Grammatical and Information-Structural Influences on Pronoun Production

H. Rohde*[,a], A. Kehler[b]

[a] *University of Edinburgh, Department of Linguistics and English Language, Edinburgh, UK*

[b] *University of California San Diego, Department of Linguistics, La Jolla, CA, USA*

Abstract

A standard assumption in psycholinguistic research on pronoun interpretation is that production and interpretation are guided by the same set of contextual factors. A line of recent research has suggested otherwise, however, arguing instead that pronoun production is insensitive to a class of semantically-driven contextual biases that have been shown to influence pronoun interpretation. The work reported in this paper addresses three fundamental questions that have been left unresolved by this research. First, research demonstrating the insensitivity of production to semantic biases has relied on referentially-unambiguous settings in which the comprehender's ability to resolve the pronoun is not actually at stake. Experiment 1, a story continuation study, demonstrates that pronoun production is also insensitive to semantic biases in settings in which a pronoun would be referentially ambiguous. Second, previous research has not distinguished between accounts in which production biases are driven by grammatical properties of intended referents (e.g., subject position) or by information-structural factors (specifically, topichood) that are inherently pragmatic in nature. Experiment 2 examines this question with a story continuation study that manipulates the likelihood of potential referents being the topic while keeping grammatical role constant. A significant effect of the manipulation on rate of pronominalization supports the claim that pronoun production is influenced by the likelihood that the referent is the current topic. Lastly, the predictions of Kehler et al.'s (2008) Bayesian analysis of the relationship between production and interpretation has never been quantitatively examined. The results of both experiments are shown to support the analysis over two competing models.

Key words: pronoun production, implicit causality, information structure

1. Introduction

Like all natural languages, English offers its speakers[1] a wide variety of referring expressions from which to choose (e.g., proper names, definite descriptions, pronouns). This observation immediately gives rise to a question: What factors will guide a speaker in determining which of these alternatives to select in a particular context so as to communicate successfully? As is well known,

*Corresponding author

Email address: hannah.rohde@ed.ac.uk (H. Rohde)

[1] Here and throughout the paper, we use the term 'speakers' to include both speakers and writers.

the choice is far from arbitrary, being constrained by such factors as assumed knowledge of the referent by the addressee, prior evocation in the discourse, and level of activation in the addressee's mental state, among others.

Undoubtedly, the most well-studied referential form in psycholinguistics is the singular, third-person personal pronoun. The recurring claim one encounters in the literature is that pronoun use requires a high degree of activation of the referent in the cognitive state of the comprehender, a concept variously referred to as being PROMINENT, SALIENT, ACCESSIBLE, IN FOCUS, or the CENTER OF ATTENTION, among other characterizations (Ariel, 1990; Gundel et al., 1993; Grosz et al., 1995; Arnold, 2001, 2010, inter alia). The picture that results is that production and interpretation are mirror images of each other: Speakers produce pronouns to denote referents that they believe to be prominent in the comprehender's mental model of the discourse at the time of utterance, and correspondingly, comprehenders interpret pronouns to refer to such entities. The remaining task for psycholinguistic research is merely to identify those factors (grammatical, semantic, information-structural, and so forth) that determine the degree of prominence of potential referents.

Numerous authors have weighed in on this question, positing a range of biases that involve both structurally-driven and semantically-driven factors (Section 2). In this paper, we examine the predictions of a Bayesian model proposed by Kehler et al. (2008) that specifies a more complex relationship between production and interpretation (Section 3). According to this model, structural and semantic factors play fundamentally different roles: Whereas pronoun production biases are determined primarily by structural factors, the interpretation process integrates these biases with semantically-driven expectations about what entity will be mentioned next – henceforth referred to as NEXT-MENTION biases – that hold independently of the form of reference ultimately chosen by the speaker. A consequence of the model is that pronoun production and interpretation processes are in fact not mirror images of each other, since pronoun interpretation utilizes a set of contextual factors that production processes ignore – an intuitively strange set of affairs.

We report on two experiments that provide evidence for the model by way of addressing three fundamental questions that have as yet been left unanswered. First, previous experiments that have offered preliminary support for the insensitivity of pronoun production to semantically-driven next-mention biases have employed unambiguous contexts. This is problematic since it stands to reason that such biases could be safely ignored by a speaker when a pronoun would not be ambiguous to begin with. Experiment 1, a story continuation study, examines production biases in a referentially-ambiguous setting, and shows that the insensitivity of production to next-mention biases extends to ambiguous situations. Second, previous work has not pinpointed the underlying source of the biases on the production side: specifically, whether they are grammatically-driven (e.g., by subjecthood), as much of the literature has assumed, or following the claims of early incarnations of Centering Theory (Grosz et al., 1995) as well as linguistic accounts (Ariel, 1990; Gundel et al., 1993; Lambrecht, 1994), are driven instead by information-structural biases (specifically, topichood) that are inherently pragmatic in nature. Experiment 2 examines this question with a story continuation study that manipulates the likelihood of potential referents being the topic while keeping grammatical role constant. In stark contrast to the lack of effect of semantically-driven next-mention biases, the results demonstrate that topicality influences pronoun production beyond what can be accounted for by grammatical role. Third, the predictions about interpretation biases made by the Bayesian analysis have not been quantitatively tested to date. The results of Experiment 1 and 2 are both shown to support the predictions of the account.

2. Background: Interpretation Biases

Previous literature has posited a variety of factors that influence how comprehenders will assign referents to pronouns. For instance, Crawley et al. (1990) have argued that comprehenders use a SUBJECT-ASSIGNMENT STRATEGY, whereby the entity mentioned from the subject position of the previous clause is, all else being equal, considered more prominent than referents occupying other grammatical roles. The Centering-based interpretation algorithm of Brennan, Friedman, and Pollard (1987) and Walker, Iida, and Cote (1994) also ranks subjects as higher than other referents on a prominence hierarchy (specifically, the *forward-looking center* list), although the hierarchy affects pronoun interpretation only indirectly, through a ranking of discourse transition types that arise from different referent assignments.[2] On the other hand, researchers such as Gernsbacher and Hargreaves (1988) and Gernsbacher et al. (1989) have argued for a FIRST-MENTION ADVANTAGE, whereby it is the first-mentioned entity that has a privileged status for subsequent reference (see also Järvikivi et al. (2005)). Other researchers, including Smyth (1994) and Chambers and Smyth (1998), have argued instead for a PARALLEL GRAMMATICAL ROLE preference, whereby the preferred referent for a pronoun will be the entity that occupies the same grammatical role in the previous clause. In this account, the prominence of the parallel entity falls out naturally from a structurally-governed Extended Feature Match process. Yet other researchers have argued for the role of semantic and world knowledge as the central factor that determines the preferred referent. According to Hobbs (1979), for example, pronouns are represented merely as free variables that get bound to referents as a side-effect of discourse-level inference processes. Finally, some researchers have posited that prominence is determined by a combination of many factors, with none being singled out as primary. According to Arnold's (2001) EXPECTANCY HYPOTHESIS, for instance, prominence is inherently probabilistic, correlated directly with the likelihood that a particular entity will be mentioned next. Numerous factors determine likelihood of next mention; the aforementioned subjecthood, grammatical parallelism, and semantic influences being among them. Despite their differences, these accounts all share the implicit assumption that the same contextual properties that license the speaker to use a pronoun are those that the comprehender will rely on to interpret it.

Another line of work has painted a more complex picture, however; one which suggests that pronoun production and comprehension are not driven by the same set of contextual factors (Stevenson et al., 1994; Miltsakaki, 2007; Rohde, 2008; Kehler et al., 2008; Fukumura and van Gompel, 2010). Stevenson et al. (1994) reported on a series of story continuation experiments that investigated pronoun biases across eight distinct context types. Particularly revealing were the results for two contexts that employed transfer-of-possession verbs as in (1).

(1) *Transfer-of-possession contexts from Stevenson et al. (pronoun condition)*
   a. John seized the comic from Bill. He _____
   b. John passed the comic to Bill. He _____

Both contexts describe events with Source and Goal referents; in (1a) the Goal occupies the subject position, whereas in (1b) it appears as the object of a prepositional phrase. Stevenson et al. found

[2] And hence, contra a claim found in the psycholinguistics literature (Chambers and Smyth, 1998, inter alia), this algorithm will not always identify the previous subject as the preferred referent.

that Goal-Source prompts like (1a) yielded significantly more continuations in which the ambiguous pronoun *He* is used to refer back to the subject than the non-subject (85%). Following a Source-Goal prompt like (1b), on the other hand, participants were equally likely to refer back to the subject and non-subject (51% bias to the subject). The fact that an object-of-PP referent at the end of the sentence – normally a relatively non-prominent position for pronominal reference – competes with the sentential subject is somewhat surprising, especially in light of the previously discussed first mention, subject assignment, and parallel grammatical role biases, all of which favor the subject. Stevenson et al. concluded that there are two types of bias at work in such examples: a thematic role bias (here, favoring Goals over Sources), and a grammatical role bias (favoring subjects over other roles). The Goal bias and the subject bias agree on the same referent in the Goal-Source condition, predicting the large percentage of Goal interpretations. The two biases compete in the Source-Goal condition, on the other hand, predicting an even distribution of assignments.

Importantly, Stevenson et al. concluded that the way in which these two biases come to affect pronoun interpretation are fundamentally different: The thematic role bias emerges via a 'top-down' predictive mechanism that determines the prominence of an entity even before a referring expression is encountered (and hence applies regardless of the form of referring expression chosen by the speaker), whereas the subject bias results from a form-specific, 'bottom-up' response to the presence of a pronoun. Evidence for the top-down nature of thematic role biases came from the fact that, in addition to the pronoun-prompt conditions in (1a)-(1b), they also tested versions that did not include the pronoun as in (2a)-(2b).

(2) *Transfer-of-possession contexts from Stevenson et al. (no-pronoun condition)*

a. John seized the comic from Bill. _____

b. John passed the comic to Bill. _____

Here, participants not only choose who to refer to first in the continuation but also the form of reference to use; the distribution of first mentions therefore provides estimates of the next-mention biases that comprehenders have after reading the context sentence. Stevenson et al. found a Goal bias in this condition that paralleled the one found in the pronoun condition, supporting the idea that the effect is independent of whether a pronoun is used. Evidence for the form-specific, 'bottom-up' nature of the subject bias, on the other hand, came from the fact that there was a greater percentage of first-mentions of the previous subject in the pronoun conditions than in the corresponding no-pronoun conditions. Taken together, these two results support the idea that the occurrence of a pronoun contributes a subject bias that operates independently of, albeit in concert with, contextually-driven, top-down next-mention biases in determining the ultimate interpretation bias for the pronoun.

3. Background: Production Biases

The foregoing data explains how semantic biases toward entity next-mention can conspire with a subject bias associated with pronouns to produce the ultimate interpretation bias in ambiguous contexts. There is one result of Stevenson et al.'s that remains unexplained by this picture, however. Recall that in their no-pronoun condition (2), participants not only chose who to refer to first, but also the form of reference to use. Across their stimulus types, they found that this choice was heavily biased towards a pronoun when the referent was the previous subject, and likewise towards a name when the referent was a non-subject. (Arnold (2001) also found correspondingly strong biases in a similar experiment.) This result may at first seem contradictory: If participants have a clear preference to use pronouns to refer to the previous subject and names to refer to non-subjects, why would the interpretation bias for the pronouns in passages like (1b) be 50-50? Hence we have evidence for a dissociation between production and interpretation.

Kehler et al. (2008) offered an explanation for the apparent contradiction by modeling the relationship between production and interpretation in terms of Bayes' Rule, as shown in (3).

$$p(\textit{referent} \mid \textit{pronoun}) = \frac{p(\textit{pronoun} \mid \textit{referent})\ p(\textit{referent})}{\sum_{\textit{referent} \in \textit{referents}} p(\textit{pronoun} \mid \textit{referent})\ p(\textit{referent})} \tag{3}$$

The term $P(\textit{referent} \mid \textit{pronoun})$ represents the interpretation bias: the probability, given that a pronoun has occurred, of it being used by the speaker to refer to a particular referent. On the other hand, the term $P(\textit{pronoun} \mid \textit{referent})$ represents the production bias: the probability, assuming that a particular entity is being referred to, that the speaker would have used a pronoun to refer to it. Bayes' Rule says that these biases are not mirror images of each other, but instead are related by the prior $P(\textit{referent})$, which represents the next-mention bias: the probability that a particular referent will get mentioned next regardless of the referring expression used.[3] Equation (3) thus explains why there is nothing contradictory about having both a strong production bias toward pronominalizing the previous subject (and not pronominalizing non-subjects) and yet a lack of a subject bias in interpretation, as long as the prior $P(\textit{referent})$ points strongly enough away from the subject referent. The roughly 50-50 distribution of references to the Source and Goal in transfer-of-possession passages like (1b) results from the fact that for the subject referent, the next-mention bias is low and the pronominalization bias is high, whereas for the non-subject referent, the next-mention bias is high and the pronominalization bias is low.[4]

On Kehler et al.'s analysis, therefore, the bottom-up interpretation bias toward subjects that Stevenson et al. posited for pronouns is actually the result of a *production* bias toward pronominalizing references to the subject. A comprehender's interpretation bias thus relies jointly on his

[3]The denominator of (3) is simply the probability that a pronoun is the form of reference chosen by the speaker ($P(\textit{pronoun})$), which can be computed by summing the numerator over all referents that are compatible with the pronoun. This term has the effect of normalizing the probabilities to 1.

[4]Arnold (2001) reported on a story continuation study that included Source-Goal transfer-of-possession contexts like (2b). She found the same production asymmetry as Stevenson et al., whereby 76% of references to the subject Source were pronominalized yet only 20% of those to the object-of-PP Goal were. However, the next-mention bias toward the Goal was an overwhelming 86%. Equation (3) would actually predict a 61% interpretation bias to the Goal for her stimuli if a pronoun prompt were to be included (Kehler et al., 2008).

estimates of the likelihood that a particular referent will be mentioned next (regardless of form of reference) and of the likelihood that the speaker would have chosen a pronoun (instead of another form of reference) to refer to that referent. The predictions of the analysis can be tested using story continuation experiments: The values for the terms on the right-hand side of equation (3) can be estimated from the data collected in no-pronoun prompt conditions (2), which will yield a prediction for $p(\textit{referent} \mid \textit{pronoun})$. If the Bayesian characterization is correct, this predicted bias should be highly correlated with the actual interpretation biases estimated directly in pronoun-prompt conditions (1) in otherwise identical contexts.

At this point an intriguing picture has emerged; one which suggests that the contributors to the ultimate interpretation bias are conditioned by different sets of contextual factors. On the one hand, the data suggest that the factors that condition the next-mention bias $P(\textit{referent})$ are primarily semantic. On the other hand, the factors that condition the production bias $P(\textit{pronoun} \mid \textit{referent})$ appear to be structural (e.g., based on grammatical role). Considering this in light of the asymmetry between production and interpretation captured by equation (3), this picture makes a striking prediction: that the speaker's decision about whether or not to pronominalize a reference will be insensitive to the semantically-driven contextual factors that in part determine the comprehender's interpretation biases. This hypothesis is surprising because it violates the intuition that speakers will pronominalize mentions of referents in just those cases in which their comprehenders would be expected to interpret the pronouns to those same referents.

The results of several recent story-continuation studies have provided preliminary support for this prediction. For instance, Rohde (2008, Experiment VII) reports on a continuation study that employed an aspect manipulation (perfective vs. imperfective) using transfer-of-possession contexts with no-pronoun prompts (e.g., *Sue handed / was handing a timecard to Fred.* _____). (The motivation for this design was a previous experiment that demonstrated that this aspect manipulation significantly influences interpretation biases, with a greater number of references to the Goal in the perfective case than in the imperfective case (Rohde et al., 2006)). As expected, the manipulation influenced next-mention biases in the predicted direction, both for the set of all continuations as well as for the subset of continuations in which the participant chose to use a pronoun. However, the aspect manipulation had no effect on rate of pronominalization; only grammatical role mattered.[5]

Other experiments have employed contexts with so-called IMPLICIT CAUSALITY (IC) verbs. Such verbs (e.g., *impress*, *admire*, *detest*, *annoy*, *congratulate*) are well known to create a bias to re-mention the causally-implicated referent, especially in an upcoming clause that provides a cause or reason (Garvey and Caramazza, 1974; Caramazza, Grober, Garvey, and Yates, 1977; Brown and Fish, 1983; Au, 1986; McKoon, Greene, and Ratcliff, 1993; Kehler, Kertz, Rohde, and Elman, 2008, inter alia). Some IC verbs are SUBJECT-BIASED, such as *impress* in (4a), because the causally-implicated referent occurs in subject position. Other verbs are OBJECT-BIASED, such as *admire* in (4b), because the causally-implicated referent occurs in object position.

[5] Arnold (2001) reported an effect of thematic role on rate of pronominalization in her study of transfer-of-possession contexts, whereby references to the Goal were pronominalized more often than references to the Source. Fukumura and van Gompel (2010) point out a number of reasons to suggest that Arnold's result is not fully conclusive, however, including the fact that her Source-Goal and Goal-Source contexts differed not only in the order of the thematic role fillers but in a number of other potentially relevant respects as well.

(4) *Implicit causality contexts with opposite-gender referents*

a. John impressed Mary. _____

b. John admired Mary. _____

The strong divergence between the next-mention biases for subject-biased and object-biased IC verbs makes the alternation particularly useful for examining whether rate of pronominalization is affected by semantic factors. Rohde (2008, Experiment V) elicited story continuations with contexts such as (4a) and (4b), as well as with non-IC context controls (*John chatted with Mary.* _____). Whereas grammatical role again influenced rate of pronominalization, there was no interaction between grammatical role and verb type, showing that the rate of pronominalization does not depend on the next-mention bias of the verb. Two experiments reported by Fukumura and van Gompel (2010) yielded highly complementary results. Their first experiment used contexts that varied IC bias with *because* prompts (*Gary scared/feared Anna after the long discussion ended in a row. This was because* _____). Their second experiment used only subject-biased IC verbs but varied the connective between *because* and *so*; these verbs are known to flip from a subject bias with *because* to an object bias when *so* is used (Stevenson et al., 1994). In both cases the manipulation affected the choice of who got mentioned next but not the rate of pronominalization. Once again, the only factor that affected pronominalization rate was the grammatical role of the referent.

A legitimate complaint that could be lodged against these studies, however, is that the contexts mention two opposite-gendered individuals, meaning that the subsequent pronoun is never ambiguous. As such, they do not demonstrate that interpretation biases do not influence the rate of pronominalization when an ambiguity is present; it is quite possible that speakers would only attend to interpretation biases in contexts in which the comprehender's ability to resolve the pronoun to the correct referent is actually in question. To address this, our first experiment examines production biases in IC contexts in a gender-ambiguous context. Establishing the lack of effect of semantic bias in such contexts not only provides an important test of the analysis, but it is also a necessary prerequisite to our Experiment 2, which will use gender-ambiguous contexts to examine whether production biases are driven primarily by grammatical or information-structural factors.

4. Experiment 1

To address the question of whether production biases are immune to semantic biases in situations in which reference is ambiguous, we conducted a story continuation study using contexts that contain competing referents of the same gender.[6] The story continuation prompts in (5) implement a 3x2 design, utilizing three types of context (containing subject-biased IC verbs, object-biased IC verbs, and non-IC verbs), as well as the now-familiar prompt manipulation in which a sentence-initial pronoun for the continuation either is or is not included.

[6] This study appeared as Experiment VI in the first author's dissertation (Rohde, 2008). It is presented here with additional analyses pertaining to the Bayesian model.

(5) *Manipulation of verb bias and continuation prompt in gender-ambiguous contexts*

a. [Subject-biased IC verb, no-pronoun] John infuriated Bill. _____
b. [Object-biased IC verb, no-pronoun] John scolded Bill. _____
c. [Non-IC verb, no-pronoun] John chatted with Bill. _____

d. [Subject-biased IC verb, pronoun] John infuriated Bill. He _____
e. [Object-biased IC verb, pronoun] John scolded Bill. He _____
f. [Non-IC verb, pronoun] John chatted with Bill. He _____

Crucial to the design is the expectation that previous results for IC-driven next-mention and interpretation biases will be replicated. That is, we expect to find more first mentions of the subject following subject-biased IC contexts than object-biased IC contexts in both the no-pronoun (next-mention bias) and pronoun (interpretation bias) conditions. (We have no predictions about the more heterogeneous set of verbs used for the non-IC context condition, which was included as a control to ensure that the production results are consistent across a broader set of context types.) Finally, the production bias associated with pronouns predicts a greater number of first mentions of the subject in the pronoun-prompt condition than the corresponding no-pronoun condition for each context type.

Replicating these effects sets the stage for testing our hypothesis about production. Specifically, if production biases are unaffected by next-mention biases, then the rate of pronominalization is predicted to be the same across (5a)-(5c). As such, our hypothesis predicts a main effect of referent position (subject vs. non-subject) on rate of pronoun production, but the effect of referent position is not expected to vary with verb bias (no referent position × verb bias interaction). On the other hand, an interaction with verb bias would suggest that speakers are in fact accounting for interpretation biases in their decisions to pronominalize, presumably favoring pronominalizations of the referent favored by the IC bias of the verb.

4.1. Methods

Participants

Twenty-eight monolingual English-speaking undergraduates from UC San Diego participated in the experiment either for extra credit in Linguistics courses or for the chance to be entered in a raffle to win a gift certificate.

Materials and Procedure

For the experimental items, each context sentence mentioned two referents in a situation described with a subject-biased IC verb, an object-biased IC verb, or a non-IC verb, as in (5). The two competing referents were of the same gender, counterbalanced between male and female names. 80 verbs were taken from Kehler et al. (2008), consisting of 40 IC verbs (20 subject-biased, 20 object-biased) and 40 non-IC verbs. The experiment consisted of one hundred items: eighty experimental items (40 IC, 40 non-IC) intermixed with twenty non-IC fillers. The additional fillers consisted of prompts with context sentences containing non-IC verbs followed by intersentential connectives, no-pronoun, or pronoun prompts.

Story continuations were collected via a web-based interface that participants could access from their own computer. Each item was presented on a page by itself with a text box in which

participants were instructed to write their continuation. The entire experiment took roughly forty-five minutes, but participants were encouraged to have an hour available so that the experiment could be completed in one session. Participants could leave the website and return at a later time by identifying themselves with an ID number. They were instructed to imagine a natural story continuation for each prompt, writing the first continuation that came to mind and avoiding humor.

In this task, participants create a mental model of the event in the context sentence and then write a continuation that reflects their expectations about where the story is going. As such, the task involves both interpretation and production (Arnold, 2001). The pronoun provided in the pronoun-prompt condition constrains the surface realization of their continuation, but their continuation depends on their expectations about how the discourse will proceed and which individual in the event will be mentioned again.

4.2. Evaluation and Analysis

Two judges, the first author of this paper and a UCSD Linguistics undergraduate, coded the first mentioned referent in each continuation (which, in the case of the pronoun-prompt condition, was the assignment of the pronoun). The judges were instructed to err on the side of categorizing a pronoun as ambiguous if the pronoun could be interpreted as plausibly coreferential with either referent, even if their own interpretation biases suggested a particular one.

To measure the effects of several predictors on the observed choice of first mention and the observed choice of referring expression, we used mixed-effect logistic regressions (Jaeger, 2008). We modeled the binary choice of first mention (subject vs. non-subject) with fixed-effect predictors for verb bias, prompt type, and the interaction between the two. For the binary choice of referring expression (pronoun vs. not), we considered only the continuations elicited in the no-pronoun condition and modeled the observed referring expressions with predictors for referent position, verb bias, and the interaction between the two. Both prompt type and referent position varied within participants and within items; verb bias only varied within participants. The referent position and prompt type predictors were centered. For verb bias, which is a 3-level predictor, we used sum coding in order to be able to test for main effects in the presence of interactions. All models contained maximal random effects structure for both participants and items, namely random intercepts plus random slopes for all predictors and their interactions (Barr et al., 2013). We report the coefficient estimate and p-value for each binary predictor (based on the Wald Z statistic; Agresti, 2002). For the sum-coded verb bias predictor and its interactions, we use likelihood ratio tests to compare mixed-effects models differing only in the presence or absence of the fixed factor that pertains to verb bias.

4.3. Results and Discussion

The results reflect a conservative analysis in which a continuation was excluded if at least one judge assessed the pronoun reference as ambiguous (15.7% of the total 2240 continuations). Also excluded were cases in the no-pronoun condition in which both referents were mentioned together in a plural pronoun or conjoined noun phrase (10.0%), neither was mentioned at all (2.1%), the mention used a referring expression other than a pronoun or name (1.9%), or the

relationship between the continuation and the prompt was not clear (2.6%). This left a dataset of 1516 continuations.

We first ask whether the context manipulation had the predicted effects on next-mention (no-pronoun condition) and pronoun interpretation (pronoun condition) biases. Figure 1 shows the rate at which participants wrote continuations about the subject across the six conditions. As predicted, the results replicated the widely reported IC bias—that the preferred entity for next mention following an IC verb is the causally-implicated one. To test for a main effect of verb bias, we conducted a likelihood-ratio test between mixed-effects models differing only in the presence or absence of a fixed main effect of verb bias. Both models included in their fixed effects an intercept, a main effect of prompt type, and an interaction between verb bias and prompt type. The likelihood-ratio test showed a main effect of verb bias ($p<0.001$, 1 d.f.). Pairwise comparisons showed that the binary predictor verb bias is a significant factor for modeling choice of next mention in the subsets containing only subject-biased and object-biased IC verbs (β=1.266, $p<0.001$) and only object-biased IC verbs and non-IC verbs (β=1.034, $p<0.001$), but not for the subset containing subject-biased IC verbs and non-IC verbs (β=0.248, $p=0.16$).

In keeping with previous results and the predictions of the Bayesian model, prompt type was a significant factor as well (β=1.068, $p<0.001$), with pronoun prompts being associated with higher rates of subject first mentions than no-pronoun prompts: 90.4% vs. 73.0% for subject-biased IC verbs; 61.6% vs. 26.4% for object-biased IC verbs; 86.1% vs. 54.3% for non-IC verbs. Again using model comparison, a likelihood-ratio test showed no evidence for a verb bias × prompt type interaction ($p=0.53$, 1 d.f.).

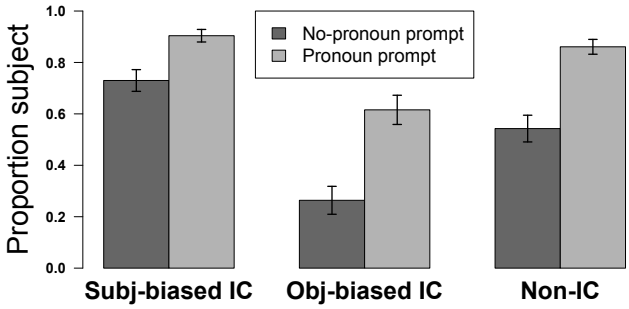

Figure 1: Proportion of continuations about the subject, by verb bias and prompt type

We now ask whether the different semantic biases that influence next mention and pronoun interpretation preferences have an effect on pronoun production. The rates of pronominalization in the no-pronoun condition are shown in Figure 2. The analysis of referring expressions replicates the previously reported bias to pronominalize references to the subject. In a logistic regression, referent position was the only significant factor in modeling the binary outcome of pronominal referring expression (β=-2.14, $p<0.001$), with more subject-referring pronouns than object-referring pronouns across all verb types: 77.5% vs. 26.6% for subject-biased

IC verbs; 80.8% vs. 21.7% for object-biased IC verbs; 85.0% vs. 16.5% for non-IC verbs. To test for a main effect of verb bias and its interaction with referent position, we again used a sum-coding numeric representation of the verb bias predictor. We conducted two likelihood-ratio tests between mixed-effects models differing only in the presence or absence of a verb bias main effect or interaction. The likelihood-ratio test showed no main effect of verb bias (p=0.56, 1 d.f.) nor a referent position × verb bias interaction (p=0.95, 1 d.f.).[7]

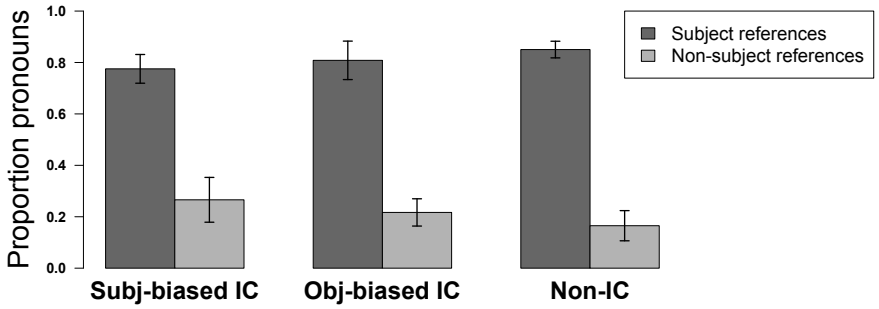

Figure 2: Rate of pronominalization, by verb bias and referent position (subject vs. non-subject)

As such, the lack of effect of semantic biases on pronoun production previously shown for unambiguous contexts is found in ambiguous contexts as well: Even when the comprehender's ability to successfully interpret the pronoun is at stake, pronoun production biases are not affected by the same semantically-driven contextual factors that have been demonstrated to influence pronoun interpretation. Counter to the intuition that speakers will pronominalize mentions of referents in just those cases in which comprehenders will be biased to assign pronouns to those same referents, it is most striking that we find the same rates of pronominalization in contexts that give rise to strongly divergent next mention biases. Instead, the evidence just surveyed suggests that grammatical role is the critical factor in determining pronoun production biases, without regard to the bias that the comprehender will bring to the interpretation process after factoring in prior, top-down next-mention expectations.

Finally, as mentioned in Section 3, we can use the data collected here to test our Bayesian Hypothesis, i.e., that equation (3) captures the relationship between pronoun production and pronoun interpretation biases. We compare the predictions of the Bayesian model and two competing models against the actual interpretation biases witnessed, for each participant and item. The first competing model is what we call the Expectancy model, according to which the interpretation bias toward a referent equals the probability that the referent gets re-mentioned. This prediction follows from the Expectancy Hypothesis of Arnold (2001), as well as the Hobbsian treatment of pronouns as unbound variables (Hobbs, 1979). The predicted interpretation bias for this model is thus estimated to be

[7] As mentioned earlier, these analyses reflect a conservative data inclusion strategy in which a continuation was excluded if at least one coder assessed it as ambiguous. The pattern of statistical significance for all results remains the same if continuations are included for which at least one coder assigned a non-ambiguous interpretation.

the next-mention bias $P(\textit{referent})$ measured in the no-pronoun condition. The second competing model is what we call the MIRROR MODEL, according to which the interpretation bias toward a referent is proportional to the likelihood that a speaker would produce a pronoun to refer to that referent. Capturing the intuition that speakers will choose a pronoun in those cases in which hearers will be biased toward the correct referent, the predicted interpretation bias for this model is estimated using the pronominalization rate $P(\textit{pronoun} \mid \textit{referent})$ measured in the no-pronoun condition. Because these values will not typically result in a valid probability distribution (i.e., the probabilities over referents will not sum to 1), the values are normalized with a scaling factor (the sum of the pronominalization rate of both referents). Finally, for the Bayesian model, the predicted interpretation bias results from combining estimates of the probabilities utilized by both of these models: the prior probability for next-mention of a referent $P(\textit{referent})$ and the probability of producing a pronoun when re-mentioning the referent $P(\textit{pronoun} \mid \textit{referent})$, scaled by the normalizing probability $P(\textit{pronoun})$ which, per equation (3), is the numerator summed over the two referents in question. We then compare these three predicted interpretation values against the observed pronoun interpretation biases $P(\textit{referent} \mid \textit{pronoun})$, as measured by the data collected in the pronoun-prompt condition.

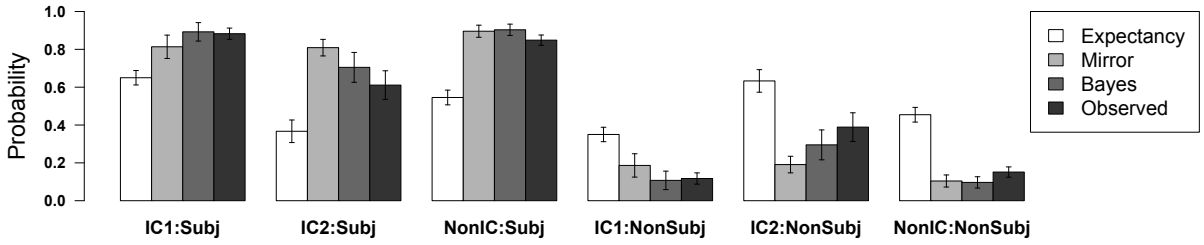

Figure 3: Estimated and observed pronoun interpretation biases across conditions and referents in Experiment 1. For example, the bars in the "IC1:Subj" group show the predicted rates at which pronouns would be interpreted as referring to the subject referent in IC1 contexts under (i) the Expectancy model, (ii) the Mirror model, and (iii) the Bayesian model; these rates can be compared to (iv) the Observed rate. Error bars represent standard error over participant means.

We expect that the predictions of all models will reveal some degree of correlation with the observed data: The Mirror model should capture the differences in biases between subject and non-subject referents, whereas the Expectation model should capture differences across context type. Crucially, however, in combining the biases captured by both models, we expect the Bayesian model to be more highly correlated than either of the other models alone. Figure 3 shows the observed and model-predicted rates at which a pronoun will be interpreted to refer to referents in the two grammatical roles across the three verb bias conditions. As can be seen, the observed values are

most consistently matched by the Bayes-derived values. [8] For the statistical analysis, we use a linear model to test the correlation between the values observed in the pronoun-prompt condition and the values under the three different models. The correlation is performed over participant and item means; each participant (or item) contributes a value for the four pronoun interpretation estimates in each of the verb bias × referent combinations. We excluded data from participants (or items) for which the Mirror- and Bayes-derived values could not be estimated—specifically if a participant's no-pronoun prompt responses for a particular verb type contained no mentions of a particular referent or no pronouns for either referent; in both cases computing the predicted probabilities of the Mirror and Bayesian models would involve division by zero.

	Obs∼Mirror	**Obs∼Expectancy**	**Obs∼Bayes**
by participants	R^2=0.59*	R^2=0.13*	R^2=0.69*
by items	R^2=0.69*	R^2=0.16*	R^2=0.74*

Table 1: Correlations between observed data and model predictions, by participants and by items. * indicates significance at or below 0.001.

The results are shown in Table 1. As predicted, whereas all of the models are correlated with the observed data, it is the Bayes-derived values that provide the closest and most consistent fit. [9]

5. Interlude: Subjecthood versus Topichood

The results of Experiment 1 and those of the previous studies surveyed agree that the semantically-driven next-mention biases that have been demonstrated to affect interpretation do not influence production. On the other hand, these results also agree that grammatical role *does* have an effect on production: Across experiments, rates of pronominalized reference to the previous subject are substantially higher than for other grammatical roles. A grammatical-role based preference is not the only possible explanation for this pattern, however. In particular, several authors have claimed that a central function of pronouns is to signal a continuation of the current TOPIC

[8] As can be seen in Figure 3, the predictions of the Bayesian and Mirror models are very close for the non-IC condition. This reflects the fact that the next-mention biases for this condition were close to 50-50 for the two referents. The two models are equivalent when the prior is uniform.

[9] As mentioned above, our account predicts that the correlation between the values produced by the Mirror model and actual interpretation biases are due to the strong effect of grammatical role on production. Hence, we would not expect to find a correlation if we analyze only one of these sets of referents (i.e., subjects or non-subjects). This prediction is borne out: When the correlation is restricted to one referent only, the R^2 values drop considerably (by participants: R^2=0.07, $F_1(1,61)$=5.377, $p<0.05$; by items: R^2=0.09, $F_2(1,69)$=8.022, $p<0.01$) compared to the values shown in Table 1 in which both referents were included. Further, the remaining significance is likely driven by the fact that not all non-subject referents were direct objects; some were objects of prepositional phrases and hence expected to have an even lower pronominalization rate than direct objects. In an analysis that excludes those items in which the non-subject was an object-of-PP (almost all the non-IC verbs as well as the IC verbs *apologize to*, *confess to*, and *stare at*), the correlation between the observed data and the Mirror model is no longer significant (by participants: R^2=0.04, $F_1(1,40)$=2.865, p=0.10; by items: R^2=0.01, $F_2(1,40)$=1.557, p=0.22).

(Ariel, 1990; Gundel et al., 1993; Lambrecht, 1994; Grosz et al., 1995, inter alia). The sentence topic, a pragmatic, information-structural concept rather than a grammatical one, is commonly (albeit informally) characterized as the constituent that expresses what the sentence is about, i.e., as denoting the entity about which an utterance is primarily intended to expand the addressee's knowledge (Strawson, 1964; Kuno, 1972; Gundel, 1974; Reinhart, 1981; Lambrecht, 1994, inter alia). While subject and topic are highly correlated in English – indeed, subject position is the canonical place for a topic to appear – the two notions cannot be conflated. Sometimes non-subjects serve the role of sentence topic (Lambrecht, 1994, p. 118); for instance the sentence *Few people amaze Brittany* intuitively predicates a property of *Brittany* and not of *few people*.

The idea that one of the functions of pronouns is to signal a continuation of the current topic – which we refer to as the TOPICHOOD HYPOTHESIS – offers an alternative explanation of the production data that we have seen in the foregoing experiments. Under this hypothesis, speakers preferentially realize a continuing topic as a pronoun, whereas comprehenders must infer the identity of the topic from properties of the discourse context, including the grammatical roles occupied by the alternatives. On this analysis, therefore, the declining rates of pronominalization we find as one moves down the grammatical obliqueness hierarchy (subjects > objects > other referents) in fixed word order languages would therefore reflect the declining *likelihood* that an entity in that position is the topic, rather than being related directly to grammatical role.[10]

So the question before us is whether production biases are really dictated by the grammatical roles that potential referents occupy or by the likelihood that a potential referent is the current topic. None of the experiments carried out to date resolve the issue, since none have manipulated the likelihood that a potential referent is the topic without also varying grammatical role. To examine this question, we can take advantage of the fact that different syntactic constructions mark the potential for topicality of their grammatical role occupants to different degrees, in some cases placing the presumed topic in a particular syntactic position (Davison, 1984; Gundel and Fretheim, 2004; Ward and Birner, 2004, inter alia). Consider the contrast between active and passive voice:

(6) a. Amanda amazed Brittany.
b. Brittany was amazed by Amanda.

Sentences (6a) and (6b) convey the same proposition. However, not only do they differ with respect to which entity is considered the default topic (the subjects Amanda in (6a) and Brittany in (6b)), but they also differ with respect to the likelihood that their respective subjects serve as the topic: Brittany is more likely to be the topic in (6b) than Amanda is in (6a). That is, whereby Amanda is merely the default topic in (6a) outside of a larger context, the promotion of another constituent to the syntactic subject (and hence, topical) position in (6b) constitutes a much stronger signal that the subject is the topic; indeed establishing a non-Agent as a topic is commonly considered to be one of the primary functions of the passive (Shibatani, 1985; Givón, 1990, inter alia).

[10] An alternative to casting the phenomenon in terms of likelihood of being the topic would be to treat topicality itself as a gradient rather than binary concept (Givón, 1983; Arnold, 2010), with grammatical subjects being more topical than objects and so forth. The proposal offered here is compatible with either possibility, and hence we will not attempt to resolve the issue further.

We can thus use the active-passive alternation to test the hypothesis that manipulating the likelihood of potential referents being the topic will influence speakers' biases to pronominalize even when grammatical position is kept constant. We do this in Experiment 2.

6. Experiment 2

Based on the idea that being the subject of a passive voice clause is a stronger indicator of topichood than being the subject of an active voice clause, we ask whether this difference has an effect on rate of pronominalization of the subject across constructions. Consider the story continuation prompts in (7), which describe the same event in the active and passive voice. Following standard terminology, we will henceforth refer to Amanda as the LOGICAL SUBJECT in all four prompts, and Amanda as the SYNTACTIC SUBJECT in (7a) and (7c) (and likewise Brittany in (7b) and (7d)). Because the verb *amaze* in (7) is known to be a subject-biased IC verb, it is Amanda who is implicated as the cause of the event in all cases.

(7) *Manipulation of topichood and continuation prompt*

a. [Active, no-pronoun prompt] Amanda amazed Brittany. _____
b. [Passive, no-pronoun prompt] Brittany was amazed by Amanda. _____
c. [Active, pronoun prompt] Amanda amazed Brittany. She _____
d. [Passive, pronoun prompt] Brittany was amazed by Amanda. She _____

If production biases are sensitive specifically to grammatical role, then the rate at which participants produce pronominalized references to the syntactic subject in the no-pronoun condition is predicted to be the same across (7a) and (7b). If the Topichood Hypothesis is correct, however, we expect the rate at which participants pronominalize references to the syntactic subject in the passive (7b) to be higher than the rate at which they pronominalize such references in the active case (7a).

Our study will also examine a set of secondary predictions. First, recall that the first-mention statistics revealed by the continuations in the no-pronoun prompt condition estimate the biases toward next mention that participants favor before they encounter any referring expression. Based on the IC biases reported in previous studies, participants are predicted to rely primarily on the verb's semantic bias in the no-pronoun condition and therefore write more continuations about the logical subject Amanda than Brittany. Second, as before, if pronoun production is conditioned by subjecthood or topichood, the Bayesian formulation predicts that providing a pronoun in each pronoun-prompt condition should pull the distribution of first mentions toward the subject position as compared to the corresponding no-pronoun condition. Third, the Bayesian analysis also predicts that, in the pronoun condition, there will be a greater number of references to Amanda (the IC congruent referent) in the active condition (7c) than in the passive one (7d), since the fact that Brittany is the surface subject of (7d) is expected to counteract the IC bias toward Amanda. Finally, we will test the correlation between the actual interpretation biases measured in the pronoun condition and the values predicted by equation (3) using data collected in the no-pronoun condition, as we did in Experiment 1.

6.1. Methods

Participants

Forty-two monolingual English-speaking undergraduates from UC San Diego participated in the experiment for extra credit in Linguistics courses.

Materials and Procedure

For the experimental items, each context sentence mentioned two referents in an event described with a subject-biased IC verb, as in (7). The two competing referents were of the same gender, counterbalanced between male and female names. Twenty verbs were taken from a set of subject-biased IC verbs that have been used in previous studies on IC: *aggravate*, *amaze*, *amuse*, *annoy*, *astonish*, *bore*, *charm*, *deceive*, *disappoint*, *exasperate*, *fascinate*, *frighten*, *humiliate*, *infuriate*, *inspire*, *intimidate*, *irritate*, *offend*, *scare*, and *surprise*. Voice and prompt type varied within participants and within items.

The experiment consisted of sixty-eight items: twenty experimental items interleaved with twenty-four transfer-of-possession items for an unrelated experiment and twenty-four additional fillers. The stimuli for the interleaved experiment contained sentences with transfer-of-possession verbs followed either by a no-pronoun prompt or an ambiguous pronoun prompt (*John brought a glass of water to a guy. (He)...*). The additional fillers consisted of context sentences containing non-IC verbs followed by intersentential connectives, blank prompts, or pronoun prompts. Story continuations were collected using the same web-based interface that was described in Section 4.1.

6.2. Evaluation and Analysis

Two judges, the first author of this paper and a Northwestern University graduate student, coded the first-mentioned referent in each continuation (which, in the case of the pronoun-prompt condition, was the assignment of the pronoun). The judges were instructed to err on the side of categorizing a pronoun as ambiguous if the pronoun could be interpreted as plausibly coreferential with either referent, even if their own interpretation biases suggested a particular one.

Mixed-effect logistic regressions were again used to measure the effects of several within-participants/within-items predictors on the observed choice of referring expression and the observed choice of first mention. For the choice to use a pronoun, we considered only the continuations elicited in the no-pronoun condition and modeled the observed referring expressions with predictors for referent position and voice. We modeled the binary choice of first mention with predictors for voice and prompt type. All predictors were centered, and all models contained random intercepts and fully crossed random slopes. As in Experiment 1, we report the coefficient estimate and p-value for each predictor.

6.3. Results and Discussion

The results reflect a conservative analysis in which a continuation was excluded if at least one judge assessed it as ambiguous. Out of all 840 continuations, 9.8% were excluded due to ambiguity or because there was no mention of either referent or because of inconsistencies suggesting that the participant misread the prompt, leaving a dataset of 758 continuations.

We first examine our predictions regarding production biases. The means for rate of pronominalization in the no-pronoun condition are shown in Figure 4. The analysis of referring expressions produced in the no-pronoun condition replicates the previously reported bias to pronominalize references to the subject. There were also more pronouns in the passive condition, but this was driven by the predicted referent position × voice interaction whereby re-mentions of passive subjects were pronominalized at a higher rate than active subjects (86.5% vs. 62.1%), but re-mentions of active and passive non-subjects were pronominalized at the same rate (24.0% vs. 22.9%). In a logistic regression, referent position and voice were significant factors in modeling the binary outcome of pronominalization, as was the predicted interaction between referent position and voice (referent position: β=-2.53, p<0.001; voice: β=0.812, p<0.05; referent position × voice: β=-1.085, p<0.01).

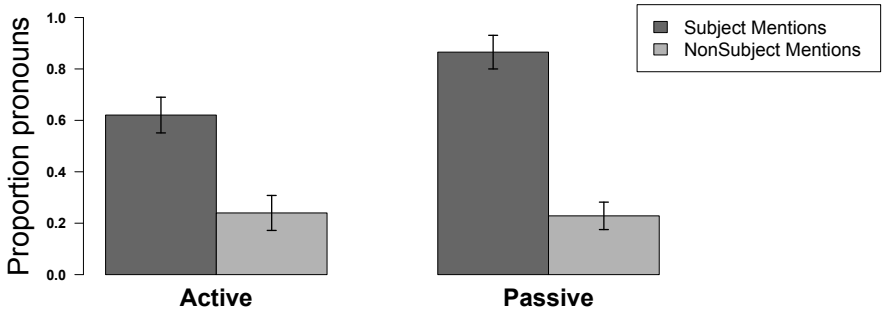

Figure 4: Rate of pronominalization, by voice and referent position (syntactic subject vs. non-subject)

Pairwise comparisons of the pronominalization rates between active and passive revealed an effect of voice for subject first mentions (β=1.918, p<0.01) and no effect of voice for non-subject first mentions (β=-0.154, p=0.76).[11]

We next consider our predictions regarding next-mention and pronoun interpretation biases. Figure 5 shows the rate at which participants wrote continuations about the syntactic subject across the four conditions. As expected, the results replicated the widely reported IC bias: Participants wrote more continuations about the causally-implicated referent (63.7%) than the non-implicated one. Since the causally-implicated entity for subject-biased IC verbs appears as the syntactic subject in the active-voice condition and the non-subject in the passive-voice condition (the logical

[11] One might note that the rate of pronominalization towards subjects in the active condition (62.1%) appears lower than for the same condition in Experiment 1 (77.5%), which used similar (although not identical) stimuli. Further analysis of the data revealed that seven of the participants in Experiment 2 never used a pronoun in any continuation; this wasn't the case for any participants in Experiment 1 nor is it typical to see in other experiments. The results after the data for these seven participants is removed are as follows:

	syntactic subject	*non-subject*
active	.761 ± .061	.283 ± .078
passive	.900 ± .058	.274 ± .061

The results are now highly consistent with the previous experiment (76.1% v. 77.5% for references to the subject in active subject-biased contexts, and 28.3% v. 26.6% for references to the object in such contexts). The results of statistical analysis with this exclusion match those reported in the main text.

subject, Amanda in (7)), the IC bias emerges as a main effect of voice: Participants wrote more continuations about the syntactic subject following an active-voice prompt (68.7%) than a passive-voice prompt (41.4%). In a logistic regression, voice was a significant factor in modeling the binary outcome of re-mention of the syntactic subject (β=-0.76, p<0.001).

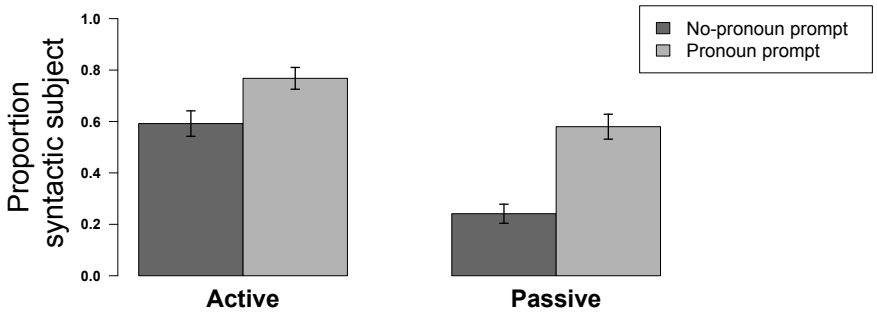

Figure 5: Proportion of continuations about the syntactic subject, by voice and prompt type

The results also confirmed our prediction that prompt type affects choice of first mention: Participants wrote more continuations about the syntactic subject following a pronoun prompt (67.2%) than a no-pronoun prompt (39.8%). Prompt type was a significant factor in modeling subject re-mention (β=-0.75, p<0.001). Furthermore, in keeping with the claim that passive voice is a stronger indicator of subject topicality than active voice, the effect of the pronoun prompt was marginally greater in the passive-voice condition than the active-voice condition (voice × prompt type interaction: β=-0.18, p=0.06).

In pairwise comparisons of the effect of prompt type across the two voice conditions, the rate of first mention of the syntactic subject was higher for the pronoun prompt than the no-pronoun prompt in the active-voice condition (76.8% vs. 59.2%: β=-1.136, p<0.001), and the effect of prompt type was even stronger in the passive-voice condition (58.0% vs. 24.1%: β=-1.885, p<0.001). Finally, a further pairwise comparison of the effect of voice in the pronoun-prompt condition confirms a stronger first-mention bias to the IC-congruent referent Amanda in the active-voice condition (where Amanda is the subject) than the passive-voice condition (where Amanda is the non-subject; 76.8% vs. 42.0%: β=-2.012, p<0.001).[12,13]

Interestingly, the re-mention rate of the logical subject was higher in the no-pronoun passive condition (75.8%) than the no-pronoun active condition (59.1%), which is an effect that is unantic-

[12] Again, the analysis reflects a conservative data inclusion strategy in which a continuation was excluded if at least one coder assessed it as ambiguous. As was the case in Experiment 1, the pattern of statistical significance for all results remains the same if continuations are included for which at least one coder assigned a non-ambiguous interpretation.

[13] Caramazza and Gupta (1979) similarly found an effect of passivization in IC contexts using a timed comprehension task. Kaiser et al. (2011) describes a complementary effect of passivization in Agent-Patient contexts.

ipated by our analysis and its antecedents.[14] We suspect that this finding is due to the optionality of the *by*-phrase in a passive construction. Indeed, Arnold (2001) had a parallel finding in her comparison of Source-Goal and Goal-Source transfer-of-possession contexts, in which Source referents were re-mentioned unexpectedly often in a no-pronoun condition for the Goal-Source case. Unlike Source-Goal sentences, in which both thematic roles are obligatory, the Source is optional in Goal-Source sentences (e.g., both *John seized the comic from Bill* and *John seized the comic* are acceptable). Arnold hypothesized that participants may have felt compelled to re-mention the Source in the continuation in order to justify its inclusion in the story. We take our similar finding for the optional logical subject in passives to support Arnold's hypothesis. Importantly, participants show no evidence of this preference in our pronoun-prompt condition, favoring the referent occupying the syntactic subject position instead.

Finally, we can again use the data collected to test our Bayesian Hypothesis, i.e., that the relationship between pronoun production and pronoun interpretation biases are as predicted by equation (3). As was done in Experiment 1, we use individual participants' data collected in the pronoun-prompt condition as our observed pronoun interpretation bias. The data from the no-pronoun prompt condition allows us to calculate estimates of the interpretation bias under the Expectancy, Mirror, and Bayesian models. We use the same scaling and exclusion strategies as in Experiment 1. In this case, the exclusion criteria eliminate the data from seven participants who used no pronouns in any of their continuations (see footnote 11). Figure 6 shows the observed and model-predicted rates at which pronouns will be interpreted to either the subject or non-subject across the two voice conditions, and Table 2 shows the linear correlation results. Although the predictions of the Mirror and Bayesian models are close, the Bayesian model again makes the most reliable predictions regarding the observed pronoun interpretation biases.[15]

	Obs∼Mirror	**Obs∼Expectancy**	**Obs∼Bayes**
by participants	R^2=0.32*	R^2=0.05*	R^2=0.35*
by items	R^2=0.49*	R^2=-0.01	R^2=0.52*

Table 2: Correlations between observed data and model predictions, by participants and by items. * indicates significance at or below 0.05.

To sum, the results confirm the predictions of the Topichood Hypothesis: Participants produced

[14] An anonymous reviewer remarks that the 59.1% figure in the active condition seems low considering the strong biases usually associated with IC verbs. However, biases previously reported in the literature are almost always collected using prompts containing 'because' which, by restricting the continuations to causal follow-ons, enhances the bias toward the causally-implicated referent (i.e., the subject for subject-biased IC verbs). Similarly, these prompts also commonly include a subject pronoun, which as we have seen will also raise the subject bias. Indeed, our result replicates the bias reported for a similar condition in Kehler et al. (2008; Experiment 3, page 33, Section 6.1.4).

[15] Footnote 9 reported on separate analyses for referents in different grammatical roles that established the lack of a correlation between the predictions of the Mirror model and the observed interpretation biases. Since our analysis predicts an effect of voice on rate of pronominalization for referents in subject position in this experiment, a similar analysis is not applicable here.

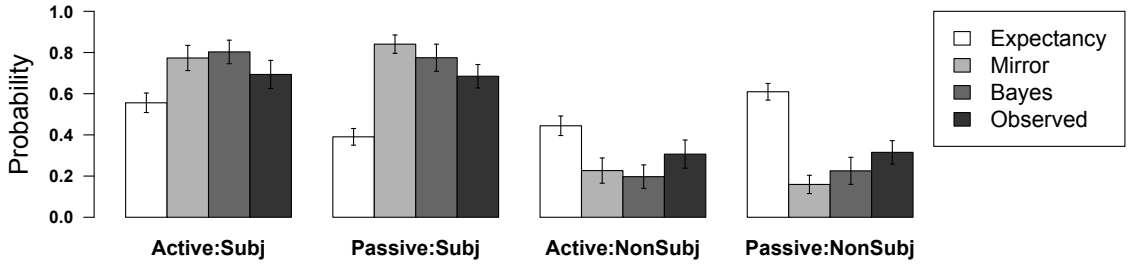

Figure 6: Observed and estimated pronoun interpretation biases across conditions and referents in Experiment 2 (participant means ± standard error)

more pronouns when referring back to the subject of a passive than the subject of an active. The results confirm our secondary predictions as well: Participants favored reference to the causally-implicated referent, they favored the syntactic subject to a greater degree when prompted with a pronoun, and their preference for re-mentioning the causally-implicated referent was lowest when prompted with a pronoun and when the causally-implicated referent was not the syntactic subject. Finally, the results again supported the predictions of the Bayesian Hypothesis relating pronoun production and interpretation.

7. General Discussion

A natural and commonly-held assumption in pronoun research has been that interpretation and production are driven by the same set of contextual factors. That is, we expect that speakers will employ pronouns in just those contextual circumstances in which the intended referent will be favored by the comprehender's own biases. A recent line of research has suggested, however, that the two are in fact dissociated. This research is instead consistent with a model in which production biases are determined by grammatical and/or information-structural factors, whereas interpretation processes integrate these with contextually-driven semantic biases concerning what entity will be mentioned next that hold independently of the form of reference chosen by the speaker. Whereas it may seem unintuitive that a producer, upon deciding whether to use a pronoun, would ignore a set of biases that will be utilized by the comprehender, this is precisely what is predicted by a Bayesian formulation of the relationship between pronoun interpretation and production.

Support for this picture has remained incomplete in three important respects, however. First, whereas the results of several studies (Rohde, 2008; Fukumura and van Gompel, 2010) have been argued to provide support for the hypothesis that rate of pronoun production is not influenced by semantic biases, these results came from experiments that employed gender-unambiguous contexts. This is an inadequate test of the hypothesis, unfortunately, since it could be the case that speakers only account for comprehenders' interpretation biases when an ambiguity is present. Experiment 1 examined production biases in gender-ambiguous contexts. Whereas a context manipulation between subject-biased IC verbs, object-biased IC verbs, and non-IC verbs yielded different interpretation biases, the context distinction did not affect rates of pronominalization. In line with the previous gender-unambiguous studies, only the grammatical role of the antecedent mattered.

The second open question was whether the subject bias in production is due to a grammatical

role driven preference or to a production rule that is sensitive to the topichood status of the referent. According to the Topichood Hypothesis, the evidence for a subjecthood bias is an epiphenomenon, resulting from the fact that subject position is the default place for the topic to reside in English. Experiment 2 examined a prediction of the hypothesis by employing a voice manipulation to vary the likelihood that the grammatical subject is the topic. A significant effect on rates of pronominalization provides support for the role of topichood beyond what can be accounted for by grammatical role, as a bias based purely on grammatical role should have yielded no difference. This finding thus fits in with a series of previous results that demonstrate that various linguistic factors purported to influence pronoun interpretation are likely epiphenomena of deeper semantic and pragmatic properties of the context (e.g. Kehler et al. (2008)).

Third, the results of both experiments were shown to support the Bayesian account of the relationship between pronoun production and interpretation put forth by Kehler et al. (2008). The predicted interpretation bias was calculated using equation (3) with estimates of production and next-mention biases collected in the no-pronoun prompt conditions. In both experiments, this prediction was highly correlated with the actual interpretation bias measured in the pronoun prompt conditions, and in fact more so than two other models represented in the literature. This is a significant finding – the first of its kind, to our knowledge – and one that is not predicted by any heuristic-driven approach to interpretation.

These results may also shed light on a set of intriguing findings reported by Arnold and Griffin (2007). Arnold and Griffin found that speakers are significantly less likely to use a pronoun when the context introduces two event participants than when it only introduces one, even when reference in the two-participant contexts is gender-unambiguous. They attribute this effect to the role of accessibility of referents in the speaker's mind, under the assumption that multiple characters in the discourse context decreases the amount of attention that the speaker can give to each, which in turn reduces the likelihood that she chooses a pronoun. The current account offers an explanation of a different sort, as the presence of multiple characters will decrease the likelihood that each is the topic, thereby reducing rate of pronominalization. Neither Arnold and Griffin's data nor ours is explained by any analysis in which the failure to pronominalize is motivated primarily by the need to avoid ambiguity.

A question that arises is how the different sorts of biases at play in our analysis relate to the notions of 'prominence', 'salience', and 'accessibility' of referents that are commonly invoked in research on pronoun interpretation. Arnold (2001), for example, argues specifically for a notion of referent accessibility that is tied directly to the comprehender's probabilistic expectation that the referent will be mentioned, i.e., what we have referred to as next-mention biases. Because this probability is indifferent to whether or not a pronoun is used, it avoids a common circularity in the literature whereby the factors posited as contributors to prominence are identified on the basis of the very pronoun interpretation patterns that the theories are attempting to explain. The results surveyed here, however, demonstrate that pronoun production is not determined by entity prominence on this definition, and likewise this probability is only one factor that determines pronoun interpretation. Fukumura and van Gompel (2010), on the other hand, argue that likelihood of reference does not influence accessibility, but instead only the structure of the previous sentence does, acting as a cue to activate particular discourse entities. It is not clear to us on this explanation, however, why pronoun interpretation is sensitive to likelihood of next mention when production is not, insofar as interpretation preferences are likewise based on accessibility. It seems to us that

notions such as prominence, salience, and accessibility may only be confusing the issue. At the end of the day, what we have is an analysis of pronoun interpretation that fits in well with what has come to be a modern view of comprehension in psycholinguistics, in which interpretation is not something that initiates when linguistic material in encountered, but is instead what happens when top-down expectations about the ensuing message come into contact with the linguistic evidence. In the case of pronouns, interpretation is the result of integrating top-down expectations about who will be mentioned next with the pronoun's linguistic function of indicating a continuation of the current topic.

This notwithstanding, numerous questions remain for the analysis offered here. First, while the results of Experiment 2 were strongly consistent with the Topichood Hypothesis, the manipulation did not completely rule out the existence of an independent role for grammatical position, as the grammatical position of the preferred topic was not varied. A study that fully crosses topichood and grammatical position would require a way of marking non-subjects as topics, which is not readily accomplished in English without introducing confounds (e.g., by fronting the non-subject with a topic-marking phrase). Other languages that suggest themselves introduce complexities as well. For instance, while Japanese has the purported topic-marker *-wa*, grammatical objects so marked are typically interpreted as contrastive topics, which introduces an additional layer of pragmatic complexity that renders them inappropriate for answering the question at hand. This issue must therefore be left for further work.

Second, as we noted in the introduction, studies of pronoun interpretation in psycholinguistics have focused almost exclusively on the singular, third person, personal pronoun, and in our building upon this work we have necessarily followed suit. One is nonetheless led to ask how the predictions of our analysis (and those of our predecessors) extend to other forms, which often bring additional complexities. Consider plural pronouns, for example. Whereas the predictions of the analysis are analogous to those for singular pronouns when the competing antecedents are all plural, plural pronouns can also refer to plural entities that are evoked from disjoint antecedents. For example, the passage *Mary gave John a ride to Bill's house. They* _____ allows for four ways of assembling the three people into groups that are each compatible with the pronoun *They*. The account proposed here requires that we be able to both evaluate the topichood status of referents that arise from grouping disjoint antecedents and estimate their next-mention biases; we are aware of no existing work that informs this question. Similarly, consider the inanimate pronoun *it*, which can not only refer to entities, but also events, entire situations, propositions, descriptions, speech acts, and so forth. What is the topichood status of referents in these ontological categories, and what are the relevant next-mention biases? Again, we are not aware of any work that helps us answer this question, or for that matter, even considers the ontological ambiguity of such anaphors in an experimental setting. As such, these questions must also be left for future work.

Finally, in keeping with the theme of the special issue, we can ask what lessons this research has to offer the computational side of the field. Whereas to our knowledge there is no work in the computational realm specifically modeling the relationship between reference production and interpretation of the sort represented by the current study, it seems fair to say that an underlying goal of natural language generation systems is to produce referring expressions that will be successfully interpreted without being unnecessarily explicit – a desideratum that implies that, in general, a pronoun should be used in just those cases in which one would expect it to be successfully interpreted. A rational strategy for such a system would be to approach the decision about

whether to pronominalize by appealing to the same set of factors that we know comprehenders will use to interpret the anaphor – an approach which, unfortunately, would require that any system for *generating* referential expressions include a system capable of *interpreting* them. As Arnold and Griffin (2007) point out with respect to the human production system, ambiguity avoidance of this sort would put a considerable filtering load on the generation system: Each referring expression under consideration by the system would need to be evaluated for interpretability with respect to competing referents in the current discourse context (a process which, to our knowledge, no existing generation system explicitly carries out). The good news is that research such as the work presented here suggests that this need not be done – at least for some phenomena, interpreters cope with ambiguity even when not accounted for in the model used by the speaker. Indeed, our work fits in with other literature that suggests that speakers do not actively seek to avoid producing expressions that give rise to temporary ambiguities (Ferreira and Dell, 2000; Arnold et al., 2004; Kraljic and Brennan, 2005, inter alia).

Acknowledgements

We thank Albert Gatt and three anonymous reviewers for useful comments and criticisms, and Roger Levy for helpful discussion. This research was supported by an Andrew W. Mellon postdoctoral fellowship to the first author. The results of Experiment 2 have been presented at the 22nd Annual CUNY Conference on Human Sentence Processing, the 2011 meeting of the German Linguistics Society, the 2011 Constraints in Discourse meeting, and the 86th Annual Meeting of the Linguistics Society of America. We thank our audiences at those meetings for useful discussions, and Meredith Larson for her help in annotation.

References

Agresti, A. 2002. *Categorical data analysis*. 2nd ed. Wiley.

Ariel, M. 1990. *Accessing noun phrase antecedents*. Routledge.

Arnold, J. E. 2001. The effects of thematic roles on pronoun use and frequency of reference. *Discourse Processes* 31: 137–162.

———. 2010. How speakers refer: the role of accessibility. *Language and Linguistic Compass* 4:187–203.

Arnold, J. E., and Z. M. Griffin. 2007. The effect of additional characters on choice of referring expression: Everyone counts. *Journal of Memory and Language* 56:521–536.

Arnold, J. E., T. Wasow, A. Asudeh, and P. Alrenga. 2004. Avoiding attachment ambiguities: The role of constituent ordering. *Journal of Memory and Language* 51:55–70.

Au, T. K. 1986. A verb is worth a thousand words: The causes and consequences of interpersonal events implicit in language. *Journal of Memory and Language* 25:104–122.

Barr, D. J., R. Levy, C. Scheepers, and H. J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3):255–278.

Brennan, S. E., M. W. Friedman, and C. J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th meeting of the association for computational linguistics*, 155–162. Stanford, CA.

Brown, R., and D. Fish. 1983. The psychological causality implicit in language. *Cognition* 14:237–273.

Caramazza, A., E. Grober, C. Garvey, and J. Yates. 1977. Comprehension of anaphoric pronouns. *Journal of Verbal Learning and Verbal Behaviour* 16:601–609.

Caramazza, A., and S. Gupta. 1979. The roles of topicalization, parallel function and verb semantics in the interpretation of pronouns. *Linguistics* 3:497–518.

Chambers, G. C., and R. Smyth. 1998. Structural parallelism and discourse coherence: A test of centering theory. *Journal of Memory and Language* 39:593–608.

Davison, A. 1984. Syntactic markedness and the definition of sentence topic. *Language* 60(4):797–846.

Ferreira, V. S., and G. S. Dell. 2000. The effect of ambiguity and lexical availability on syntactic and lexical production. *Cognitive Psychology* 40:296–340.

Fukumura, K., and P. G. van Gompel. 2010. Choosing anaphoric expressions: Do people take into account likelihood of reference? *Journal of Memory and Language* 62:52–66.

Garvey, C., and A. Caramazza. 1974. Implicit causality in verbs. *Linguistic Inquiry* 5:459–464.

Gernsbacher, M. A., and D. J. Hargreaves. 1988. Accessing sentence participants: The advantage of first mention. *Journal of Memory and Language* 27:699–717.

Gernsbacher, M. A., D. J. Hargreaves, and M. Beeman. 1989. Building and accessing clausal representations: the advantage of the first mention versus the advantage of clause recency. *Journal of Memory and Language* 28: 735–755.

Givón, T. 1983. Topic continuity in discourse: An introduction. In *Topic continuity in discourse: A quantitative, cross-language study*, ed. T. Givón, 1–42. Amsterdam: John Benjamins.

———. 1990. *Syntax: a functional typological introduction.* Amsterdam: John Benjamins.

Grosz, B. J., A. K. Joshi, and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21:203–225.

Gundel, J. 1974. The role of topic and comment in linguistic theory. Ph.D. thesis, University of Texas at Austin. Reprinted in Outstanding Dissertations in Linguistics Series, Garland Publishers, 1989.

Gundel, J. K., and T. Fretheim. 2004. Topic and focus. In *The handbook of pragmatics*, ed. L. R. Horn and G. Ward, 175–196. Oxford: Basil Blackwell.

Gundel, J. K., N. Hedberg, and R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse. *Language* 69(2):274–307.

Hobbs, J. R. 1979. Coherence and coreference. *Cognitive Science* 3:67–90.

Jaeger, T. F. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* (Special issue on Emerging Data Analysis) 59:434–446.

Järvikivi, J., R. van Gompel, J. Hyönä, and R. Bertram. 2005. Ambiguous pronoun resolution: contrasting the first mention and subject preference accounts. *Psychological Science* 16:260–264.

Kaiser, E., D. Cheng-Huan Li, and E. Holsinger. 2011. Exploring the lexical and acoustic consequences of referential predictability. In *Anaphora processing and applications*, ed. Iris Hendrickx, Antonio Branco, Sobha Lalitha Devi, and Ruslan Mitkov, 171–183. Lecture Notes in Artificial Intelligence, Vol. 7099. Heidelberg: Springer.

Kehler, A., L. Kertz, H. Rohde, and J. Elman. 2008. Coherence and coreference revisited. *Journal of Semantics* 25: 1–44.

Kraljic, T., and S. E. Brennan. 2005. Prosodic disambiguation of syntactic structure: For the speaker or for the addressee? *Cognitive Psychology* 50:194–231.

Kuno, S. 1972. Functional sentence perspective. *Linguistic Inquiry* 3(3):269–320.

Lambrecht, K. 1994. *Information structure and sentence form: Topic, focus, and the mental representations of discourse referents.* Cambridge: Cambridge University Press.

McKoon, G., S. Greene, and R. Ratcliff. 1993. Discourse models, pronoun resolution, and the implicit causality of verbs. *Journal of Experimental Psychology* 19:1040–1052.

Miltsakaki, E. 2007. A rethink of the relationship between salience and anaphora resolution. In *Proceedings of the 6th discourse anaphora and anaphor resolution colloquium*, ed. A. Branco, 91–96. Lago, Portugal.

Reinhart, T. 1981. Pragmatics and linguistics: An analysis of sentence topics. *Philosophica* 27(1):53–94.

Rohde, H. 2008. Coherence-driven effects in sentence and discourse processing. Ph.D. thesis, UC San Diego.

Rohde, H., A. Kehler, and J. Elman. 2006. Event structure and discourse coherence biases in pronoun interpretation. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society.* Vancouver, July 26-29, 2006.

Shibatani, M. 1985. Passives and related constructions: A prototype analysis. *Language* 61:821–848.

Smyth, R. J. 1994. Grammatical determinants of ambiguous pronoun resolution. *Journal of Psycholinguistic Research* 23:197–229.

Stevenson, R., R. Crawley, and D. Kleinman. 1994. Thematic roles, focusing and the representation of events. *Language and Cognitive Processes* 9:519–548.

Strawson, P. F. 1964. Identifying reference and truth values. *Theoria* 30:96–118.

Walker, M. A., M. Iida, and S. Cote. 1994. Japanese discourse and the process of centering. *Computational Linguistics* 20(2):193–232.

Ward, G., and B. Birner. 2004. Information structure and non-canonical syntax. In *The handbook of pragmatics*, ed. L. R. Horn and G. Ward, 153–174. Oxford: Basil Blackwell.