RUNNING HEAD: Thematic roles affect reference form

# Predictability affects production: Thematic roles affect reference form selection

Elise C. Rosa and Jennifer E. Arnold

University of North Carolina at Chapel Hill

Key words: pronouns, reference form, utterance production, predictability, thematic roles

Address all correspondence to: Jennifer E. Arnold, UNC Chapel Hill, Dept. of Psychology, Davie Hall #337B, CB #3270, Chapel Hill, NC 27599-3270, Email: jarnold@email.unc.edu

ABSTRACT

Speakers use pronouns and zeros when referring to information that is topical, recently mentioned, or salient in the discourse. Although such information is often predictable, there is conflicting evidence about whether predictability affects reference form production. This debate centers on the question of whether reference form is influenced by the predictability of certain thematic roles. While some (Arnold, 2001) argue that referents in certain thematic roles are more likely to be pronominalized, others (Fukumura & van Gompel 2010; Rohde & Kehler, 2014) argue predictability does not play a role in determining referential form. We tested this puzzle in three experiments, using both a richly contextualized production paradigm, and two versions of the standard story-completion paradigm. In all experiments we manipulated the predictability of pairs of characters using Goal-Source verbs. In all three experiments, we found that speakers used more reduced referring expressions when talking about the Goal referent as opposed to the Source. These results show that thematic role does affect both perceived predictability and the speaker's choice of reference form.

The selection of an appropriate referring expression is an important component of successful communication. For example, in relating a story about a villain, you need to make multiple decisions about how to refer to him. You would likely use a descriptive expression such as *Bob* or *this creepy guy* upon first mention, and when referring to him again, might choose a more reduced expression such as "*he*", or even an elliptical or "zero" form, such as *and then Ø drew a knife.*

It is well established that speakers use reduced forms (pronouns and zeros) under particular discourse conditions, such as when the referent has been recently mentioned, or was in the grammatical subject position of the last sentence (Ariel, 1990; Arnold, 1998, 2008, 2010; Brennan, 1995; Givon, 1983; Gundel et al., 1993). One hypothesis is that recently and prominently mentioned things tend to be topical to the current discourse segment, and that reduced forms are selected on the basis of the topicality of the referent (e.g., Givon, 1983; Kehler et al., 2008; Kehler & Rohde, 2013; van Rij et al., 2013).

Yet an unresolved debate concerns the role of predictability in reference form. Given a particular discourse context, comprehenders have consistent expectations that some characters are more likely to be mentioned again, meaning that they are relatively more predictable as referents. For example, the sentences in (1) depict events in which people tend to assume that one participant is the more likely cause of the event (e.g., Brown & Fish, 1983; Hartshorne, O'Donnell, & Tenenbaum, 2015). If a causal statement includes a pronoun ("because he…"), participants tend to interpret the pronoun as coreferential with the implicit cause (Stevenson et al., 1994). Similarly, the sentences in (2) depict transfer-of-possession events, in which readers tend to expect that a subsequent event will mention the receiver of the object (Rohde & Kehler, 2014; Stevenson et al., 1994).

1a). The butler blamed the chauffeur because he…. (murdered someone).

1b). The butler impressed the chauffeur because he… (figured out the case).

2a). The butler gave the threatening note to the chauffeur and he… (turned it in to the police).

2b). The butler received a ticking bomb from the chauffeur and he … (chucked it into the river).

In discourses like these, the predictability of referents is identified with their thematic role in the event. The thematic role is determined by the verb, and represents the semantic role of the participants in an event. In 1, the stimulus role is the expected continuation (the chauffeur in 1a, the butler in 1b), while the experiencer is not. In 2, the goal is the expected continuation (the chauffeur in 1a, the butler in 1b), while the source is not.

Critically, the effects of thematic roles on referential predictability are closely tied to the relationship between the two utterances (Ehrlich, 1980; Kehler, 2002; Kehler & Rohde, 2013; Stevenson et al., 2004). In the implicit causality sentences, people only expect the causal character to be mentioned if they expect the speaker to produce an utterance about the cause of the first event. This expectation is created by the connector *because* in (1), but if the sentence continued *so he*..., expectations would reverse (Ehrlich, 1980; Stevenson et al., 1994). In (2), the expectation of the goal reference is conditioned on the expectation that the speaker will describe the result of the first event (Stevenson et al., 1994).

The question we are concerned with here is what speakers do in production. Following one of the sentences above, would a speaker choose the pronoun *he,* or use the more explicit

descriptive noun? Critically, this question is debated, and there is conflicting evidence in the literature.

One view is that thematic roles do affect reference form production. Arnold (1998, 2001) proposed that referential predictability contributes to discourse accessibility (see also Givon, 1983), and thereby increases the speaker's likelihood of using pronouns and zeros. Arnold's Expectancy Hypothesis suggests that predictability comes from numerous sources, including the fact that grammatical subjects are more likely to be mentioned again than non-subjects, and that recently mentioned entities are more likely to be mentioned than less recent entities (Arnold, 1998; 2010). In addition, Arnold (2001) presented results from a story-continuation experiment, in which participants were asked to invent continuations for passages that included a critical transfer-of-possession prompt, e.g. *Lisa gave the leftover pie to Brendan.* Results revealed that when participants referred to the second character, they used pronouns more often for goals than sources. A corpus analysis confirmed that goals are more likely to be mentioned again than sources.

Kaiser, Li, and Holsinger (2011) also report that thematic roles influence pronoun usage, using prompts with agent and patient roles, such as *Mary slapped Lisa…*, and *Lisa was slapped by Mary*. However, they also argue that thematic roles are not linked to predictability, based on the observation that the strong patient preference for pronoun use in their data was not mirrored by an equally strong preference to continue talking about the patient.

By contrast, several studies have reported the opposite, that thematic roles do not influence the speaker's choice of referential form (Kehler et al., 2008; Fukumura & van Gompel, 2010; Rohde & Kehler, 2014). All of these studies also used a story-continuation methodology. For example, Fukumura and van Gompel (2010) examined implicit causality verbs, with prompts

such as *Gary scared Anna…,* or *Gary feared Anna*…. Across the board, these studies found that verb biases influenced the likelihood that subjects would mention one character or another. That is, they affect the referential predictability of the character. However, in none of these experiments did the verb bias influence pronoun use. Instead, participants followed the first-mentioned/subject bias, using pronouns when they mentioned the first character (e.g. *Gary*), and names when they mentioned the second (e.g. *Anna*). This has led to the claim that topicality is the sole determinant of pronoun selection (Kehler & Rohde, 2014; Fukumura & van Gompel, 2010).

However, there are several reasons to reconsider the question of whether thematic roles affect referential form, and how both are related to predictability. First, predictability has widespread effects on other aspects of language production, such as acoustic form (e.g., Lieberman, 1963). Second, there are numerous differences between the Arnold (2001) goal/source study and the studies that found no effects on referential form. Third, all of the previous studies used a story-continuation methodology, which has different task demands than natural language production. This paper therefore examines this question systematically, using both picture-description and story-continuation tasks.

**Why might predictability matter?**

Predictability plays a central role in current theories of both language production and comprehension. The probability of words and phrases plays a central role in probabilistic models of language processing (e.g., Jurafsky, 1996; Seidenberg & MacDonald, 1999; Hale, 2001), and there is extensive evidence that predictability affects language comprehension (e.g., Altmann & Kamide, 1999, van Berkum et al., 2005; Staub & Clifton, 2006; Levy, 2008).

Even more relevant to our question, it is well established that predictability affects language production, by modulating word and vowel duration and consonant cluster reduction. Over time this results in shortening of words that are frequent (Zipf, 1936) or tend to be produced in probable contexts Piantadosi (2011). The token-by-token pronunciation of words also varies by context. Both function and content words tend to be produced in a more reduced manner when they are predictable in context, as measured by vowel duration (Bell et al., 2003), word duration and final consonant deletion (Gregory et al., 1999; Jurafsky et al., 2000).

The relationship between the probability of a word and how it is pronounced has been formalized in the Probabilistic Reduction Hypothesis (Bell et al., 1999; Gregory et al., 1999; Jurafsky et al., 1998), which states that the probability of a word in context predicts its degree of reduction. Information-theoretic models more broadly suggest that linguistic form is related to the information expressed by the linguistic item (Aylett & Turk, 2004; Levy & Jaeger, 2007; Jaeger, 2006): According to this view, the forms of low-information words are reduced, and the forms of high-information words are lengthened, such that the overall stream of information content is uniform. This accounts for multiple levels of language: overall speech rate (Aylett & Turk, 2004), phonemic production (Son & van Santen, 2005); the use of optional function words (Jaeger, 2006; Levy & Jaeger, 2007); and contraction production (*you are* vs. *you're*, (Frank & Jaeger, 2008)). When extended to referential expressions, another type of language form variation, this hypothesis suggests that less predictable, high-information concepts will be produced with longer forms than highly predictable, low-information concepts. Indeed, word duration has also been shown to be sensitive to thematic role (Kaiser, Li, & Holsinger, 2011).

When we consider the selection of pronouns vs. other forms, the relevant level of predictability is the referent itself – that is, how likely is it that a particular entity will be

mentioned? This contrasts with most work on word pronunciation, which has mostly focused on the probability of words (but see Gahl & Garnsey, 2004). Thus, the question is whether the predictability of **reference to a particular entity** affects the likelihood of using a pronoun to refer to that entity.

There are many factors that might influence the predictability of a referent featuring in the upcoming discourse. For example, salient events such as a person falling down may lead to the expectation that the event will be mentioned. Here we examine predictability as it relates to the thematic roles, which influence predictions about next-mention likelihood.

**Methodological issues**

Despite the support for predictability effects in language, there is mixed evidence in the literature for thematic role effects on reference production. We consider several methodological issues that may explain differences amongst existing studies.

***Verb type.*** Most prior studies have investigated the effect of thematic role on pronoun use by examining implicit causality verbs (Fukumura & van Gompel, 2010; Rohde & Kehler, 2014; Kehler et al. 2008 Exp. 3; Kehler & Rohde, 2013). By contrast, the one study that did find an effect of thematic role used transfer verbs (Arnold, 2001). These verbs differ in many ways. Many of the implicit causality studies used experiencer/stimulus verbs, which denote psychological states (such as *admired* and *blamed*), and are less imageable than discrete actions (*handed, gave*). The stimulus role is considered the implicit cause of these events, but the experiencer may be accessible due to the focus on that person's mental state.

***Controlling for grammatical role effects.*** It is well established that speakers tend to use pronouns when referring to the grammatical subject. Transfer verbs provide a good test case for predictability effects, because they allow for the effects of thematic roles to be separated from

grammatical role. Some transfer verbs place the goal in subject position (get, receive, take) while other transfer verbs place the goal in the non-subject position (give, hand, send). For example, in "*Bob handed the threatening note to Sue*", the grammatical subject is the source, and the non-subject is the goal. In a sentence like "*Larry got the romantic note from Ellen"*, the subject is the goal and the non-subject is the source.

Given the known subjecthood bias, we would expect pronouns to be used for both Bob and Larry. If the predictability of goals also affects pronoun use, we would expect relatively more pronouns for Larry than Bob, and for Sue than Jamie, on top of the subject effect. This pattern was observed by Arnold (2001), except that the goal/source difference only emerged for the non-Subject references, in which speakers used around 18% pronouns for non-subject goals, but only 7% pronouns for non-subject sources. By contrast, pronoun use was at ceiling for reference to the subject. This finding is consistent with both a strong role for grammatical role, and a contributing effect of thematic roles.

By contrast, Rohde (2008, Exp. VII) compared goal and source thematic roles across different grammatical roles. She also examined pronoun use following transfer verb prompts such as *John handed a book to Mary*…. However, in her stimuli the goal was always the non-subject, and the source was always the subject. Participants were more likely to provide goal continuations than source continuations, but they used more pronouns for the subject than object. Rohde (see also Kehler et al.,2008) focuses on the discrepancy between the continuation bias and the rate of pronoun use. That is, if speakers tend to talk about Mary more than John, but use pronouns for John more than Mary, it suggests that pronoun use is influenced by more than pure predictability estimations. However, this design does not allow for a more fine-grained test of

thematic role effects, which may result in a relative difference between goal and source references, holding grammatical role constant.

***Task demands in the story continuation paradigm.*** Without exception, the effect of thematic roles on pronoun use has been examined with a story continuation paradigm in every study published (to our knowledge). In standard story continuation studies, participants read a probe sentence (or sentences), and generate a continuation to the story (Arnold, 2001; Fukumura & van Gompel, 2010; Kaiser et al., 2011; Kehler et al., 2008; Rohde & Kehler, 2014; Stevenson et al., 2004). The advantage to this paradigm is that it allows authors to tightly control the linguistic context, using items that are usually unrelated to one another, which is intended to ensure independence across items. It also allows researchers to simultaneously test two questions: 1) which character is more likely to be mentioned in the continuation? (as a measure of predictability), and 2) what referential form is used?

Yet there is also reason to believe that the story continuation paradigm has task demands that differ from normal language production, and these demands may interfere with the researcher's ability to detect thematic role effects on reference form choice production (see Arnold, 2013, for a critique of this paradigm). Continuation studies require participants to do three things: 1) understand the probe sentence, 2) invent a response, and 3) formulate the utterance. Both the process of interpretation and the invention of a new event are mentally taxing, perhaps leaving relatively few resources to plan their upcoming phrases.

Critically, this task delays the generation of the message, which may be necessary for coherence-based effects on production. Language production models generally agree that speakers first retrieve non-linguistic representations of concepts, then words, which are fit into sentence frames (Dell 1986; Levelt 1989; Roelofs, 1992). In natural speech situations, people

generally have some idea of the upcoming concepts they'd like to mention. Thus, even though utterance planning is often at least partially incremental (Levelt, 1989; Ferreira & Swets, 2002), the activation of the intended message may already be in place. Pre-activation of the message would allow the speaker to know the coherence relation between adjacent utterances, and support integration of the two utterances.

In continuation tasks, however, the message for the second sentence can only be generated after understanding the first sentence. This means that the coherence relation between the two utterances may not be activated until fairly late in the response process. Nevertheless, participants may begin to formulate the utterance in parallel with message generation. If so, their choices about reference form will necessarily be driven by information that is available already in the context (e.g., subjecthood), and predictability effects will not be apparent until the second utterance and coherence relation have been planned. If the story continuation paradigm delays the activation of coherence relations, it may not be well suited to finding thematic role effects.

If conceptual integration between utterances is important for predictability effects, we may also expect to see stronger effects of thematic roles within a richer discourse context. All the studies that found no effect of thematic role on reference form have used single-sentence contexts. By contrast, Arnold (2001) used a three-sentence context that ended with the critical transfer verb probe, e.g.: *There was so much food for Thanksgiving, we didn't even eat half of it. Everyone got to take some food home. Lisa gave the leftover pie to Brendan.* It may be that having a stronger discourse context provides the conceptual support for generating predictions about who will be mentioned next earlier in the response process.

***Current study goals and methodological approach***

This paper reports the results of a systematic investigation of whether thematic roles influence reference production. We examined transfer verbs, which have been shown to influence the use of reduced expressions (Arnold, 2001). Our primary goal was to test whether this effect was real, given widespread claims that thematic roles do not affect pronoun use (Kehler et al., 2008; Kehler & Rohde, 2013; Fukumura & van Gompel, 2010). Our secondary goal was to examine the relationship between referent predictctability and thematic roles, and to consider possible mechanisms by which predictability might affect reference production.

Given the methodological concerns about the sentence continuation paradigm described above, we introduced a novel picture-description paradigm, and compared it with sentence-continuation studies, while varying task demands associated with utterance production.

Figure 1. Characters in the event-retelling paradigm (from left-to-right: The butler and the maid, Sir Barnes, the chef, Lady Mannerly, the chauffeur).

This study introduces an innovative event-retelling paradigm that we designed to address some of the shortcomings of the story-continuation paradigm described above, such as the impoverished discourse context. Participants were asked to describe a series of picture pairs, which together told the story of Clue-like murder mystery. The participant was given the role of tabloid photographer, and asked to describe "their pictures" to a detective who was trying to solve a murder mystery. This task was designed to be engaging and interesting for task participants, and encourage them to develop a richer discourse representation. The story featured three main male characters (Sir Barnes, the chauffeur, and the butler), and three female (Lady Mannerly, the chef, and the maid; see Figure 1). The characters' behaviors and actions were consistent with their real-world roles.

The storyline was divided into pairs of sentences, which described actions that took place involving two of the characters (in the critical stimuli items) or one to three characters (in the filler items). This paradigm had several advantages. First, it utilized a typical trial-by-trial experimental structure, while still retaining the coherence of a naturalistic storytelling situation. In order to increase the participant's ability to conceptualize the story as a whole, they previewed all pictures before beginning the picture-description task. Second, it allowed us to manipulate the linguistic context for the subject's utterances. In each trial, two pictures were presented. The detective (an experimenter) described picture 1, allowing us to control the linguistic form of the context sentence. The subject described picture 2. Third, we were able to control the content of the participant's responses through the pictures, such that the continuation mentioned either the goal or source characters. Fourth, our paradigm allowed us measure the latency between the picture onset and the participant's response, which provided a measure of response difficulty.

A potential limitation to our story-telling paradigm is that the items are not entirely independent of one another, since together they tell a story. However, the benefits from this property were judged to outweigh the non-independence of items. Recent work has demonstrated that even with unrelated stimuli, subjects pick up on experiment-wide patterns and often change their behavior over the course of the experiment (Fine et al., 2013). We took care to consider this effect by including trial order as a fixed effect in the models.

Our study focuses on the speaker's choice between reduced forms (both pronouns and zeros), compared with more explicit names or titles. We group pronouns and zeros, given that they both are used for highly salient discourse entities. However, the production of zeros in our data was extremely low, so the effects reported here are due to variation in pronoun production.

Our experiment design also took care to avoid ceiling or floor effects in the data. The use of reduced forms is influenced by much more than just thematic and grammatical roles, and can be heavily influenced by details of the task and instructions. Moreover, participants may individually adopt modes of speaking in which they use either only pronouns or only names. If so, it compromises our ability to detect effects of the linguistic context, in that there is no variation in responses. To avoid these problems, we excluded participants who performed at floor or ceiling (for a similar convention, see Filmer, Mattingly, Dux, 2015; Buschkuehl et al., 2014), by only including data from participants who at least two pronouns or zeros and at least two names in the critical items. Former studies (Fukumura & van Gompel, 2010; Kehler & Rohde, 2013) did not use this exclusion criterion, leaving open the possibility that some of the participants included were not using any referential variation. If the semantic predictability effect is fairly small, including such participants might might mask the thematic role effect.

We also sought to avoid ceiling or floor effects by manipulating gender ambiguity in our utterances. It is well known that speakers use pronouns more often when there is only a single referent in the context that matches the pronoun's gender (e.g., *Bill called Sue and she...*) than when the pronoun would be ambiguous (*Amy called Sue and she…*; Arnold & Griffin, 2007; Fukumura et al., 2010). Depending on task-specific biases, this may lead to either ubiquitous use of pronouns for different-gender contexts, or ubiquitous use of names for same-gender contexts. Since we did not know ahead of time how participants would respond to our task, we included half same-gender and half different-gender contexts.

Experiment 1 describes our in-person event-retelling experiment. In Experiments 2 and 3, we compare the results on experiment 1 to two story-continuation experiments, to assess the role of task demands on thematic role effects in reference production. In order to examine the relationship between thematic roles and predictability in our materials.

***General study design***

The stimuli for this study were designed with verbs that describe transfer events, called Goal-Source verbs. Examples 3 and 4 depict two sample items. The underlined character is the one who is pictured in the second image in Experiment 1, and thus the character who is continued. Half the critical items (N=12) continued with the Goal, and half (N=12) continued with the Source. Within each condition, half the continuations occurred in a context where the referent was in subject position, and half in non-subject position.

(3) Goal continuation

a. Subject position: <u>*Sir Barnes*</u> *got a backrub from Lady Mannerly.*

b. Non-subject position: *Lady Mannerly gave a backrub to* <u>*Sir Barnes*</u>*.*

(4) Source continuation

a. Subject position: *The chef handed a cookbook to the maid.*

b. Non-subject position: *The maid took a cookbook from the chef.*

In sum, there were four conditions in the experiment: (1) Goal, Subject reference; (2) Goal, non-Subject reference; (3) Source, Subject reference; and (4) Source, non-Subject reference.


*Figure 2.* Sample trial from Experiment 1

In the event-retelling paradigm (Experiment 1) participants heard the prompt sentences and saw the pictures, as in Figure 2. In the accompanying rating studies they read the sentences

and saw the pictures. In the sentence completion paradigms (Experiments 2 and 3) participants read the prompt sentences, without pictures.

The stimuli were arranged following a Latin Square design, where each participant was exposed to each item in only one condition, but saw all conditions across different items. This design also encouraged variation in the pattern of reference across the items, which helped discourage participants from falling into a repetitive pattern of responses. The filler items helped develop the storyline and added variety to the kinds of sentence structures encountered.

***General analytic approach***

The same analytic approach was used for all the experiments. Any adjustments to this approach will be discussed in detail in the analysis section of each experiment's description. Generalized linear mixed-effects models were used to account for the dependencies in the repeated measures. We used a logistic regression (SAS proc glimmix) for analyses of dichotomous outcomes, and a mixed-effects linear regression (SAS proc mixed) for analyses of continuous outcomes.

We first built a control model, including a random participant intercept and control predictors in four categories: a) list, as a control of experiment design; b) overt connectives, as another measures of discourse connectivity, c) other trial-by-trial predictability measures (target mention likelihood and relatedness, as measured by the rating studies), and d) measures of production difficulty (see table 1 for details). Any control variables that had a t-value of greater than 1.5 were retained for the critical models. The inclusion of control variables achieved two goals. Most importantly, it allowed us to test the critical predictors within a well-characterized model of reference form, in which other relevant predictors were controlled. The secondary

purpose of the control predictors was to guide speculation about the mechanisms underlying reference form, which is postponed until the general discussion.

Table 1. Control variable descriptions

| | **Control variable** | **Description** |
|---|---|---|
| DESIGN VARIABLE | List | Which list participants was run on (A or B) |
| DISCOURSE CONNECTIVITY | Connective | Identified whether participant used a word such as (and, then, next) to begin |
| TRIAL-BY-TRIAL PREDICTABILITY MEASURES | Likelihood of mention (Exp. 1 only) | Measure of how likely the designated referent was judged to be[1] |
| | Relatedness z-score (Exp. 1 only) | Measure of how related the two events were judged to be[2] |
| MEASURES OF CONCEPTUAL COMPLEXITY AND/OR FORMULATION DIFFICULTY | Mention other person | Accounted for whether another character was mentioned in the same clause[3] |
| | Referent on right (Exp. 1 only) | Indicated whether continuation referent was on the right in the first picture. Participants typically scan a picture left-to-right, so the left picture may be processed sooner and available earlier. |
| | Word count (Exp. 1 only) | Measure of how many words the participant used in the utterance |
| | Verb codability (Exp. 1 only) | Measure of consistency in verb chosen for a particular picture; hypothesized to reflect ease of verb retrieval[4] |
| | Disfluency (Exp. 1 only) | Coded as 1 if participants were disfluent at the beginning of the utterance |

In order to avoid over-fitting our model, we first built a model that included only the critical predictors (subject/nonsubject, goal/source, gender, and order), plus any control variables that were significant at $t > 1.5$. Order was included as a fixed effect. Order was the same as item

[1] Calculated from the first rating study
[2] Calculated from the second rating study
[3] What constituted a clause was determined by the clause coding schema in Appendix B
[4] For example, if 18/20 participants described an action as shooting and 2/20 describe the action as loading a gun, the responses that used "shooting" got a score of 18/20 for the codability of item, and the responses that used "loading" got a score of 2/20 for that item.

in this experiment, as the items were presented in the same order to all participants, to preserve the story nature. Goal/Source, gender, and Subject/Non-Subject were centered by coding them as 0/1 and grand-mean centering.

The main effects of semantic and grammatical role were the focus of the analyses. In order to check whether these effects were qualified by interactions, we then added the interactions Subject*Goal, Goal*Gender, Subject*Gender, and Subject*Goal*Gender in a second model, which only retained those control variables that were significant. Here we report just the interactions model; any control variables not listed were not significant predictors in either the control or main effects models.[5] Any other interaction terms in addition to the four described above, specific to a particular model, will be explained in the analysis section of that experiment.

All models included random intercepts for participants. Random slopes for participants by subject/non-subject and goal/source were included when possible (i.e. when the model converged and the effects were not estimated to be zero). Since item/order was included as a fixed effect in all critical models it was not utilized as a random intercept in any model.

## Experiment 1: In-person study

### Method

#### *Participants*

32 undergraduates completed the task for class credit. 10 participants were excluded for using fewer than two pronouns or zeros and two were excluded for being non-native English speakers. This left 20 participants in the analysis.

[5] Note that in the models where there were no significant interactions, the pattern of significance for the main effects was the same in both the model with interactions and the main-effects-only models, so for simplicity we only present the model with interactions.

*Materials and Design*

Participants viewed pairs of pictures that were depictions of the sentence pairs described above. Participants heard a description of the first picture in each pair, produced by a lab confederate, and then provided their own description of the second. The stimuli were divided across two lists, such that all participants saw the same pictures, but heard one of two versions of the critical prompt sentences. Experiment 1 stimulus pictures can be found on the supplementary materials website[6].

Preliminary rating studies

Two rating studies were conducted using the picture stimuli for Experiment 1. Twenty participants completed each of the rating studies.

*Next-Mention Biases.* A rating study determined participants' next-mention biases. Participants viewed the first sentence and picture of each of the 53 stimuli and filler pairs, and selected which character they thought would be more likely to be talked about next. The probability of choosing the goal was 71%, supporting the predictability of goals. The probability of choosing the Subject was 54%, suggesting that subjects were not considered more predictable than non-subjects. A logistic regression with a random intercept for subjects confirmed these findings, revealing a main effect of thematic role $t(19)=3.71$, $p=.0015$, and no effect of subjecthood ($p > 0.1$).

This data was used to calculate the variable Target Likelihood, which refers to the average likelihood that the target character (i.e., the Goal or Source character that was featured in the continuation) was judged to be more likely to be talked about next. This predictor was used in analyses of the main experiments below.

[6] jaapstimuli.web.unc.edu

*Relatedness and Predictability.* Another rating study examined how related and predictable the participants thought the events in the pairs of pictures were. A different set of participants viewed the 53 sentences and pairs of pictures. Using a 7-point scale, they rated the pairs for: (1) how related the second event was to the first and (2) how predictable it was based on the first.

Goal continuations were judged to be more related (mean rating = 5.47) than the Source continuations (mean rating = 4.81), as confirmed by a main effect of thematic role $F(1,475)=20.22$, $p<.0001$. There was no main effect or interaction with subjecthood on relatedness (all $p$'s > 0.5). This data was used to calculate a Relatedness score for each item by calculating a z-score within each participant for each item, then creating an average z-score across participants for each item.

Goal continuations were also judged to be more predictable (mean rating = 3.825) compared to Source continuations (mean rating = 3.57). This effect nearly reached significance, $F(1,475)=3.70$, $p=.055$. There was no main effect or interaction with subjecthood on event predictability (all $p$'s > 0.5). Note that this metric reflects the predictability of the second event overall, and not specifically the predictability of reference to the target character, which was measured in the first rating study. Since target predictability was a more robust measure, this was used as a predictor in subsequent analyses.

*Procedure*

Participants were brought into the lab and seated at a computer. Participants were consented and completed an optional participant questionnaire, and then were shown a narrated background slideshow. The slideshow told them that they were a tabloid photographer, and described the family they had been secretly taking photographs of. It then told them that a

murder occurred while they were at the house, and they were going to review the photographs they had taken to help a detective solve the crime. The participants were introduced to the characters in the pictures, and then were shown all their pictures, in order. They then completed a sample item with the experimenter. The experimenter explained that the detective, who would arrive shortly, would describe the first picture in the pair. After that, the participant should say what happened next, using the second picture as a guide.

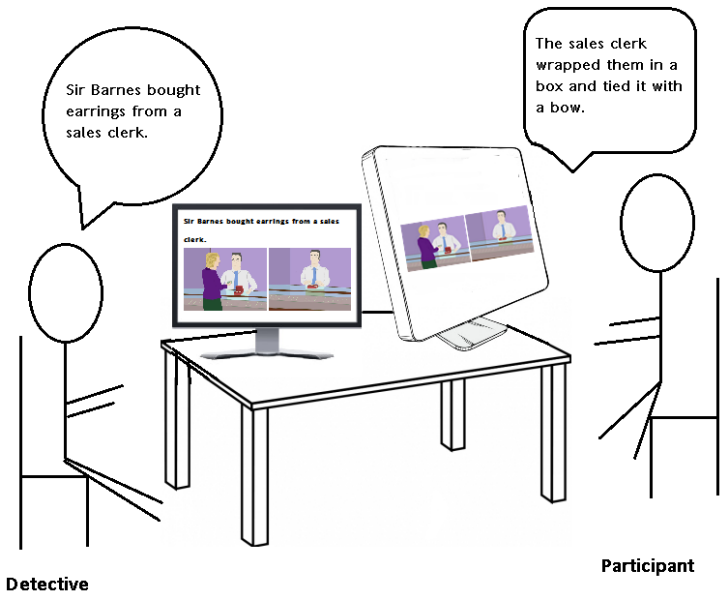


*Figure 3*. Experimental set-up from Experiment 1

The detective then entered the room and asked the participant to recount who the family was they had been photographing. Then the audio recorder was turned on and the detective sat down at her own computer. The two computers were placed back-to-back, and the participant's monitor was large enough that the participant and detective could not see one another. The detective and participant then began looking at the pairs of pictures together. The detective

would describe the first picture using a script, and the confederate would then say what happened next, by referring to the second picture displayed on her computer. Both pictures in the pair appeared at once on the screen, to encourage participants' conception of them as a coherent set. After the participant described the second picture the detective would then advance the pairs of pictures on both computers simultaneously. A depiction of this set-up can be seen in Figure 3.

When the detective and participant had described all the events, the detective then asked the participant who had been murdered, who had committed the crime and with what weapon, and why. The detective then told the participant they could both come out of character. The detective proceeded with further debriefing questions about the participants' familiarity with the Clue game.

*Analysis*

*Response coding*

Participants needed to refer to the character pictured in the second picture of each pair for the item to be included. Given the very high consistency of ratings between the original coder and the re-coders for Experiments 2 and 3, which had been run first, no double-coding was performed for this experiment.

56 trials were excluded from the final analysis, leaving 424 trials in the analysis. These were evenly divided among Subjects (203 items) and non-Subjects (221 items) and Goals (218 items) and Sources (206 items). 24 trials were excluded for being about non-human referents, 24 were excluded as the wrong character was referred to, one was excluded for being a plural Subject, one was excluded for using 'who' as the Subject, and six were excluded due to mechanical issues (two pictures were advanced instead of one; the picture was advanced too soon, etc.).

We also coded several control variables: 1) Use of a connective (after, afterwards, and, and then, next, now, then, after that), which indicates increased use of the discourse and conceptualization of the two events as a unit (Arnold & Griffin, 2007); 2) Whether participants mentioned the other character in their continuation; 3) Any disfluencies at the onset of the response.

*Audio data coding*

The audio data were analyzed with Praat to measure latency to begin speaking, defined as the end of the detective's speech to the beginning of the subject's response. These time points were coded both by two undergraduate research assistants, to check for reliability. When the latencies for the two coders were more than 10% different from one another (n=100, 23% of total), the first author (ER) either selected the coding she thought was correct, or if she thought both were incorrect, she supplied the correct latency.[7] For the rest of the cases, the latencies of the first research assistant were used.

*Statistical modeling*

Following the analytical procedure described above, we examined three dependent measures: 1) choice of reduced referential expressions (pronouns vs. zeros) vs. names, 2) whether the response included a connective word (*and, then, after that,* etc.), and 3) latency data. In the latency model, the control variable of Relatedness correlated with both Verb Codability ($r(422)=-.085$, $p=.081$) and Likelihood ($r(422)=.198$, $p<.0001$). Each of these variables was tested individually, in isolation of the others, in the model, and if it was significant it was retained. The control variables included are described above in Table 1.

[7] Many of these cases were ones in which the offset of the detective's speech was subtle, or one of the coders misapplied the rule concerning disfluent fillers (*um* was to be included in the latency period, rather than as the onset of the speech).

## Results

*Semantic and grammatical role effects on pronoun/zero production*

The critical question was whether participants would use more pronouns/zeros to refer to the Goal of the prior sentence as compared to the Source. Indeed, participants used more reduced forms when referring to prior-Goals as compared to prior-Sources, as can be seen in Figure 4 and Table 2. As was expected, they also used more pronouns and zeros when referring to Subjects of the prior sentence as compared to non-Subjects. There were no interactions.


*Figure 4.* Proportion of pro/zero used by semantic and grammatical roles in prior sentence in Experiment 1

Table 2. Pronoun/zero rate model (including interactions). Experiment 1 predictor variables, control variables, interaction terms, and random effects.

| | Variable | Estimate | Error | *t*-value | *p*-value |
|---|---|---|---|---|---|
| **Critical Predictors** | Goal vs. Source | 0.97 | 0.29 | 3.32 | 0.0036 |
| | Subject vs. non-Subject | 1.40 | 0.32 | 4.38 | 0.0003 |
| | Same gender vs. Different gender | -0.30 | 0.24 | -1.22 | 0.22 |
| | Order | -0.017 | 0.008 | -2.03 | 0.04 |
| **Interaction terms** | Goal *Subject | 0.76 | 0.49 | 1.55 | 0.12 |
| | Goal*Gender | -0.23 | 0.48 | -0.47 | 0.64 |
| | Subject*Gender | -0.006 | 0.48 | -0.01 | 0.99 |
| | Subject*Gender*Goal | 0.32 | 0.97 | 0.33 | 0.74 |
| **Control** | Use of connective word | 0.88 | 0.30 | 2.93 | 0.004 |
| | Likelihood | --- | --- | --- | --- |
| **Random Effects** | Participant | * | | | |
| | Participant by Subject vs. Non-Subject | * | | | |
| | Participant by Goal vs. Source | * | | | |

*Note.* T-values for predictor variables indicate their significance. Control variables with *t*-values >1.5 in the control model were included in the main effects model and their values in the main effects model are given here. Dashed lines for control variables indicate the variable was not significant in the control model and thus was not included. Random effects are noted with asterisks if included.

*Connective use effects*

Use of a connective was hypothesized to be an indicator of the participants' use of the discourse context. If participants conceptualized the two events as a related unit, it may encourage them to think about the relationships between the two, potentially increasing their use of words like *and then*, *next*, or *after that* to emphasize their connection. Participants in Experiment 1 used far more connectives (194) than participants in the other experiments with comparable numbers of participants (11 in Experiment 2, 9 in Experiment 3). This was likely due to both the verbal modality of response and possibly their use of the discourse context.

Contrary to expectations, we found that participants used more connectives (*after, afterwards, and, next, now, then*), when talking about Sources; see Table 3. This effect was not qualified by any interactions. The control predictor Referential form was also significant, indicating that connectives were more likely on trials with reduced referential expressions.

Table 3. Connective use model (including interactions). Experiment 1 predictor variables, control variables, interaction terms and random effects.

| | **Variable** | **Estimate** | **Error** | ***t*-value** | ***p*-value** |
|---|---|---|---|---|---|
| **Critical Predictors** | Goal vs. Source | -0.77 | 0.30 | -2.58 | 0.01 |
| | Subject vs. non-Subject | 0.56 | 0.34 | 1.66 | 0.11 |
| | Same gender vs. Different gender | -0.42 | 0.28 | -1.50 | 0.13 |
| | Order | -0.01 | 0.009 | -1.44 | 0.15 |
| **Interaction terms** | Goal *Subject | 0.31 | 0.56 | 0.56 | 0.57 |
| | Goal*Gender | 0.49 | 0.56 | 0.89 | 0.37 |
| | Subject*Gender | -0.10 | 0.56 | -0.19 | 0.85 |
| | Subject*Gender*Goal | 0.62 | 1.11 | 0.55 | 0.58 |
| **Control Variables** | Referent on right in picture | 0.67 | 0.30 | 2.21 | 0.03 |
| | Relatedness | --- | --- | --- | --- |
| | Disfluency | -1.22 | 0.46 | -2.65 | 0.008 |
| | Referential form (pro/zero vs name) | 0.66 | 0.35 | 1.92 | 0.05 |
| **Random Effects** | Participant | * | | | |
| | Participant by Subject vs. Non-Subject | * | | | |
| | Participant by Goal vs. Source | Not positive definite | | | |

*Note.* T-values for predictor variables indicate their significance. Control variables with *t*-values >1.5 in the control model were included in the main effects model and their values in the main effects model are given here. Dashed lines for control variables indicate the variable was not significant in the control model and thus was not included. Random effects are noted with asterisks if included.

*Latency effects*

If Goal continuations are easier to plan and produce, we would expect to see an effect on participants' latencies to begin speaking. Indeed, utterance initiation latencies were shorter when

the participant was referring to a Goal as opposed to a Source. There was no such effect of referring to the Subject versus the non-Subject, and no interaction between the two. Latency data can be seen in Figure 5 and Table 4.


*Figure 5.* Latency to begin speaking by grammatical and semantic roles in prior sentence in Experiment 1

Table 4. Latency model (including interactions). Experiment 1 predictor variables, control variables, interaction terms, and random effects.

| | Variable | Estimate | Error | *t*-value | *p*-value |
|---|---|---|---|---|---|
| **Critical Predictors** | Goal vs. Source | -0.06 | 0.02 | -3.27 | 0.001 |
| | Subject vs. non-Subject | -0.007 | 0.02 | -0.40 | 0.70 |
| | Same gender vs. Different gender | 0.02 | 0.02 | 1.09 | 0.28 |
| | Order | 0.0012 | 0.0006 | 1.98 | 0.05 |
| **Interaction terms** | Goal *Subject | 0.001 | 0.035 | 0.03 | 0.97 |
| | Goal*Gender | 0.04 | 0.037 | 1.17 | 0.24 |
| | Subject*Gender | -0.007 | 0.036 | -0.19 | 0.85 |
| | Subject*Gender*Goal | 0.12 | 0.07 | 1.66 | 0.10 |
| **Control Variables** | Referent on right in picture | 0.07 | 0.2 | 3.56 | 0.0004 |
| | Relatedness | -0.05 | 0.015 | -3.02 | 0.002 |
| | Disfluency | 0.21 | 0.03 | 6.92 | <.0001 |
| | Ease of verb retrieval | -0.08 | 0.03 | -2.33 | 0.02 |
| | Referent form choice (pro/zero vs name) | --- | --- | --- | --- |
| | Likelihood | --- | --- | --- | --- |
| **Random Effects** | Participant | * | | | |
| | Participant by Subject vs. Non-Subject | * | | | |
| | Participant by Goal vs. Source | Estimated to be zero | | | |

*Note.* T-values for predictor variables and interaction terms indicate their significance. Control variables that were significant in the main effects model were included, and T-values indicate their significance in this model. Dashed lines for control variables indicate the variable was not significant in the main effects model and thus was not included here. Random effects are noted with asterisks if included.

*Post-experiment questionnaire*

A post-experiment questionnaire was used for three purposes. First, we assessed participants' prior experience with the Clue game, since we hypothesized that familiarity would support their ability to create richer mental models. We confirmed that most participants (18/20) were familiar with the game. Second, we used the questionnaire to motivate participants' engagement with the task. Participants were told at the beginning of the experiment that at the end they would try to help the detective figure out who the culprit was, which they were asked in

the post-experiment questionnaire. All 20 participants correctly identified Sir Barnes as having been the character that was killed, and the motive as having something to do with Sir Barnes' affair with the maid. Third, we asked participants what they thought the goal of the experiment was. Participants reported several possibilities but none identified the specific linguistic manipulations.

Discussion

Experiment 1 established that speakers do indeed use reduced forms for goal referents more than source referents, supporting prior claims that thematic roles influence reference form production. As expected, participants also used more pronouns or zeros when talking about Subjects of the prior sentence than non-Subjects. A rating study demonstrated that goal-continuations were considered more likely than source-continuations, and also that the events were considered to be more related to each other in the goal continuations. This suggests that the preference to use reduced forms for goals is related to their predictability.

These findings raise questions about why predictability might affect reference production. One hypothesis we considered was that both pronouns and zeros are more likely in situations where production planning is easier. If it is easy for the speaker to activate representations of the events to be described, it may encourage both earlier speech and greater connection between the utterances. Consistent with this, we found that goal utterances were easier to plan than source utterances, in that they required shorter latencies to respond. However, when latency was added to the model for reference form choice, it did not independently predict reference form choice. This may indicate that utterance planning is not directly related to reference form choice. However, our task was not designed to allow a fine-grained measure of planning time. Participants had previewed the pictures, and further were able to begin planning

their response during the detective's description of the first picture, such that the latency was only a rough measure of planning difficulty. Thus, further work is needed to precisely specify the relationship between latency and reference production.

A second hypothesis we considered was that goal continuations would lead participants to perceive greater connectivity between the two events, perhaps due to a focus on the end state of the event. We tested this by analyzing the presence of connective words (*then, and, after*, etc.). Surprisingly, participants used more connectives when talking about Sources of the prior sentence as opposed to Goals. However, the presence of a connective moderated the goal effect on pronoun use, in that connectives were associated with the use of pronouns and zeros. We speculate that participants may have provided explicit connectives precisely because the Source continuation was less expected. That is, perhaps the use of a connective helped establish a relationship between two events that flowed less naturally. Participants used far more connectives in this experiment than in Experiments 2 and 3, probably due to the verbal nature of their response, but also perhaps reflecting a greater use of the discourse context. Nevertheless, it is clear that the thematic role effect is not explained by the production of connectives.

Most importantly, however, this study confirmed that thematic roles do indeed affect reference form. This finding stands in contrast to several published studies. This raises questions about what accounts for the differences. As described above, our event-retelling task differed from previous paradigms on two important dimensions. First, participants were not required to invent a continuation, and instead described a picture that they were already familiar with. Second, our items together told a story, increasing the ability to represent the discourse context. We examined each of these differences in Experiments 2 and 3.

## Experiment 2

*Motivation*

Experiment 2 was conducted using the same linguistic material as Experiment 1 but with the written story-continuation paradigm, to determine whether the same effects could be found with the standard methodology. We followed the method of Fukumura and van Gompel (2010, experiment 1), in which participants were told they had to begin their continuation with the underlined character.[8]

Method

*Participants*

20 participants were included in the analysis; an additional 26 completed the task but were replaced. Of these 46 participants, 10 were undergraduates who were reimbursed with course credit and 36 were recruited via Amazon Mechanical Turk and received monetary compensation. The Amazon Mechanical Turk participants needed to be native English speakers with a HIT approval rate greater than or equal to 98%, with at least 5000 approved HITs. The undergraduate participants needed to be native English speakers, have normal or correct-to-normal vision, and couldn't have participated in a similar experiment in the lab. 8 participants were excluded for using fewer than 2 proper names, 17 for using fewer than 2 pronouns or zeros, and 1 because we had collected enough data for even numbers on each list.

*Materials, Design, and Procedure*

The experimental trials consisted of the first in each pair of sentences from the storyline, which included 24 critical items and 29 fillers. The sentences were presented to participants with

[8] Another experiment, reported in Rosa (2015), used the same story-continuation paradigm, but did not tell participants who to mention. The numerical patterns in this experiment were very similar to those reported here, but the frequency of continuing with the goal character was so strong that there were very few source-continuation items, which hindered our ability to assess the effect statistically.

a computerized survey through Qualtrics. Participants were instructed to provide a plausible continuation about the character that was underlined in the first sentence, which they typed in the box. Although the stimuli sentences were identical to the context sentences from Experiment 1, this task did not include any instructions to place the task within a murder-mystery narrative. Participants were presented with one of the two lists created, allowing them to see each item in one of the two conditions, but both conditions across items. An example item from Experiment 2 is shown in Figure 6:

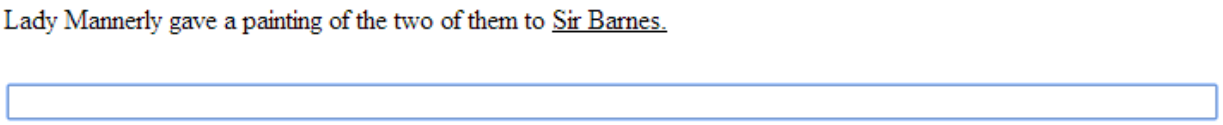


*Figure 6.* Sample trial from Experiment 2

*Analysis*

*Response coding*

We coded the grammatical subject of the first clause of each response, where a clause could be either a main or subordinate clause. Responses were coded for both a) choice of referring expression (pronoun/zero or proper name) and b) role of referent in the prior sentence (Subject or non-Subject and Goal or Source). Items were excluded if participants referred to more than one person at once (e.g., *Then they put the groceries away*), they did not refer to the character that was underlined, or they referred to someone's possession or body part as the subject (e.g., *Sir Barnes' back was sore*).

The first author coded the data and then one of two undergraduate research assistants re-coded the data, blind to the original coding. Of the 480 items there were six in which the original and re-coding did not match, or 1.25% of the data. The first author determined one of these to be

about the wrong referent, and excluded the other five on the basis that the responses were ambiguous, and they were excluded. Following the criteria mentioned above, fifty-five items were excluded from the final analysis, leaving 425 items (223 Goal items and 202 Source items; 223 Subject items and 202 non-Subject items.

Responses were also coded for use of connectives (*if, then, and, so,* etc), and for whether the non-designated character was referred to in the same clause as the designated referent.

The chef, who was pictured as female in experiment 1, was interpreted as male by most of the participants. Likewise, the sales clerk was intended to be male but most participants interpreted him as a female. We changed the coding of gender for the items with the chef or the sales clerk to reflect this.

*Coherence relations coding*

The relationships between the prompts and the continuations given were also examined. Given that certain coherence relations support Goal continuations (next-event mentions) and others support Source continuations (explanations or motivations for the events), it was important to code for and analyze the types of continuations participants provided to consider all possible contributors to referential form. Using Rohde's coding schema[9] two undergraduate research assistants independently coded each continuation. The seven categories they used were elaboration, explanation, occasion, parallel, result, violated expectations, and background. After coding all the items, the RAs then compared their ratings. On 147 of the 426 total they had coded items differently, so these items were discussed until they had reached an agreement about the appropriate coding.

Two of continuation types (result, occasion) describe events that occur as a result of another event or after it temporally, and thus are more consistent with Goal as opposed to Source

[9] We are very grateful to Hannah Rohde for sharing her coding schema with us for this project.

continuations. We therefore collapsed the continuation codings into two groups: a) Occasion/Result and b) Other. Of the Goal continuations, 120 were coded as Occasion or Result, and 103 were coded as a different continuation type. Of the Source continuations, 81 were coded as Occasion or Result and 121 were coded as a different type.

*Statistical modeling*

The data were analyzed following the general analytic approach outlined above, except that coherence continuation was added as a control variable. We also examined interactions between coherence continuation and the critical Goal/Source variable.

## Results

*Semantic and grammatical role effects on pronoun/zero production*

Similar to Experiment 1, participants used more reduced forms when referring to goals of the prior sentence as opposed to sources. As expected, participants used more pronouns or zeros when referring to Subjects of the prior sentence as compared to non-Subjects (see Figure 7 and Table 5). In addition, the Goal effect was qualified by a marginal interaction with continuation type that approached significance (see Figure 8 and Table 5). Contrasts revealed the marginal interaction was due to a significant difference in pronoun/zero use for Occasion/Result continuations between Goal and Source items $t(379)=3.50$, $p=0.0005$. No difference in pronoun/zero use was seen for Other continuations between Goal and Source items $t(379)=0.91$, $p=0.37$ (see Figure 8).

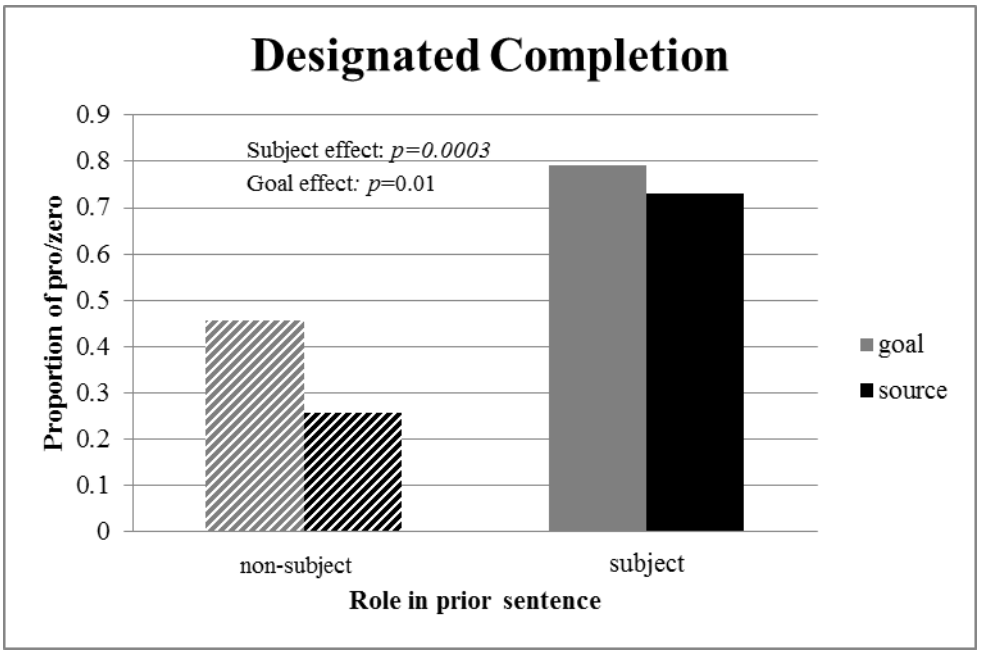


*Figure 7.* Proportion of pro/zero used by semantic and grammatical roles in prior sentence in Experiment 2

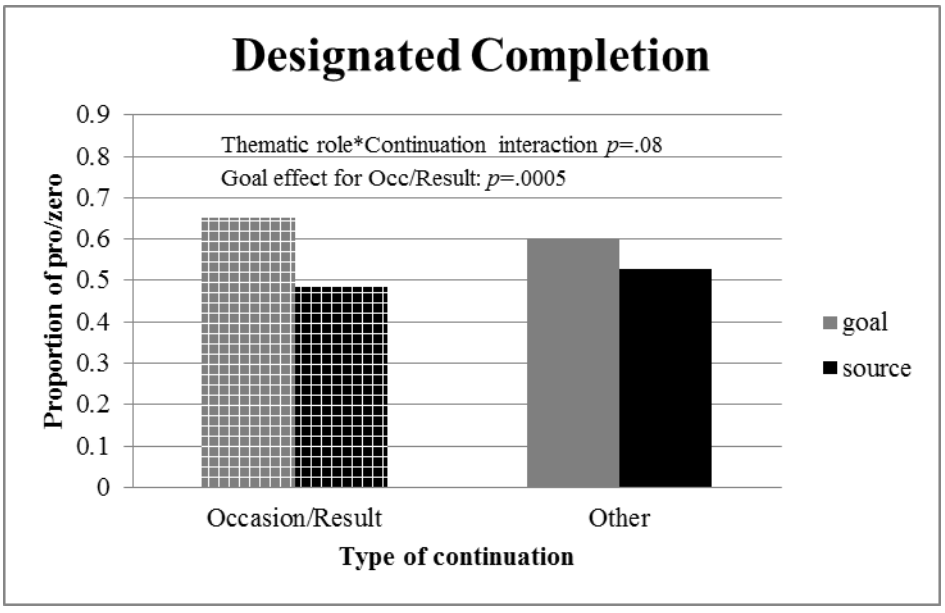


*Figure 8*. Proportion of pro/zero used by semantic roles and continuation type in prior sentence in Experiment 2

Table 5. Pronoun/zero rate model (including interactions). Experiment 2 predictor variables, control variables, interaction terms, and random effects.

| | Variable | Estimate | Error | *t*-value | *p*-value |
|---|---|---|---|---|---|
| **Critical Predictors** | Goal vs. Source | 0.81 | 0.31 | 2.60 | 0.01 |
| | Subject vs. non-Subject | 3.00 | 0.37 | 8.19 | <.0001 |
| | Type of Continuation (Occ/Result vs other) | 0.50 | 0.32 | 1.58 | 0.11 |
| | Same gender vs. Different gender | -1.47 | 0.34 | -4.25 | <.0001 |
| | Order | 0.03 | 0.01 | 2.16 | 0.03 |
| **Interaction terms** | Goal *Subject | -0.42 | 0.58 | -0.72 | 0.47 |
| | Goal*Gender | -0.14 | 0.65 | -0.22 | 0.82 |
| | Subject*Gender | 0.85 | 0.62 | 1.37 | 0.17 |
| | Goal*Continuation type | 1.03 | 0.59 | 1.74 | 0.08 |
| | Subject*Gender*Goal | 1.17 | 1.22 | 0.96 | 0.34 |
| **Control** | Mention other person | 0.82 | 0.34 | 2.41 | 0.016 |
| | Use of connective word | --- | --- | --- | --- |
| **Random Effects** | Participant | * | | | |
| | Participant by Subject vs. Non-Subject | * | | | |
| | Participant by Goal vs. Source | Estimated to be zero | | | |

*Note.* T-values for predictor variables and interaction terms indicate their significance. Control variables that were significant in the main effects model were included, and T-values indicate their significance in this model. Dashed lines for control variables indicate the variable was not significant in the main effects model and thus was not included. Random effects are noted with asterisks if included.

## Discussion

Experiment 2 found the predicted thematic role effect. Participants produced more pronouns or zeros when referring to prior-Goals than prior-Sources. They also showed the expected effects of producing more pronouns or zeros for prior Subjects as opposed to prior non-Subjects, and in the different-gender condition. Consistent with previous effects (Arnold, 2001; Rohde, Kehler, & Elman, 2007) coherence relations somewhat modulated this effect: when the

coherence relation was consistent with the thematic bias (i.e. a prior-Goal in an Occasion/Result continuation), more reduced forms were used than when it was inconsistent.

Critically, this experiment demonstrated that the thematic role effect is robust to experimental paradigm. Despite the concerns that we had about the story-continuation paradigm, it appears that it is possible to detect thematic role effects using this method. However, there was one way in which this experiment differed from the standard method, in that the items all referred to the same cast of six characters, and together told a story. Experiment 3 tested whether this relatedness was necessary to observe the thematic role effect on reference production.

## Experiment 3

### *Motivation*

Experiment 3 was conducted to determine whether the thematic effect found in Experiments 1 & 2 was dependent upon the experimental items being related to one another. It may have been the case that the semantic predictability effect was due to the fact that participants were able to build a mental model of the events as a whole, freeing up mental resources to utilize predictability information. Experiment 3 was identical to Experiment 2, except it eliminated the repeated mention of people and items.

### Method

### *Participants*

57 participants completed the task on Amazon Mechanical Turk, all for a monetary reward. The same inclusion criteria were used as for the M-Turk participants in Experiment 2. 37 participants were excluded, leaving a total of 20 participants for whom data was analyzed. Of

the 37 who were not analyzed, five were excluded for doing an earlier version of the experiment, three were excluded for providing meaningless continuations, 25 were excluded for using fewer than two pronouns or zeros, and four were excluded for using fewer than two names.

*Materials, Design, and Procedure*

The design and procedure was identical to Experiment 2. The only difference was that the names/occupations and objects were changed, however, such that none were repeated across items. Names were replaced by another common name of the same gender (selected from a list of the most popular male and female names in 1958), and occupations were replaced by another common occupation. All attempts were made to replace occupations with other occupations that typically are gender-specific, to preserve the same and different gender makeup of the stimuli. Objects that had been mentioned in the original items were replaced with other common objects such that none were repeated. An example is *Michael received a painting of the two of them from Mary.*

*Analysis*

*Response Coding*

The inclusion criteria were identical to those of Experiment 2. Given the very high consistency of ratings between the original coder and the re-coders for Experiment 2, no double coding was performed for this experiment. Fifty-eight items were excluded from the final analysis, leaving 422 items (211 in the Subject condition and 211 non-Subject condition; 222 in the Goals condition and 200 in the Sources condition).

The coding procedure was identical to Experiment 2, except that for the coherence relations coding we only coded for the binary distinction that was used for the analysis (Occasion/Result vs. other). One undergraduate research assistant coded all the data and the first

author (ER) coded a sample of 20% of the data (84 items) to check for consistency. The two coders disagreed on 6 of the 83 items, or 7%. Four of these six were determined to be coded correctly by the research assistant, and the other two were determined to be coded correctly by the first author.

Results

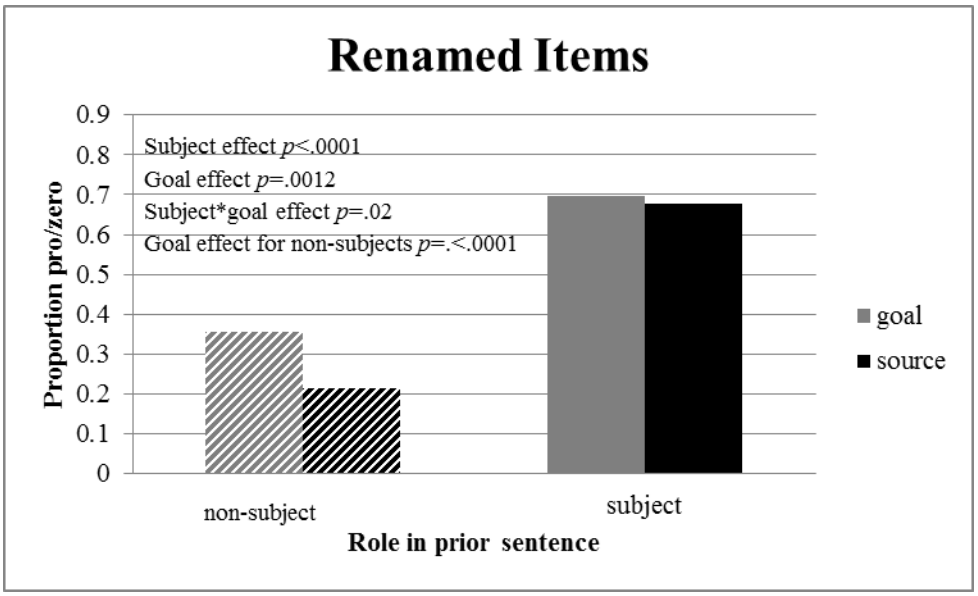


*Figure 9*. Proportion of pro/zero in Exp. 3, by thematic and grammatical role in prior sentence.

Just as in experiments 1 and 2, participants used more pronouns/zeros to refer to both Goals and prior subjects (see Table 6). These effects were qualified by a Goal by Subject interaction (see Figure 9), and a Goal by gender interaction (see Figure 10). Contrasts in the model suggested that the interaction between Subject and Goal was due to an effect in the non-Subject condition $t(373)=4.12$, $p<.0001$, and no effect in the Subject condition $t(373)=1.01$, $p=0.31$. The Goal by gender interaction, as seen in Figure 12, was revealed by contrasts to be due to a significant thematic role effect in the same gender condition $t(377)=3.50$, $p=.0005$, and no difference in the different gender condition $t(377)=0.97$ $p=.33$.

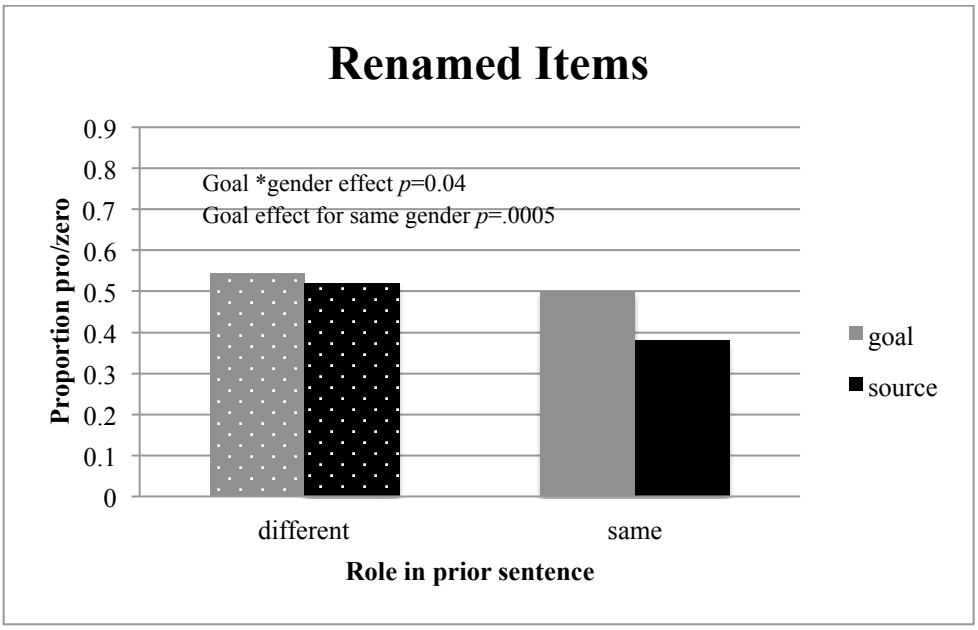


*Figure 10*. Proportion of pro/zero in Exp. 3, by gender thematic role in prior sentence.

Table 6. Pronoun/zero rate model (including interactions). Experiment 3 predictor variables, interaction terms, and random effects.

| | **Variable** | **Estimate** | **Error** | ***t*-value** | ***p*-value** |
|---|---|---|---|---|---|
| **Critical Predictors** | Goal vs. Source | 1.09 | 0.33 | 3.27 | 0.0012 |
| | Subject vs. non-Subject | 3.24 | 0.40 | 8.16 | <.0001 |
| | Same gender vs. Different gender | -0.93 | 0.30 | -3.05 | 0.002 |
| | Type of continuation (Occ/Result vs other) | -0.12 | 0.37 | -0.33 | 0.74 |
| | Order | 0.05 | 0.01 | 4.46 | <.0001 |
| **Interaction terms** | Goal *Subject | -1.37 | 0.60 | -2.28 | 0.02 |
| | Goal*Gender | 1.32 | 0.61 | 2.16 | 0.03 |
| | Subject*Gender | 0.66 | 0.59 | 1.11 | 0.27 |
| | Subject*Gender*Goal | -1.90 | 1.19 | -1.59 | 0.11 |
| | Goal*Type of Continuation | 0.11 | 0.65 | 0.17 | 0.86 |
| **Random Effects** | Participant | * | | | |
| | Participant by Subject vs. Non-Subject | * | | | |
| | Participant by Goal vs. Source | Estimated to be zero | | | |

*Note.* T-values for predictor variables and interaction terms indicate their significance. No control variables were significant in the main effects model and thus none were used here. Random effects are noted with asterisks if included.

Discussion

As found in Experiments 1 & 2, participants referred to prior Goals with more pronouns or zeros as compared to prior Sources. As expected, participants continued to prefer using reduced forms when referring to Subjects of the prior sentence as opposed to non-Subjects, and in the different-gender condition. As the experimental items were no longer related to one another, this effect could not be due to any coherent, overarching model of the events.

Experiment 3 revealed a few interactions that were not observed in Experiment 2. The goal-bias was stronger for the non-subject condition, compared with the subject condition. In addition, it was stronger in the same-gender condition than the different-gender condition. These findings underscore the fact that the thematic role effects are relatively weak. It may be that the lack of a coherent story here diminished the ability of participants to build a robust mental representation of the events, weakening the thematic role effect further. The weakness of the thematic role effect may also explain why the goal effect was not modulated by the coherence relations (unlike in Experiment 2).

Nevertheless, it is notable that in at least some conditions we observed the same tendency to use a higher rate of pronouns and zeros for prior goals, compared with prior sources. This finding argues against the hypothesis that a story-like context is required to observe the effect.

## General discussion

The findings from our experiments were strikingly consistent: In all three studies, we found that thematic roles do influence referential form. Across two paradigms and with different sets of materials, participants consistently used more reduced referring expressions to refer to the

goal character than the source character. This supports the goal bias reported by Arnold (2001), and contrasts with studies that failed to find thematic role effects.

We considered the possibility that the failure of earlier studies to find thematic role effects was due to the use of the story-continuation method, which we suspected was ill-suited to finding coherence-driven effects on reference form. Yet we found robust evidence against this claim, in that the same goal bias affected pronoun/zero production across all three experiments, regardless of experimental paradigm. The effect emerged regardless of how interactive the task was, whether participants' responses were written or spoken, and whether or not the experimental items were related to each other.

We also observed the expected interaction between coherence relation and thematic role in experiment 2 (although it was only marginal), where the goal bias was strongest in the Occasion/Result continuations. This supports the "end-state" claim (Stevenson et al., 1994), that is, that a sentence about the result or next event encourages a focus on the goal. Importantly, this suggests even in a story continuation paradigm, participants are able to generate a continuation and activate the coherence relation early enough for this to affect the selection of a referential form.

Nevertheless, we also observed some evidence that thematic role effects are stronger in a paradigm with a robust discourse context, as in our experiments where the items told a story. In Experiment 1 (event-retelling) and Experiment 2 (sentence completion), the items centered on the same characters and had a murder-mystery theme. In these experiments, we observed the goal bias across the board in all conditions. When the story context was eliminated in experiment 3, the goal bias was mitigated by interactions with subjecthood and gender. In both cases, the goal bias emerged only in the condition that discouraged pronoun use, i.e. the non-

subject condition and the same-gender condition. This suggests that while a story-like context may not be essential for thematic role effects on reference form, it may be helpful.

***What explains the contrast with previous reports?***

The current results provide an empirical contribution to the literature on reference production, because they stand in contrast to numerous claims that thematic role biases do not affect the production of pronouns (Fukumura & van Gompel, 2010; Kehler & Rohde, 2013; Kehler et al., 2008). There are several possible explanations for this, stemming from differences between this study and previous ones.

The main difference is that the current set of studies utilized transfer verbs, while most other studies have examined implicit causality verbs. We therefore speculate that the thematic role effect may be restricted to transfer verbs, but further work is needed to test this hypothesis. One way in which the classes of verbs differ is in their telicity. Transfer verbs are telic (they have an endpoint), whereas the experiencer/stimulus verbs used here are atelic (they do not have an endpoint). Transfer events, and possibly telic events more generally, may be easier to conceptualize, leading to a stronger discourse model.

Another possibility is that detecting thematic role effects requires specific experimental conditions. Despite the robustness of the goal bias on reference production across tasks, overall the thematic role effect was small. Moreover, reference form choice in our tasks was modulated by numerous other factors, such as grammatical role, gender ambiguity, and several control variables. This highlights the importance of examining thematic role effects against the backdrop of strong control of other predictors.

We also observed substantial variation among tasks, and across participants. Moreover, individual participants ranged widely in their use of pronouns, and many were excluded for lack

of variability in their responses. If the features of the task elicit performance that is at ceiling (~100% pronoun/zero use) or floor (~0% pronoun/zero use), the effect cannot be detected. Thus, in order to test any effect – including thematic roles -- it is important to find the "sweet spot" of variability in a particular task. In our study, we helped achieve the right balance in variability by adopting stricter exclusion criteria for participants than previous studies, including only participants who used some variation in their expressions. We examined how our analyses would be affected without this criterion by taking a sample of the first 20 participants in Experiment 2, regardless of whether they exhibited variation. In this sample, the thematic role effect disappeared.

On the other hand, there may be some concern that we replaced a high number of participants in our studies, especially experiments 2 and 3. This could be a problem if our sample did not represent the general population. However, we argue that it is important to use strict inclusion criteria in order to sample participants who are trying to follow the instructions of the task. This is especially important in the increasingly popular "crowd-sourcing" approach to data collection, using Amazon mechanical turk. Some of these participants may be less engaged in the study than a live subject performing in front of a live experimenter. Indeed, there is no guarantee that participants read the stimulus sentence carefully in the mechanical turk studies. Consistent with this, five subjects were excluded for meaningless responses. Lack of engagement would decrease the participant's sensitivity to the discourse context, and thus decrease variability in responses.

A final difference between this and previous studies is that we examined the production of pronouns and zeros together, as opposed to just pronouns. We chose to combine these because they play a similar discourse function. In addition, excluding zeros from analysis would

lead to an unrepresentative view of reference production, in that zeros are likely to be used when the referent is particularly accessible. Nonetheless, this analytical choice was not necessary to observe the goal bias on reference form. Zeros represented 2.16% of the reduced forms overall (7.8% in Exp. 1, 0% in Exp. 2, and 0.48% in Exp. 3), and thus were not a major portion of our data.

***What mechanism underlies thematic role effects?***

The empirical results of this paper establish that speakers use reduced forms more often for goals than sources. This question is theoretically interesting because it demonstrates one way in which reference form is influenced by the predictability of referents. Numerous authors have claimed that goal arguments are more referentially predictable than source arguments, especially in a context where the following sentence is expected to provide information about a following event (e.g., Arnold, 2001; Kehler et al., 2008; Stevenson et al., 2004). We found confirming evidence that the goal character in our stimuli was perceived as more predictable, in the rating study (Exp. 1) where participants chose the goal as the character likely to be mentioned next. Further evidence comes from a companion story-completion study (Rosa, 2015), in which the choice of who to mention in the continuation was left up to the participant. In this study, continuations overwhelmingly mentioned the goal character. However, when we included likelihood and relatedness as trial-by-trial metrics, in none of our analyses were these measures significant on top of the effects of thematic role and subjecthood.

However, we also found that referential predictability patterns with other properties of the stimuli. Another rating study found that the events involving the goal character were rated as more predictable than events involving the source character, and that the two events were also rated as more related. These properties are conceptually highly related to the predictability of the

character him or herself. Thus, any or all of these properties may underlie the thematic role effect.

It is also important to point out that predictability is not the only determinant of reference form. In our data, there was also a strong tendency to use pronouns and zeros more for subjects than non-subjects. Arnold (1998, 2001) has suggested that subjects are more predictable than non-subjects, and Brennan (1995) has also suggested that speakers continue talking about referents in subject position more than other referents. However, our rating studies did not find that predictability or relatedness varied between subject and non-subject continuations. We consider three possible interpretations of this finding.

One possibility is that subjecthood is unrelated to predictability. If so, subjecthood and thematic roles may influence referential form for different reasons. A second possibility is that predictability is calculated in a dynamic way, as the utterance unfolds. Early in the utterance, subjects may be perceived as highly predictable, and thus represented in a relatively accessible way. By the end of the utterance, the coherence relation may shift predictability toward the goal, but the accessible discourse representation remains. A third possibility is that subjecthood by itself is a weak indicator of predictability, but it tends to co-occur with other indicators of topicality in natural discourse, like repeated mention and topicalization. Further work on the timecourse of discourse representations is needed to test these possibilities.

Even if predictability explains the thematic role effect, a critical question is what processing mechanisms explain this. For a reader or listener, predictability corresponds to the listener's ability to anticipate the upcoming input. For a speaker, the implications of predictability are not as straightforward. Speakers plan their utterances ahead of time, which means that they do not need to "predict" their utterances per se.

One hypothesis we considered was that predictability in production corresponds to the ease of planning. Information that is predictable to comprehenders is also redundant with the context. This may support the production processes required to conceptually plan an utterance and formulate it linguistically. If processing facilitation leads to fluency, it may support the use of reduced expressions (Arnold, under review; Arnold & Watson, 2015; Arnold & Nozari, under review).

If planning facilitation underlies the thematic role effect, we should see evidence that goal continuations are easier to plan than source continuations. Indeed, speakers in Experiment 1 initiated their utterances more quickly in goal continuations. However, if utterance planning is the primary determinant of reduced reference forms, we might also expect to see a relationship between onset latency and the production of pronouns/zeros. Yet we found no such relationship. Nor did we observe the influence of other measures of planning difficulty: whether or not participants were disfluent was not related to referential form, nor was the measure of verb codability, which indexed the ease of planning and producing the verb. This evidence is not definitive, though, because the latency measure was a somewhat rough measure of participants' planning. The latency was measured from the end of the detective's speech until the beginning of the participant's speech. However, participants had, at that point, been examining the pair of pictures and had heard the description of the first event. Their planning, therefore, had been going on for several seconds, and they had already been able to examine the pictures and establish the relationship between them.

Another possibility is that predictability affects reference form by directing the attention of discourse participants toward particular referents. When something is predictable, speakers instantiate a more-accessible representation of that referent in their mental model. This helps

speakers achieve the goal of cooperative communication, by anticipating what the addressee considers predictable and thus accessible. This mechanism would be broadly consistent with claims that salience in the discourse context increase the likelihood of using reduced expressions (Ariel, 1990; 2001; Chafe, 1976, 1994; Gundel et al., 1993), and with the observation that speakers in Experiment 1 used reduced forms more often in trials with an overt connective, which signaled that they were aware of the connection between the utterance and the prior discourse context. However, further work is needed to isolate the mechanism behind thematic role effects.

## Conclusion

In conclusion, the current study has found that thematic roles do play a role in determining referential form, in that speakers used pronouns and zeros for reference to goal characters more than source characters. We also found that goals were perceived as more predictable than sources in a rating study, and in Experiment 1 target likelihood at the item level increased the use of pronouns and zeros. This suggests that predictability does indeed affect reference form, contrary to claims that it does not.

**Author Note**

This work was funded by NSF grant 1348549 to J. Arnold, and the development of the visual materials was supported by the Stephenson/Lindquist fund at UNC Chapel Hill. Many thanks to Michaela Neely, Grant Huffman, Bryan Smith, Anita Simha, and Taylor Beard for their help collecting and coding the data.

**References**

Altmann, G.T.M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247-264.

Ariel, M. (1990). *Accessing noun-phrase antecedents*. London: Routledge.

Arnold, J. E. (1998). *Reference Form and Discourse Patterns*. Dissertation, Stanford University.

Arnold, J. E. (2001). The effect of thematic role on pronoun use and frequency of reference continuation. *Discourse Processes, 31*, 137-162.

Arnold, J.E. (2008). Reference Production: Production-internal and Addressee-oriented Processes. *Language and Cognitive Processes*. *23*(4), 495-527.

Arnold, J.E. (2010). How speakers refer: the role of accessibility. *Language and Linguistic Compass, 4,* 187-203.

Arnold, J.E., & Griffin, Z. (2007). The Effect of Additional Characters on Choice of Referring Expression: Everyone Competes. *Journal of Memory and Language*. *56*(4), 521-536.

Arnold, J.E., & Nozari, N. (under review). The effects of utterance planning and stimulation of left prefrontal cortex on the production of referential expressions.

Arnold, J.E. & Watson, D.G. (2015). Synthesizing meaning and processing approaches to prosody: Performance matters. *Language, Cognition, and Neuroscience, 30,* 88-102.

Aylett, M., Turk, A. (2004). The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence and Duration in Spontaneous Speech. *Language and Speech. 47*(1), 31-56

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. *The Journal of the Acoustical Society of America. 113*(2),1001.

Brennan, S.E. (1995). Centering attention in discourse. *Language and Cognitive Processes, 10,* 137-167.

Brown, R., & Fish, D. (1983). The psychological causality implicit in language. Cognition, 14 (3), 237-273.

Buschkuehl, M., Garcia, L. H., Jaeggi, S. M., Bernard, J. A., & Jonides, J. (2014). Neural Effects of Short-Term Training on Working Memory. *Cognitive, Affective & Behavioral Neuroscience*, *14*(1), 147–160.

Dell, G.S. (1987). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*(3), 283-321.

Erlich, K. (1980). Comprehension of pronouns. *Quarterly Journal of Experimental Psychology, 32,* 247–255.

Ferreira, F., & Swets, B. (2002). How incremental is language production? Evidence from the production of utterances requiring the computation of arithmetic sums. *Journal of Memory and Language, 46*(1), 57-84.

Filmer, H.L., Mattingly, J.B., & Dux, P.E. (2015). Object substitution masking for an attended and foveated target. *Journal of Experimental Psychology: Human Perception and Performance. 41*(1), 6-10.

Fine, A.B., Jaeger, T.F., Farmer, T.A., Qian,T. (2013). Rapid Expectation Adaptation during syntactic comprehension. *PLoS One, 8*(10).

Frank, A. & Jaeger, T.F. (2008). Speaking rationally: Uniform information density as an optimal strategy for language production. *Proceedings of the Annual Meeting of the Cognitive Science Society.*

Fukumura, K., Hyönä, J, & Scholfield, M. (2013). Gender affects semantic competition: the effect of gender in a non-gender marking language. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *39*, 1012-1021

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

Gahl, S., & Garnsey, S. M. (2004). Knowledge of grammar, knowledge of usage: syntactic probabilities affect pronunciation variation. *Language, 80*(4), 748-775.

Givon, T. (1983). Topic continuity in discourse: An introduction. In T. Givon (Ed.), *Topic continuity in discourse: A quantitative cross-language study.* (pp. 1-41). Amsterdam: John Benjamins.

Gregory, M.L., Raymond, W. D., Bell, A., Fosler-Lussier, E., and Jurafsky, D. (1999). The effects of collocational strength and contextual predictability in lexical production. In CLS-99, Chicago. University of Chicago.

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

Hale, J. (2001). A probabilistic earley parser as a psycholinguistic model. In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics,* Pittsburgh, PA.

Hartshorne, J.K., O'Donnell, T.J., Tenenbaum, J.B. (2015). The causes and consequences explicit in verbs. *Language, Cognition & Neuroscience, 30*(6), 716-734.

Jaeger, T. F. (2006). Redundancy and Syntactic Reduction in Spontaneous Speech. PhD thesis, Stanford University, Stanford, CA

Jurafsky, D. (1996). A probabilistic model of lexical and syntactic access and disambiguation. *Cognitive Science, 20*, 137–194.

Jurafsky, D., Bell, A., Fosler-Lussier, E., Girand, C., & Raymond, W. D. (1998). Reduction of English function words in Switchboard. *Proceedings of ICSLP-98,* Sydney.

Kahn, J. M. and Arnold, J. E. (2012). A processing-centered look at the contribution of givenness to durational reduction. *Journal of Memory and Language*, 67(3):311–325.

Kaiser, E., Li, D., Holsinger, E. (2011). Exploring the lexical and acoustic consequences of referential predictability. In I. Hendricks, A. Branco, S. Lalitha Devi, & R. Mitkov. (Eds.) *Anaphora Processing and Applications, Lecture Notes in Artificial Intelligence,* Vol. 7099. Heidelberg: Springer, 171-183.

Kehler, A. (2002). *Coherence, Reference, and the Theory of Grammar*. Stanford, CA:CSLI Publications.

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

Kehler, A., Rohde, H. (2013). A Probabilistic Reconciliation of Coherence-Driven and Centering-Driven Theories of Pronoun Interpretation. *Theoretical Linguistics, 39*, 1-37.

Kehler, A. & Rohde, H. (2014). Pronominal reference and inferred explanations: a Bayesian account. *RefNet Workshop on Psychological and Computational Models of Reference Comprehension and Production*, Edinburgh, UK. .

Levelt, W.J.M. (1989). *Speaking: From intention to articulation*. Cambridge: The MIT Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition, 106*(3), 1126-1177.

Levy, R. and Jaeger, T. F. (2007). Speakers optimize information density through syntactic reduction. In *Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems*.

Lieberman, P. (1963). Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech, 6*, 172-187.

Piantadosi, S. T., Tily, H., and Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences of the United States of America, 108*(9):3526–3529.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition, 42*, 107-142.

Rohde, H. (2008). *Coherence-driven effects in sentence and discourse processing*. Dissertation: UCSD.

Rohde, H., Kehler, A. (2014). Grammatical and information-structural influences on pronoun production. *Language, Cognition, and Neuroscience*, *29*(8), 1-16.

Rohde, H., Kehler, A., Elman, J.L. (2007). Pronoun interpretation as a side effect of discourse coherence. In *Proceedings of the 29th Annual Conference of the Cognitive Science Society.*

Seidenberg, M.S. & MacDonald, M.C. (1999). A probabilistic constraints approach to language acquisition and processing. *Cognitive Science, 23*, 569-588.

Son, R.J.J.H, & van Santen, J.P.H. (2005). Duration and spectral balance of intervocalic consonants: A case for efficient communication. *Speech Communication, 47*, 100-123.

Staub, A. & Clifton Jr., C. (2006). Syntactic Prediction in Language Comprehension: Evidence From Either…or. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32*(2), 425-436.

Stevenson, R.J., Crawley, R.A., Kleinman, D. (1994). Thematic roles, focus, and the representation of events. *Language and Cognitive Processes, 9*, 519-548.

van Berkum, J.J., Brown, C.M., Zwitserlood, P., Kooijman, V., Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(3), 443-467.

van Rij. J., van Rijn, H., & Hendriks, P. (2012). How WM load influences pronoun interpretation. In N. Rußwinkel, U. Drewitz & H. van Rijn (eds.),*Proceedings of the 11th International Conference on Cognitive Modeling*, Berlin: Universitaetsverlag der TU Berlin.

Zipf, G. (1936). *The Psychobiology of Languag*e. London: Routledge.