature explored above. Cohn-Gordon et al. (2018) note that such a model poses grave problems of computational tractability. To design an RSA-based image captioning system, the authors propose a solution to this efficiency problem whereby probabilistic inference occurs incrementally, at the character level: for each stage of inference, the space of possible utterances is constrained to the finite and very small set of possible orthographic characters. Production probabilities of utterance given an intended communicated world state (i.e. one particular from a set including distractor images) are built up from inferring the production probabilities of sequences of characters. The authors present evidence that this algorithm performs even better than one where production probabilities are derived sequentially over a finite set of lexical items.

Another solution to the issue of computational tractability is proposed by Monroe et al. (2017), who design an RSA-based model of pragmatic competence in a reference game whereby a listener must identify an interlocutor's intended referent from three possible color patches. Using a preliminary web-based norming experiment, the authors create a space of possible utterances from the linguistic decisions made my human participants in this reference game. However, the computational model of pragmatic competence designed by the authors for this reference game does not involve reasoning over all possible utterance alternatives elicited from the corpus. Instead, the model samples a finite subset of these possible utterances in pragmatic inference. According to the authors, this improves the efficiency of the model in a way that yields "a satisfactory compromise between effectiveness and computation time" (331).

[10]The reader my verify this by experimenting with cost terms in the attached code (see the previous footnote).

6 Pyscholinguistic research on alternatives

In this section, I review three recent studies from the psycholinguistics literature on the activation of alternatives in pragmatic inference.

6.1 Nicolae and Sauerland (2015): Exclusivity interpretations with *or* and *either-or*

Nicolae and Sauerland (2015) report that "the difference in strength [between *or* and *either-or*] only arises when the two forms are both used; in isolation both disjunctions exhibit the same level of exclusivity" (2016: 1). In one between-subjects experiment, participants who were asked to rate the likelihood of an exclusivity reading given sentences containing *or* behaved no differently from participants in a separate experimental condition who were asked to rate the likelihood of an exclusivity reading with *either-or*. However, in a second experiment, where participants saw critical trials containing both *or* and *either-or*, there was an observed strength asymmetry between the two lexical items.

To explain these results, the authors claim that absent exposure to *either-or*, participants contrast *or* with its scale-mate *and*, which leads to a pragmatically-derived exclusivity inference that is truth-conditionally equivalent to the semantic exclusivity entailment of *either-or*. However, in a context where *either-or* has been made salient, participants make *either-or* (rather than *and*) the object of contrast with *or*. In subsequent experiments reported in the paper, the authors report that this empirical finding holds cross-linguistically (i.e. in German in addition to English).

6.2 Rees and Bott (2018): Scalar inference primed by alternatives

Rees and Bott (2018) present three experiments designed to assess the extent to which scalar inferences may be primed, either via exposure to putative utterance alternatives or by presenting contexts where participants derive scalar inferences before exposure to a target trial. Their experimental paradigm involved exposing participants to visual contexts (computer-generated cards) matched with utterances whose contents were statements about what was on that card.

There were three priming conditions: a weak prime condition - for example, an utterance of *Some of the letters [on the card] are K* when in fact all of the letters were K; a strong prime condition - for example, an utterance of *Some of the letters [on the card] are K* when some but not all of the letters were K; and an alternative prime condition - for example, an utterance of *All of the letters [on the card] are K* when all of the letters were K. Across three implicature paradigms (*some & all* as well as cardinal numerals, and *ad hoc* existential statements), the researchers found that the strong prime and alternative prime conditions led to higher rates of implicature calculation[11] than did the weak prime condition, but that

[11] 'Implicature rate' was operationalized from the following linking hypothesis: on target trials, participants saw an utterance with the paradigm's weak scalar item (e.g. *some of the letters are L*) and told to select between a). a card for which the utterance was a weak descriptor (e.g. where all of the letters on the card were L); or b). an option of saying that there was a "Better Picture" than the card presented on the screen. Choices of "Better Picture" were assumed to associate with implicature calculation.

there was no significant difference between these two conditions.

The authors assume that in both the strong prime and alternative prime conditions, the stronger alternative utterance on the target trial has been primed: in the case of the strong prime condition, this activation is due to the fact that the stronger lexical alternative (e.g. *all)* must be activated to achieve the strengthened meaning (e.g. *some but not all*); in the case of the alternative prime condition, this activation is due to the fact that the participant sees the stronger lexical item. The authors conclude that these results provide evidence that while activating alternatives in context may prime implicature behavior, it is not possible to independently prime "the usage mechanism" on which pragmatic inference proceeds.

6.3 Degen and Tanenhaus (2015): Upper-bounded *some* inferences primed by cardinal numerals

Degen and Tanenhaus (2015) present three experiments related to pragmatic inference with *some*, but I restrict my discussion here to their Experiment 2, from which it was concluded that priming participants with small natural cardinal numbers (e.g. 2, 3, 4) decreases naturalness ratings for *some*. To demonstrate this, the authors construct the following experimental paradigm: on a computer screen, participants see a gumball machine containing 13 gumballs, from which some number of gumballs comes out. On critical trials, this event was accompanied by a statement of the form *You got some (of the) gumballs*; however, in non-control condition, participants were also exposed to utterances of the form *You got N (of the) gumballs*. For high values of N, the presence of these utterance alternatives did not affect naturalness ratings of *You got some (of the) gumballs* relative to the control condition; however, for low values of N, there was a negative effect on naturalness of *some* relative to the control.

From this, the authors find support for the notion that "listeners rapidly adjust expectations about the types of utterances that speakers will produce in a particular situation" (672). Though "*some* contextually competes with many other alternatives, like number terms" the context-sensitivity of these alternatives relative to, e.g., *all* suggests that number terms "are not lexicalized alternatives" to *some* (676).

References

Chierchia, G. (2006). Broaden your views: Implicatures of domain widening and the ''logicality'' of language. *Linguistic Inquiry*, 37:535–590.

Cohn-Gordon, R., Goodman, N. D., and Potts, C. (2018). Pragmatically informative image captioning with character-level inference. In *Human Language Technologies: The 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Stroudsburg, PA. Association for Computational Linguistics.

Collins, J. (2017). *Structure-Sensitive Interpretation: A Case Study in Tagalog*. PhD thesis, Stanford University, Stanford, CA.

Degen, J. and Tanenhaus, M. K. (2015). Availability of alternatives and the processing of scalar implicatures: A visual world eye-tracking study. *Cognitive Science*, 40(1):172–201.

Fox, D. and Katzir, R. (2011). On the characterization of alternatives. *Natural Language Semantics*, 19(1):87–107.

Gazdar, G. (1979). *Pragmatics: Implicature, Presupposition and Logical Form*. Acad. Press.

Goodman, N. D. and Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, 5(1):173–184.

Grice, H. P. (1975). Logic and conversation. In Cole, P. and Morgan, J. L., editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, New York.

Hirschberg, J. B. (1985). A theory of scalar implicature (natural languages, pragmatics, inference).

Horn, L. (1972). *On the Semantic Properties of Logical Operators in English*. PhD thesis, UCLA.

Horn, L. (1989). *A Natural History of Negation*. University of Chicago Press.

Horn, L. and Abbott, B. (2012). <the, a>: (in)definiteness and implicature. In Kabasenche, W. P., O'Rourke, M., and Slater, M. H., editors, *Reference and Referring*. MIT Press, Cambridge, MA.

Katzir, R. (2007). Structurally-defined alternatives. *Linguistics and Philosophy*, 30(6):669–690.

Matsumoto, Y. (1995). The conversational condition on horn scales. *Linguistics and Philosophy*, 18(1):21–60.

Monroe, W., Hawkins, R. X. D., Goodman, N. D., and Potts, C. (2017). Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Nicolae, A. and Sauerland, U. (2015). A contest of strength: or versus either–or. In *Proceedings of Sinn und Bedeutung 20 (SuB 20)*.

Potts, C. (2015). Presupposition and implicature. In Lappin, S. and Fox, C., editors, *The Handbook of Contemporary Semantic Theory*, pages 168–202. Wiley-Blackwell, 2 edition.

Rees, A. and Bott, L. (2018). The role of alternative salience in the derivation of scalar implicatures. *Cognition*, 176:1 – 14.