Recent *Iterated Response* (IR) models conceptualize pragmatic reasoning as a recursive process in which abstract speakers and listeners reason about each other to increase communicative efficiency and enrich the meanings of the utterances they hear in context-dependent ways [7, 6, 3, 2]. IR models generally separate pragmatics from incremental utterance processing. However, there is substantial evidence that pragmatic processing is incremental. To address this, we develop an IR model that is incremental in the sense that the pragmatic reasoning takes place during utterance production and comprehension. We present computational experiments which demonstrate the difference between incremental and global pragmatics, observe that the former approximates the latter under reasonable conditions, and apply the incremental model to an empirical observation: the asymmetry between adjective-noun and noun-adjective languages in over-informative referential behaviour [8]. We briefly discuss the anticipatory implicatures arising from contrastive modifiers [9], which we analyze in detail in the associated paper.

Global Pragmatics Our starting point is the *Rational Speech Acts* (RSA) model [2, 4]. This is a global model in the sense that the pragmatic reasoning is defined over complete utterances. Speakers are conditional distributions of the form $P(u|w)$ while listeners are of the form $P(w|u)$, for an utterance u and world state w.

(1) $L_0^{\text{UTT}}(w|u) \propto [\![u]\!](w)$

(2) $S_1^{\text{UTT-GP}}(u|w) \propto \exp\left(\log(L_0^{\text{UTT}}(w|u)) + \mathit{cost}(u)\right)$

(3) $L_1^{\text{UTT-GP}}(w|u) \propto S_1^{\text{UTT-GP}}(u|w)$

Given an utterance u, the literal listener L_0^{UTT} evenly distributes probability mass to the worlds compatible with u. Intuitively, $S_1^{\text{UTT-GP}}$ prefers utterances which are not only true but best convey to L_0^{UTT} which world the speaker is in. $L_1^{\text{UTT-GP}}$ can use this information to draw more refined inferences. For instance, suppose we are in a setting with three referents, a red dress (R1), a blue dress (R2), and a red hat (R3), and that the language is composed of three utterances, *dress*, *red dress*, and *red object*, with their expected semantics. For L_0^{UTT}, *red object* is equally likely to refer to R1 or R3. However, $S_1^{\text{UTT-GP}}(\textit{red object}|\text{R3}) = 1$ (with all costs set to 0). In turn, on hearing *red object*, $L_1^{\text{UTT-GP}}$ will infer that the referent is most likely R3 ($L_1^{\text{UTT-GP}}(\textit{red object}|\text{R3}) = 0.8$).

We take *cost* to be a language model, either estimated from data (higher probability to attested utterances), derived from a grammar (higher probability to grammatical utterances), or simply assigning longer utterances more cost. $[\![\cdot]\!]$ is an interpretation function mapping utterance and world pairs to $\{0, 1\}$.

Incremental Pragmatics Speakers and listeners use language incrementally. To approximate this process, we represent utterances as sequences of words and allow RSA-style reasoning to happen at each word, moving left to right. To do this, we first define an incremental semantics: for any partial sequence c and set of referents W, $[\![c]\!](w) \in [0, 1]$ is the number of true full-utterance extensions of c in w divided by the number of possible extensions of c into full utterances that are true of any world in W. Where c is a full utterance, $[\![c]\!](w) \in \{0, 1\}$ as in the global model; where c is a partial utterances, $[\![c]\!]$ represents the biases created by c. We can then define an incremental literal listener L_0^{WORD} and incremental pragmatic speaker and listener S_1^{WORD} and L_1^{WORD}:

(4) $L_0^{\text{WORD}}(w|c, \mathit{word}) \propto [\![c + \mathit{word}]\!](w)$

(5) $S_1^{\text{WORD}}(\mathit{word}|c, w) \propto \exp\left(\log(L_0^{\text{WORD}}(w|c, \mathit{word})) + \mathit{cost}(\mathit{word})\right)$

(6) $L_1^{\text{WORD}}(w|c, \mathit{word}) \propto S_1^{\text{WORD}}(\mathit{word}|c, w)$

Returning to the example given above, and assuming 0 cost on all words, an incremental pragmatic speaker will prefer *red* as a first word when conveying R1 (the red dress): $S_1^{\text{WORD}}(\textit{red}|c = [], w = \textit{R1}) = 0.571$. However, if R3 is the intended referent, the speaker *must* begin their utterance with *red* (since *hat* is not an available global utterance in this example, and accordingly, *hat* is not an available word). As such, the pragmatic listener will infer from *red* that the referent is most likely R3 ($L_1^{\text{WORD}}(\text{R3}|c = [], \textit{red}) = 0.64$). This is the approach we apply in our full paper to modeling the anticipatory implicatures observed by [9].

Finally, from the word level agent S_1^{WORD}, we can obtain $S_1^{\text{UTT-IP}}$, an utterance-level speaker which performs incremental pragmatics:[1]

(7)
$$S_1^{\text{UTT-IP}}(u|w) = \prod_{i=1}^{n} S_1^{\text{WORD}}(u_i|c = [u_1 \dots u_{i-1}], w)$$

[1]We use [] for the empty list, $u[n]$ for the nth element of a list u, and $u[:n]$ for the sublist of u up to but not including $u[n]$.

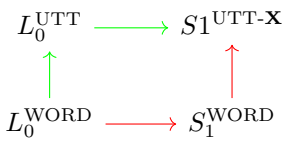

Figure 1: A diagram showing two ways of constructing an utterance level pragmatic speaker from L_0^{WORD}. The green path is to obtain a literal listener over full utterances and then perform pragmatics, which gives rise to $S_1^{\text{UTT-GP}}$ while the red path is to construct a word-level pragmatic speaker from L_0^{WORD} and then use this to define an utterance-level pragmatic speaker, $S_1^{\text{UTT-IP}}$.

While $S_1^{\text{UTT-GP}}$ makes pragmatic calculations on the basis of whole utterances, $S_1^{\text{UTT-IP}}$ makes incremental pragmatic decisions about each choice of word, which gives rise to a distribution over utterances.

The relation between global and incremental pragmatics We show that $S_1^{\text{UTT-GP}}$ and $S_1^{\text{UTT-IP}}$ define distinct distributions, even when costs are fixed at 0. For instance, in the global model, $S_1^{\text{UTT-GP}}(u = \textit{red dress}|w = \text{R1}) = 0.5$ is greater than $S_1^{\text{UTT-GP}}(u = \textit{dress}|w = \text{R1}) = 0.25$. In contrast, in the incremental model, $S_1^{\text{UTT-IP}}(u = \textit{red dress}|w = \text{R1}) = 0.38$ is smaller than $S_1^{\text{UTT-IP}}(u = \textit{dress}|w = \text{R1}) = 0.42$. The reason for the $S_1^{\text{UTT-IP}}$ result is that saying *dress* ensures the termination of the utterance (given the set of utterances that are available in this example), which therefore has probability 1 at the next timestep, while saying *red* leaves two options, *dress* and *object*. Thus, $S_1^{\text{UTT-GP}}$ and $S_1^{\text{UTT-IP}}$ are not only different, but differ in their predictions about what utterance is optimal. Our paper will contain a fuller exploration of the relationship between these speaker agents, arguing that $S_1^{\text{UTT-IP}}$ approximates globally pragmatic behaviour. We now briefly consider a case where incremental pragmatics provides an explanation of a phenomenon where global pragmatics does not seem to suffice.

Over-informative Referring Expressions (REs) It has been observed [1, 5] that when generating referring expressions, humans often provide more information than necessary to refer unambiguously. For instance, [8] shows that English speakers often use redundant color terms (e.g., *the red dress*) in a scene with only a single dress, where the shorter utterance *dress* would suffice. However, [8] also notes that Spanish speakers are less likely to over-describe with the analogous RE, *el vestido azul*, in the same situation. This difference is a challenge for non-incremental pragmatic accounts, since, *ceteris paribus*, we would expect semantically equivalent Spanish and English REs to have the same production probability.

Using incremental pragmatics, we model the English case as follows: let the referents be a red hat (R1) and a blue hat (R2), and the possible utterances be *dress*, *red dress*, *hat* and *blue hat*, with the obvious semantics. Assume a cost of 1.0 for all words but the *stop* token, and that an utterance's cost is the sum of the cost of its words. Then $S_1^{\text{UTT-GP}}(\textit{dress}|\text{R1}) = 0.731 > S_1^{\text{UTT-GP}}(\textit{red dress}|\text{R1}) = 0.268$, while $S_1^{\text{UTT-IP}}(\textit{dress}|\text{R1}) = S_1^{\text{UTT-IP}}(\textit{red dress}|\text{R1}) = 0.5$. The increase in mass on the overinformative RE *red dress* in $S_1^{\text{UTT-IP}}$ is a result of incremental processing: the decision between *red* and *dress* is made on the basis of informativity, and both words are equally informative. However, if *red* is chosen, the subsequent, now over-informative word *dress* has to follow, since *red* on its own is not an utterance.

In the Spanish case, let our utterances be *vestido*, *vestido rojo*, *sombrero* and *sombrero azul*, with the same costs as before. Then $S_1^{\text{UTT-GP}}(\textit{vestido}|\text{R1}) = 0.731 > S_1^{\text{UTT-GP}}(\textit{vestido rojo}|\text{R1}) = 0.268$, and $S_1^{\text{UTT-IP}}(\textit{vestido}|\text{R1}) = 0.731 > S_1^{\text{UTT-IP}}(\textit{vestido rojo}|\text{R1}) = 0.268$. The difference here from the English case is that, when choosing the word to follow *vestido*, the incremental pragmatic speaker has no need to say *rojo* rather than *stop*, since the goal of communicating the referent has already been completed by *vestido*, and so chooses the less costly option, *stop*.

In summary, we define a notion of incremental pragmatics, and show that it differs in meaningful ways from global pragmatics. We plan to explore the relation of global and incremental pragmatics summarized in figure 1, including the case of anticipatory implicatures [9] in more detail in our paper.

References

[1] Paul E Engelhardt, Karl GD Bailey, and Fernanda Ferreira. Do speakers and listeners observe the gricean maxim of quantity? *Journal of Memory and Language*, 54(4):554–573, 2006.
[2] Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998, 2012.
[3] Michael Franke. *Signal to Act: Game Theory in Pragmatics*. ILLC Dissertation Series. Institute for Logic, Language and Computation, University of Amsterdam, 2009.
[4] Noah D. Goodman and Andreas Stuhlmüller. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–184, 2013.
[5] Theo Herrmann and Werner Deutsch. *Psychologie der objektbenennung*. Huber, 1976.
[6] Gerhard Jäger. Game theory in semantics and pragmatics. pages 2487–2425.
[7] Gerhard Jäger. Game dynamics connects semantics and pragmatics. In Ahti-Veikko Pietarinen, editor, *Game Theory and Linguistic Meaning*, pages 89–102. Elsevier, Amsterdam, 2007.
[8] Paula Rubio-Fernández. How redundant are redundant color adjectives? an efficiency-based analysis of color overspecification. *Frontiers in psychology*, 7:153, 2016.
[9] Julie C Sedivy. Implicature during real time conversation: A view from language processing research. *Philosophy compass*, 2(3):475–496, 2007.