# Aspects of a Theory of Pronoun Interpretation

Andrew Kehler
*University of California, San Diego*

Hannah Rohde
*University of Edinburgh*

Submitted 2013-09-03

We thank our commentators for their valuable thoughts on our paper. The contributions represent a wide variety of perspectives, and cover a fair bit of ground. Hence, we cannot respond to everything that we would like to. Instead, we organize our comments around a few consistent and important themes.

**On Processing Models and Their Complexity** Several commentators ask how our proposal fits within a larger theory of language interpretation and processing, so let us start there. Following Rohde, Levy & Kehler (2011) and others, we envision our results situated in a *strongly incremental* theory, one in which both sentence-level and discourse-level decisions are made utilizing a broad variety of information sources on a moment-by-moment basis. This comes with two important ramifications. First, it predicts a particular time course of processing, whereby linguistic expressions influence context at the time they are encountered, and hence many only affect the interpretation of a subsequent pronoun indirectly. For instance, in interpreting a passage like *Amanda fired Brittany and she immediately filed a lawsuit*, the comprehender will have a particular set of biases toward coherence and next mention after the first clause (presumably favoring Explanation and Brittany respectively, since *fired* is an object-biased IC verb), which then get updated when *and* is encountered, and then again when the pronoun *she* is encountered. Crucially, the updates proceed along opposite paths in these two cases. On the one hand, *and* updates $P(CR)$ to increase the likelihood of the coherence relations with which it is most compatible (Occasion, Result, Parallel) and to decrease it for others (e.g., Explanation will drop to near zero), which cascades to update $P(\textit{referent})$. On the other hand, *she* updates $P(\textit{referent})$ to become $P(\textit{referent} \mid \textit{pronoun})$ (by including the production probability $P(\textit{pronoun} \mid \textit{referent})$) which, in pulling the distribution toward the subject referent, cascades to update $P(CR)$ to increase the likelihood of subject-biased relations per the results of Rohde & Kehler (2008). The second ramification is that interpretation does not stop there; these probabilities will continue to be updated as subsequent linguistic material is encountered through the end of the clause. This answers Arnold's and de Hoop's question about how an ultimate

interpretation is identified – in the example above, the difference between follow-ons such as *immediately filed a lawsuit* and *immediately hired Susan* will continue to affect the probability distributions for both coherence relation and pronoun reference. Whereas the results of our studies cannot fully establish this on-line processing profile, it is the one that most readily explains why we find dramatically different biases in passage completions with/without connective prompts and with/without pronoun prompts.

Whereas Bayesian belief revision (think graphical models) is not the only way to perform such updates, it does provide a mathematically principled basis for bringing predictive, 'top-down' expectations into contact with 'bottom-up' linguistic evidence. The idea of using Bayes' Rule in this way is not new. It is common in computational speech recognition systems, for instance, to use Bayes' Rule to compute the probability of a word $P(\textit{word} \mid \textit{signal})$ by estimating what is essentially an acoustic production probability $P(\textit{signal} \mid \textit{word})$ and a prior $P(\textit{word})$, the latter of which is based on a language model that utilizes local context to generate expectations about the ensuing word regardless of the acoustic signal. Use of context is how a system can select among homophones such as *too, to,* and *two*, for instance. In the case of reference, pronouns similarly produce a signal that, while placing constraints on the ultimate interpretation, may be ambiguous and hence require contextual information to fully resolve. The use of Bayes' Rule thus allows us to posit a relatively uniform (albeit underconstraining) analysis of pronouns, yet explain why the biases associated with them appear to vary so dramatically across contexts.

The foregoing discussion highlights a key property of our model, specifically the rich causal structure among the different factors. This stands in opposition to the sort of 'bag of cues' models often found in the pronoun literature, which use a discriminative model (e.g., regression) to estimate $P(\textit{referent} \mid \textit{pronoun})$ directly. It is important to distinguish between updates to mental representations of context that are specific to pronouns from those that happen anyway during the normal course of discourse interpretation. We have seen characterizations of our work as showing that factors like aspect, connectives, and coherence relations should be added to (and modeled alongside) existing sets of other factors (subjecthood, parallelism), but this misses the point. The relationship between aspect and pronoun interpretation, for instance, is highly indirect: Aspect influences the part of event structure that comprehenders will focus on as they interpret the next clause, in turn affecting expectations about coherence and next mention, which in turn affects pronoun interpretation. All but the last step occur whether or not a pronoun appears in the input, and are expected to affect other aspects of language interpretation as well.[1] Note also that aspect is predicted to matter only for context types in which

1 For another example of how aspect affects prominence in event structure, see Ferretti, Kutas &

entity prominence varies significantly across event structure (recall that transfer of possession is special in this regard); for many other verb types (e.g., motion verbs; *John ran / is running toward Bill*) there is less reason to expect it to make much of a difference. Hence, aspect cannot take place alongside context type as a pronoun interpretation 'cue'.

We understand de Hoop to reject probabilistic approaches, arguing that a model such as ours cannot account for the fact (if it is one) that pronoun interpretation is not just a matter of (high) probability but is instead categorical. As we have recently argued, however, this doesn't account for the fact that any reasonable interpretation model will be one in which probabilistic update continues beyond the occurrence of ambiguous forms, and one might imagine that utterances typically carry enough content so as to allow their references (as well as other loci of ambiguity) to be disambiguated. It is then but a small step to get from high probability to categorical commitment. De Hoop opts instead for a symbolic model that is similar to other cue-based models except that it utilizes a set of constraints within an OT-based optimization procedure, but we see problems. First, the model requires that the factors participate in a separate optimization process specifically triggered by the presence of a pronoun, rather than exert their influence on context naturally during the course of language interpretation (which, as we have argued, for most factors will happen whether or not a pronoun occurs). Second, the idea that one could model all of the relevant biases and their interactions as categorical constraints seems untenable. For instance, de Hoop has a constraint representing implicit causality (IC), but we know that IC-driven biases differ in their strength across verbs and are tied specifically to Explanation relations. Both subject- and object-biased IC verbs tend to have strong object biases when a Result relation is operative, which would require another constraint. Transfer-of-possession verbs have equally strong biases toward nonsubjects when an Occasion relation is operative, requiring yet another constraint. And so forth. The fact is that all contexts have gradient biases associated with them that need to be accounted for; we see no benefit to modeling a subset of them with separate categorical constraints, nor in restricting the manner in which they are integrated by placing them on a strict dominance hierarchy. Finally, the analysis predicts an alignment between the speaker's goal of optimizing form given meaning (in our terms, $P(\textit{form} \mid \textit{meaning})$, that is, $P(\textit{pronoun} \mid \textit{referent})$) and the comprehender's goal of optimizing meaning given form ($P(\textit{meaning} \mid \textit{form})$, that is, $P(\textit{referent} \mid \textit{pronoun})$). Our data show, however, that these terms are not aligned: Because comprehenders factor in the prior $P(\textit{referent})$ but speakers do not, speakers will sometimes choose to use a (less optimal) name even when a (more optimal) pronoun would have been biased to the correct referent, and likewise choose

McRae (2007) for a demonstration that imperfective verbs (*was skating*) prime typical locations for the events they denote (*arena*) whereas their perfective variants (*skated*) do not.

a pronoun to refer to referents that go against the comprehender's interpretation bias. This latter point may not concern de Hoop since subsequent context will usually disambiguate the reference, but consider cases like (1):

(1) Norm lent his car to his brother's girlfriend. He doesn't own one.

Informants we have polled find this passage to be confusing even after they are done reading it; they typically ask how Norm could lend someone a car if he doesn't own one. This is despite the fact that interpreting *He* to refer to Norm's brother results in a perfectly coherent passage (*Norm lent his car to his brother's girlfriend. His brother doesn't own one.*). It would seem that incremental referential biases toward Norm lead the comprehender up the referential garden path, and the WORLD KNOWLEDGE constraint fails to save the day. As in the domain of parsing, probabilistic accounts are capable of predicting the existence of such garden paths (e.g. through pruning of low-probability alternatives), whereas it is not clear to us how this could be captured by a system like de Hoop's.[2]

Several commentators (Arnold, van Berkum, Kaiser) also raised the issue of where the complexity resides in a 'simple' theory of pronoun interpretation of the sort we advocate, pointing out that the complexity is not being eliminated, but instead 'moved'. This is correct. To be clear, our point here was to argue against the tendency for researchers to attribute all of the (invariably complicated) patterns found when analyzing the data to the particular form being studied itself, and hence build this complexity into their theories. We are very much in line with van Berkum's thinking in terms of effortful versus automatic processing, although as he mentions, the specifics are an empirical matter. At a general level, it is obvious that the computations that ultimately affect pronoun interpretation, like all of language understanding, are complex in the sense that computation in the brain is complex. On the other hand, our theory of pronoun interpretation is comparatively simple: Pronouns take the prior biases toward entity mention $P(\textit{referent})$ and convert them to interpretation biases $P(\textit{referent}|\textit{pronoun})$ by factoring in the production bias $P(\textit{pronoun}|\textit{referent})$, which

2 We need to correct the record on one issue, with respect to de Hoop's statement that in Stevenson et al.'s passage completion studies "the reference of the pronouns was coded by the experimenters", and further that "clearly, the experimenters coded their own pronoun interpretations". We spoke a bit loosely in our article; Stevenson et al. actually had participants code their own references in their pronoun condition, but then had to use judges anyway as some codings were clearly incorrect. For our own experiments, the coding process similarly always included judges who were blind to the experimental hypothesis. Further, the judges were instructed to categorize a pronoun as ambiguous if the pronoun could be interpreted as plausibly coreferential with either referent, even if their own biases suggested a particular one (e.g. *Sue gave a book to Mary. She then left.*); this was done precisely to avoid inadvertently measuring the judges' own biases. This always led to a non-trivial set of completions being classified as ambiguous and thus held out from analysis. These details appear in the original papers, but did not in the target article, per footnote 4.

itself appears to be conditioned on a limited set of contextual factors. Yes, there is considerable complexity behind calculating $P(\textit{referent})$; indeed we've argued that certain aspects of it have a particular, linguistically-motivated causal structure. But those aspects are relevant to predictions about the ensuing message that occur anyway rather than about linguistic form, and hence belong in a theory of anticipating subsequent mention rather than of pronouns.

Arnold also comments on this issue, stating:

> KR13 suggest that a theory of pronoun interpretation ought to be simple. I disagree. The primary goal of any theory is that it should be accurate.

We're a bit lost as to the source of disagreement here. Nowhere in our paper do we suggest that simplicity should trump accuracy, nor do we see any cases in which our experimental results reflect such a tradeoff.[3] Whereas Arnold is correct when she says "there is no reason to expect that the machinery of the human mind will be simple", that is different from arguing that this complexity should be inherited into one's theory of pronouns (or of any other linguistic phenomenon, for that matter). She points out that the question we posed at the outset – i.e., if pronoun interpretation were as complex as some theories suggest, why would a speaker ever use one – implicitly assumes a mechanism of language production that is heavily guided by audience design, noting evidence to the contrary. We agree, and in fact that was one of our main points: Our results suggest that speakers do not take into consideration the comprehender's interpretation biases when choosing to use a pronoun.

All this leads us to the question of how (and how well) brains compute these probabilities. Clearly they aren't doing everything that would be required to do to a complete job. For example, Kaiser is quite right to inquire about the space of predictive inferences comprehenders consider in support of coherence establishment. To take (a simplified version of) van Berkum's example, upon reading *Norman threw the precious vase toward the concrete wall*, it seems reasonable to think that a comprehender might make a prediction about it breaking and, from that, predict that the discourse might say something about that next. It is also possible that the vase is saved by a large stray dog that jumps up and grabs it, or that it explodes in midair because it had a stick of dynamite inside. Does the comprehender proactively model the probabilities of such eventualities, along with the infinitude of

3 On the issue of accuracy, Rohde & Kehler (2013) present a comparison of the actual interpretation biases measured for the voice experiment discussed toward the end of the target article with the predictions of the Bayesian model, a 'Mirror Model' in which interpretation biases are computed from normalized production probabilities, and an Expectancy Model that equates interpretation biases with next-mention biases. The results showed that the predictions of the Bayesian model are more highly correlated with the actual biases than the other two models.

others? Presumably not.[4] Arnold makes a similar point with respect to production probabilities, casting doubt on whether comprehenders actually activate potential referring expressions for each entity in the discourse model. Now we do know that comprehenders must activate linguistic alternatives to some extent; the processing that underlies the recognition of Gricean implicatures requires that speakers not only model the possible meanings of what was said, but also what else could have been said for those same meanings (consider Grice's (1975) well known example *X is meeting a woman this evening*, which implicates that he is not meeting his wife, mother, sister, or close platonic friend, even though *a woman* this evening is compatible with those referents). That notwithstanding, Arnold is right that there must be limits: We can't model every possible way that entities can be referred to when computing $P(\textit{pronoun} \mid \textit{referent})$. Thus the question appears to be not whether such prediction takes place, but how much is done, to what extent the computation is effortful versus automatic, and how well the results approximate the true values for the different terms in Bayes' Rule.

**Methodologies** Several commentators remarked on the use of passage completions in our work. For instance, van Berkum points out that it seems wise to complement the task with others that are representationally less rich (reading time, ERPs, etc), and Arnold similarly stresses the importance of understanding the processing mechanisms underlying reference. We agree, and would of course welcome work of this sort. It is not clear that using on-line techniques to fully test the predictions of the Bayesian approach will be straightforward, however. Recall that according to our analysis, the bias contributed by a pronoun is not the bias toward referents measured after the pronoun is encountered, but instead the difference between the biases immediately before and after the pronoun. Hence we need a way to measure the next-mention biases before the pronoun is encountered as well as the interpretation biases after. As we discussed in the conclusion to our article, an advantage of using passage completion is that it allows these measurements to be taken using simple prompt manipulations. We nonetheless agree with van Berkum, Arnold, and others who point out that relying only on this methodology might cause one to miss the details of processing unfolding in time, and we certainly don't suggest that the methodology is sufficient for answering all of our questions.

We nonetheless do see one of the contributions of our work as guiding researchers on where to look and what to control for in such studies. One cannot just dump

4 We are less sympathetic with the idea that people perform 'good enough' processing as discussed by Kaiser, as we don't see how the richness of the data we have presented could be compatible with the obvious candidates for shallow methods, i.e. grammatical role heuristics and the like. Furthermore, per van Berkum, the jury is out on the question of to what extent prediction is the result of deliberate, effortful computation, as opposed to automatic, effortless, and therefore inevitable processing.

a collection of stimuli down, proclaim them to be 'pragmatically unbiased' with no controls for coherence or next-mention expectations, and expect to draw valid conclusions from the data about what biases are associated with pronouns. If anything is clear from this work, it is that there is no such thing as a context with no pragmatic bias (Kehler, Kertz, Rohde & Elman 2008). Whereas there may be reasons to think that the methodology is not fully ecologically valid (Ariel, de Hoop, Arnold), the fact that it has produced a large body of highly consistent results, supporting predictions made by hypotheses with independent linguistic motivation, suggests to us that something is going right.

Drawing on the issue of ecological validity, several commentators suggest instead that corpus studies are a more appropriate methodology. Whereas we are strong proponents of consulting attested data in constructing linguistic theories, we're less sympathetic to the idea that such studies would prove particularly informative for our goals, and even believe that in some ways they can prove misleading. One obvious problem is that one doesn't get to control for a large set of contextual factors that need to remain fixed before conclusions can be drawn. Another is that whereas such studies can tell you what patterns exist, they can't tell you *why*. We elaborate on these in turn, focusing on points made by Ariel and Arnold respectively.

Ariel presents an analysis of a corpus data sample that leads her to conclude that our lab-collected data is not representative of actual attested data.[5] However her data set is very small, as she admits, and there were no apparent controls over other contextual factors nor are the results of statistical significance tests provided. We want to be clear that the biases we measure are not associated specially with the verb in the context sentence but instead with the entire context; hence, there are many other factors to consider besides verb type. For instance, as we mentioned in our article, our analysis makes the prediction that the more strongly the context singles out the current topic (e.g., through repeated mention, a marked syntactic construction, or what have you), the less effect coherence-driven expectations will have on pronoun interpretation. Hence, an object-biased IC verb might be predicted to show a subject bias in multi-clause contexts that have clearly established the subject to be the topic. It also predicts that if the cause of an event introduced by an IC verb is already given in the context, that will reduce expectations that a cause will ensue (Simner & Pickering 2005), which will change the operative biases toward next mention and pronoun interpretation. Rohde, Kehler & Elman (2007) showed that event arguments matter: Contexts with 'abnormal' objects (*John gave a bloody*

5 Ariel focuses on examples with clauses conjoined by *because*, suggesting that we did the same, but in fact none of the studies we discussed in detail used stimuli with *because* prompts. This was by design, since we wanted to see the full range of coherence relations employed, and did not want the complexities introduced by using subordinate clauses (e.g., binding theory constraints that would restrict the ability to use a pronoun in certain grammatical positions) to confound the studies.

*meat cleaver to Bob*) lead participants to write more Explanations than normal ones (*John gave a book to Bob*), which could affect biases. And so on. The point is that simple counts on naturally-occurring data are insufficient for concluding that the results found in the lab are invalid in absence of a wide variety of controls. Ariel is further concerned that sentences with IC verbs and two proper name arguments are unnatural, but we see no reason to believe that people do not have native intuitions about such prompts (and indeed, quick Google searches for phrases like *Obama praised Clinton* and *Obama despises Romney* yield many examples).

Arnold similarly appeals to corpus data to question our claims. For instance, she argues that there are potential problems with our proposal about how $P(\textit{referent})$ is calculated, claiming that our view "ignores the fact that grammatical role also correlates with next-mention likelihoods", which she bases on a text analysis that revealed that entities mentioned from subject position are more likely to be mentioned again than nonsubjects. Whereas this leads her to conclude that the estimate of $P(\textit{referent})$ is likely to be wrong in our model, we again worry about this line of reasoning. As we all know, statistical correlation does not imply causation, and hence finding a correlation between subject position and next-mention biases does not impute a direct role for grammatical position. For instance, it seems likely that coherence-driven biases for many context types will yield a next-mention bias toward the Agent mentioned in the previous clause. That is, for many contexts we expect the Agent to be prominent throughout event structure, with the result being that many discourses will tell of the Agent doing something and then doing something else (Occasion, Result), tell of the Agent doing something because she had done something before (Explanation), or provide additional details that involve a re-mention of the Agent (Elaboration). In light of the fact that Agents commonly appear as grammatical subjects, how could a corpus study distinguish between a pragmatic explanation such as this and one based on grammatical role? It can't, at least not without a careful exploration of the underlying causes of the pattern. The need to distinguish between grammatical and pragmatic biases is one of the primary reasons why we sought contexts in which the next-mention biases were known to point strongly away from the subject, such as events introduced by Source-Goal transfer-of-possession verbs and by object-biased IC verbs. Now maybe it will turn out that grammatical role does in fact impact next mention biases independently of other factors – our data doesn't prove that it doesn't. But merely demonstrating a correlation between subjecthood and next mention in naturally occurring data doesn't either.

Arnold similarly questions our conclusions about the factors that influence pronoun production biases, claiming that our analysis "ignores the fact that parallelism does affect production, for both subject and nonsubject references" – in other words, keeping all else equal, speakers are more likely to produce a pronoun if the an-

tecedent is in a parallel grammatical position. Her support for this is based on an analysis of referring expressions that participants used in a story-telling experiment (Arnold, Bennetto & Diehl 2009). Again, however, we see issues here. First, the data that are crucial to this claim are those referring expressions that appear in nonsubject positions, because only in those cases do a subject/topic bias and a grammatical parallelism bias make different predictions (when the antecedent is the previous subject, both agree that subject references should usually be pronominalized). Whereas Arnold took care to include nonsubject references in the aforementioned quote, the data reported in her paper reveals that pronouns and zeroes in nonsubject positions were used more often to refer to subject referents than nonsubject referents. Second, we want to be clear that we did not claim that one won't ever find a bias toward grammatical parallelism in the data; we argued instead that positing an independent parallelism bias is unnecessary for explaining such patterns. We offered several reasons why one might find a correlation between pronoun use and parallelism despite the lack of a specific bias toward it: (i) the fact that information structure in Parallel and other Resemblance-based coherence relations usually necessitate that unaccented referring expressions corefer with parallel entities for independent reasons[6], (ii) the fact that in certain sentences with multiple pronouns (e.g., *John met Bill. He liked him.*) a subject bias will predict parallelism for both subjects and objects (since binding constraints prohibit a non-reflexive object pronoun from coreferring with the previous subject once subject-to-subject reference is established), and (iii) sometimes semantics will independently favor a parallel referent. Countering our claim would therefore require that these alternative causes be controlled for, but there is no indication in Arnold et al. (2009) that they were. Our point is that whereas finding statistical correlations in data is an important part of theory development, such correlations are not themselves the theory, but merely the data that theories need to explain.

**Cross-Linguistic Applicability** Several commentators rightfully ask whether the theory extends to other forms of reference and to other languages. Kaiser, for instance, inquires as to whether it applies to the most reduced forms in all languages. This is a good question, one that needs to be investigated. Obviously, it would be strange if speakers of English behaved in accordance with the Bayesian analysis and those of other languages did not. As we move to other types of reference,

6 Webber notes that when a Resemblance relation (including Instantiation in the Penn Discourse Treebank) holds at the clause level rather than among specific predicates and parallel entities, referring expressions are not subject to these constraints, as supported by several cases of intrasententially referring pronouns in the PDTB data she provides. This is correct; the aforementioned constraints only pertain to referring expressions that are specifically related to parallel elements in the previous clause by the operative focus/presupposition partition.

however, the situation becomes more complicated. For instance, as Kaiser alludes to, demonstratives in English can refer not only to entities but also events, situations, propositions, descriptions, and speech acts (Webber 1991). The pronoun *it* is ontologically ambiguous in a similar way, as well as having bound variable and pleonastic interpretations. Further work is necessary to determine the applicability of the analysis to these more complicated referential settings (Rohde & Kehler 2013).

Having said that, we should be clear about what exactly the predictions of the theory are. Importantly, the Bayesian analysis itself is agnostic about what the linguistic properties of particular anaphors are and how they give rise to the production biases associated with them; as Arnold mentions, it only cares about the distribution $P(\textit{expression} \mid \textit{referent})$. What the analysis does predict is that the more discriminating the production biases associated with a particular form are, the less the prior next-mention bias will affect its interpretation. This appears to fit with Kaiser's findings on German demonstratives; if demonstratives have a stronger production bias than personal pronouns (in this case, toward grammatical objects), then the analysis predicts her finding that interpretation will not be modulated by coherence relations to the same degree. The kinds of switch-reference languages that Huang mentions would be the limiting case. In many such languages, anaphors receive one marking if they continue subject reference, and another if they do not. This means that (barring speaker error) the production probability for an SS-marked anaphor is essentially one if the referent is the previous subject and zero otherwise, and hence Bayes will produce a zero interpretation bias for a nonsubject referent no matter how large the next-mention bias toward the nonsubject might be. On the other hand, the analysis would have something to say about DS-marked anaphors in contexts in which more than one nonsubject was available. Speaking more generally, insofar as we understand the distinction, we imagine that the Bayesian analysis may offer a way to quantify Huang's categorization between 'syntactic' and 'pragmatic' languages. Syntactic languages would correspond to ones with strong production biases (i.e., in which the production biases associated with forms vary considerably across possible referents), which then render next mention biases less influential, whereas pragmatic languages would correspond to ones with weaker production biases (i.e., in which production biases are similar across possible referents), raising the influence of next mention biases. This characterization would turn his syntactic/pragmatic language distinction into a continuum, as well as allow for the possibility that a single language would have referential forms of both sorts.

Finally, several commentators also remark on division of labor effects in language systems, whereby the space of possible referents for particular referring expressions (e.g., strong pronouns, demonstratives) appear to be affected by the preferred interpretations of competing forms (e.g., weak pronouns). The standard Gricean explanation for such effects is that by using a more marked/prolix form

instead of a less marked/prolix one, the speaker conversationally implicates that the referent is not the one that would have been preferred by the less marked/prolix form. This results in a complementary distribution whereby, for example, weak pronouns prefer subject referents and strong pronouns and demonstratives prefer nonsubjects.

The Bayesian analysis does not in itself directly capture the existence of division of labor effects, although such effects can independently affect the production biases fed to Bayes' Rule. This may be just as well, since such effects don't always show up where one might expect. As Kaiser notes, for instance, her results on Finnish and Estonian demonstrate that the division of labor between personal pronouns and demonstratives is not symmetric. Similarly, experiments by Ueno & Kehler (2010a,b, 2011) show that whereas null and overt pronouns in Japanese differ in their biases (with null pronouns being more strongly biased to the previous subject), both in fact display a subject bias and hence do not stand in complementary distribution. The Givenness Hierarchy of Gundel, Hedberg & Zacharski (1993) offers a Gricean explanation of division of labor effects across a variety of referential form types, but as argued by Kehler & Ward (2006), it predicts a much broader range of implicatures brought about by referential form choice than is actually witnessed.

We would similarly argue against Huang's neoGricean explanation of the difference between accented and unaccented pronouns in English (Kehler 2005; Kehler et al. 2008). For one, accented pronouns are not felicitous in all of the contexts one would expect on the Gricean analysis. For example, whereas accent is licensed in contexts in which there is contrast in the context of semantic parallelism (*Mitt narrowly defeated Rick, and Newt absolutely trounced HIM*) or contrast with expectation (*John pushed Bill and HE fell*), it is considerably less felicitous in cases in which neither of these is true:

(2) a. John ran over to Bill. He said hello.

b. ?? John ran over to Bill. HE said hello.

Consistent with known biases for motion contexts, John is the preferred referent for *He* in (2a). Using an accented pronoun to switch reference to Bill instead is odd in (2b) however, absent a context that makes it surprising for Bill to say *hello*.

Second, the Gricean story doesn't explain why accented pronouns would ever be necessary in referentially unambiguous contexts:

(3) a. # Condi Rice admires Donald Rumsfeld, and George W. Bush absolutely worships her. [=Rumsfeld?]

b. Condi Rice admires Donald Rumsfeld, and George W. Bush absolutely worships HER. [=Rice]

Here accent is required on the pronoun (passage (3a) reads as if the speaker is confused about Rumsfeld's gender), but there is no alternative gender-compatible referent to shift preferences from. Lastly, the fact that accent is required even if a name is used demonstrates that constraints on accent placement apply to the referent rather than the referring expression:[7]

(4) a. # Condi Rice admires Donald Rumsfeld, and George W. Bush absolutely worships Rice.

b. Condi Rice admires Donald Rumsfeld, and George W. Bush absolutely worships RICE.

Kehler et al. (2008) describe how these facts and others can be explained through the interaction of coherence relations and information structural constraints on the placement of accent that apply to any type of referring expression. This argues not only against a Gricean explanation of accented pronouns, but any treatment that considers accented pronouns to be anything but ordinary pronouns that happen to receive accent for independent reasons. For instance, accented pronouns should not take their place as a distinct type of referential form within unidimensional theories of referential systems, such as Gundel et al.'s Givenness Hierarchy or Ariel's Accessibility Theory.

**On Being Less than a Complete Theory** Finally, a number of contributors stress that what we offer should not be considered to be a complete theory of pronouns and/or reference. We certainly don't argue with that, and to be clear that wasn't our goal. There are no doubt many, many existing questions that this analysis is in no position to answer, as well as new ones that the analysis itself evokes. Van Berkum, for example, speaks of the roles of social factors, conversational structure, intentionality, and emotions, among other influences. His insightful comments about there being multiple layers of communication to model are right on the money.[8] Arnold is similarly concerned about disfluency, distraction, working memory, gesture, eye gaze, and the like. All of these and many more factors might

7 Huang claims that unlike our model, his model predicts a contrast between accented pronouns and names (see discussion of examples 15b-c in his commentary) "though the contrast is not one at the level of reference but one at some other levels". We don't understand the contrast that Huang is referring to here, nor how his system could predict the fact that accent is not only possible in these examples, but required.

8 Staying closer to home, Kaiser's and Webber's discussions highlight the fact that even the proper inventory of coherence relations remains in dispute. Not to mention the fact that a move from monologue to conversation would require appeal to theories of coherence of quite a different sort, ones that require predictions about the interlocutors' goals, intentions, and conversational moves, as van Berkum points out.

ultimately affect pronoun usage. Hence we weren't just hedging when we said "Importantly, our analysis should not be construed as claiming that no factors other than those we have discussed influence pronominal reference", although perhaps we were guilty of understatement.

We are reminded of a quote from Carl Sagan, who once said "if you wish to make an apple pie from scratch, you must first invent the universe". So be it here as well: To have a truly adequate theory of reference, you first have to have an adequate theory of everything else. That's the problem that comes with trying to model individual pieces of an interconnected dynamical system. This, however, brings us back to our point about the importance of establishing the causal structure among the factors that influence reference. Do we want a theory of pronouns that, in van Berkum's words, resembles a FORTRAN program running on a big mainframe, full of if-then-else statements to deal with factors like gesture, emotions, and disfluency? Of course not. Those phenomena need not be built into a theory of pronouns, they just affect things that pronouns (and other linguistic phenomena) are sensitive to. Van Berkum describes interesting experimental results that (based on our understanding of his description) predict an effect of mood on pronoun interpretation, but only because it affects predictive processing generally, which in turn affects next-mention expectations, which in turn affects pronouns. It's similarly not hard to imagine how gesture and eye gaze might serve as evidence of topicality. This is what we mean when we say that the pronoun data can be complex without pronouns themselves being complex, if what's going on around them is complex. Indeed, what's going on around them is *very* complex in ways that no doubt go far beyond modeling coherence in short monologues.

We therefore see our analysis as representing initial steps toward the development of a linguistically-motivated, causal structure among factors that affect pronoun behavior, with the hopes of inspiring movement away from 'bag-of-cues' accounts and the idea that a single notion of accessibility can explain both production and interpretation. We fully expect that not all of the detailed predictions we make will hold up across a more complete variety of referential form types and languages, and welcome work that establishes that. Indeed, we hope that the strong stances taken in our contribution will open new lines of investigation and inspire further work to this end. We once again thank the commentators for their thoughtful feedback on our article.

## References

Arnold, Jennifer E., Loisa Bennetto & Joshua J. Diehl. 2009. Reference production in young speakers with and without autism: Effects of discourse status and processing constraints. *Cognition* 110. 131–146.

Ferretti, Todd. R., Marta Kutas & Ken McRae. 2007. Verb aspect and the activation of event knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33. 182–196.

Grice, H. P. 1975. Logic and conversation. In Peter Cole & Jerry Morgan (eds.), *Speech acts*, 41–58. New York: Academic Press.

Gundel, Jeanette K., Nancy Hedberg & Ron Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2). 274–307.

Kehler, Andrew. 2005. Coherence-driven constraints on the placement of accent. In *Proceedings of the 15th conference on semantics and linguistic theory (salt-15)*, 98–115. CLC Publications.

Kehler, Andrew, Laura Kertz, Hannah Rohde & Jeffrey L. Elman. 2008. Coherence and coreference revisited. *Journal of Semantics* 25. 1–44.

Kehler, Andrew & Gregory Ward. 2006. Referring expressions and conversational implicature. In *Drawing the boundaries of meaning: Neo-gricean studies in pragmatics and semantics in honor of laurence r. horn*, 177–193. Amsterdam/Philadelphia: John Benjamins.

Rohde, Hannah & Andrew Kehler. 2008. The bidirectional influence between coherence establishment and pronoun interpretation. In *21st Annual CUNY Conference on Sentence Processing (Poster Session)*. University of North Carolina at Chapel Hill.

Rohde, Hannah & Andrew Kehler. 2013. Grammatical and information-structural influences on pronoun production. *Language and Cognitive Processes*, to appear.

Rohde, Hannah, Andrew Kehler & Jeffrey L. Elman. 2007. Pronoun interpretation as a side effect of discourse coherence. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. Nashville, TN, August 1-4, 2007.

Rohde, Hannah, Roger Levy & Andrew Kehler. 2011. Anticipating explanations in relative clause processing. *Cognition* 118. 339–358.

Simner, Julia & Martin J. Pickering. 2005. Planning causes and consequences in discourse. *Journal of Memory and Language* 52. 226–239.

Ueno, Mieko & Andrew Kehler. 2010a. Grammatical and pragmatic biases in japanese pronoun interpretation. Poster presented at the 23rd Annual CUNY Conference on Human Sentence Processing.

Ueno, Mieko & Andrew Kehler. 2010b. The interpretation of null and overt pronouns in japanese: Grammatical and pragmatic factors. In *Proceedings of the 32nd annual conference of the cognitive science society*, 2057–2061. Portland, OR.

Ueno, Mieko & Andrew Kehler. 2011. Implicit causality biases in japanese pronoun interpretation. Poster presented at the 24th Annual CUNY Conference on Human Sentence Processing.

Webber, Bonnie Lynn. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes* 6(2). 107–135.